ISSN 1913-8989

COMPUTER AND INFORMATION SCIENCE

Vol. 1, No. 2
May 2008

Canadian Center of Science and Education

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

Contents

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

An OAIS Based Approach to Effective Long-term Digital Metadata Curation

Arif Shaon (Correspondence Author)

Centre for Advanced Computing and Emerging Technologies (ACET), The University of Reading

PO box 68, Whiteknights Campus, Reading, UK

E-mail: a.b.s.shaon@rdg.ac.uk

Andrew Woolf

Science And Technology Facilities Council (STFC)

Rutherford Appleton Laboratory, Didcot, UK

E-mail: a.woolf@rl.ac.uk

The research is jointly financed by the University of Reading and the Science And Technology Facilities Council (STFC)

Abstract

Metadata has the proven ability to provide information necessary for successful long-term curation of digital objects. However, without curation metadata itself may deteriorate in terms of its quality and integrity over time. Therefore, a digital curation process needs to incorporate the curation of metadata along with that of data in order to ensure the accurate description of data over time. Unfortunately, no comprehensive method for effective curation of metadata for long periods of time is known to exist at present. Even the Reference Model for Open Archival Information System (OAIS), despite being the most comprehensive and widely adopted framework for long-term data preservation, fails to address the requirements of long-term metadata curation in a comprehensive and unambiguous manner. This paper presents an approach to efficiently curating digital metadata over the long-term that is achieved through articulating the metadata curation related ambiguities of the OAIS Reference Model. The approach essentially involves the use of a "Metadata Curation Model", which is a specialised edition of the "Data Management" module of the OAIS Reference Model, dedicated to the purpose of long-term metadata curation.

Keywords: Metadata, Curation, Preservation, OAIS

1. Introduction & Motivation

Exponential increases in computing power and communication bandwidth have resulted in a dramatic rise in the volume of generated and published data within various complex information domains. This increasingly large volume of data needs to be preserved and made highly available (i.e. curated) over substantially long-periods of time in order to assist in high quality future research and experiments in both same and cross-discipline environments, as well as other productive uses of the data. However, the rapid growth of related technology and increased flexibility in their use also create a significant imbalance between the capacity for data generation and (long-term) data curation as the former is advancing significantly faster than the latter. And this imbalance in effect, poses the major challenge toward ensuring efficient and continued use of valuable data resources with their quality and integrity intact over the long-term.

Under the challenges set by the task of successful long-term data curation, the word 'Metadata' (i.e. further information about data) is becoming increasingly prevalent, with a growing awareness of the role that it can play in capturing information necessary for efficient functioning of different curation operations, such as data preservation and provenance tracking. For data preservation in particular, metadata can be used to record information required to reconstruct or at the very least understand the reconstruction process of digital resources on future technological platforms. However, without curation, metadata itself may deteriorate in terms of its quality and integrity over time. Therefore, a digital curation process needs to incorporate curation of metadata along with data in order to ensure an accurate description of data over time.

Over the past few years, several organised and arguably successful endeavours (e.g. NEDLIB, 2000 and CEDARS, 2002) have been made in order to find an effective solution for successful long-term data preservation. However, the territory of long-term metadata curation, although increasingly acknowledged, has yet to be conquered. In fact, in

most digital preservation or curation motivated workgroups and projects, the necessity of long-term metadata curation is deemed secondary, mainly due to the lack of awareness of the criticality of the problem. Even the Reference Model for Open Archival Information System (OAIS, 2002), despite being the most comprehensive and widely adopted framework for long-term data preservation, fails to address the requirements of long-term metadata curation in a comprehensive and unambiguous manner. As a result, no comprehensive method for effective curation of metadata for long periods of time is known to exist at present. This paper presents an approach that aims to fill the void for an efficient strategy for curating digital metadata over the long-term. The approach involves the use of a "Metadata Curation Model", which is a specialised version of the "Data Management" module (OAIS, 2002) of the OAIS Reference Model, dedicated to the purpose of long-term metadata curation.

2. Metadata Defined

In light of its acknowledged role in the organisation of and access to networked information and importance in long-term digital curation, metadata may be defined as structured and standardised information that is crafted specifically to describe a digital resource, in order to aid the intelligent and efficient discovery and retrieval of that source, accurate verification of its integrity (e.g. provenance tracking) as well as its apposite use and effective preservation over time. In the context of digital preservation, information about the technical processes associated with a data preservation technique is an example of metadata.

3. Digital Metadata Curation Defined

As mentioned earlier, a digital curation process needs to incorporate curation of metadata along with data in order to ensure an accurate description of data. It is therefore necessary to have an understanding of the underlying notion of the term "Digital or Data Curation" before attempting to define Metadata Curation. The phrase "Digital Curation" has different interpretations within different information domains. For example, in the museum domain, which is one of the oldest curation environments, data curation covers three core concepts – data conservation, data preservation and data access. Access to data or digital information in this sense may imply preserving data and making sure that the people to whom the data is relevant can locate it - that access is possible and useful. Another interpretation of the phrase "Data Curation" may be the active management of information, involving planning, where re-use of the data is the core requirement (Macdonald and Lord, 2002).

Therefore, from a generic standpoint, long-term data or digital curation can be defined as the continuous activity of managing, improving and enhancing the use of a digital object (i.e. data) as well as its preservation over its life cycle and over time for current and future generations of users. This is to ensure that the suitability of a data object is sustained for its intended purpose or range of purposes, and it is available with its quality and integrity intact for efficient discovery and apposite re-use over the long-term.

In light of the above construal of digital curation, the term "Metadata Curation" may be defined as an inherent part of a digital curation process for the continuous management (which involves creation and/or capturing as well as assuring overall integrity of metadata amongst other things) and preservation of metadata records over their life cycles. This is primarily to ensure the suitability of metadata for aiding the long-term curation of a digital resource that it refers to, by facilitating the intelligent and efficient discovery and retrieval of that resource, along with accurate verification of its integrity (e.g. provenance tracking), its apposite (re-) use and effective preservation over the long-term.

4. Principal Requirements of Effective Digital Metadata Curation

The efficacy of metadata curation significantly relies upon satisfying a number of requirements. Although metadata curation requirements may be quite different according to the type of data described, the information outlined below attempts to provide a general overview of the main requirements.

- ***Metadata Standards*:** The very first step to successful long-term data and metadata curation is to employ a curation-aware metadata standard(s) or format that provide(s) necessary elements to capture sufficient information about both a data object and its associated metadata. Examples of such information include Representation Information (RI), annotations made to both data and metadata, information about changes made to data and metadata, amongst other things. Of particular note is the Representation Information about a data object, which is defined as the information required to enable access to the preserved digital object in a meaningful way (OAIS, 2002). The use of RI can be recursive, especially in cases where meaningful interpretation of one RI element requires further RI. This recursion continues until there is sufficient information available to render a digital object in a form the user base can understand.
- ***Metadata Preservation:*** Metadata curation requires metadata to be preserved along with data in order to ensure its accurate description over time. Therefore, it is necessary to devise or use a suitable long-term metadata preservation strategy that is also flexible for addition of further requirements.

- ***Metadata Quality Assurance:*** A long-term curation process should effectively ensure well-structured metadata records with an accepted (by the community or domain concerned) level of quality, notwithstanding any forms of evolution or changes in related technology or metadata requirements of the organisation(s) concerned. In general, quality of a metadata record is measured by the degree of consistency with and/or accuracy in reference to the actual dataset and conformance to some agreed standard(s). Therefore, appropriate quality assurance procedures or mechanisms need to be in place to eliminate any quality flaws in a metadata record and thereby ensure its suitability for its intended purpose(s).
- ***Metadata Versioning*:** It is essential for a metadata curation system to be able to distinguish between metadata in different states, which arise and co-exist over time by suitably versioning metadata information.
- ***Metadata Annotation:*** Annotation is a widely practiced means of adding value to data as well as establishing collaborative links between data producers and users. An efficient long-term metadata curation strategy will, therefore, need to facilitate annotation of both data and metadata (by data consumers/users), preserve and curate annotations made over the long-term.
- ***Audit Trailing & Provenance Tracking***: A metadata curation process should ensure recording of information with the required granularity, and facilitate necessary means of tracking any significant changes (e.g. provenance change) to both data and metadata over their life cycles and determining their quality and integrity.
- ***Metadata Search-ability***: Metadata needs to remain available and searchable to the potential users of the data objects or resources that they describe in order to aid the appropriate use of those data objects or resources over time. Therefore, search-ability of metadata in convenient and efficient manners is regarded as a crucial factor in successful long-term metadata curation, hence a part of its remit.
- ***Metadata Policy:*** A set of broad, high-level principles that form the guiding framework within which the metadata curation can operate, must be defined.
- ***Access Constraints & Control:*** Appropriate security measures should be adopted to ensure that the metadata records have not been compromised by unauthorised sources, thereby ensuring overall consistency in the metadata records.

5. The OAIS Reference Model as a Framework for Long-term Metadata Curation System

In the complex realm of digital preservation and curation, the Reference Model for an Open Archival Information System (OAIS) (OAIS, 2002) is perhaps the only standardised effort that attempts to provide answers to virtually all questions related to long-term digital curation and preservation. Unsurprisingly, it has proliferated rapidly through the digital preservation community and has been explicitly adopted by, or at least informed, many prominent digital preservation initiatives (e.g. NEDLIB, CEDERS, etc.). However, a problem with the OAIS model is the variable detail of the answers that it provides. While the OAIS provides very unambiguous and in-depth answers to some questions, others are only touched upon, in some cases requiring one to even assume the answers. Of particular note is the considerable level of ambiguity in description of the metadata curation technique presented in the model.

When attempting to build a metadata curation system for an OAIS-compliant digital archive, one faces the problem of having very vague definitions of different metadata curatorial functions. In fact, in order to outline the functional requirements of a metadata curation system, it would be necessary to extract and in some cases assume the definitions of such functions from some generalised description of related preservation functions as described within the OAIS model. As an example, the Data Management entity (Figure 1) of the OAIS model can be considered. This (according to OAIS, 2002) aims to provide necessary functions for populating, maintaining and accessing Descriptive Information, i.e. metadata about the digital objects being preserved. Based on the definition of this entity, one would naturally infer that "maintaining" metadata essentially includes curating it. However, when it comes to implementing such an entity in a curation system, owing to very vague definitions in the OAIS, one would also need to make further assumptions to develop a complete list of concretely defined functional requirements of metadata curation (e.g. metadata versioning, validation and so on).

Furthermore, while the OAIS model provides a generic overview of the ingest function, it does not however describe how and at what stage syntactic and semantic validity of metadata should be checked during the ingest process. Also, the OAIS does not specify how one would go about ensuring interoperability and coherence of metadata irrespective of the formats that it is submitted in, i.e. if any form of translation or mapping between different formats would be required, how and at what stage it should be done.

In addition, the model provides a macro view in which information objects are migrated to a future technological platform as part of the preservation measure employed, but it does not mention anything pertinent to the fact that metadata itself may need to be migrated to newer formats, versions and platforms to ensure its longevity and usability. In terms of handling updates and changes to both data and metadata, the OAIS model does not seem to take into

account (at least not in direct terms) how the system would facilitate annotating both data and metadata as well as storing, searching and retrieving the annotations made to data and/or its metadata.

The answer to this predicament is that the OAIS model is part of the user requirements that is only intended to serve as the starting point for the development process for an OAIS archive. The OAIS specification, in fact, clearly states that it is not a guideline for an implementation or a design. Consequently, building a dedicated system for metadata curation needs further specific requirements to be added on to the OAIS model, as described in the following section.

6. The Metadata Curation Model

As underlined in the previous section, the OAIS model contains certain ambiguities about metadata curation related functions. Therefore, having to build a dedicated OAIS-compliant system for metadata curation needs further specific requirements to be added on to the OAIS model, especially to its "Data Management" entity, which is essentially responsible for managing (i.e. curating) metadata in an OAIS archive. In other words, a specialised edition of the OAIS model is required for the design and implementation of an efficient long-term metadata curation system. The Metadata Curation Model (MCM) attempts to fulfil that requirement by articulating the metadata curation related ambiguities of the OAIS model and refining its "Data Management" entity, and thus making it metadata curation focused. From this perspective, the MCM should be regarded as *an OAIS-based solution to long-term metadata curation*.

Furthermore, long-term metadata curation requires a model that is efficient and precise in reflecting all core requirements of metadata curation (see section 4) as well as being extensible and adaptable to incorporate any future requirements. The Metadata Curation Model presented in this section endeavours to accomplish these objectives.

6.1 Overview of the Model

The curatorial functions designed in the MCM include metadata ingest, metadata versioning, metadata quality assurance, annotation of data and metadata, preservation of metadata, access to (e.g. querying, searching) preserved metadata, migration of metadata to new formats and tracking provenance of metadata. In addition, the use of a curation-aware metadata format (see section 4) is also incorporated into the design of the model and is essential for efficient and optimal execution of its curatorial functions.

The model is only focused on the curation of metadata and does not assume the responsibility of curation of the data that the metadata describes. As Figure 2 illustrates (compared to Figure 1), the model can be seen and implemented as one of the functional entities of the OAIS reference model.

A brief description of each of the entities in Figure 2 is given as follows:

6.1.1 Data Ingest

The **Data Ingest** entity directly refers to the **Ingest** (Figure 1) entity of the original OAIS model, with one significant addition. In short, the entity provides functions to accept **Submission Information Packages (SIPs** – Note 1**)** from Producers, extracts metadata and its corresponding meta-metadata from the actual digital object and prepares them for preservation and curation. Metadata together with its corresponding meta-metadata constitute a **Metadata Submission Package (MSP)**, which is an addition to the original OAIS design of the Ingest entity. At this stage, meta-metadata in the MSP may contain various information about a corresponding metadata record, such as information about metadata creator(s), metadata publisher(s), metadata format (e.g. Dublin core), metadata provenance, existing annotations made to the metadata and so on. It is not necessary for meta-metadata to have further information as both metadata and its corresponding meta-metadata are assumed to be in the same format (e.g. XML, Text) and changes made to meta-metadata should only be reflective of changes made to the metadata, which it refers to. Therefore, curation mechanism(s) applied to metadata should also suffice for its corresponding meta-metadata without the need for any further information. The MCM requires a curation-aware metadata format (Section 4) as the underlying format of the metadata and meta-metadata in a MSP.

In addition, the Data Ingest module is responsible for assigning unique session identifier to each data/metadata submission request, thus enabling data objects and their corresponding metadata records to accurately reference each other from their respective entities (i.e. **Archival Storage** for data objects and **Metadata Curation** entity for metadata record), during the submission. This is particularly useful when a curation system is dealing with multiple data submission requests at the same time as it eliminates the risk of a metadata record referencing a data object that it does not describe or vice versa. The Data Ingest module could also have suitable means for checking or scanning submitted files for corruption and virus infection.

6.1.2 Archival Storage

As with the Data Ingest entity, the **Archival Storage** is also a direct reference to the Archival Storage entity of the original OAIS model. Functions within this entity include receiving digital objects from the data ingest module and adding them to permanent storage, managing the storage hierarchy and migrating preserved digital objects to new media or platforms.

6.1.3 Preservation & Curation Planning

As a revised version of the OAIS-defined **Preservation Planning** module, the **Preservation & Curation Planning** entity monitors and provides periodic recommendations for both data and metadata preservation to ensure they remain accessible to the User Base (Note 2) over the long-term. These recommendations cover preservation techniques, metadata standards, curation policy and so on. This entity is also responsible for developing detailed Migration plans, software prototypes and test-beds to enable implementation of successful migration of both data and metadata.

6.1.4 Metadata Curation

The **Metadata Curation** entity essentially represents the Metadata Curation Model within an OAIS system/archive by implementing a range of different functions to efficiently curate metadata over the long-term. This entity also elaborates the vaguely defined metadata curation related functionalities (Section 5) of the Data Management entity as outlined in the original OAIS model and presents a complete list of suitably defined functional requirements of metadata curation. Figure 3 takes a closer look at the Metadata Curation entity in Figure 2.

The **Metadata Ingest** entity in Figure 3 is essentially the passageway for metadata to curation and preservation. In short, the entity receives a Metadata Submission Package (MSP) or **Metadata Update Package** (MUP – Note 3) from the Data Ingest entity, isolates meta-metadata from metadata and finally forwards them both to the **Metadata Quality Assurance** (QA) entity for validation. The isolation of metadata from its corresponding meta-metadata at this stage is only essential if they both are not adhering to the same format. The MCM supports two generic ways in which metadata in an MSP can be captured and ingested into the system:

- As pre-created manually (i.e. by human), i.e. the SIP contains pre-created metadata files.
- Using a combination of automatic (i.e. extracted from the data object using tools external to the system) and manual (i.e. created at the time of the SIP submission through the submission interface, which could be a web form or a standalone tool provided by the curation system) metadata creation methods to ensure adequacy and accuracy of metadata and thereby minimising metadata creation costs and efforts.

A similar approach should be applicable to creation of meta-metadata and ingesting it into the system. Figure 4 illustrates the functions of the Metadata Ingest entity.

The primary task of the **Receive MSP/MUP** function (Figure 4) within the Metadata Ingest module is to receive Metadata Submission Packages from the Data Ingest entity or **Metadata Update Packages (MUP)** from the Administration entity (section 6.1.6) and put them forward for quality assurance in the Metadata Quality Assurance entity. On receipt and if necessary, an MSP or MUP is disassembled into its constituent metadata and meta-metadata, both of which are then fed into the QA entity for validation.

This function also receives the outcomes of quality assurance operations (returned by the QA entity) on metadata and meta-metadata and informs their source entity (i.e. Data Ingest in case of MSP and **Administration** for MUP) accordingly. In the case of a MSP, should either metadata or meta-metadata fail to pass any of the quality checks, i.e. if the QA returns negative results, a re-submission request for the MSP is sent to the Data Ingest entity, which then forwards the report to the Producer entity. In any case, a full report detailing the outcomes of different functions of the Metadata Ingest entity, such as MSP/MUP disintegration and quality assurance, is sent to the Administration entity. The report sent to the Administrator entity is also used (along with a relevant session identifier) by the Archival Storage (Section 6.1.2) entity to ensure that only data objects with valid metadata records are stored for curation.

The **Extract Meta-Metadata** function in effect assists the Receive MSP/MUP function in extracting meta-metadata from metadata in a MSP/MUP if necessary (see above). This function is also responsible for subjecting meta-metadata to different quality checks through the QA entity. In fact, the tasks of handling Metadata and its associated meta-metadata are isolated and allocated to Receive MSP/MUP and Extract Meta-Metadata function respectively for overall greater efficiency but could well just be handled by one function if design simplicity is desired.

The **Metadata Quality Assurance (QA)** entity (Figure 3) is responsible for ensuring overall quality and validity of submitted metadata. Figure 5 presents the functions of the QA entity.

The **Metadata Crosswalking** function (Figure 5) ensures that submitted metadata and meta-metadata conform to the format(s) supported by the curation environment. This essentially involves translating or transforming metadata in unsupported format(s) to format(s) that is/are supported. As there is always the danger of potential data loss in mapping between metadata in different formants, i.e. “metadata cross-walking”, the need for such an operation will largely depend on the related policy of the system. For example, if the system's policy was to support only one particular metadata format, there would be no need for any metadata crosswalk. However, it would also imply that any existing metadata (in non-supported format(s)) would need to be rewritten in the supported format in question before it could be accepted by the system, which might be deemed impractical where a large amount of metadata is involved. More importantly, it would result in sacrificing interoperability with other related curation systems and metadata

formats.

A solution to this problem would be to maintain suitably formulated pre-created mapping rules for the metadata translation within the repository in order to minimise data loss. However, such a solution would also imply that every time any of the supported metadata formats was to undergo any significant changes, the mapping rules for that format would need to be updated, which would essentially require detailed examination of the new version of the format in question to ensure accuracy. In other words, the archiving or curatorial body concerned would need to have a dedicated team who would be responsible for monitoring changes in supported metadata formats and calculating/updating mapping rules. While this should not be a problem if the body in question has the necessary financial and technical resources to facilitate this over the long-term, an organisation with comparatively lower or limited curation budget might not find it cost-effective.

Nevertheless, while successfully transformed metadata and meta-metadata are passed on to the **Structural Validation** function, failure in metadata cross-walking results in both metadata and meta-metadata being discarded. In both cases, the final outcome of the transformation is reported to the **Generate Quality Assurance (QA) Result** function, which then re-directs it to the Metadata Ingest (Figure 4) entity. Metadata and meta-metadata in supported format(s), however, bypass this function and go straight to the Structural Validation function.

The **Structural Validation** function (Figure 5) checks syntactical or structural validity of metadata records (and associated meta-metadata) against the corresponding metadata format(s). Ideally, this function should be fully automated. This function also sends a report to the Generate QA Result indicating the outcome of the validation. While structurally invalid metadata and meta-metadata are discarded, structurally valid metadata records are forwarded to the Semantic Validation function.

The **Semantic Validation** function (Figure 5) facilitates semi-automatic ways of checking whether the values assigned to the elements in structurally valid metadata records comply with the actual content of the data object. This function could also include metadata cleansing in order to remove any noise or anomalies in metadata records (e.g. correction of spelling mistakes, grammatical errors) and thereby maintain the desired level of consistency across all metadata records being preserved. This function is also expected to make use of some controlled vocabulary (Note 4), if applicable, in order to check semantic validity of values in metadata records. Ideally, a curation system would maintain a controlled vocabulary server for the system's principal metadata format. For other supported metadata formats, however, the system could maintain a database of information (e.g. server URL, port number, etc.) required to connect to and use the appropriate vocabulary server. Alternatively, the users submitting metadata records could be provided with the facility for specifying such information at the time of submission.

In case of extremely erroneous and inconsistent metadata or meta-metadata records, which fail semantic validation, the function will be forced to discard both metadata and its corresponding meta-metadata and (as with the Structural Validation function) send an appropriate report to the Generate QA Result function. A semantically valid metadata record makes its way to the **Record Quality Assurance Event Info**, where associated meta-metadata is updated with information about different QA operations that the metadata has been subjected to.

The **Generate Quality Assurance (QA) Result** function (Figure 5) collates information about the outcomes of different QA processes, such as cross-walking, structural validation and semantic validation, generates report based on it and sends the report to the Metadata Ingest entity (Figure 4). The information collected by this function is also used as the QA Event Info (Note 5), which is recorded in the meta-metadata associated with the metadata records.

The **Record Quality Assurance Event Info** function records information about different QA processes (e.g. cross-walking) that metadata is subjected to, in its corresponding meta-metadata. QA Event info includes description of a process, changes made to metadata, tools used, date and time of the occurrence of the process and so on. QA Event info is essentially obtained from **Generate QA Result** function. Updated Meta-metadata is forwarded to the **Generate Metadata Versioning Package** function.

The **Generate Metadata Versioning Package** is the final QA stage for both metadata and meta-metadata before they are ingested into the Versioning entity (Figure 6). For successfully validated (and cross-walked if necessary) metadata and associated meta-metadata, this function obtains Representation Information (RI) for both the digital object (if it is not already included in the metadata) and its corresponding metadata from a trusted Representation Information Registry and updates both the metadata record (s) and its associated meta-metadata respectively with it. The task of acquiring RI for a digital object could also be performed at the time of, or before, data ingest and more practically before the metadata ingest as that is when file format and/or rendering software related information is computed (ideally using a suitable tool/software, such as DROID – Note 6) for the object. File format and software related information (e.g. extension name or rendering software name) is normally what is used by RI repositories to determine RI for digital objects. An example of such RI repository is the PRONOM technical registry (PRONOM, 2007). The use of a trusted repository for RI ensures authenticity and accuracy of the RI obtained, which in the long run ensures accurate

interpretation and use of the digital object in question.

Successfully validated and updated (with RI and QA event info) metadata records and meta-metadata collectively form a **Metadata Versioning Package (MVP)**, which is then forwarded to the **Metadata Versioning** entity. The structure of an MVP at this stage should be similar to that of an MSP.

The **Metadata Versioning** module as depicted in Figure 3 is responsible for assigning unique version numbers to metadata records (both newly submitted and updated versions of existing records) to represent their states at particular times. Figure 6 pictorially presents different functions of the Metadata Versioning Entity.

The **Receive MVP** function (Figure 6) accepts MVPs from the Quality Assurance function. For MVPs consisting of separate files containing metadata and meta-metadata, this function feeds the file containing metadata into the **Process Metadata Versioning** function, while the associated meta-metadata file moves across to the **Record Versioning Info** function and waits for the metadata to be versioned. An MVP consisting of a single file containing both metadata and meta-metadata is sent directly to the Process Metadata Versioning function.

The **Process Metadata Versioning** function (Figure 6) performs a version check on the metadata received from the Receive MVP function. In the case of a modified instance of an existing metadata record, this entails assigning a unique version identifier to the edited record as well as establishing and updating relationships between this version and other co-existing versions in the database.

In effect, the **Process Metadata Versioning** function performs the versioning task in collaboration with the **Assign Version Number** function. In case of a failure in accurately assigning version identifiers to metadata records, a failure report is sent to the Administration entity. Successfully versioned metadata records, however, move on to the **Generate Metadata Preservation Package** function (Figure 6).

The **Record Versioning Info** function (Figure 6) receives the newly assigned version number for a metadata record in transition and adds it to the meta-metadata of the record. Updated Meta-Metadata is then forwarded to the Generate Metadata Preservation Package function.

The **Generate Metadata Preservation Package** function (Figure 6) begins with accepting a metadata record and its corresponding meta-metadata from the Assign Version Number and the Record Versioning Info functions respectively. Subsequently, it creates **Metadata Preservation Package (MPP)** with updated metadata records and its meta-metadata, which is then forwarded to the **Metadata Management** entity for preservation.

The **Metadata Management** entity in Figure 3 can be regarded as the heart of the curation model as it is responsible for satisfying perhaps the most significant requirement of long-term metadata curation - the actual preservation and management of metadata. This entity is essentially responsible for executing the final phase of metadata's journey from ingest to storage. Figure 7 represents the functions of the Metadata Management entity.

The **Receive MPP** function (Figure 7) is primarily responsible for storing metadata records and associated meta-metadata in the database. This function begins with acquiring an appropriate unique storage identifier (or reference information) and version history (particularly important for updated data objects) for the data object that a MPP (received from the Metadata Versioning entity) refers to, from the Administration entity. The acquired data object identifier and version history are attached to the metadata record in the MPP (elaborated in the following paragraph) at a later stage. During a digital curation process, a metadata record may be required to provide accurate reference to or accurately identify the particular version of a data object that it describes, especially when queried by a Consumer. Ideally, this is facilitated by assigning automatically generated unique identifier(s) or reference(s) to valid data objects (i.e. the ones that have passed the necessary validation steps) and attaching the identifier(s) to their corresponding metadata records before they are stored in the designated storage media in the Archival Storage (Figure 2) and the Metadata Management entities respectively. This method of uniquely identifying data objects in a curation system is particularly useful for enabling search engines that execute user-submitted queries for (a) specific data object(s) against their metadata records, to accurately link each metadata record (returned in a search result) to the particular version of a data object that it is associated with. On the other hand, information about the version history of a data object is required to track changes and provenance of that object. Figure 8 provides an overview of the workflow between the Data and Metadata storing functions of a curation system. It should be noted that Figure 8 only presents the primary functions responsible for storing data and metadata and assumes the incorporation of other (i.e. intermediate) functions and/or entities (e.g. Metadata Validation) that the primary functions may depend on, in the system.

In general, after a data object has been successfully stored in the Archival Storage, its storage identifier is passed on to the Administration entity, which then stores it in the relevant data/metadata submission session. In addition, the function responsible for storing the data object also generates a detailed version history of the data object, which (i.e. version history) is also forwarded to the Administration entity to be stored in the relevant submission session. Conversely, for data objects that fail to validate and/or to be stored, the session contains a failure report. Therefore,

acquisition of the data object identifier for a MPP (that corresponds to the data object) is achieved by querying the Administration entity based on the relevant session identifier for the data/metadata submission. For invalid data objects, the Receive MPP function terminates by sending a failure report to the **Generate Report** function (Figure 7), which subsequently forwards the report to the Administration entity. Depending on the related policy of a curation system, a MPP at this stage may either be removed from the system or held temporarily until the Producer re-submits a valid data object for the MPP or the session expires or until a certain pre-defined period of time, whichever is the earliest.

In the case of a valid data object, the Receive MPP function acquires the corresponding identifier and version information of the data object from the Administration entity and then disintegrates the MPP into constituent metadata and meta-metadata. The metadata is subsequently stored along with its corresponding data object identifier and its version history in the database, while the meta-metadata is stored with the metadata versioning info attached to it during the versioning process in the Metadata Versioning entity. From the implementation perspective, the data object identifier and metadata versioning info could be stored in their corresponding columns of the metadata record table and meta-metadata table in the database respectively. The data object identifier and metadata versioning information are mainly used by the **Access** entity (see section 6.1.5) to identify and provide access to an appropriate data object and metadata record respectively, when a Consumer (Note 7) queries for the respective objects. The format of the database can be any known database format, such as relational, XML, and object oriented, whichever is deemed suitable by the curatorial organisation or body.

The **Administer Database** function (Figure 7) is responsible for creating and updating schema or table definition of the metadata database as well as any other database administration related task(s) as required. More importantly, this function performs migration of metadata with the help of the **Metadata Migration** function in order to keep pace with changes in related technology and formats. This function also conducts periodic checking of metadata in collaboration with the **Periodic Quality Assurance (QA)** function. This function entails periodically evaluating metadata to ensure its quality for intended purpose or a range of purposes and updating the metadata (if required) based on the outcome of the evaluation. This function is carried out in accordance with the curation policy imposed by the Administration entity.

Updated metadata resulting from Periodic QA or the Metadata Migration function is fed into the **Generate Metadata Update Package (MUP)** function (Figure 7), which retrieves corresponding meta-metadata from the database and uses them both to generate Metadata Update Package. Generated MUPs are fed into Metadata Ingest entity to be eventually stored in the Metadata Management entity.

The **Perform Queries** function (Figure 7) receives queries about metadata stored in the database from a Consumer via the Access entity, searches the database based on the queries and returns the result set to the Access entity, where the result set is presented to the Consumer.

The **Generate Report** function receives reports from the Administer Database function about different activities that it conducts, such as Database Updates, Periodic QA and Metadata Migration. These reports are sent to the Administration entity for reviewing and assessment purposes. This function is also responsible for notifying a Producer via the Administration entity (Figure 7 and 8) of a success or failure of storage of metadata and meta-metadata extracted from a Metadata Preservation Package that was received (by the Receive MPP function) from the Metadata Versioning entity. In addition, the Generate Report function responds to report requests from the Administration entity about other processes or functions of the Metadata Management entity.

6.1.5 Access

The **Access** entity is another adaptation of an OAIS defined module - the "Access" module in this case. This entity has been re-designed for the curation model with a view to reflect the role that metadata plays in searching and retrieving digital objects (that it refers to) under preservation in the Archival Storage. In effect, this entity is responsible for facilitating search-ability and tracking provenance of metadata that are core requirements of long-term metadata curation

6.1.6 Administration

The **Administration** entity is an adaptation of the Administration entity of the OAIS model (OAIS, 2002), with a number of added features, such as receiving metadata updates and annotations made to either data or metadata in the form of **Annotation Submission Packages (ASPs –** Note 8**)**, dealing with errors in metadata and digital objects reported by the Consumer; and generating **Metadata Update Package (MUP** – Note 9**)** (Figure 3) for curation and preservation. In effect, it is the Administration entity through which the MCM facilitates annotation of both data and metadata – a core requirement of long-term metadata curation. Of particular note is the approach employed by this entity (and by the MCM as a whole) for curating annotation that allows annotation to be made to both digital objects and its corresponding metadata as an external entity and treats it in isolation as part of existing metadata records of the object and its meta-metadata respectively. The advantage of this approach over the one that allows annotation to be

embedded or attached in the actual data object or metadata is that the former does not cause any violation of the edits related legal rights (e.g. Copyright) associated with the digital object while retaining the ability of the latter to make the annotation available to the consumer in a convenient manner. Typically, the users of the system would be provided with an annotation interface, which would allow them to select any particular context(s) of the digital object or the metadata record of their interest and add annotation(s) to that context(s). The interface would also facilitate searching, displaying and editing annotations made to the digital objects under preservation.

6.2 Applicability of the Model

The Metadata Curation Model may be applicable to any OAIS-based information preservation system or archive as well as any long-term data curation system, where metadata is preserved separately to the actual digital resource that it is associated with. In general, the model is applicable to any organisation that is responsible for making digital resources available over the long-term and actively acknowledges the role that metadata can play in efficiently fulfilling that responsibility. Moreover, the MCM is equally applicable to both the organisations that are looking to build new curation systems, and those aiming to incorporate curation-related functions into their existing non-curation focused systems. This is facilitated by the modular architecture of the MCM that enables it or any of its entities to be easily integrated into any existing metadata system to make it curation-aware. In addition, the model (or any of its components) may be extended or customised to incorporate domain/system specific functions and accommodate future curation requirements. The case-study below illustrates potential use of the Metadata Curation Model in the Science and Technology Facilities Council (STFC, 2007) data portal (Note 10).

The STFC operates for the UK research community several large scale scientific facilities that all generate large quantities of data. While the STFC provides a common way for discovering and accessing these multi-disciplinary data through a web-based data portal, there is currently no comprehensive measure in place to curate and preserve these data over the long-term. Without proper and efficient curation and preservation, these data could potentially become obsolete due to fast changing technologies and data formats. Therefore, considering the current status and increasingly large volumes of data managed, the STFC could benefit from an efficient long-term curation system and hence makes an ideal use case for the Metadata Curation Model.

A close inspection of the STFC's data management architecture (Figure 9) suggests that it would not be too difficult, at least in theory, to implement an efficient curation system for the STFC data.

The present data management architecture enables users to manage (e.g. edit, store) their data files on file servers at Cambridge and London through the central Storage Resource Broker (SRB) software and database at STFC (Blanshard, Tyer, Calleja, Kleese and Dove, 2004). The architecture also facilitates (via the Metadata Editor) creation of metadata about the data to make it discoverable to the users via the data portal. In order to transform the present architecture into a long-term data curation focused architecture, the first step would be ensuring availability of adequate, appropriate and good quality metadata about the data. This would require appending necessary metadata curation elements to the currently employed metadata format, i.e. the STFC Scientific Metadata Model (SSMM), which currently lacks the ability to record sufficient information (e.g. data/metadata provenance, Representation Information, meta-metadata) required for efficient long-term curation. Modification to the metadata format would in turn require the modification to the metadata database schema, which is based on the SSMM format.

The next step would involve the implementation of the Metadata Curation Model, which would incorporate the features of the data portal and the metadata editor as well as other curation features, such as provenance tracking and data/metadata annotation amongst other things. Therefore, a revised version of the data management architecture (Figure 10) would replace the STFC data portal and the metadata editor with the metadata curation component as it supersedes them. The implementation of the metadata curation model would also require implementation and/or employment of other services, such as a Representation Information repository and controlled vocabulary server.

The final and most challenging step would be developing a long-term preservation archive for the data. This step would require an in-depth assessment of the SRB and existing data storage mechanisms to determine whether it would be feasible (in terms of costs and effort required) to extend them to incorporate long-term preservation features or they would need to be replaced with more suitable technologies (therefore marked with "?" sign in Figure 10).

Moreover, in order to evaluate and demonstrate the underlying concepts of the MCM, a web-services based prototype system has been developed. The prototype system is available online at http://www.metadata-curation.co.uk

7. Conclusions

Efficient and effective long-term metadata curation is a key component of successful preservation, apposite enrichment and sustained accessibility of digital information in the long term. Unfortunately, no comprehensive method for effective curation of metadata for long periods of time is known to exist till date. The Metadata Curation Model aims to meet the necessity of an efficient metadata curation approach by combining the best features of existing long-term digital preservation strategies (i.e. the OAIS model) with a considerable degree of innovation. However, there is still a

great deal of scope for further advancement, as the suitability and efficiency of the MCM may only be accurately measured when implemented and tested in a fully-operational long-term digital curation system. Nevertheless, the approach presented in this paper may be regarded as a conceptually complete and scalable solution for long-term metadata curation that would benefit any discipline concerned with long-term data curation.

References

Blanshard, L., Tyer1, R., Calleja, M., Kleese, K. and Dove, M.T. (2004). *Environmental Molecular Processes: Management of Simulation Data and Annotation*, Proceedings of the UK e-Science All Hands Meeting 2004, © EPSRC Sept 2004, ISBN 1-904425-21-6, [Online] Available: http://archive.niees.ac.uk/documents/AHM_dataman_2004.pdf

CEDARS, (2002). CEDARS Project. [Online], 2002, Available: http://www.leeds.ac.uk/cedars/index.html (November 4, 2007)

Macdonald, A. and Lord, P. (2002). *Digital Data Curation Task Force Report of the Task Force Strategy Discussion Day*, November 2002, [Online] Available: http://www.jisc.ac.uk/uploaded_documents/CurationTaskForceFinal1 (January 15, 2008)

NEDLIB, (2000). Networked European Deposite Library. 2000 [Online] Available: http://nedlib.kb.nl/ (November 4, 2007)

OAIS, (2002). *Reference Model for an Open Archival Information System (OAIS)*, CCSDS Blue Book. Issue 1. January 2002, [Online] Available: http://public.ccsds.org/publications/archive/650x0b1.pdf (January 3, 2008)

PRONOM, (2007). The technical Registry PRONOM, The National Archive, 2007, [Online] Available: http://www.nationalarchives.gov.uk/pronom/ (January 17, 2008)

SRB, (2007). Storage Resource Broker, 2007 [Online] Available: http://www.sdsc.edu/srb/index.php/Main_Page (January 20. 2008)

STFC, (2007). The Science & Technology Facilities Council (STFC), 2007, [Online] Available: http://www.scitech.ac.uk (January 20, 2008)

Notes

Note 1. A SIP contains three objects - data to be preserved, its associated metadata and information about the metadata itself, i.e. meta-metadata.

Note 2. The term “User Base” encompasses all identified potential consumers (e.g. human, software application etc.) to whom curated metadata is beneficial in terms of accurate interpretation and proper utilisation of the digital object that the metadata describes and/or refers to. The User Base is essentially an adaptation of the Designated Community as defined in the OAIS reference model (OAIS, 2002).

Note 3. A Metadata Update Package consists of existing metadata records and their corresponding meta-metadata with any significant changes made to them. Changes to existing metadata records occur due to amendments submitted by producer and/or as a result of different curation related activities, such as metadata migration.

Note 4. A standardised and structured list of pre-defined values for different elements within a metadata record that conforms to some agreed standard(s). These pre-defined values also represent true knowledge organisation schemes that define the metadata concept, specify the scope and the relationships among the concepts.

Note 5. Information regarding different quality assurance functions or processes within a curation system, such as metadata crosswalking, structural validation and semantic validation, that a metadata record has to pass through before it is declared valid. This information includes time of the function execution, changes it makes to the record, tools used and so on.

Note 6. DROID (Digital Record Object Identification) is an automatic file format identification tool developed by the National Archives, UK- http://droid.sourceforge.net/wiki/index.php/Introduction (4 November 2007).

Note 7. The role played by those persons or client systems that find preserved information of interest and access that information in detail (OAIS, 2002).

Note 8. An Annotation Submission Package is comprised of annotation made to a digital object and metadata about the annotation, e.g. name and affiliation of annotation, date annotation made, part of the digital object it refers to, type of annotation and so on.

Note 9. A Metadata Update Package consists of existing metadata records and their corresponding meta-metadata with any significant changes made to them. Changes to existing metadata records occur due to amendments submitted by producer and/or as a result of different curation related activities, such as metadata migration.

Note 10. STFC Data Portal - http://tiber.dl.ac.uk:8080/

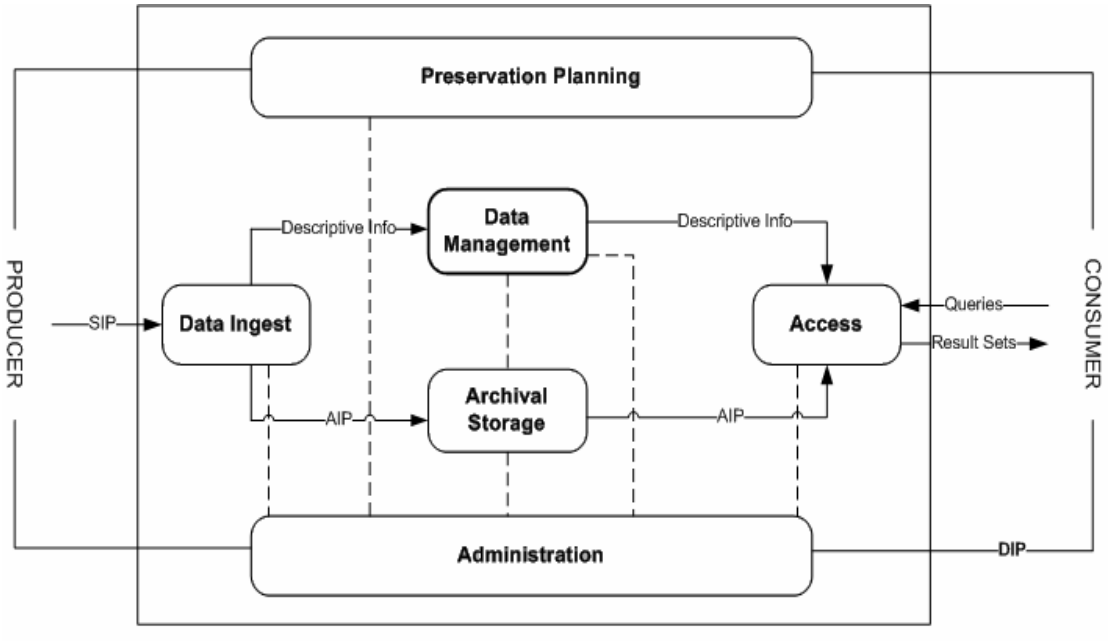


Figure 1. Functional Entities of the OAIS Reference Model [1]

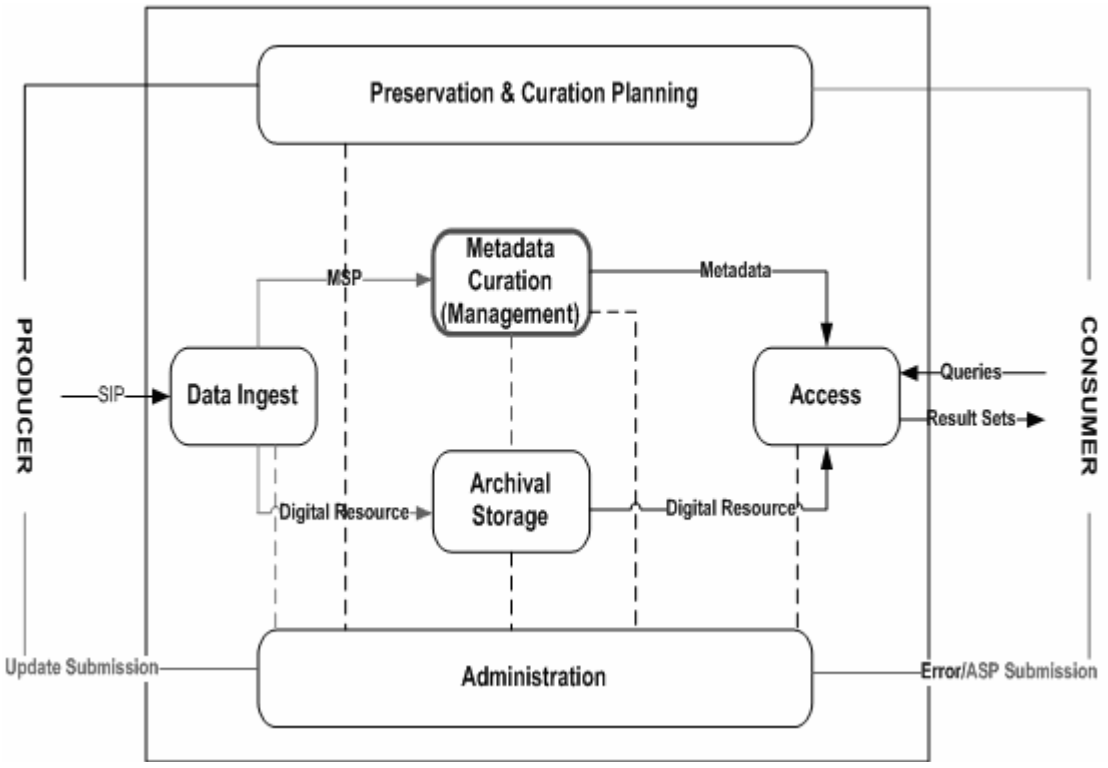


Figure 2. The Metadata Curation Model embedded in the OAIS Reference Model (highlighted in red)

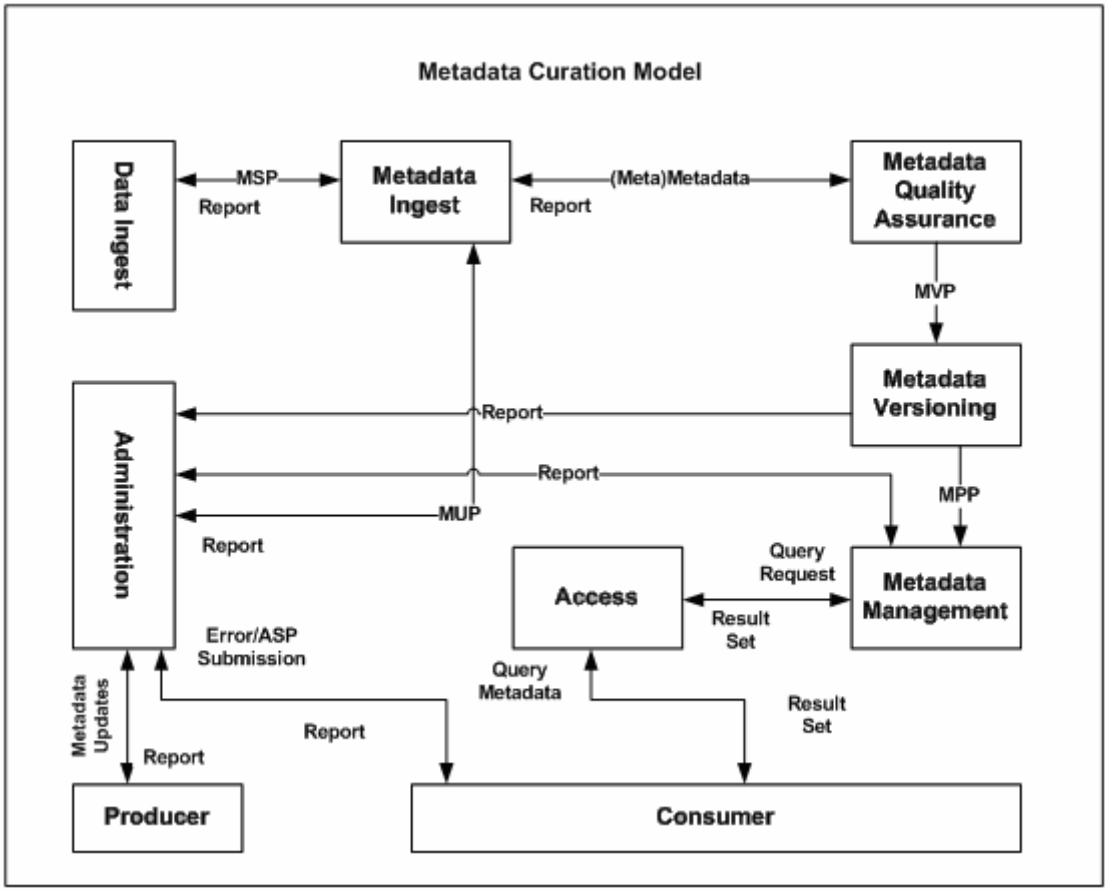


Figure 3. Functional Entities of the Metadata Curation Model

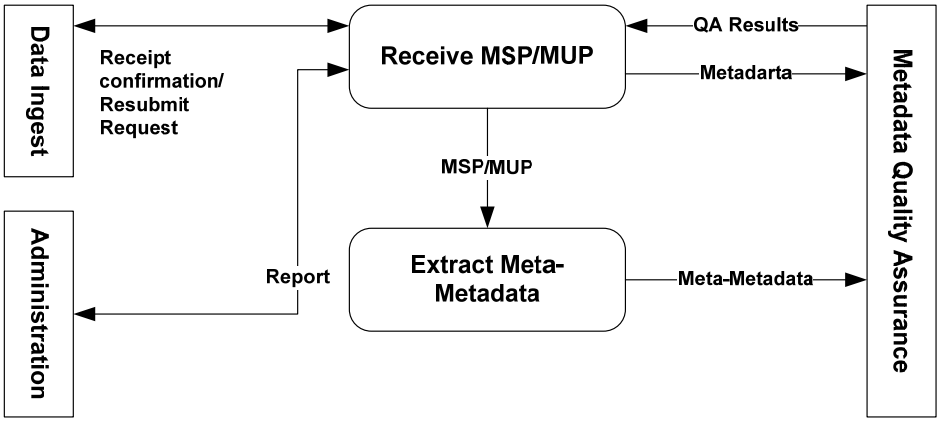


Figure 4. Functions of the Metadata Ingest Entity

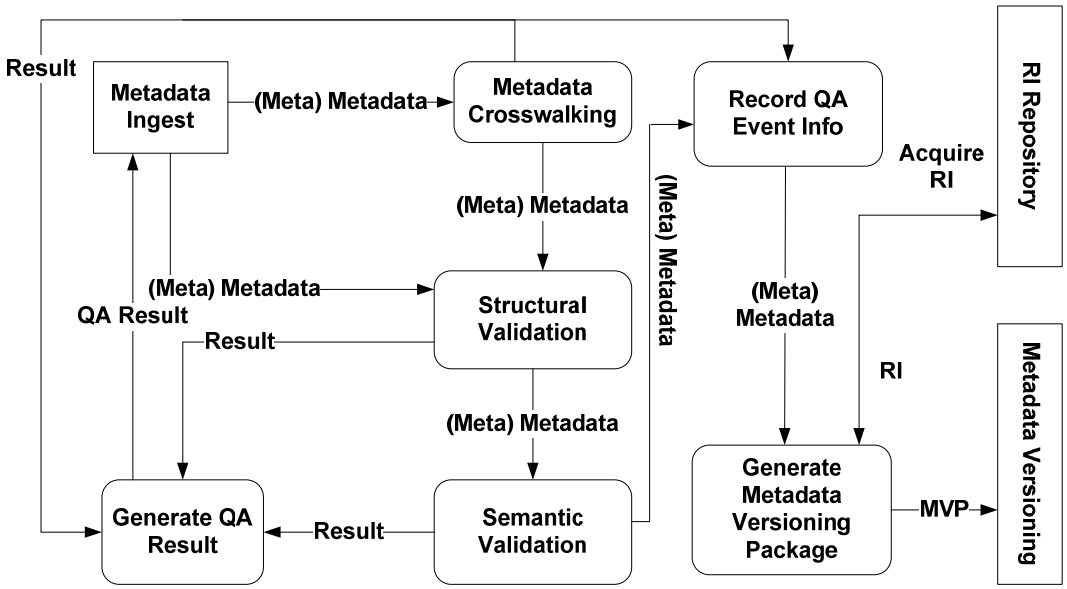


Figure 5. Functions of the Metadata Quality Assurance Entity

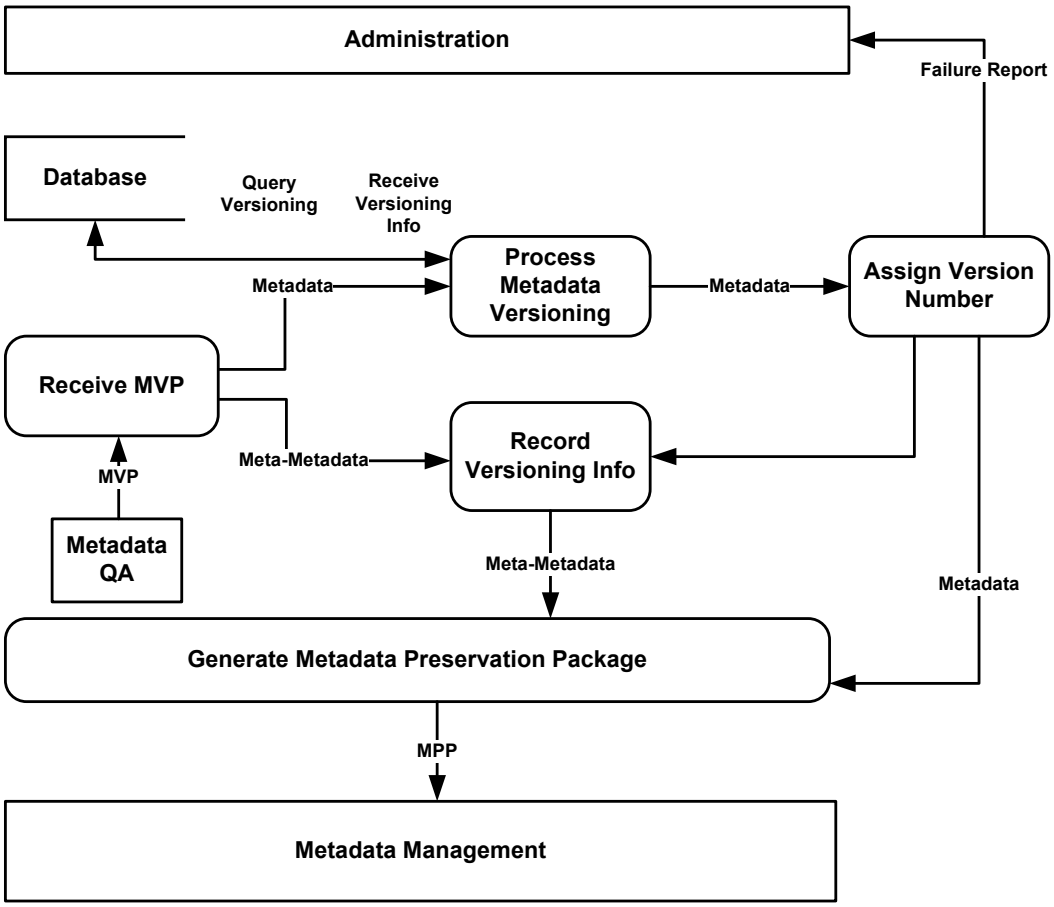


Figure 6. Functions of the Metadata Versioning Entity

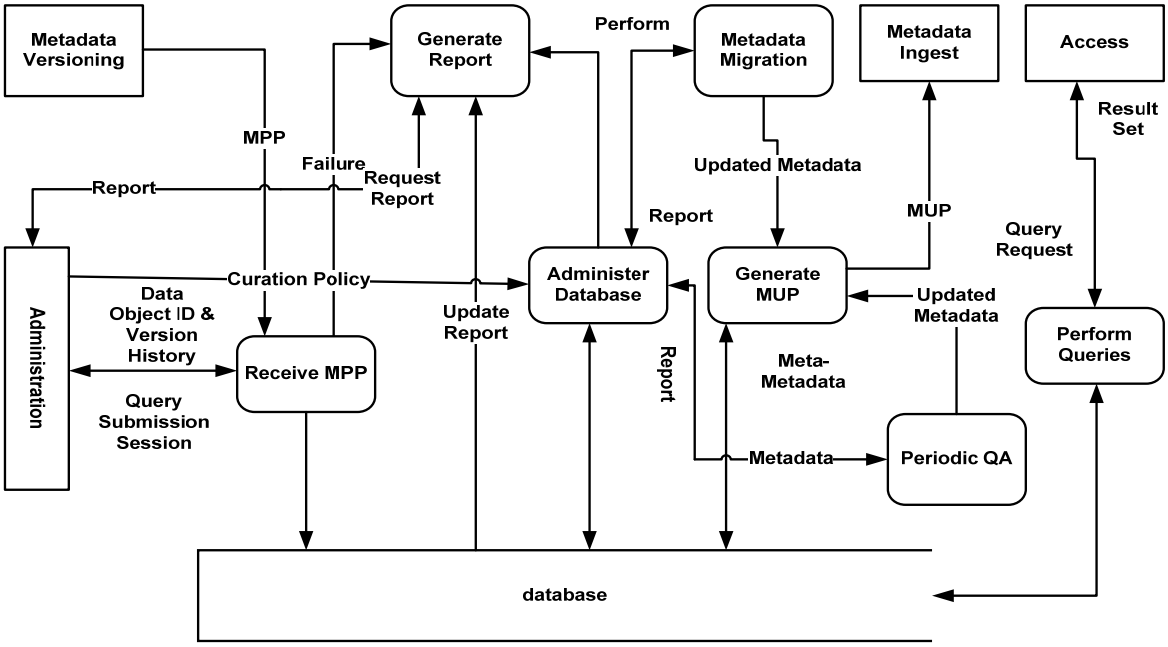


Figure 7. Functions of the Metadata Management Entity

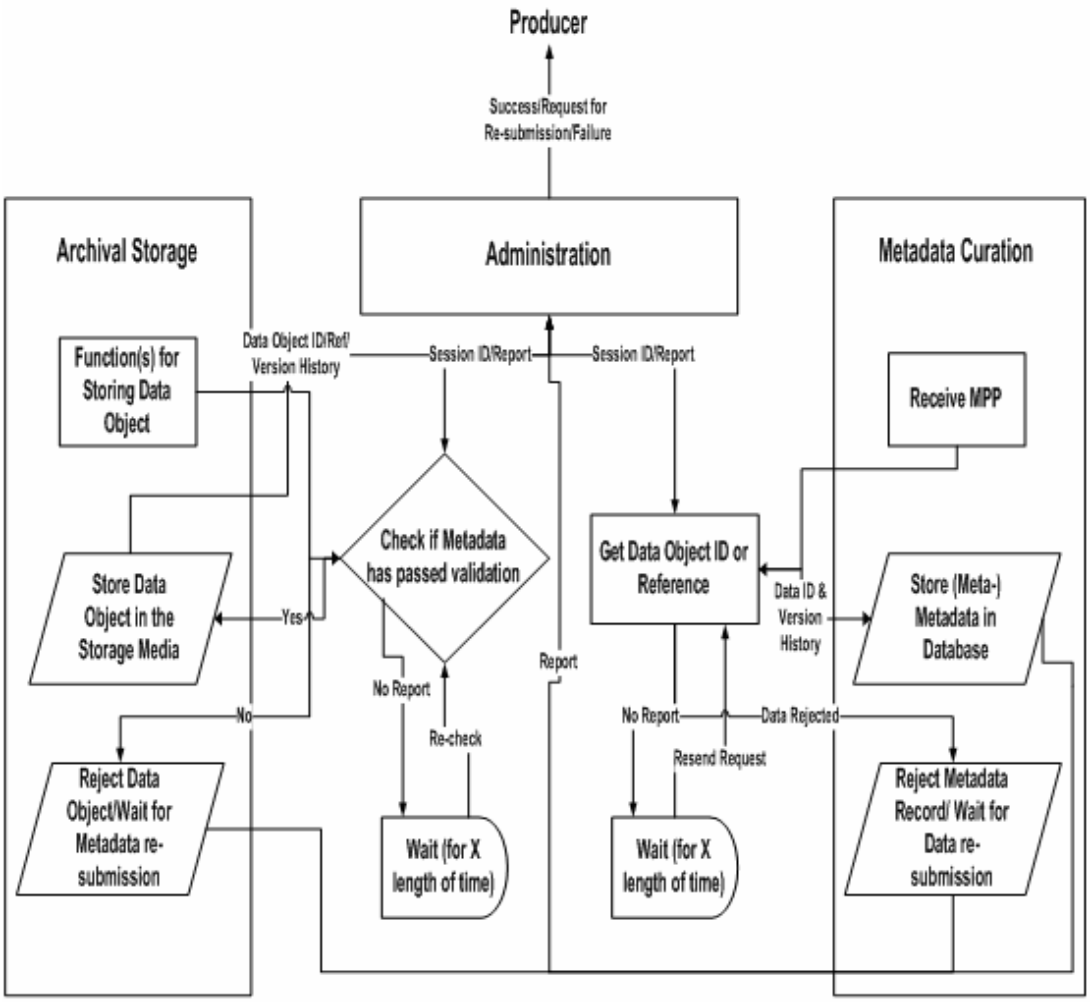


Figure 8. An Overview of the Data and Metadata Storing Process in a Curation System

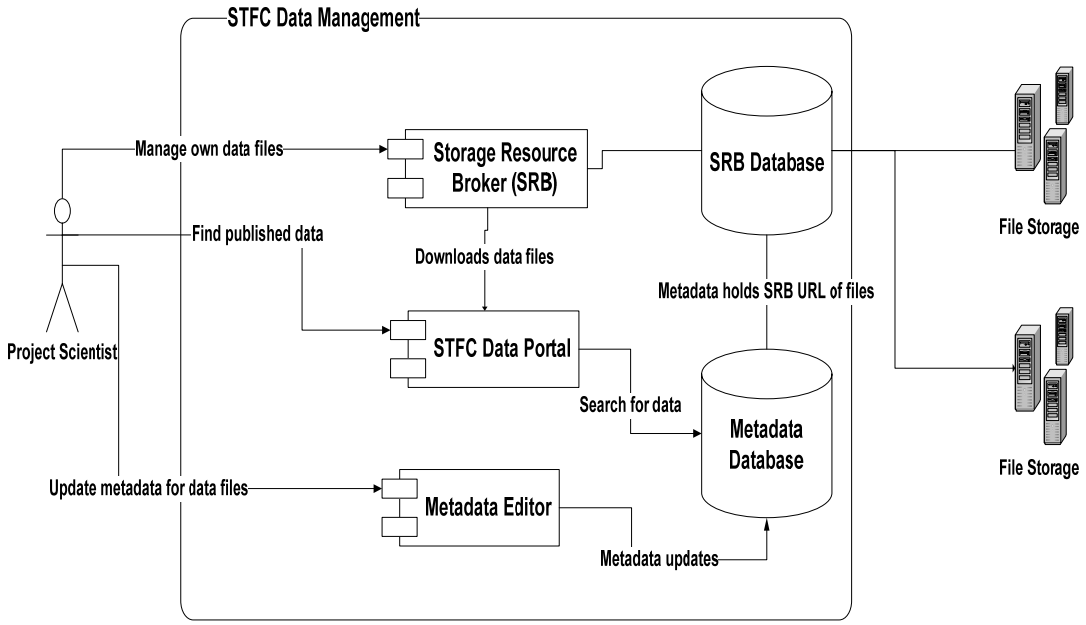


Figure 9. The STFC Data Management Architecture (Source: Blanshard, Tyer, Calleja, Kleese and Dove, 2004)

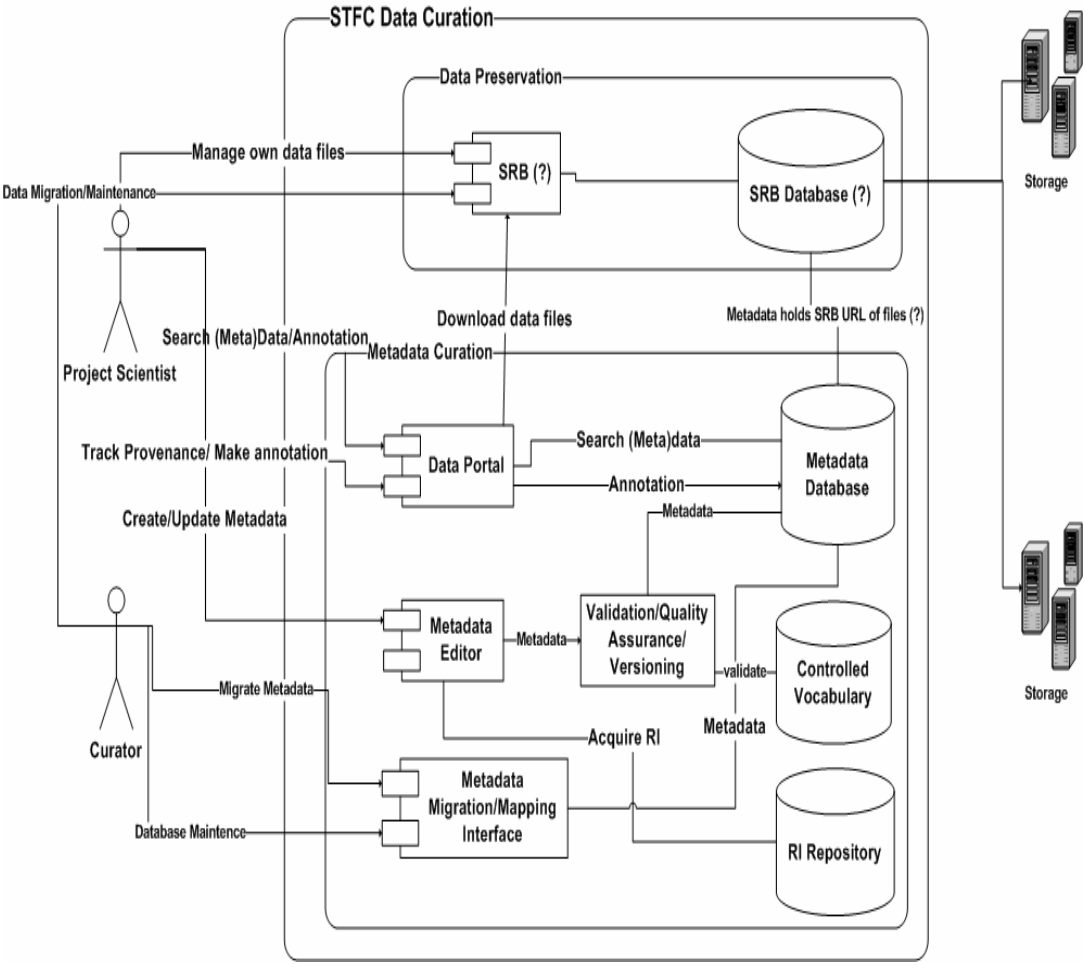


Figure 10. A Revised Version of the STFC Data Management Architecture with the Long-term Curation Features

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

Research on Decision Forest Learning Algorithm

Limin Wang

College of Computer Science and Technology

Jilin University

Changchun 130012, China

E-mail:wanglim@jlu.edu.cn

Xiongfei Li

College of Computer Science and Technology

Jilin University

Changchun 130012, China

E-mail: lxf@jlu.edu.cn

Abstract

Decision Forests are investigated for their ability to provide insight into the confidence associated with each prediction, the ensembles increase predictive accuracy over the individual decision tree model established. This paper proposed a novel "bottom-top" (BT) searching strategy to learn tree structure by combining different branches with the same root, and new branches can be created to overcome overfitting phenomenon.

Keywords: Decision Forest, BT, Overfitting

1. Introduction

Decision tree based methods of supervised learning represent one of the most popular approaches within the AI field for dealing with classification problems. They have been widely used for years in many domains such as pattern recognition, data mining, signal processing, etc. The overfitting phenomenon is a persistent problem in using decision trees for classification. In existing decision tree based classification approaches a fully trained tree is often pruned to improve the generalization accuracy even if the error rate on the training data increases (Krzysztof, 2002). In the last decade multiple approaches have been studied to overcome this problem. Promising results were achieved using ensembles of multiple classifiers, which is under the assumption that "two (or more) heads are better than one." The decisions of multiple hypotheses are combined in ensemble learning to produce more accurate results. This type of learning algorithms are called Decision Forests include Random Forests (RFs) (Lariviere et al, 2005), Random Split Trees (RSs) (Breiman, 2001), and Bootstrap Aggregating (Bagging) (Pino-Mejias et al, 2008). These trees can be formed by various methods (or by one method, but with various parameters of work), by different sub-samples of observations over one and the same phenomenon, by use of different characteristics. This strategy is based on the observation that, decision tree classifiers can vary substantially when a small number of training samples are added or deleted from the training set.

With the ability of Decision Forests to increase predictive accuracy over the individual model established, the ensembles are investigated for their ability to provide insight into the confidence associated with each prediction. Test samples invariably include new variation that was not in the training set and the ability of the model to accurately predict these samples may be limited. Decision tree inducers are unstable in that resultant trees are sensitive to minor perturbations in the training data set. Largely for this reason, decision trees are widely applied as the base learners in ensemble learning schemes. Some ensemble algorithms have implemented modified decision tree inference algorithms in order to generate diverse decision forests.

2. Details of Some Related Ensemble Schemes

Bootstrap aggregation (Bagging) is a common way of introducing variation into individual trees in a forest. When unstable learning algorithms are applied as base inducers, a diverse ensemble can be generated by feeding the base learner with training sets re-sampled from the original training set. Each decision tree in a bagged decision forest is generated from a bootstrapped sample of the original training dataset. The subsample is formed by random independent selection of samples from initial sample. The probability of selection is identical to each sample. The volume of

sub-sample is set beforehand. After the construction of a tree by way of analysis of given sub-sample, the selected observations return into initial sample and the process repeats the given number of times. This process generates a forest where each tree has been trained on a slightly different dataset which hopefully reduces the trees tendency to overfit the training set.

AdaBoost algorithm is referred to as an arcing (adaptive re-sampling and combining) algorithm. AdaBoost iteratively alters the probability over instances in the training set while performing the re-sampling. It works very well when the data is noise free and the number of training data is large. But during construction of the second tree, more attention is given to those objects which have a bigger error, with the purpose to reduce it. When noise is present in the training sets, or the number of training data is limited, AdaBoost does not perform as well as Bagging. The fundamental difference between bagging and AdaBoost is that while bagging is non-deterministic, AdaBoost is deterministic and iterative.

Random forests are a recent addition to the set of available decision forest methods and essentially extend bagging to include bagging the columns (variables) of the data matrix as well as the rows (samples) of the data matrix. Each decision tree in a random forest is trained on a distinct and random data set re-sampled from the original training set, using the same procedure as bagging. This additional randomization step generates significantly more diversity in the forest and additionally provides a significant speed improvement in tree construction time as compared to a Bagged model. While selecting a split at each internal node during the tree growing process, a random set of features is formed by either choosing a subset of input variables or constructing a small group of variables formed by linear combinations of input variables. Random forests have achieved "right"/"wrong" predictive accuracy comparable to that of AdaBoost and much better results on noisy data sets. Breiman also claimed and showed that AdaBoost is a form of random forest (algorithm).

3. Construction of Forests

We proposed a new approach of forest generation based on "bottom-top" (BT) searching strategy. In each stage of BT searching procedure, a branch is found to best fit current test sample, so these branches with the same root can construct a decision tree and more complex structures are created. The proposed method can help to build the shortest tree with maximal accuracy possible and thus, the tree does not need to be pruned to obtain good generalization.

Traditionally, attribute selection measure $Gain(T, A_k)$ generally based on information theory, serves as a criterion in choosing among a list of candidate attributes at each decision node, the attribute that generates partitions where samples are distributed less randomly, with the aim of constructing the smallest tree among those consistent with the data.

$$Gain(T, A_k) = E(T) - E_{A_k}(T) \tag{1}$$

where

$$E(T) = -\sum_{i=1}^{n} \frac{n(C_i, T)}{|T|} \log_2 \frac{n(C_i, T)}{|T|} \tag{2}$$

and

$$E_{A_k}(T) = \sum_{v \in D(A_k)} \frac{|T_v^{A_k}|}{|T|} E(T_v^{A_k}) \tag{3}$$

$n(C_i,T)$ denotes the number of samples in the training set T belonging to the class C_i, $D(A_k)$ denotes the finite domain of the attribute A_k and $T_v^{A_k}$ denotes the cardinality of the set of objects for which the attribute A_k has the value v.

To classify a new sample, having only values of all its attributes, we start with the root of the constructed tree and follow the path corresponding to the observed value of the attribute in the interior node of the tree. This process is continued until a leaf is encountered. Finally, we use the associated label to obtain the predicted class value of the instance at hand.

Correspondingly, given a test sample and in which attribute A_k has the value A_k (p), then $T_v^{A_k}$ in Eq.(3) is redefined as $T_V^{(Ak(p))}$.

The precise algorithm proposed in this manuscript goes as follows:

– Run BT searching procedure for the whole training data, branches with different roots are created.

– Combine these branches with the same root to construct decision forest.

– Discard the branch that do not correspond to a state in any of cross validation parts (the idea is that such search states are not common and would not generalize well.

– If test sample does not match any branch, build a new branch for it.

References

Krzysztof Gra and bczewski, Wl odzisl aw Duch. (2002). *Heterogeneous Forests of Decision Trees*. In proceedings of the International Conference on Artificial Neural Networks. 504-509.

Lariviere Bart, Van Den Poel Dirk. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*. 29(2). 472-484.

Breiman, L. (2001). Random forests. Machine Learning. 45(1). 5-32.

Pino-Mejias Rafael, Jimenez-Gamero. (2008). Reduced bootstrap aggregating of learning algorithms. *Pattern Recognition Letters*. 29(3). 265-271.

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

Feature Selection in Extrusion Beltline Moulding Process Using Particle Swarm Optimization

Abdul Talib Bon (Corresponding author)

Department Informatique – Laboratorie L3i, Pole Sciences et Technologie

Universite de La Rochelle, 17042 La Rochelle, Cedex 1, France

Tel: 60-12-766-5756 E-mail: talibon@gmail.com

Jean Marc Ogier

Department Informatique – Laboratorie L3i, Pole Sciences et Technologie

Universite de La Rochelle, 17042 La Rochelle, Cedex 1, France

Tel: 33-05-4645-8215 E-mail: jean-marc.ogier@univ-lr.fr

Ahmad Mahir Razali

School of Mathematical Sciences, Faculty of Science and Technology

Universiti Kebangsaan Malaysia, 86000 Bangi, Malaysia

Tel: 60-17-888-6805 E-mail: mahir@pkrisc.cc.ukm.my

Ihsan M. Yassin

Faculty of Electrical Engineering

Universiti Teknologi MARA, 40450 Shah Alam, Malaysia

Tel: 60-17-257-6295 E-mail: ihsan_yassin@yahoo.com

Abstract

Optimization is necessary for the control of any process to achieve better product quality, high productivity with low cost. The beltline moulding process is difficult task due to its low defects, making the material sensitive to reject. The efficient beltline moulding process involves the optimal selection of operating parameters to maximize the number of production while maintaining the required quality limiting beltline surface damage. In this research, objective is to obtain optimum process parameters, which satisfies given limit, minimizes number of defects and maximizes the productivity at the same time. A recently developed optimization algorithm called particle swarm optimization is used to find optimum process parameters. Accordingly, the results indicate that a system where multilayer perceptron is used to model and predict process outputs and particle swarm optimization is used to obtain optimum process parameters can be successfully applied to beltline moulding process through Particle Swarm Optimization (PSO). Results obtained are superior in comparison with Genetic Algorithm (GA) approach.

Keywords: Beltline moulding, Parameters, Particle swarm optimization, Proces

1. Introduction

Particle Swarm Optimization (PSO) is a recently proposed algorithm by R.C Eberhart and James Kennedy in 1995, motivated by social behaviour of organisms such as bird flocking and fish schooling. PSO algorithm is not only a tool for optimization, but also a tool for representing socio cognition of human and artificial agents, based on principles of social psychology. PSO as an optimization tool provides a population-based search procedure in which individuals called particles change their position or state with time. In a PSO system, particles fly around in a multidimensional search space. During flight, each particle adjusts its position according to its own experience, and according to the experience of a neighbouring particle, making use of the best position encountered by itself and its neighbour. Thus, as in modern Gas and memetic algorithms, a PSO system combines local search methods with global search methods, attempting to balance exploration and exploitation.

The PSO Algorithm shares similar characteristics to Genetic Algorithm, however, the manner in which the two algorithms traverse the search space is fundamentally different. Both Genetic Algorithms and Particle Swarm Optimizers share common elements:

i. Both initialize a population in a similar manner.

ii. Both use an evaluation function to determine how fit (good) a potential solution is.

iii. Both are generational, that is both repeat the same set of processes for a predetermined amount of time.

Particle Swarm has two primary operators: Velocity update and Position update. During each generation each particle is accelerated toward the particles previous best position and the global best position. At each iteration a new velocity value for each particle is calculated based on its current velocity, the distance from its previous best position, and the distance from the global best position. The new velocity value is then used to calculate the next position of the particle in the search space. This process is then iterated a set number of times or until a minimum error is achieved.

This study has presented beltline moulding process by using multilayer perceptron modelling and particle swarm optimization. A multilayer perceptron model of beltline moulding was used to determine the optimal number of hidden units to represent the model and particle swarm optimization was used to minimize the Mean square error (MSE) between the actual output and the modelled output. Two different test cases illustrated that the combined multilayer perceptron and particle swarm optimization system is capable of generating optimal process parameters and can be used successfully in the parameters selection optimization of beltline moulding. Particle swarm optimization is also proved to be an efficient optimization algorithm. For the test cases it yielded optimal parameter around 100 iterations, which take only a little time with today's computers.

2. Literature Review

PSO shares many similarities with evolutionary computation techniques such as Genetic Algorithms (GA). The system is initialized with a population of random solutions and searches for optima by updating generations. However, unlike GA, PSO has no evolution operators such as crossover and mutation. In PSO, the potential solutions, called particles, fly through the problem space by following the current optimum particles.

Each particle keeps track of its coordinates in the problem space which are associated with the best solution (fitness) it has achieved so far. (The fitness value is also stored.) This value is called *pbest*. Another "best" value that is tracked by the particle swarm optimizer is the best value, obtained so far by any particle in the neighbours of the particle. This location is called *lbest*. When a particle takes all the population as its topological neighbours, the best value is a global best and is called *gbest*.

The particle swarm optimization concept consists of, at each time step, changing the velocity of (accelerating) each particle toward its p*best* and *lbest* locations (local version of PSO). Acceleration is weighted by a random term, with separate random numbers being generated for acceleration toward *pbest* and *lbest* locations. In past several years, PSO has been successfully applied in many research and application areas. It is demonstrated that PSO gets better results in a faster, cheaper way compared with other methods. Another reason that PSO is attractive is that there are few parameters to adjust. One version, with slight variations, works well in a wide variety of applications. Particle swarm optimization has been used for approaches that can be used across a wide range of applications, as well as for specific applications focused on a specific requirement.

Multilayer perceptron models which are developed for a better understanding of the effects of beltline moulding process and the resultant quality of beltline can be combined with optimization methods in order to determine optimum control parameters for different objectives such as minimizing manufacturing cost or maximizing productivity. Evolutionary computation algorithms such genetic algorithms and particle swarm optimization are usually utilized for optimization of multilayer perceptron based models. Tandon et al (2002) optimized machining parameters in end milling to minimize machining time by combining a feed forward neural network force model with particle swarm optimization.

3. Methodology

Instead of mutation PSO relies on the exchange of information between individuals, called particles, of the population, called swarm. In effect, each particle adjusts its trajectory towards its own previous best position, and towards the best previous position attained by any member of its neighbourhood (Kennedy.J,1998)**.**

The particles evaluate their positions relative to a goal (fitness) at every iteration, and particles in a local neighbourhood share memories of their "best" positions, then use those memories to adjust their own velocities, and thus subsequent positions. The original formula developed by Kennedy and Eberhart was improved by Shi and Eberhart (1998) with the introduction of an inertia parameter, w, that increases the overall performance of PSO. The best previous position (i.e. the position corresponding to the best function value) of the i-th particle is recorded and represented as $Pi = (p_{i1}, p_{i2}, \ldots, p_{iD})$, and the position change (velocity) of the i-th particle is $Vi = (v_{i1}, v_{i2}, \ldots, v_{iD})$. The particles are manipulated according to the following equations (the superscripts denote the iteration):

$$V_i^{k+1} = \chi\left(\omega V_i^k + c_1 r_{i1}^k \left(P_i^k - X_i^k\right) + c_2 r_{i2}^k \left(P_g^k - X_i^k\right)\right), \quad (1)$$

$$X_i^{k+1} = X_i^k + V_i^{k+1} \quad (2)$$

where i = 1,2,….,N, and N is the size of the population; χ is a constriction factor which is used to control and constrict velocities; ω is the inertia weight; c_1 and c_2 are two positive constants, called the cognitive and social parameter respectively; ri1 and ri2 are random numbers uniformly distributed within the range [0,1]. Eq. (1) is used to determine the i-th particle's new velocity, at each iteration, while Eq. (2) provides the new position of the i-th particle, adding its new velocity, to its current position. The performance of each particle is measured according to a fitness function, which is problem {dependent. In optimization problems, the fitness function is usually identical with the objective function under consideration.

The first term on the right hand side of Eq. (1) is the previous velocity of the particle, which enables it to fly in search space. The second and third terms are used to change the velocity of the agent according to *pbest* and *gbest*. The iterative approach of PSO can be described as follows:

Step 1: Initial position and velocities of agent are generated. The current position of each particle is set as *pbest*. The *pbest* with best value is set as *gbest* and this value is stored. The next position is evaluated for each particle by using Eq. (1) and (2).

Step 2: The objective function value is calculated for new positions of each particle. If a better position is achieved by an agent, the *p*best value is replaced by the current value. As in Step 1, *gbest* value is selected among *p*best values. If the new *gbest* value is better than previous *gbest* value, the *gbest* value is replaced by the current *gbest* value and stored.

Step 3: Steps 1 and 2 are repeated until the iteration number reaches a predetermined iteration number.

Success of PSO depends on the selection of parameters given in Eq (1). Shi and Eberhardt (1998) studied the effects of parameters and concluded that *c1* and *c2* can be taken around the value of 2 independent from problem. Weighting function *w* is usually utilized according to the following formula,

$$w = w_{\max} - \frac{w_{\max} - w_{\min}}{iter_{\max}} x iter \quad (3)$$

where:

$w_{\max}$: initial weight

$w_{\min}$: final weight

$iter_{\max}$: maximum iteration number

iter : current iteration number

Eq. (3) decreases the effect of velocity towards the end of search algorithm, which confines the search in a small area to find optima accurately. The velocity update step in PSO is stochastic due to random numbers generated, which may cause an uncontrolled increase in velocity and therefore instability in search algorithm. In order to prevent this, usually a maximum and a minimum allowable velocity is selected and implemented in the algorithm. In practice, these velocities are taken as [-4.0,+4.0].

The role of the inertia weight, *w* is considered important for the PSO's convergence behaviour. The inertia weight is employed to control the impact of the previous history of velocities on the current velocity. Thus, the parameter w regulates the trade-off between the global (wide-ranging) and the local (nearby) exploration abilities of the swarm. A large inertia weight facilitates exploration (searching new areas), while a small one tends to facilitate exploitation, i.e. fine tuning the current search area. A proper value for the inertia weight, *w* provides balance between the global and local exploration ability of the swarm, and, thus results in better solutions.

4. Results and Discussions

4.1 Modelling the data using MLP

This section shows the details of the MLP modelling process of defects models. 43 data points were collected from the experiments. Initially, the dataset consisted 14 variables, but parameters Cutter and Looper were removed because it carries no informational value. Therefore, inputs for the MLP consisted of 12 variables.

For the defects model, the output is the Mean Square Error, MSE of actual versus modelled defects. MLP uses tangent-sigmoid activation function in the hidden layer, and linear activation function in output layer. This combination of activation functions can approximate any function (with a finite number of discontinuities) with arbitrary accuracy, provided that the hidden layer has enough units (H.Demuth and M.Beale, 2005).

Regularization was used to avoid over-fitting, since data points are not enough to use Early Stopping method. The MLP weights initialization was performed using the NW algorithm to improve convergence speed. To implement

regularization, training was performed using 'trainbr'. It is important to note that the performance function for the 'trainbr' algorithm was the Sum Square Error (SSE) performance function, but MSE was used to guide the PSO optimization.

Both input and output data were preprocessed prior to training so that the model is numerically robust and rapidly converge (M.Norgaard et al., 2000). The normalization is transformed so that the mean is removed ($\mu = 0$), and the standard deviation is 1 ($\sigma^2 = 1$). The rescaling is done so that inputs and outputs reside between -1 and 1. This step is important so that the inputs are properly scaled for the transfer function used in the hidden and output layers.

The tests were performed to determine the optimal number of hidden units to represent both the defects and 'time' models. The results are presented in Section 0.

4.2 MLP Modelling Results

This section describes the experiments performed to determine the optimal number of hidden units to represent the model. For this purpose, the number of hidden units is varied from 1 to 20, and the model was evaluated each time the number of hidden units is changed.

The MLP training was performed for 500 epochs (cycles) while the SSE performance function was used to evaluate the convergence of the MLP each time a hidden unit is added or removed. The optimal hidden layer size was found to be 6 for both defects model. The MLP training results for the defects model is shown in Figure 1.. The SSE comparisons for different hidden layers and the optimal MLP structures were found. The modelling results for defects is shown in Figure 1.

4.3 Feature selection using Particle Swarm Optimization (PSO)

The PSO was used to select the three best inputs to explain the input-output relationship of defects model. A ranking-based system was used to select the best features. Using this system, the value of each particle in the swarm represents the importance of each feature. During optimization, the three best-ranked features were used to train the MLP.

The objective of the PSO is to minimize the MSE fitting error between the actual output and the modelled output. If the features are discriminative, the generalization error should be small since the MLP approximation is close to the actual output. If the features are not discriminative, the model approximation should be poor (indicated by high MSE values).Three experiments were performed to:

- Determine the swarm (population) size required for PSO to converge.
- Determine the best combination of features to minimize production defects.
- Determine the best combination of features to minimize the machine adjustment time.

The minimum population size required to converge is 5 for defects. Therefore, the population of 10 was chosen to sufficiently model both problems. Refer Table 2.

5. Conclusions

From the Figure 2 we can conclude for population has 5 individuals convergence from generation number zero to 20 and best fitting on generation number 9 with best fitting value is 6.268. Meanwhile, population has 10 individuals convergence from generation number 13 to 20 and best fitting on generation number 20 with best fitting value is 1.314.

Furthermore, GA gave improvement when population has 15 individuals convergence from generation number 5 to 20 and best fitting on generation number 15 and 18 with best fitting value is 1.31. But population has 20 individuals gave the better on best fitting value to 1.309 on generation number 16.

The fitting value much better compare to others number of population.

The PSO was used to optimize the input values for the MLP. 6 units were used in the hidden layer. Both the defects and time models were tested. The objective of the PSO is to minimize either the number of defects or the manufacturing time. The fitness is calculated as number of defects or manufacturing time.

- Function $|defects|$ and $|time|$ were used as fitness functions.
- Both should yield 0 as best values.

Since the input should be within certain bounds, any value outside the range of [+1, -1] was clamped during preprocessing. The optimization was performed for 100 generations, with 10 particles for each population. Linearly decreasing inertia weight was used to ensure good convergence. The inertia weight starts with 1 and was decreased after each iterating until it reaches zero. The population test for the defects model is shown in Table 3 and Figure 3.

6. References

Eberhart, R.C and J.Kennedy (1995). A new optimizer using particle swarm theory. *Proceeding of the Sixth Symposium on Micro Machine and Human Science*. IEEE Service Center, Piscataway, NJ,

39-43.

H. Demuth and M. Beale.(2005). *MATLAB Neural Network Toolbox v4 User's Guide*. Natick, MA.: Mathworks Inc.

Kennedy, J. (1998). *The Behavior of Particles. Evol. Progr. VII* , 581-587.

M. Norgaard, O. Ravn, N. K. Poulsen, and L. K. Hansen.(2000). *Neural networks for modelling and control of dynamic systems: A practitioner's handbook*. London: Springer.

Shi, Y. H., Eberhart, R. C. (1998). A Modified Particle Swarm Optimizer*, IEEE International Conference on Evolutionary Computation*, Anchorage, Alaska.

Y. Shi and R. C. Eberhart. (1998). A modified particle swarm optimizer, *IEEE International Conf. on Evolutionary Computation*. Anchorage, Alaska: IEEE Press, 69-73.

Shi, Y. H., Eberhart, R. C. (1998). Parameter Selection in Particle Swarm Optimization. *The 7th Annual Conference on Evolutionary Programming,* San Diego, USA.

V. Tandon, H. El-Mounayri, H. Kishawy. (2002). NC End Milling Optimization Using Evolutionary Computation. *International Journal of Machine Tools and Manufacture*, Vol 42, 595-605.

Table 1. MLP structure results for defects

Number of hidden units	Training SSE	Squared Weights	Effective number of parameters
1	5.617	74.1111	8.40997
2	2.92256	22.1726	15.8015
3	0.335414	56.3065	24.0872
4	0.347128	43.4658	24.6359
5	0.295796	37.9359	25.9243
6	0.273642	40.4999	27.443
7	0.275927	39.8439	27.3663
8	0.275453	43.1597	27.1874
9	35.2039	163.285	127.0000
10	0.27734	41.9922	27.0322
15	0.280599	38.9334	27.0754
20	0.27452	42.8866	27.1995

Table 2. Summary of results for population size for defects model

Population Size	Fitness (MSE)	Features Selected		
		Feature 1	Feature 2	Feature 3
5	0.9461	1	11	12
10	0.9587	1	8	11
15	1.0314	1	11	12
20	0.9855	1	6	11

Table 3. Optimization results (defects model)

Particles	Screw	Pulling	Line speed	MSE
5	34.1610	5.6179	5.0299	0.0349
10	28.5224	4.7581	4.9684	0.0649
15	34.6320	4.8848	5.1947	0.0396
20	32.3356	6.2272	5.0496	0.0311

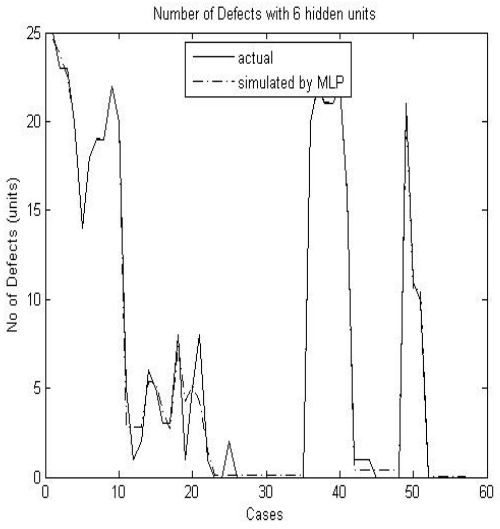


Figure 1. Modelling results for defects model with 12 variables

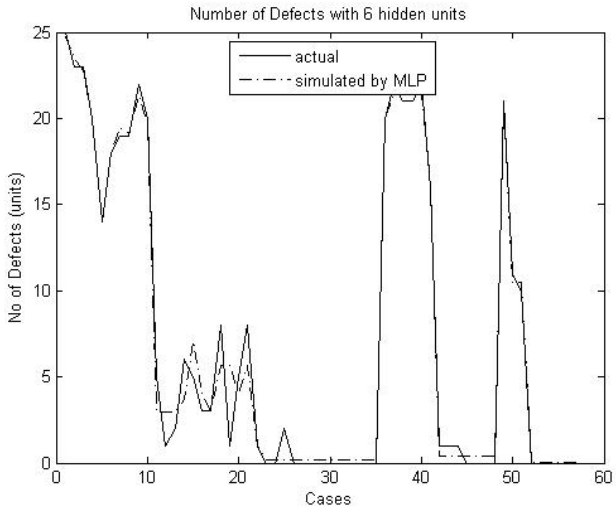


Figure 2. Best fitting model for defects

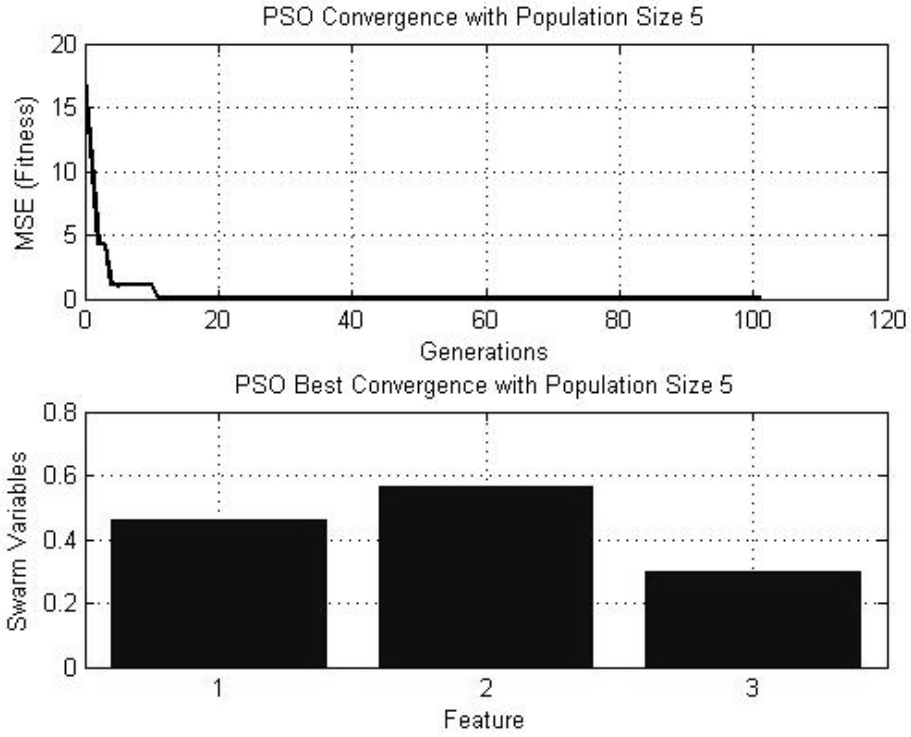


Figure 3. The defects optimization results using 5 particles

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

A Prediction Model of China Population Growth

Hao Zhang, Chao Wang & Xiumin Zhang

College of Information and Communication Engineering

Harbin Engineering University

Harbin 150001, China

E-mail: boshaos66@163.com

Abstract

In this article, we develop a prediction model of China population growth, and notice that the modified index curve is a sort of curve that possesses the growth limit K in the statistics, which is the same with the problem of China population growth. Considering influences of urbanization, population aging and sex proportion, we adopt and improve the modified index model, and add some coefficients to adjust the mathematical equation according to the data. The computation results show that in a short period, the population quantity will increase slowly and approach a fixed value, and in a long term, with the influence of population aging, some factors will put up periodic fluctuations.

Keywords: Improved modified index model, Data fitness, Sex proportion, Urbanization, Population aging

1. Introduction

The population growth prediction mainly includes the prediction of total population and birthrate, death rate, sex proportion. In China, many special problems still exist in the prediction of population growth, such as national population policy, education degree, economic environment and human ideas. All population predictions are implemented based on certain data, and available data generally include sex proportion, death rate and birthrate. But these data are always incomplete and random, so in this article, we select and deal with these data and introduce some authorized data, and overcome the difficulties of incomplete information and accuracy.

The population growth is influenced by many aspects, in this article, we mainly consider following aspects.

(1) The influence of sex proportion. The difference of sex proportion is larger and the negative influence to the total population is larger.

(2) The influence of village population urbanization. The urbanization of village population cannot but induces the change of village and urban population structure and influences the development tendency of total population.

(3) The influence of birthrate to the total population. The national policies have large influences to the birthrate, and influence the change tendency of the total population.

(4) In the long-term population growth prediction, some influencing factors put up tendency fluctuation, which makes the population growth bring the element of tendency fluctuation, and some factors present periodic fluctuation, which makes the population growth bring the element of periodic fluctuation, and some factors present occasional fluctuation, which make the population growth bring the element of random fluctuation.

In this article, we establish the model based on above factors which are presented in the equation through the mathematical expression, and approximated and weakened to a certain extent in the middle and short term model.

2. Symbol explanations

t: The year needed to be predicted (In this article, the year of 1984 represents t=1 which is the initial value.).

e_{it} : The proportion of the influence factor i in the t'th year (i=1 represents the city, i=2 represents the town, and i=3 represents the village.).

p_{it} : The sex proportion in the t'th year.

k_i : The association degree of city population, town population and village population with the total population.

3. Establishment of the Model

3.1 The model of middle and short term

According to the actuality of China and the development of population model, we develop the equation based on the modified index, and the equation is as follows:

$y_t = \mathrm{K} + ab^t - \sum_{i=1}^{3} e_{it} k_i (p_{it} - 1)^2$,

and to explain problems, we evolve the above equation and we can obtain this equation.

$y_t = \mathrm{K} + ab^t - e_1 k_1 (p_{1t} - 1)^2 - e_2 k_2 (p_{2t} - 1)^2 - e_3 k_3 (p_{3t} - 1)^2$

First, we briefly introduce the modified index model, and the general form of the modified index curve is

$\hat{Y}_t = \mathrm{K} + a\, b^t$.

Where, K, a and b are unknown constants, K>0, a≠0, 0<b≠1.

Supposed that every part has m periods, and the sums of various observation value are respectively S_1, S_2 and S_3, i.e.

$s_1 = \sum_{t=1}^{m} Y_t$, $s_2 = \sum_{t=m+1}^{2m} Y_t$, $s_3 = \sum_{t=2m+1}^{3m} Y_t$.

So, we can obtain

$$S_1 = mK + ab + ab^2 + \cdots + ab^m = mK + ab(1 + b + b^2 + \cdots + b^{m-1})$$

$$S_2 = mK + ab^{m+1} + ab^{m+2} + \cdots + ab^{2m} = mK + ab^{m+1}(1 + b + b^2 + \cdots + b^{m-1})$$

$$S_3 = mK + ab^{2m+1} + ab^{2m+2} + \cdots + ab^{3m} = mK + ab^{2m+1}(1 + b + b^2 + \cdots + b^{m-1}) .$$

And we can obtain solutions.

$$b = \left(\frac{s_3 - s_2}{s_2 - s_1} \right)^{\frac{1}{m}}$$

$$a = (s_2 - s_1) \frac{b-1}{b(b^m - 1)^2}$$

$$K = \frac{1}{m} \left[s_1 - \frac{ab(b^m - 1)}{b-1} \right]$$

We respectively consider the city population, town population and village population, and $e_{1t} k_1 (p_{1t} - 1)^2$, $e_{2t} k_2 (p_{2t} - 1)^2$, $e_{3t} k_3 (p_{3t} - 1)^2$ represent the sex population of city, town and village in turn and the influence of village population urbanization to the distribution of the population.

In the following text, we will concretely explain the function of every item and their influences to the development of the population.

3.1.1 Explanations of e_{it} and its computation method

e_{it} represents the weight of the influence of the population in the structure i occupying the total influences of city, town and village to the development of the population, i.e. we can confirm the influencing degree of city, town and village in every year to the development tendency of total population according to the available data. For example, in a certain year, e_{1t}, e_{2t} and e_{3t} are respectively 0.4, 0.3 and 0.4.

e_{it} = the sum of female and male in i in the t'th year/ the sum of female and male in the t'th year.

The urbanization of village population is also embodied in e_i because in the urbanization process of village population, changes of various weight e_1, e_2 and e_3 will certainly occur. For example, above various weights are respectively changed to 0.4, 0.4 and 0.2, which indicates that when village population enters into the city and town, the proportion of city and town population will increase, i.e. the influence of city and town to the development tendency of the total population increase.

So e_i represents the influence of city, town and village to the total population development tendency of China, and fully embodies the influence of village population urbanization to the development tendency of the total population in China.

Because of the middle and short term model, the urbanization of village population can not achieve saturation, so we approximate e_i as the linear function.

According to the data from 2001 to 2005 in China population sampling data, we fit the data to the function which takes t as the variable, and the fitted equations are:

$e_1 = 0.1278 + 0.0066t$

$e_2 = -0.0734 + 0.011t$

$e_3 = -0.0177t + 0.94857$.

3.1.2 Explanations of p_{it} and its computation method

p_{it} represents the sex proportion in city, town and village. According to the bearing age and other information embodied in the date, we can confirm values of p_{it} from 2001 to 2005, and fit the development tendency in the middle and short term, and we also fit these changes to the linear function. But up to 2001, the sex proportion has gone to be stable, so we can take it as the constant which is the value fitted on the curve in 2010. We use the tool of Mathematica to fit the data, and the fitted equations are as follows:

p_{1t}=0.800885+0.0104723t , (t<23)

$p_{2t} = 1.13934 - 0.0071676t$, (t<23)

$p_{3t} = 1.38723 - 0.0174645t$, (t<23)

When t>23, p_{it} are respectively $p_{1t}(23)$, $p_{2t}(23)$ and $p_{3t}(23)$.

3.1.3 Explanations of $(p_{it}-1)^2$

Because the sex will influence the total population, so we introduce the influencing factor $(p_{it}-1)^2$ which represents the difference among sex proportion in every year, and the sex proportion difference are larger, the negative influence to the total population is more obvious and the sex proportion is more closed to 1, so the influence of the sex proportion to the total population is smaller. We select $(p_{it}-1)^2$ to describe the deviation degree of sex proportion of city, town and village with 1. Table 1 shows values of $(p_{it}-1)^2$ with time from 2001 to 2005.

3.1.4 Explanations of k_i

We use the parameter k_i to denote the association degree of city, town and village with the total population, which is relative with human education degree, concept, living level and other factors. k_i is computed through the statistical data from 2001 to 2005 of China Statistics Bureau and the equation. The rapid change occurred in the data in 2003 because of the large sized SARS and the reason that Chinese thought the Yang year went against the birth. So we select data of 2002, 2003 and 2005 to compute k_i.

The computation method is as follows:

$$128453 = 128513.7977 - 0.262 \times 0.00034203k_1 - 0.126 \times 0.0001206k_2 - 0.613 \times 0.00305k_3$$

$$129988 = 130237.1933 - 0.258 \times 0.001132724k_1 - 0.154 \times 0.000266555k_2 - 0.588 \times 0.001306106k_3$$

$$130756 = 131044.088 - 0.005592994 \times 0.277k_1 - 0.171 \times 0.000448304k_2 - 0.000303747 \times 0.552k_3 .$$

From above equation group, we can obtain

k_1=429433

k_2=3.10905×10^0

k_3=-14585.6.

3.1.5 Confirmations of parameter k, a and b

Because we suppose the China population policy in middle and short term is stable, so its influence to the population growth tendency is stable, for example, the family planning policy and other factors are embodied in the coefficient a and b. The method to compute parameters is as follows:

$$b = \left(\frac{\sum_3 y_t - \sum_2 y_t}{\sum_2 y_t - \sum_1 y_t} \right)^{\frac{1}{n}}$$

$$a = (\sum\nolimits_2 y_t - \sum\nolimits_1 y_t)\frac{b-1}{b(b^n-1)^2}$$

$$K = \frac{1}{n}\left[\sum\nolimits_1 y_t - ab\frac{(b^n-1)^2}{b-1}\right].$$

So we can obtain

$$b = 7\sqrt{\frac{892585-838497}{838497-765078}} = 0.95729$$

$$a = (838497-765078)\times\frac{0.95729-1}{0.95729\times\left[(0.95729)^2-1\right]^2} = -47246.9$$

$$K = \frac{1}{7}\times\left[765079-(-47264.9)\times 0.95729\times\frac{(0.95729)^7-1}{0.95729-1}\right] = 149129.6.$$

From the equation

$$y_t = K + ab^t - e_1k_1(p_{1t}-1)^2 - e_2k_2(p_{2t}-1)^2 - e_3k_3(p_{3t}-1)^2,$$

we compute and figure the Figure 1 and Figure 2, and the Figure 1 denotes the tendency of population growth when t<23, and the Figure 2 denotes the tendency of population growth when t>=23.

3.2 The model of long term

The population model of long term is got based on the modification and perfection to the model of middle and short term which is as follows:

$$y_t = K + ab^t - \sum_{i=1}^{3} e_i k_i (p_{it}-1)^2 .$$

Where, e_i goes to the balance in the long-term process, and according the status of developed country and the development tendency of China, we thought e_i can be approximated as a constant, we take that e_1 is 0.4, e_2 is 0.2 and e_3 is 0.4.

P_{it} is the sex proportion, and we find that it shakes in a certain value through large of data, and we take it as 1.03 in the long-term prediction.

In the long-term prediction, the influence of replacement rate to the population is more obvious. So we add $e^{-0.09t}$ to embody its influence in the equation. We confirm that the value of c is 200000 according to the influence of the aging population.

Therefore, we put forward the long-term model as follows.

$$y_t = K + ab^t - \sum_{i=1}^{3} e_i k_i (p_{it}-1)^2 - ce^{-0.09t}$$

In the long-term prediction, some influencing factors put up tendency fluctuation, which makes the population growth bring the element of tendency fluctuation, and some factors present periodic fluctuation, which makes the population growth bring the element of periodic fluctuation, and some factors present occasional fluctuation, which make the population growth bring the element of random fluctuation.

4. The solution of the Model

The computation results are seen in Table 3.

The prediction data are seen in Table 4.

We compute the standard deviation S and the mean absolute percent error (MAPE) utilizing the data in above tables and review the prediction effect of the model.

$$S_y = \sqrt{\frac{1}{n-1}\sum_{t=1}^{21} e_t^2} = 0.168083737$$

$$MAPE = \frac{1}{21}\sum_{t=8}^{22}\frac{|e^t|}{y_t} = 0.000001356$$

References

China National Bureau of Statistics. (2006). *China Statistical Yearbook of 2006.* Beijing: China Statistics Press.

Jiang, Qiyuan. (1993). *Mathematical Model.* Beijing: China Higher Education Press.

Li, Xiaofeng. (2006). The Model of Population Prediction by Modified Index Curve. *Journal of Jingmen Technical College.* No. 6. p. 85-88.

Wang, Moran. (2003). *Matlab and Science Computation.* Beijing: Electric Industry Press.

Xu, Annong. (2004). *Mathematical Experiment.* Beijing: Publishing House of Electrics Industry.

Yang, Qifan. (2005). *Mathematical Modeling.* Beijing: Machine Industry Press.

Table 1. Values of $(p_{it}-1)^2$ with time from 2001 to 2005

$(p_{it}-1)^2$ / t	$(p_{it}-1)^2$	$(p_{it}-1)^2$	$(p_{it}-1)^2$
2001	0.000000106	0.000001021	0.003639929
2002	0.00034203	0.000120592	0.003054233
2003	0.000000111	0.000029677	0.00306567
2004	0.001132474	0.000266555	0.001306106
2005	0.005592994	0.000448304	0.000303747

Table 2. Data from China Statistical Yearbook

year	time	total population	growth rate	first order difference proportion of population	sum of three stages
year	t	y_t	y_t/y_{t-1}	$\Delta y_t/y_{t-1}$	$\sum y_t$
1984	1	104357			765078
1985	2	105851	0.01431624		
1986	3	107507	0.01564463	1.108433735	
1987	4	109300	0.01667798	1.082729469	
1988	5	111026	0.0157914	0.96263246	
1989	6	112704	0.01511358	0.972190035	
1990	7	114333	0.01445379	0.97079857	
1991	8	115823	0.01303211	0.914671578	
1992	9	117171	0.01163845	0.904697987	
1993	10	118517	0.01148748	0.99851632	
1994	11	119850	0.01124733	0.990341753	838497
1995	12	121121	0.01060492	0.953488372	
1996	13	122389	0.01046887	0.997639654	
1997	14	123626	0.01010712	0.97555205	
1998	15	124761	0.00918092	0.917542441	
1999	16	125786	0.00821571	0.9030837	
2000	17	126743	0.00760816	0.933658537	
2001	18	127627	0.00697474	0.923719958	892585
2002	19	128453	0.00647198	0.93438914	
2003	20	129227	0.00602555	0.937046005	
2004	21	129988	0.00588886	0.983204134	

Table 3. Computation results

Year	Actual data	Prediction data	Relative errors (%)
1991	115823	116120	0.25
1992	117171	117306	0.11
1993	118517	118507	0.008
1994	119850	119714	0.11
1995	121121	120919	0.16
1996	122389	122113	0.22
1997	123626	123287	0.27
1998	124761	124431	0.26
1999	125786	125538	0.19
2000	126743	126598	0.11
2001	127627	127600	0.021
2002	128453	128537	0.065
2003	129227	129398	0.13
2004	129988	130175	0.14
2005	130756	130856	0.07

Table 4. Prediction data

Year	Prediction data
2006	132004
2007	132736
2008	133436
2009	134107
2010	134748
2011	135362
2012	135949
2013	136511
2014	137048
2015	137562

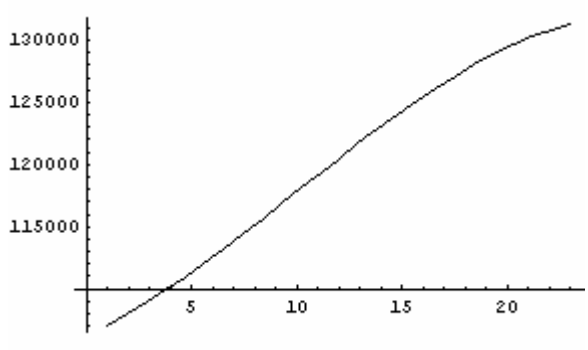


Figure 1. The Tendency of Population Growth When t<23

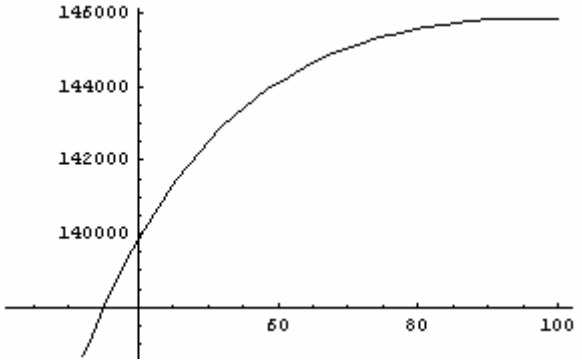


Figure 2. The Tendency of Population Growth When t>=23

Computer and Information Science

Vol. 1, No. 2

May 2008

www.ccsenet.org/journal.html

Investigating the Human Computer Interaction Problems with Automated Teller Machine (ATM) Navigation Menus

Kevin Curran & David King
Intelligent Systems Research Centre
Faculty of Engineering
University of Ulster
Northern Ireland, UK
E-mail: kj.curran@ulster.ac.uk

Abstract

The Automated Teller Machine has become an integral part of our society. Using the ATM however can often be a frustrating experience. How often have some of us experienced the people in the queue in front of you reinserting their card for another transaction. Why does this happen? Is there a design flaw in the user interface? It seems that many ATM navigation menus are not as intuitive or as efficient as they could be. This paper examines a variety of UK Bank ATM navigation menus and proposes a best of breed ATM menu.

Keywords: Human computer interaction, HCI, ATM interface design, Computer applications

1. Introduction

ATM stands for; Automated Teller Machine. It is also referred to as a cash machine, a cash dispenser and 'the hole in the wall' among other names. The ATM is an electronic computerized telecommunications device that allows financial institutions (e.g. bank or building society) customers to directly use a secure method of communication to access their bank accounts. The ATM is a self-service banking terminal that accepts deposits and dispenses cash. Most ATM's also let users carry out other banking transactions (e.g. check balance). ATM's are activated by inserting a bank card (cash or credit card) into the card reader slot. The card will contain the customers account number and PIN (Personal Identification Number) on the cards magnetic stripe. When a customer is trying to withdraw cash for example, the ATM calls up the banks computers to verify the balance, dispenses the cash and then transmits a completed transaction notice.

The idea for an ATM originally was to simply replace or reduce the workload of a bank teller (i.e. the person in the bank who gives out money to customers). The ATM would help reduce banks overheads as wages would be decreased. As for who created the first ATM or where it was first used is a topic of much debate. Basically what answer you get when the question 'who invented the ATM?' is asked depends on who you ask. Miller (2006) presents the facts as he knows it about the history and invention of the ATM. The notion of having a bank machine which automatically dispensed cash to customers came about in the 1930's. A Turkish born inventor working in America called George Simijan started building an earlier and not-so-successful version of an ATM in the late 1930's. He registered the related patents. Simijan came up with the idea of a 'hole-in-the-wall' machine which would allow customers to make financial transactions. However, at the time this idea was well ahead of its time and was met with great doubt. Simijan registered 20 patents related to the device and persuaded an American bank to trial it. However, after 6 months the bank reported little demand in the service and it was withdrawn. It was not until the 1960's that the idea of the ATM was looked at again. John Shepherd-Barron, an inventor from the UK, had an idea in the 1960's for a 24/7 cash dispenser. At the time Shepherd-Barron was the managing director of a company called De La Rue Instruments which today still manufactures cash dispensers. People who believe John Shepherd-Barron invented the ATM argue that the worlds first ATM was installed outside a north London branch of Barclays in 1967.

In 1965 a Scottish man called James Goodfellow was given a project to develop an automatic cash dispenser. Goodfellow was a development engineer with a UK company called Smiths Industries Ltd. He designed a system which accepted a machine readable encrypted card and had a numerical keypad used to enter a PIN. This design is covered in patents in both the UK and USA among other countries. This patent still describes the basic ATM function 40 years later (i.e. the design was patented in 1966). Goodfellow's machines were marketed by Chubb Ltd and installed throughout the UK during the late 1960's and early 1970's. Don Wetzel, then the Vice President of Product Planning of the American Corporation Docutel, claims he applied for a patent on an ATM in 1968. In fact some people believe Wetzel to be the inventor of the ATM. However, an ATM design patented in 1973, stating the Docutel Corporation as the assignee, states John D White as the inventor. White claims he started working on ATM system in 1968 and he

installed the first ATM in 1973. This machine was called the 'Credit Card Automatic Currency Dispenser'. Evidence suggests it was White who received the patent and not Wetzel. There is also a statement in the patent which supports the idea of the modern ATM – "Both the original code and the updated code are scrambled in accordance with a changing key". This is basically what happens today. ATM's have security keys programmed into them. The code changes and is scrambled to prevent access to credit and ATM card numbers between the ATM, the bank and the network processor. It is clear that the topic of ATM invention is quite a controversial one. However, the combined effort of all the inventors surely has helped create today's ATM. Anyone who worked on ATM design from the 1930's until today has contributed something to the modern ATM designs. The purpose of this research is to investigate existing ATM design and to design a 'best of breed' ATM user interface design.

2. Interacting with ATMs

Although ATM's provide an extremely useful service to banks customers, at times they can be very frustrating to use and therefore there is a lot of room for improvement in the interface design. The interface enables communication between the user and the machine. Therefore good user interface design is imperative for high usability levels. Often there are problems or inconveniences experienced when using an ATM. Some of these problems include:

- Waiting in the queue to use the ATM. If users ahead of you in the queue experience difficulties in using the machine, this will increase the time waiting in the queue.
- Inability to see the ATM screen well. This depends on the location of the ATM in relation to the position of the sun. At times it can be difficult to view the contents of the ATM menu.
- Wrongly inserting the ATM card. This problem is more common with new ATM users who are not familiar with their new card and the ATM.
- Getting the required amount of money. Some ATM's may not offer the user the required amount of money they want on the initial cash withdrawal screen. The user will then have to use a few more key strokes to select the required amount (e.g. to withdraw £50 the user might have to select the 'other amounts' option then type in '50' using the keypad and then press 'enter').
- Understanding how to perform operations. Some ATM users find the instructions on how to perform operations quite difficult to understand.
- Often the ATM card is returned to the user while further operations are required (e.g. the card is returned once the user requests a sum of cash. However the user may want to do further transactions; such as check balance or top-up a mobile phone). This will lead to the customer having to re-insert their ATM card, further increasing their time spent at the ATM.
- On some ATM machines the menu options are not aligned with their corresponding menu key. An example of this is illustrated in Figure 1.

Although the sums of money £10 to £100 are not aligned with the related keys, most users will be able to determine what keys are to be used to select the required sum of money. However, if a user wanted to select the 'Other Amounts' option; what button is to be pressed? There is obvious reason for confusion here. It is evident that problems exist with the use of ATM's. Some of these problems are unavoidable (e.g. an ATM running out of money) but solutions exist for others. This research paper focuses on the user interface design problems. ATM navigation menus could be improved considerably to make ATM's more usable.

As technology increases the ATM interface should evolve to take advantage of the new technological innovations. This has happened to a certain extent over the years. However, it is clear that most of today's ATM interfaces do not have the desired high level of usability they should. The modern ATM should be flexible, expressive and easier to use. As mentioned earlier ATM's were introduced in the UK in the late 60's and early 70's. ATM's can now be found in shops, hotels and airports among other places. There was a major design problem when ATM's were first introduced (Dix et al., 1998). During a transaction the ATM dispensed cash to the customer before returning the customers card. This resulted in customers not collecting their card from the ATM. This design problem has now been rectified. The customers' card is returned before cash is dispensed. There have been improvements in the usability of ATM's over the years but there is still a lot of room for improvement. The modern ATM is much more than a simple cash dispenser. Standard UK ATM's offer relatively basic services including cash withdrawals; balance checks and the ability to top-up pay-as-you-go mobile phones. ATM's in different countries (such as USA and Japan) tend to offer advanced services which include cash deposits, cheque deposits, paying bills, purchasing tickets (e.g. train, concert) and purchasing stamps.

The design of an ATM should not only include its inherent usability but also its perceived usability'. This is just one version of possible problems encountered when using (or trying to use an ATM). It reinforces the problems that exist with ATM use. Another typical problem, which was already mentioned earlier, is when an ATM returns the customers

card prematurely i.e. the user still has additional transactions to make. This problematic process is as follows (say the customer wants to with draw cash and then check their balance):

- Insert card
- Enter PIN
- Choose transaction option (Withdraw cash)
- Select/Enter amount of cash to be withdrawn
- Receipt? (yes/no)
- Card ejected from ATM
- Take cash
- Re-insert card
- Enter PIN
- Choose transaction option (Balance Enquiry)
- Return card

This shows how using an ATM can be frustrating. Human computer interface is a term used to describe the interaction between a user and a computer; in other words, the method by which a user tells the computer what to do, and the responses which the computer makes (Heathcote, 2000). (Preece, 1994) also states Human-Computer Interaction (HCI) is about designing computer systems that support people so that they can carry out their activities productively and safely. This can be summarised as 'to develop or improve the safety, utility, effectiveness, efficiency and usability of systems that include computers'. If ATM's were more usable then they would become more effective and efficient machines as users would find them easier to use. This would cause the users to spend less time using the machines and to carry out more efficient transactions. This would be very desirable as it would lessen waiting times in a queue to use an ATM's services. This research paper is concerned with the usability of ATM's; to investigate why existing ATM's user interfaces (navigation menus in particular) have problems and to design a proposed 'best of breed' ATM menu system with excellent usability. Preece (1994) explains usability is concerned with making systems easy to learn and easy to use. Poorly designed computer systems can be extremely annoying to users. This point is particularly relevant. ATM's, at times, can be extremely annoying to use for many reasons which were mentioned earlier. In order to produce computer systems with good usability HCI specialists strive to understand the factors that determine how people operate and make use of the computer technology effectively; develop tools and techniques to help designers ensure that computers systems are suitable for the activities for which people will use them and achieve efficient, effective and safe interaction both in terms of individual human-computer interaction and group interactions.

The last point is relevant for ATM design as users want their banking interactions to be as quick as possible. However, using an ATM's services is very personal (especially with the development of ATM crime) so the group interactions can be ignored in this case. A good interface design can help to ensure that users carry out task when the using the system:

- Safely – this is important for safety-critical software systems; such as software for a jumbo jet for example.
- Effectively – the user get what they want from the system e.g. if an ATM user requests £100 cash, the user should get this and not £50.
- Efficiently – this is the main point concerned with this research paper. If the ATM menu's were improved this would make ATM use more efficient. For example users don't want to spend 5 minutes trying to find the correct way to insert their cash card and type their PIN and the amount of cash they want and then eventually leave without remembering to extract their cash card.
- Enjoyably – systems should be attractive and inviting. Generally if a system is effective and efficient to use, it should also be enjoyable to use as a consequence. However additional effort could be made in ATM interface design to make ATM's more enjoyable to use such as making the screens and menus more colourful and have images for example. A lot of ATM's still just have a black background screen with illuminated text, which is quite dull.

Well designed systems can improve systems significantly. They can improve the output of employees, improve the quality of life and make the world a safer and enjoyable place. An ATM is a service a bank offers to its customers. There are two factors which contribute to the usage of a particular ATM. These are location and the usability of the ATM. Obviously location is the major factor. If an ATM is conveniently located then it will be used a lot. If an ATM is easy to use then this will encourage customers to use the ATM. Many people may have preferences over other ATM's and if they had the choice would use their preferred ATM all the time. All in all, the greater usage a banks ATM

receives, the more potential there is for the bank to make profit. This is why a bank or building society should not under-estimate the importance of good ATM interface design.

Preece (1994) states that 'the best user interface design guidelines are guidelines in a true sense: high level and widely applicable directly principles'. The following principles can be applied widely:

- Know the user – This can often be difficult to achieve, especially when a diverse population of users has to be accommodated or when the users can only be anticipated in the most general terms. This is particularly true for ATM user interface design as this system has a wide range of users from teenagers to pensioners.
- Reduce Cognitive Load – This concerns designing so that users do not have to remember large amounts of detail. Again this is very relevant for ATM user interface deign. The ATM system should be easy to use and users should remember how to understand how to use the system.
- Engineer for errors – a system should be designed to accommodate inevitable user error. If the user makes an error while using the system the system should be able to recover. Engineering for errors includes taking forcing actions to try and prevent users from making errors initially, providing good error messages, and using reversible actions to apply users to correct their own errors.
- Maintain consistency and clarity – Consistency emerges from standard operations and representations and from using appropriate metaphors that help to build and maintain a user's mental model of a system. For example the 'desktop' in a PC is an appropriate metaphor of a work desktop in an office. ATM user's interfaces generally use consistent language e.g. withdraw cash, PIN services etc. However, different banks offer different ATM user interfaces. It would be ideal if there was a universal ATM user interface design, or at least a standard design in each country.

A number of studies have already been carried out regarding ATM's. Most of these studies however have focused on ATM use in relation to the age of users and user disabilities (such as blindness). Adams and Thieben (1991), Mead et al. (1996), Rogers et al. (1997) and Rogers and Fisk (1997) concentrate on ATM use in relation to the age group of the users. Mankze et al. (1998) focuses on ATM usability by the blind while Hone et al. (1998) focuses on modes of control for ATM's including voice control. Rogers et al. (1994) say that they have been informed by banking staff that training is not necessary for ATM's because they are inherently user friendly. This statement however is often not true as many people find ATM's difficult to use, never mind the elderly users and users who are disabled in some way (for example blind). There has also being significant research done on ATM usability and user behaviour. (Hatta and Liyama (1991), El Haddad and Almahmeed (1992), Burford and Baber (1993), Rugimbana and Iversen (1994), Mead *et al*. (1996), Pepermans *et al*. (1996), Rogers *et al*. (1996, 1997), Rogers and Fisk (1997) but none propose a best of breed system. This research is concerned with usability of ATM's. Each ATM investigated (one from each bank e.g. Bank of Ireland, First Trust etc) is evaluated and measured by efficiency (transaction times). This is done using 'mock-up' ATM prototypes which are direct replicas of the Bank's ATM menu designs.

3. ATM System Design

Here we look at the design of the proposed ATM 'best of breed' menu system in relation to the potential users who could use the system. Figure 2 shows a sequence diagram for a complete operational ATM system. The proposed 'best of breed' ATM system does not need to worry about factors such as, insufficient cash or invalid card, as it only concentrates on simulating an ATM navigation menu system.

Existing ATMs menus will need to be mapped out. This is done by visiting each ATM and using the ATM, while at the same time drawing out the menu systems. The Bank/Building Society ATMs which will be visited are First Trust, Ulster Bank, Bank of Ireland, Northern Bank and Nationwide as these appear to be the most commonly used ATMs in the city. To speed up the process of capturing the ATM menus a template was used as illustrated in Figure 3.

Figure 3 represents a standard ATM screen and selection keys enabling the ATM menus to be drawn out quickly and more importantly accurately. It is crucial that each ATM menu system is mapped out accurately as these correspond exactly to the implemented version of the menu systems on the PC. The aim of transferring the ATM menu systems onto a VB program is to simulate the use of the actual ATM systems. Therefore, the performance of each ATM can be determined. When the real world ATM menu systems were drawn out with the aid of the template illustrated in Figure 3, the menu structures needed to be designed before they could be implemented in VB.NET. Once the menu structures were designed and the different levels determined, this made the implementation stage an easier process. In the diagrams/tree structures (see Figure 4) - each box represents a particular menu screen. Due to lack of space, we only include one menu tree structure for the Bank Of Ireland ATM.

The user can only move onto another menu screen after an input i.e. choosing an option. For example, a typical Bank of Ireland ATM transaction may be:

- User inserts card as prompted
- User enters 4 digit PIN as prompted
- User opts to withdraw £20 from menu selection
- The user opts to receive an advance slip when prompted
- The user is asked to take card and wait for cash and receipt.

To measure the transaction performance of the various ATMs, a VB program was created to simulate each ATM's real world menu structure to replicate ATM transactions in the lab. The ATM simulation program is used to test and monitor each of the ATMs performance. The user is presented with a collection of buttons. The user will click on the required button, taking the user to that particular banks ATM simulator.

The ATMs simulation user interface 'shell' remains consistent for each ATM i.e. the main screen, the selection keys and the keypad. This is illustrated in Figure 5. The eight selection keys (either side of the ATM screen) will be used to make user selections from the menu. The 'Insert Card' button will be used to simulate the user inserting the ATM card into the machine. The keypad containing the digits 0-9 and the keys 'Cancel', 'Clear' and 'Enter' is standard for all the various types of ATMs investigated in this research. The button 'Back to Main Menu' is simply to take the user back to the main screen illustrated in Figure 5.

The 'Best of Breed' ATM menu system (called OptiATM which means optimal ATM menu) also uses the standard interface shown in Figure 5. The system is used to run transaction performance tests on the existing ATM menu designs (as well as the 'best-of-breed' OptiATM). When using an ATM the machine often takes time to process data such as 'processing card' and 'contacting bank/building society' etc. The ATM Simulator will not simulate these processing time periods. However, this will not corrupt the transaction performance test results as it will be consistent for all the ATM simulations. As these processing time periods will not be represented by any of the ATM simulations, the transaction performance test results will be accurate as they are all relative. Another important factor to note is as follows; many inconveniences can occur when using an ATM, such as – ATM has run out of cash, user enters PIN incorrectly, error in reading card - to name a few. The ATM Simulator will not simulate these situations. When using the ATM simulator, 'perfect' transactions will be simulated i.e. the ATM reads the users card without error, the user enters the correct PIN and the user has sufficient funds. The ATM Simulator has a welcoming screen allowing the user to select which ATM to simulate. This is illustrated in Figure 6. Each of the Bank logos are buttons. The user simply clicks on a particular button to go to simulate that particular bank. At this stage, all the buttons take the user to all the existing ATM simulations.

Obviously it is not possible to physically insert an ATM card when using the ATM Simulator. Therefore an 'insert card' button has been created to simulate inserting the card into the ATM. Once this button is clicked by the user, the next menu screen appears. This of course, is the screen which prompts the user to enter their PIN (Personal Identification Number). The requirements of the ATM Simulator were simply to simulate the use of an ATM machine. Therefore this is what the ATM Simulator does. When the users enter their PIN, there are no comparison algorithms or checks to confirm that the PIN entered was indeed correct. As mentioned earlier, the ATM Simulator is just a tool to enable transaction performance tests to be conducted on ATM menu designs.

To enter a PIN, the user simply clicks four digits on the keypad shown in Figure 7. The PIN can be any four digit number so long as it ends with the digit '1'. When this number is clicked, a click event is triggered which takes the user to the next screen. However this action only occurs in the Bank of Ireland, Ulster Bank and Northern Bank ATMs. The other two ATM designs (First Trust and Nationwide) require the user to press the 'enter' key to confirm the PIN entered is correct. In this case, the click event is triggered when the 'enter' key is clicked in the ATM Simulator. These two different actions are reflected in the real world ATM designs.

To select an option from the menu, the user simply clicks on the select key adjacent to the menu option displayed on screen. This is the same when using an ATM in real-life, only the user presses the selection key with their finger. An example of a user selection is illustrated in Figure 8. An Ulster Bank ATM user may want to withdraw cash. Therefore, once the user simulates inserting their card and entering their PIN, the user will click on the selection button adjacent to the option 'Withdraw cash'. This will then present the user with different cash withdrawal options.

Each existing ATM has been designed and implemented to reflect the real-life counterparts. Therefore each ATM, as illustrated in the design chapter using the ATM menu tree structures, will have different menus presenting the user with different options. Using the Bank of Ireland ATM, when the user enters their PIN, they are automatically presented with cash withdrawal options as well as some additional options. This is illustrated in figure 9. Using the Ulster Bank ATM to withdraw cash was already illustrated in Figure 8 earlier. Using the Northern Bank ATM system, simulating user's

options is illustrated in Figure 10. The initial menu shown here is only of course displayed to the user when the insertion of the ATM card and PIN entered is simulated previously.

As you can see from Figure 10, the user has three options when using the Northern Bank ATM i.e. withdraw cash, withdraw cash with a receipt and display or print balance. Figure 11 show the options a user has when using the Nationwide ATM. Once the user enters their ATM card and PIN number, they are presented with the options; request statement, balance enquiry and cash withdrawal. Figure 11 also shows the subsequent corresponding screens.

First Trust also had its own unique menu design and layout. It was essential that each of the ATMs investigated (i.e. Bank of Ireland, Ulster Bank, First Trust, Nationwide and Northern Bank) be implemented correctly. It was essential that each ATM simulation on the PC directly represented the corresponding banks real life ATM menu design and layout to produce accurate and reliable transaction performance test results.

4. OptiATM

Over a period of days, we observed customers using the ATMs. The number of users that were seen re-inserting their ATM cards was recorded against the total number of users seen using the ATM.

The data in Figure12 is graphical represented in Figure 13. This data may indicate that Ulster Bank has the higher usability issues while Northern Bank has the least. However, many factors have to be considered when analysing this data. If the ATM observations were carried out again the results could be a lot different. The main factor which will affect these results is:

- The individuals who use the ATM when the observations are made i.e. are the customers novice, intermediate or expert users. These levels of expertise may be determined by the age of the user for example.

The data in Figure 12 and Figure 13 may be described as insignificant for analytical purposes. However, it still provides an insight into the problems of ATM HCI issues. The data collected reinforces the fact that users regularly have to reinsert their ATM cards to carry out further transactions. It was found out that the main reason why bank customers used and ATM was to withdraw money. This may seem obvious but it was important to make this assumption concrete. Out of the 217 ATM users observed (covering all 5 banks), 202 users said that the main reason why they use an ATM is to simply withdraw cash. The remaining 15 ATM users said they mainly used an ATM to check their bank account balance. This data is illustrated in Figure 14. However, these ATM customers also said that they usually follow up this initial transaction with an additional transaction of withdrawing cash.

This data illustrated in Figure 14 will be useful when designing the menu system for the 'Best-of-Breed' ATM system OptiATM i.e. it would be useful to list the most frequently used options first for example. It is also now clear that a reoccurring problem of ATM use is that customers have to reinsert their cards to carry out additional transactions. The reasons customers gave for reinserting their ATM cards are illustrated in Figure 15.

Simply observing existing ATM usage and asking ATM user's questions did not provide enough information to help create the proposed 'Best-of-Breed' OptiATM. As mentioned in an earlier chapter, each ATMs (i.e. Bank of Ireland, Ulster Bank. First Trust, Nationwide, Northern Bank) menu system was mapped out and implemented to create the ATM Simulation program. Each ATM was performance tested by three different users and an average of the times was recorded. The different performance tests were to (1) Withdraw £20 (i.e. a standard amount presented to customer); (2) Withdraw £20 with receipt; (3) Withdraw £300 (i.e. another amount); (4) Withdraw £300 with receipt; (5) Check Balance on screen; (6) Print Balance and (7) Check balance and then withdraw £20. These performance tests cover the range of functions offered by a standard ATM and give a good indication of each ATMs overall performance. Due to lack of space we simply show in Figure 16, the average transaction performance times when simulating using a Bank of Ireland ATM.

There are a few issues that were highlighted when running the tests using the Bank of Ireland ATM simulator. The Bank of Ireland ATM is the only system which offers the user immediate cash withdrawal options after the customer enters their PIN. This suggests that the Bank of Ireland ATM designers recognise that cash withdrawals is a primary transaction necessity for its customers. This point was illustrated in Figure 14 earlier. We did notice with the Nationwide ATM that the menu options do not remain consistent. When the user is asked for example; 'Would you like a receipt with this transaction?' the 'yes' and 'no' options are the bottom left and bottom right options respectively. These are selected using the selection keys. However, when the user is asked; 'Would you like another service?' the 'yes' and 'no' keys are not placed here. Alternatively they are both placed on the right hand side of the screen. 'Yes' is selected using the second from bottom key on the right hand side, while 'No' is selected using the bottom key on the right hand side. This is an issue as consistency is one of the key factors in designing good, usable interfaces.

We compared each of the ATMs in a series of transaction performance tests carried out in order to highlight which ATMs perform better than others for certain transactions so as to help identify the best and worst features of each ATM. The results feed into the 'Best-Of-Breed' OptiATM as it contains all the optimum features of existing ATMs and none

of the poorly performing features. Figure 17 shows the average time each ATM took to withdraw £20. The reason why the withdrawal of £20 was chosen as a performance test is that it is a standard withdrawal amount offered by all the ATMs and it is a common transaction for ATM customers.

As you can see from Figure 17, the Bank of Ireland ATM has the fastest cash withdrawal time for standard amounts of cash. This is so because once the user enters their PIN, they are automatically given the option to withdraw cash without the need of any additional keystrokes. The Ulster Bank, Nationwide and Northern Bank ATMs are all relatively close in performance times with the Ulster Bank edging it. It is clear that the First Trust ATM has the worst performance. This is due to First Trust giving the user a lot of information and additional prompts. Ideally the 'Best-of-Breed' OptiATM will have the Bank of Irelands fast cash withdrawal feature incorporated into its menu design.

All five ATMs were evaluated and the results used to influence the 'Best-of-Breed' OptiATM. Again, lack of space prevents us detailing all the individual scenarios. By using the performance test results, the OptiATM should include the features from the ATMs which yield the fastest and most efficient results for each transaction. Ideally the 'Best-of-Breed' OptiATM would have all the best features of the banks ATMs for each particular type of transaction; however it may not be possible for all the different transactions. Each of the ATM menus are systems, meaning they are all inter-related and inter-connected. This means that the 'Best-of-Breed' OptiATM is not able to incorporate the best feature of one particular ATM without keeping some of its less efficient features. Compromise is required when designing the 'Best-of-Breed' OptiATM system.

A 'Best-of-Breed' OptiATM should out-perform existing ATMs however, this does not resolve the problem of users having to reinsert their card after making a cash withdrawal. This problem may occur because of the design and layout of the ATM menu system or the ATM user is not given the option of carrying out another transaction after withdrawing cash. The only possible reason why ATMs don't offer the user the option of another transaction after withdrawing cash is the fact that user might simply take their cash and forget about their ATM card, thus leaving it in the ATMs card slot. Doing this would be both inconvenient and a security risk (as people could obtain the users bank card). Therefore the Best-of-Breed' OptiATM should offer the user 'Do you want another transaction?' when withdrawing cash, but at the same time overcome the problem of making sure that user's cannot leave their ATM card behind. There are some ATMs which operate differently from the ones investigated here. High street banks/building societies (such as the existing ATMs investigated) operate in the following way (1) Insert card, (2) User enters PIN and carries out transaction/s required and (3) User takes card.

Portable ATM's or Independent Convenience Cash Dispensers work in a different manner in that (1) User inserts card; (2) Card is read and user is instructed to remove card and (3) User enters PIN and carries out transaction/s required. This enables customers to be offered 'another transaction' after withdrawing cash. Using this method of operation, customers could withdraw cash and then be prompted 'Would you like another transaction?' This way, once the customer takes their cash, it is not possible to forget their card – as they already took it before they began their transaction. However, this creates a security problem on its own. What if the user just takes their cash and walks off without responding to the prompt 'Would you like another transaction?' Could the next user simply use the previous users account and withdraw cash? A fail-safe would be in place to ensure that this could not happen. If the user does not respond within a given time period e.g. 5 seconds for example, the session ends. In fact this fail-safe would be in place whenever a user is using the ATM at any given time.

The OptiATM is designed to be more efficient and easier to use than the existing ATM systems investigated. Figure 18 shows the user options displayed to the user when the PIN is entered. The OptiATM initial menu screen tries to incorporate all the main user options. This inevitably reduces transaction times. The user is always prompted after a transaction asking the question 'Would you like another transaction?' This is also the case when withdrawing cash, eliminating the problem of having to reinsert their card for another transaction. However, for this to work without problems, the user has to remove card (after details are read of course) before carrying out a transaction. To test whether OptiATM is a 'best-of-breed' ATM menu design capable of out-performing existing ATMs, we put it through a series of tests as illustrated in Figure 19.

Figure 19 shows that the OptiATM is just a little slower at withdrawing £20 than the Bank of Ireland ATM. This is so because the user has to remove their card before beginning their transaction. However, even taking this into consideration, OptiATM still out performs the other four ATMs.

Figure 20 again shows that the OptiATM is just a little slower at completing this transaction. Again this is due to the fact that the user has to remove the ATM card before continuing with the transaction. Therefore this is an acceptable result. The benefits of the added facility allowing the user to carry out another transaction after withdrawing cash, outweighs the fact that the OptiATM is out performed by both Ulster Bank and the Bank of Ireland.

Figure 21 highlights that the OptiATM can withdraw £300 i.e. other amounts of cash, faster than any of the other ATMs. This is so because the user can enter the required amount of cash on the main option screen.

Figure 22 demonstrates the OptiATM design yields the fastest time to firstly check balance and then withdraw £20. OptiATM is an improvement on the tested real world ATM systems. In only two of the seven transaction performance tests carried out did the OptiATM not have the fastest transaction time.

Figure 23 illustrates, although the OptiATM design was beaten in (and only slightly) only two tests, it still has an overall better performance than the best performing existing ATM system.

The fact that the OptiATM design has the added facility of eliminating the need to reinsert ATM cards, while at the same time improving overall performance, reinforces that the OptiATM menu design is an improved 'best-of-breed' ATM menu system.

ATM manufacturers have demonstrated several different technologies which as of yet have not gained worldwide acceptance. These include:

- Biometrics for security purposes i.e. the authorization of transactions is based on the scanning of fingerprints, the eye, face etc.
- Ability to print 'items of value' such as traveller's cheques.
- Customer specific advertising on the ATM.

Some of the examples above are potentially the way forward for ATM's. However, Banks and other providers of ATM services have to determine if these advancements are feasible. They have to ask the questions are they financially feasible and how will customers react to the changes. Although the main topic for this research paper is the problems with ATM interface design, there are many issues with the use of ATM's in today's world. The main issue is security. This can be divided into 2 broad categoreies which are physical security of the ATM and transactional security. Early ATM security focused on making ATM's safe from physical attack. ATM's were basically safes with dispenser mechanisms. It has been recorded that thieves have stole entire ATM's and its housing in an attempt to steal its cash. However, modern ATM physical security focuses on denying the use of the cash inside the ATM to thieves. Using a technique such as dye markers dyes the cash and potentially the thieves, making the cash unusable and increasing the chances of the criminals being caught. Sensitive data in ATM transactions are encrypted. However there are always problems with data security. 'Phantom withdrawals' are a major problem with ATM's. This is when money is withdrawn from a customer's account using an ATM without the customer being aware. Neither the bank nor the customer admits liability for the withdrawals. Many fraud experts believe dishonest insiders (i.e. bank workers) are responsible for phantom withdrawals. Card cloning is also another major ATM security problem. It is possible to clone ATM user's cards by installing a magnetic card reading device over the ATM's real card slot. This is able to store information such as the card number. Once the criminal has this, the card can be cloned onto a second card. Then all that is needed is the ATM users PIN. The criminal can gain this by simply observing the user enter the PIN or by placing a video camera near the ATM recording the user's PIN's being entered. Banks are working on measures to try and counteract card cloning. The use of smart cards for ATM's, as they cannot be easily copied by un-authenticated devices is one potential countermeasure. Banks are also attempting to make the outside of their ATM's tamper proof. Stealing customers ATM cards is a low-tech form of fraud. The user's PIN can be observed by 'shoulder surfing' and a second criminal can then physically steal the customers card. Also, there have been cases reported were ATM users have been 'mugged' after using an ATM machine. ATM users are vulnerable as an observing criminal will believe a user will have cash.

5. Conclusion

The main objective was to design a 'best-of-breed' ATM menu system. This was achieved in the form of the OptiATM. As demonstrated, the OptiATM menu design, out performs and is a more usable and efficient system than the existing ATMs investigated. The OptiATM system was designed to resolve the problem of users having to reinsert their ATM cards to carry out another transaction and to speed up transaction times. The system could help improve user's basic everyday ATM transactions however the OptiATM system is basic in that the functions and services they offer. Many advanced ATM machines offer an abundance of additional services including cash and cheque deposits, ability to pay bills at terminal, top-up pay as you go mobile phone and purchasing tickets such as train or concert tickets. ATMs have become part of the modern world's infrastructure. We expect ATMs for convenience as much as we expect a good transport service. However, as the services offered grow, the ATM menu designs will become more complicated. This may lead to the systems becoming even more confusing for users and harder to choose. It is recommended that ATM designers consult extensively with ATM users to help them design and create easy-to-use and efficient ATM systems.

References

Adams, A. S. and Thieben, K. A. (1991), Automatic teller machines and the older population. *Applied Ergonomics*, 22**,** 85 -90.

Bennett, S. McRobb, S. Farmer, R. (1999) *Object-Oriented Systems Analysis and Design*, McGraw-Hill.

Burford, B. C. and Barber, C. (1993) ,A user-centered evaluation of a simulated adaptive autoteller, in S. A. Roberston (ed.) *Contemporary Ergonomics*, London, UK: Taylor and Francis Ltd, 117-122.

Dix, A.J. Finlay, J.E. Abowd, G. D. Beale, R. (1998) *Human-Computer Interaction Second Edition,* Prentice Hall Europe.

El-Haddad, A. B. and Almahmeed, M. A., (1992), ATM banking behaviour in Kuwait: a consumer survey, *International Journal of Bank Marketing*, 10**,** 25 - 32.

Hatta, K. and Liyama, Y., (1991), Ergonomic study of automatic teller machine operability. *International Journal of Human Computer Interaction*, 3**,** 295-309.

Heathcote, P.M. (2000) *A-Level Computing 4th Edition*, Payne-Gallway Publishers Ltd.

Miller, A.W, 2006 [Online], *Who invented the ATM machine? Automated Teller History,* Available at http://www.atmmachine.com/atm-inventor.html [Accessed 20/11/2006].

Hone, K. S., Graham, R., Maguire, M. C., Baber, C. and Johnson, G. I. (1998), Speech technology for automatic teller machines: an investigation of user attitude and performance. *Ergonomics*, 41**,** 962-981.

Mankze, J. M., Egan, D. H., Felix, D. and Krueger, H. (1998), What makes an automated teller usable by blind users? *Ergonomics*, 41**,** 982-999.

Mead, S., Walker, N. and Cabrera, E. F. (1996), Training older adults to use automatic teller machines. *Human Factors*, 38**,** 425 - 433.

Pepermans, R., Verleye, G. and Van Capellen S. (1996), `Wallbanking', innovativeness and computer attitudes: 25 - 40 year-old ATM-users on the spot. *Journal of Economic Psychology*, 17**,** 731 - 748.

Preece, J. (1994) *Human-Computer Interaction*, Addison-Wesley Publishing Company.

Pressman, R. (1997) *Software Engineering, A Practitioner's Approach 4th Edition*, McGraw-Hill.

Rogers, W. A. and Fisk, A. D. (1997), ATM design and training issues. *Ergonomics in Design*, January 1997, pp. 4 - 9; http://hfes.org/Publications/TOC/EID-93-97.html

Rogers, W. A., Cabrera, E. F., Walker, N., Gilbert, D. K. and Fisk, A. D. (1996), A survey of automatic teller machine usage across the adult lifespan. *Human Factors*, 38**,** 156 -166.

Rogers, W. A., Gilbert, D. K. and Cabrera, E. F. (1997), An analysis of automatic teller machine usage by older adults: a structured interview approach. *Applied Ergonomics*, 28**,** 173 -180.

Rugimbana, R. and Iversen P. 1994, Perceived attributes of ATMs and their marketing implications. *International Journal of Bank Marketing*, 12**,** 30 - 35.

Sommerville, I. (2004) *Software Engineering Edition 7*, Pearson Education Limited.

The RadioHead, (2005) [Online], *The ATM Experience*,

Available at http://theradiohead.blogspot.com/2005/02/atm-experience.html [Accessed 25/11/2006].

Van Vlient, H. (2000*) Software Engineering, Principles and Practice, 2nd Edition*, Wiley & Sons.

Figure 1. Typical ATM Menu

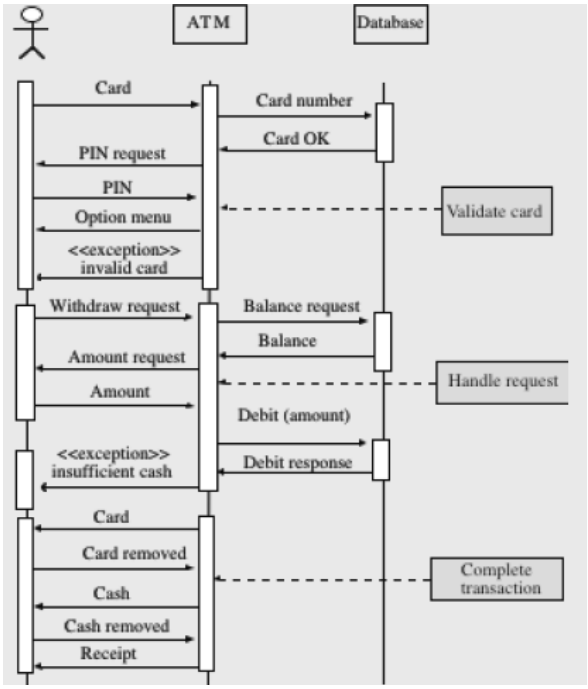


Figure 2. Sequence Diagram of an ATM System (Sommerville, 2004)

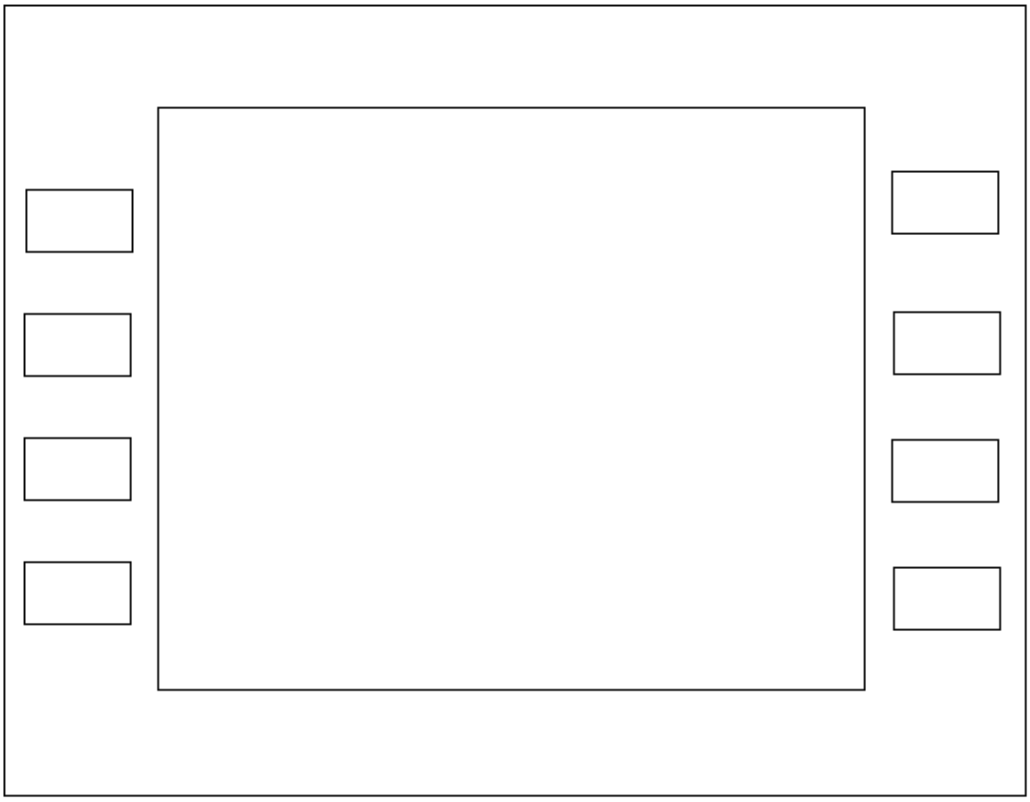


Figure 3. ATM menu screen template

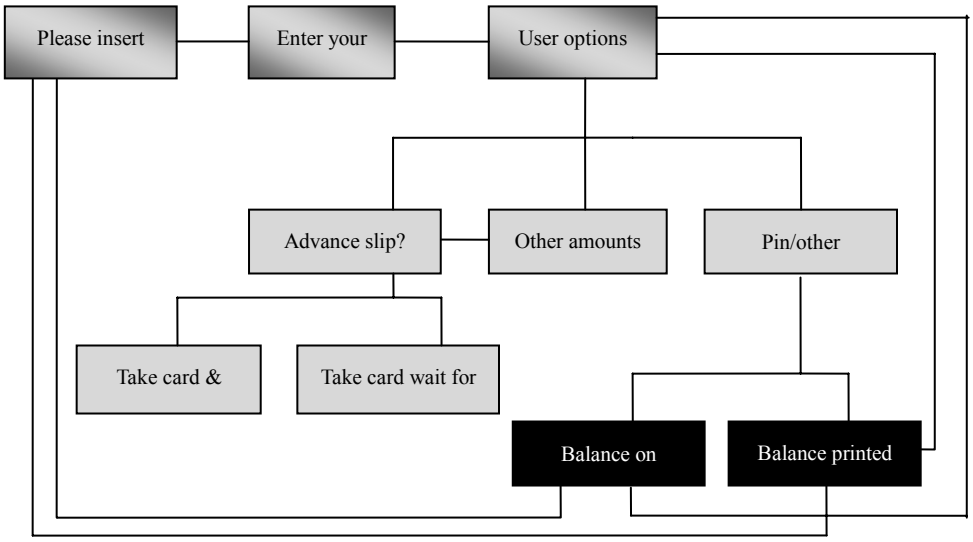


Figure 4. BOI ATM Menu Tree Structure

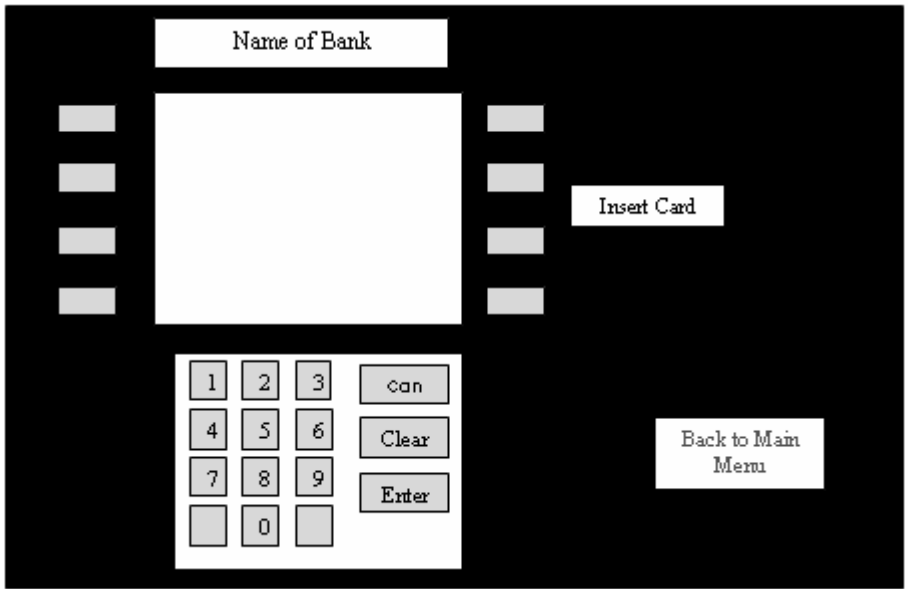


Figure 5. ATM Simulator Interface Design

Figure 6. ATM Simulator Main Screen

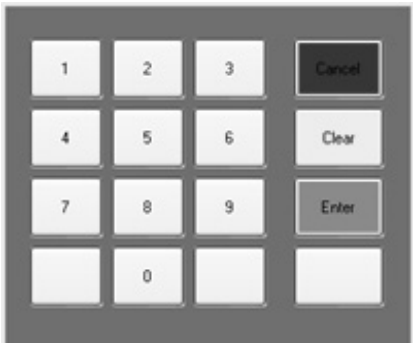


Figure 7. ATM keypad

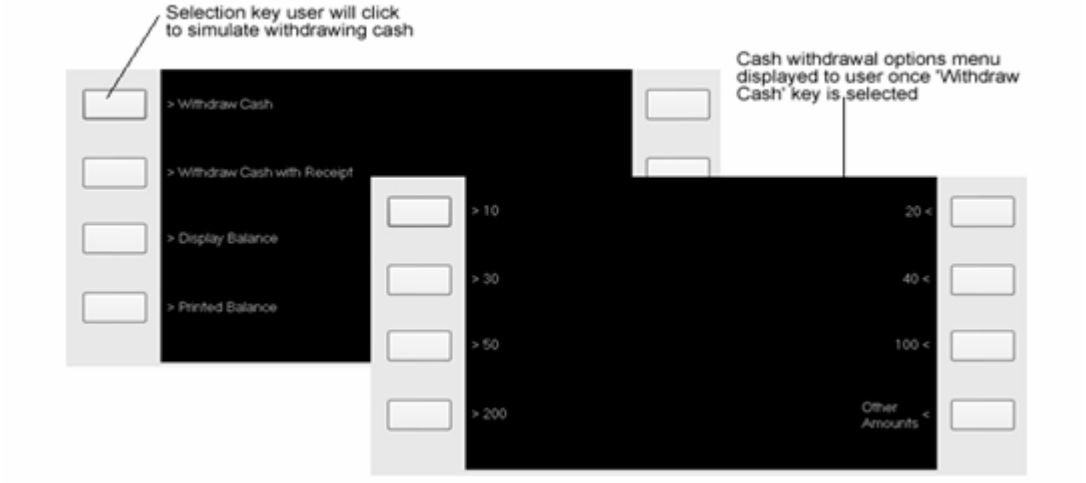


Figure 8. Ulster Bank ATM - Withdraw Cash

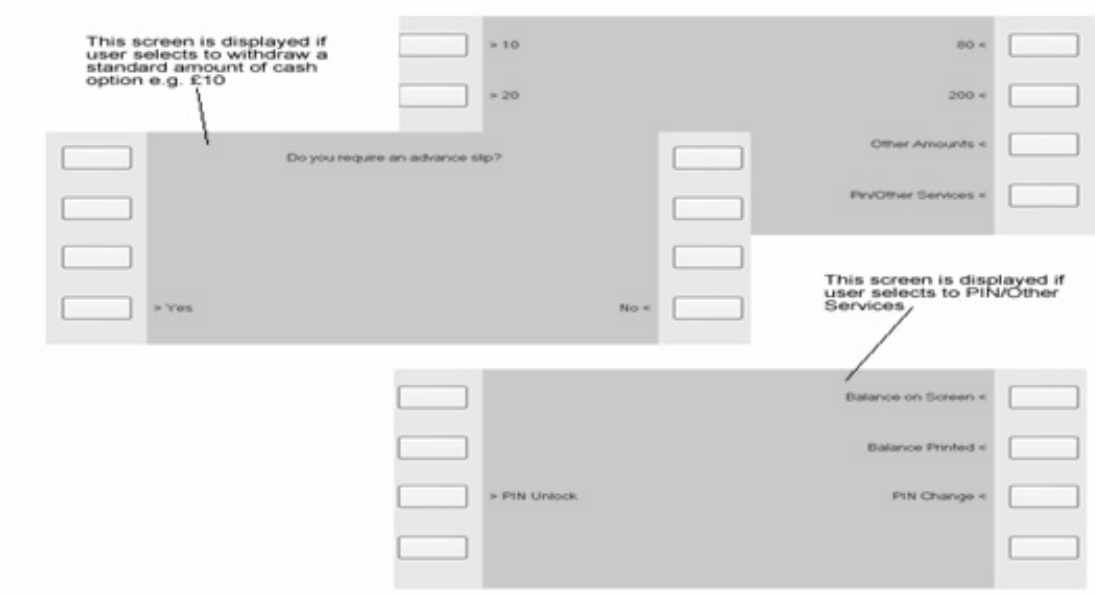


Figure 9. Bank of Ireland ATM User Options

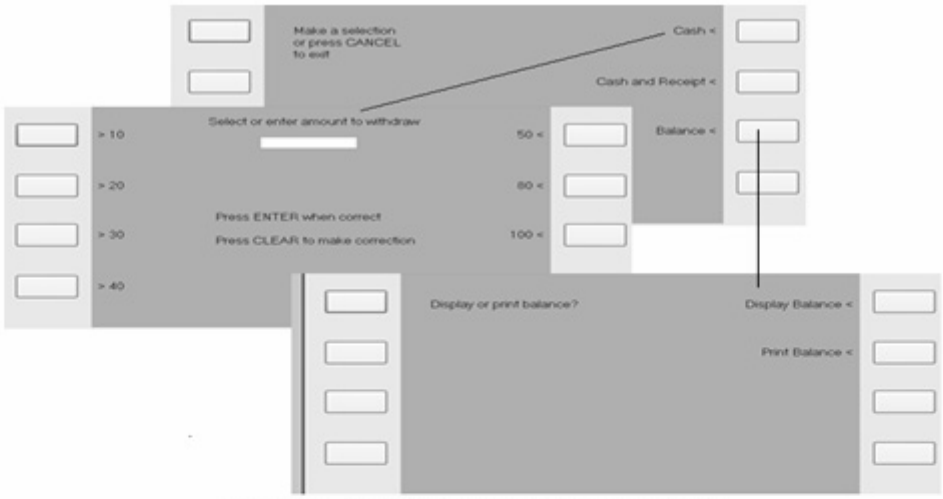


Figure 10. Northern Bank ATM User Options

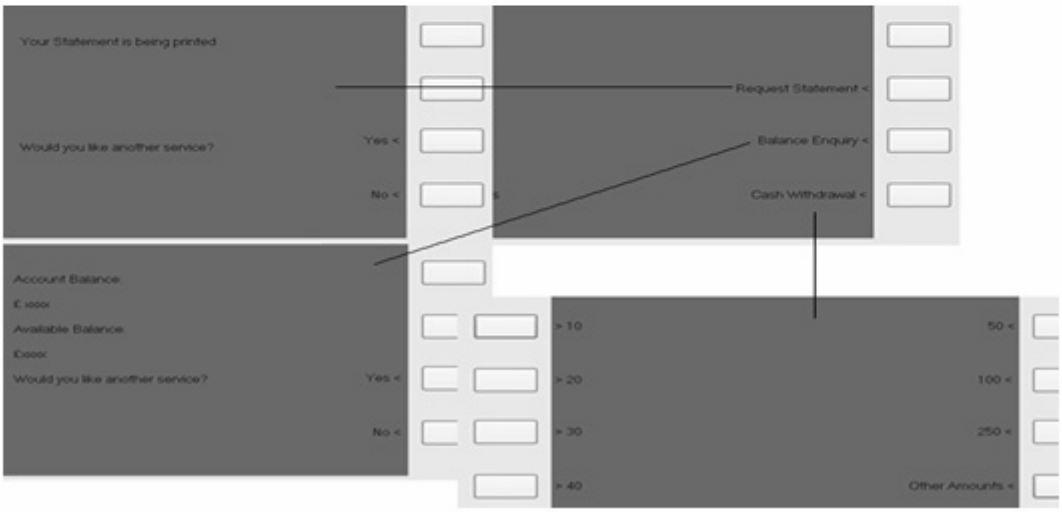


Figure 11. Nationwide ATM User Options

ATM	Total number observed using ATM	Total number observed reinserting card	% of reinserted cards
Bank of Ireland	41	4	9.76
Ulster Bank	54	6	11.11
First Trust	62	2	3.23
Nationwide	33	3	9.09
Northern Bank	27	1	3.70

Figure 12. ATM Customers observed reinserting card

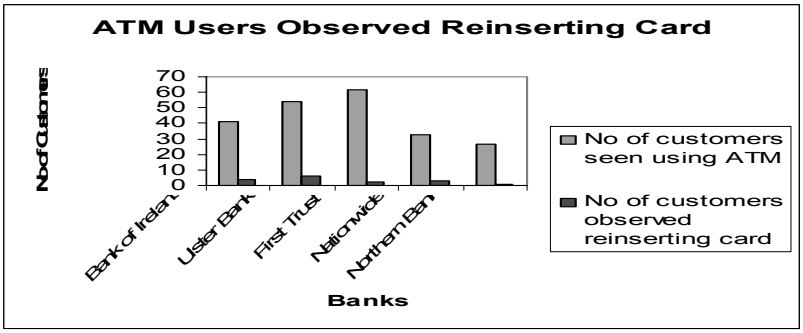


Figure 13. ATM users observed reinserting card

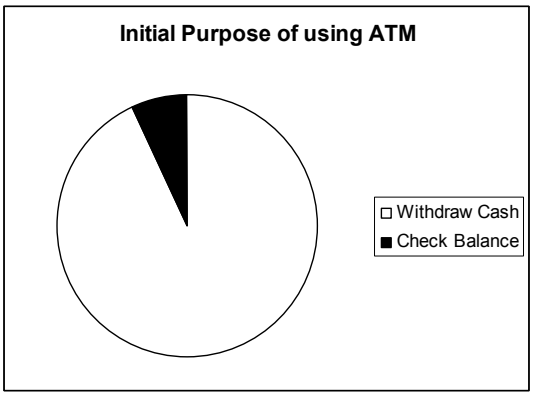


Figure 14. Main purposes of using an ATM

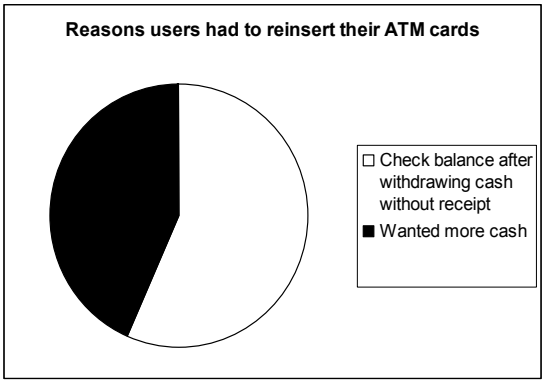


Figure 15. Reasons Users had to reinsert Card

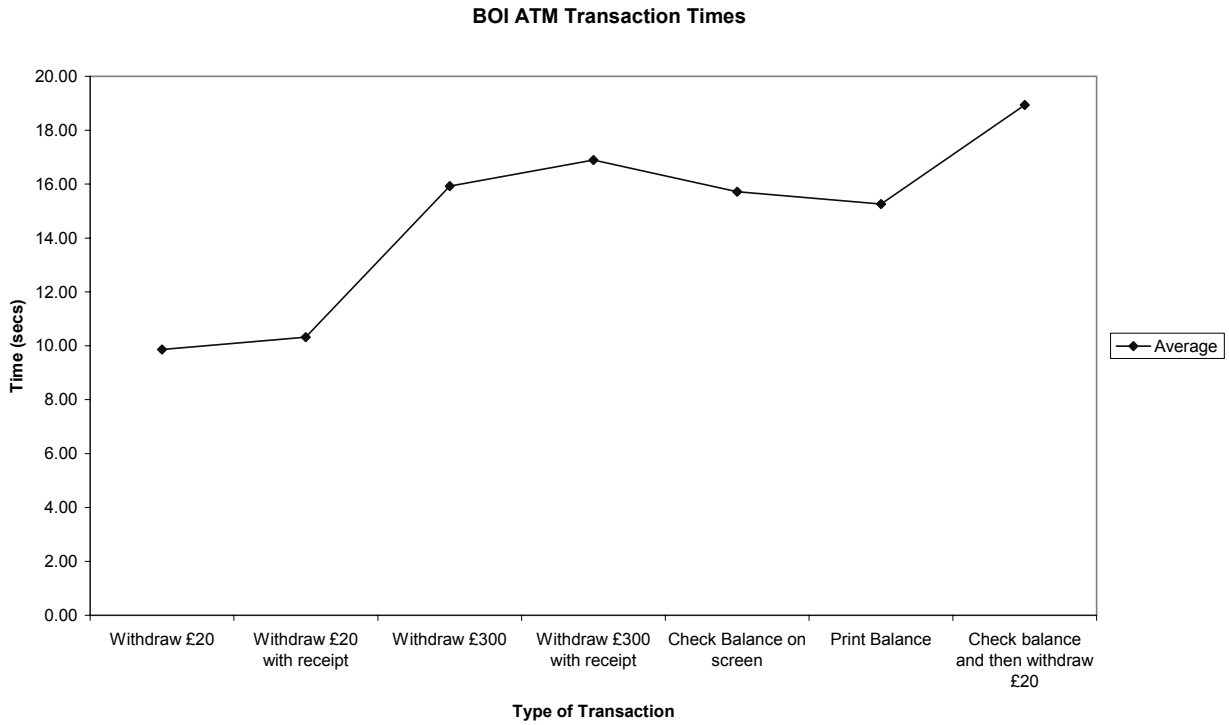


Figure 16. Bank of Ireland Transaction Performance Times

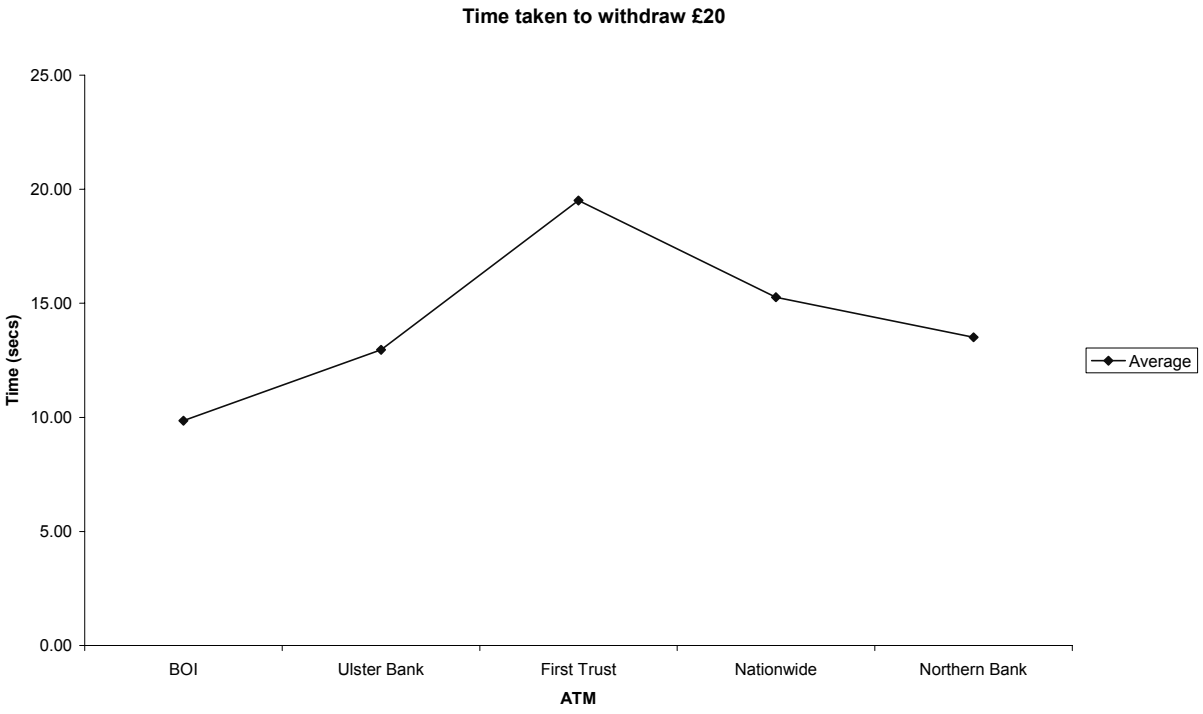


Figure 17. Times taken to withdraw £20

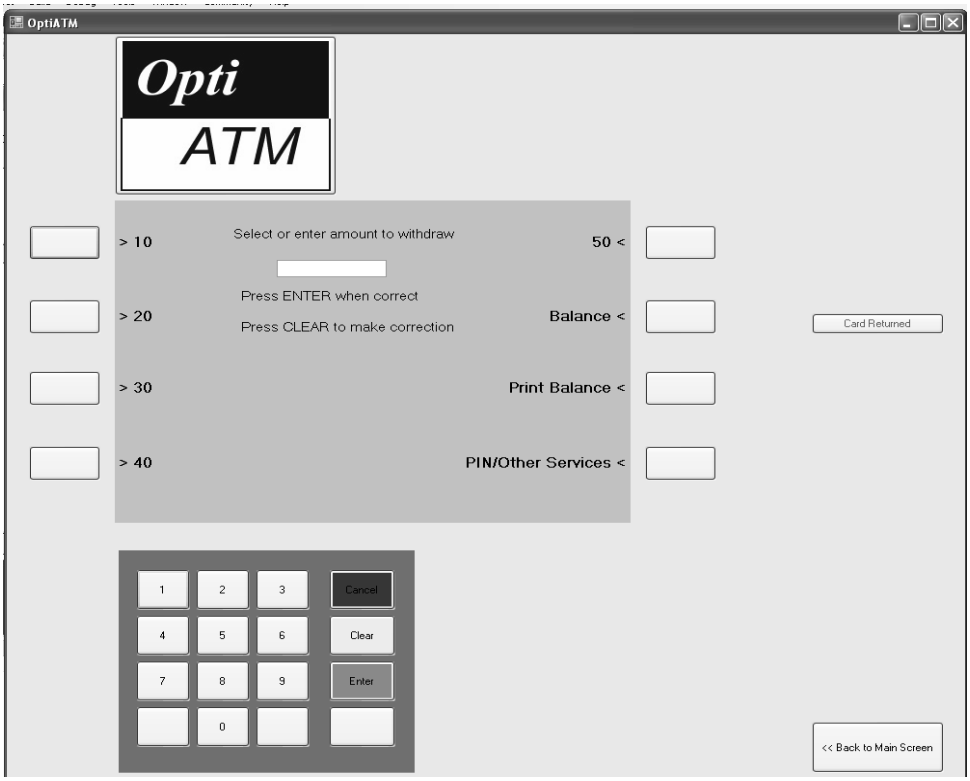


Figure 18. OptiATM User Options

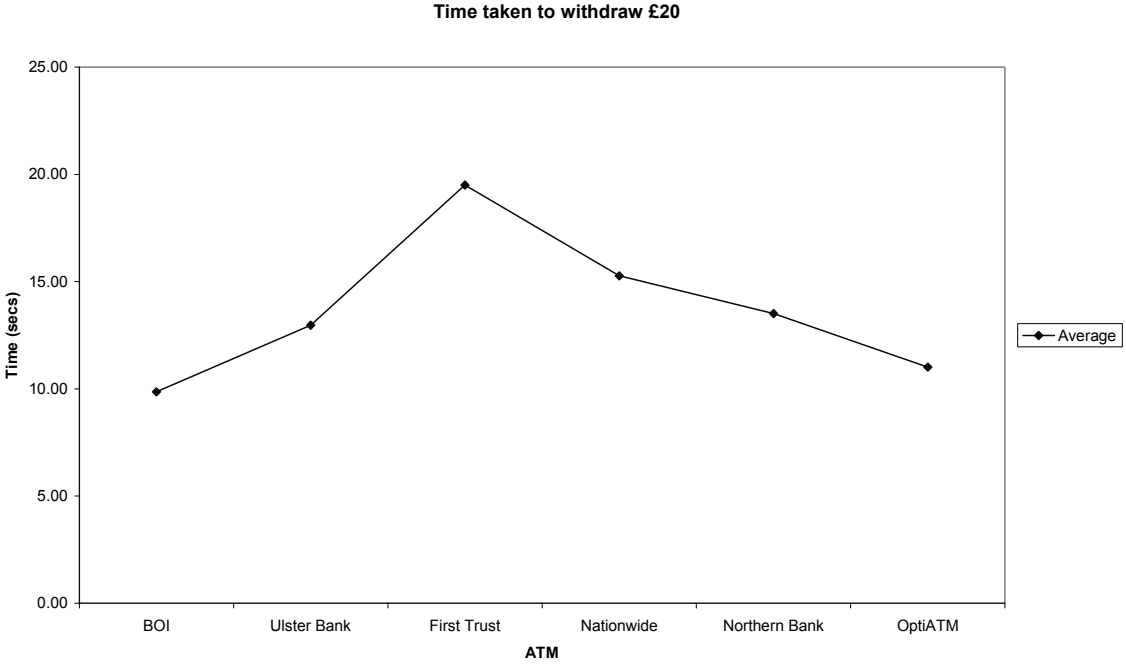


Figure 19. Comparing OptiATM - Withdraw £20

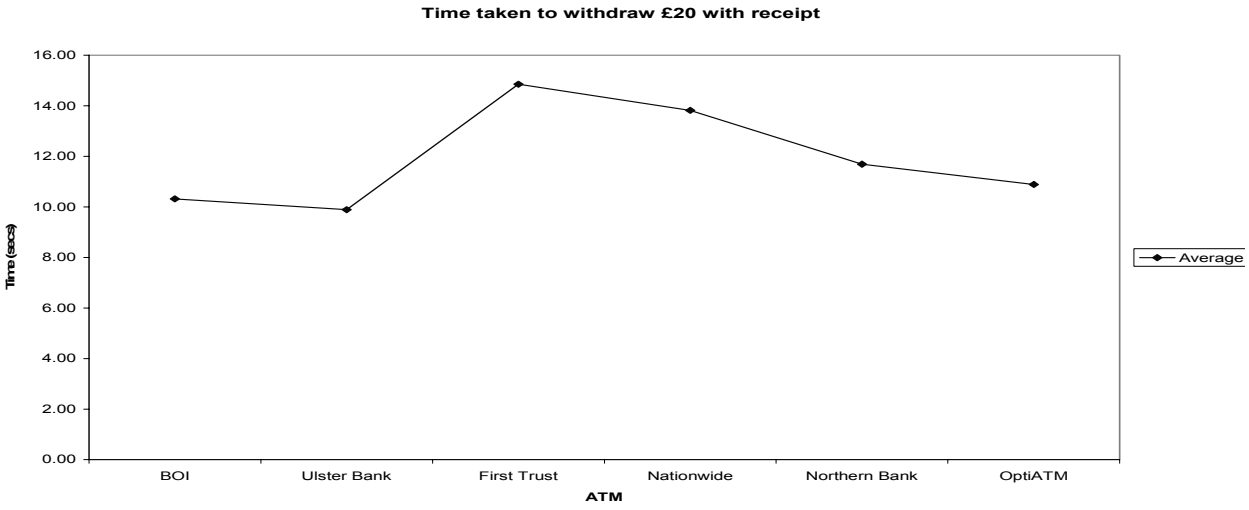


Figure 20. Comparing OptiATM - Withdraw £20 with receipt

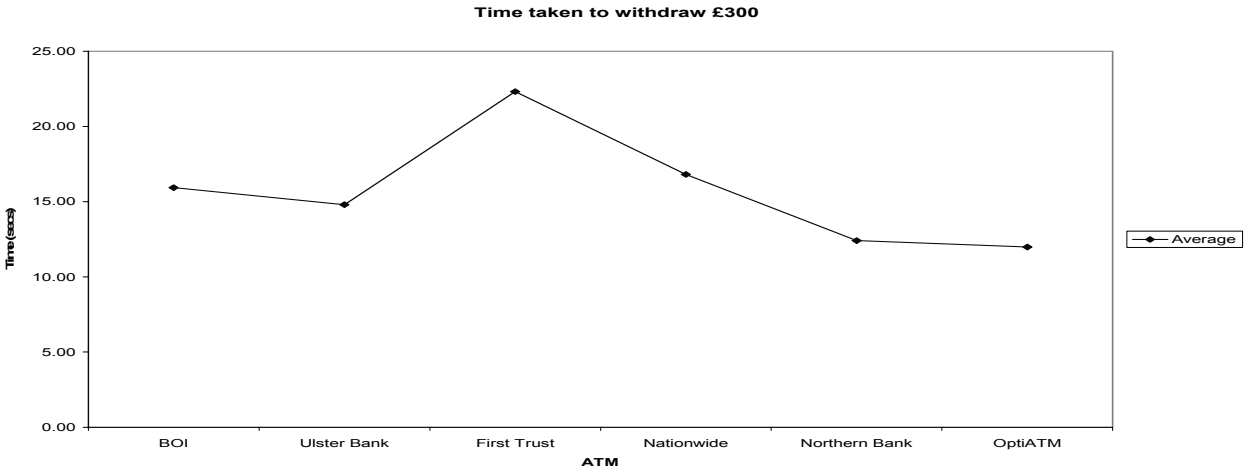


Figure 21. Comparing OptiATM - Withdraw £300

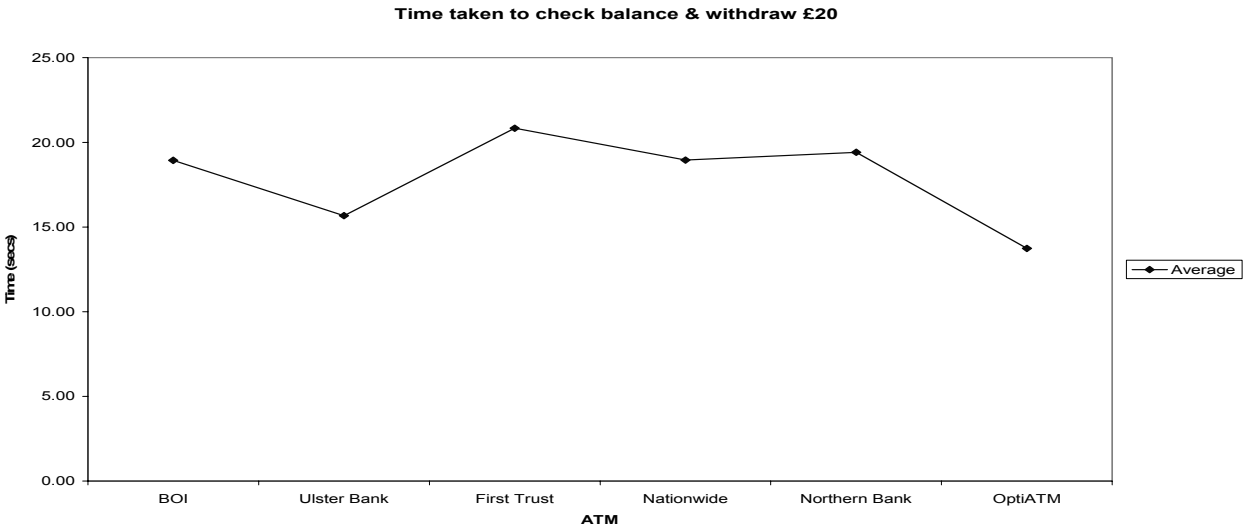


Figure 22. Comparing OptiATM – Check balance & withdraw £20

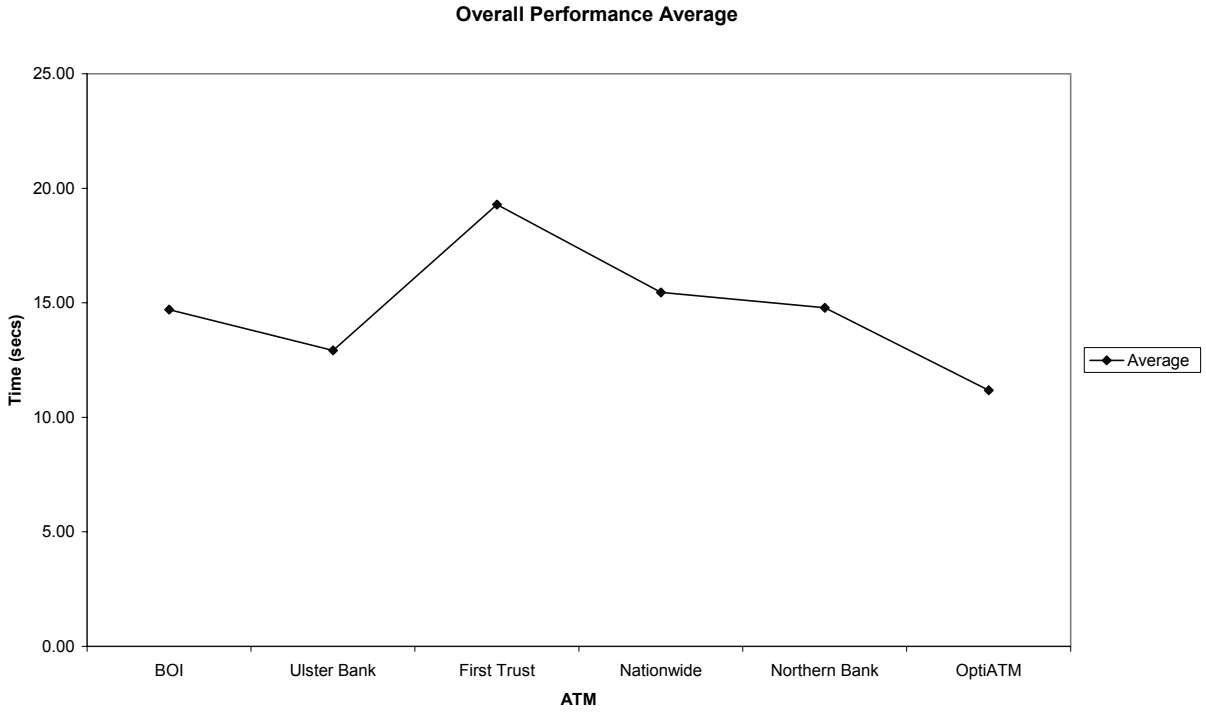


Figure 23. Average of overall transaction times for each ATM system

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

Application of Particle Swarm Optimization Algorithm to Electric Power Line Overhaul Plan

Jia Liu, Yang Li (Corresponding author) & Liqun Gao

School of Information Science and Technology

Northeastern University

3-11 Wen Hua Road, Shenyang 110004, China

E-mail: liyang@ise.neu.edu.cn

Abstract

Electric power line overhaul plan is an important issue on power system and engineering practice. As particle swarm optimization is to be a new intelligent algorithm. It is gradually applied into power system these years. This paper provides a relative mathematical model to solve the problems in power line overhaul. Particle swarm optimization algorithm has advantages of less parameters setting and highly optimizing speed. Take these advantages; we make an optimal order of electric power line overhaul. It is proved that the application of particle swarm optimization algorithm to electric power line overhaul plan is feasible.

Keywords: Electric power line overhaul plan, Particle swarm algorithm, Optimization algorithm, Power system

1. Introduction

With rapid expand of power system scales and implementation of electricity market reformation. To make the power system operate on a security, economic, stable and reliable way becomes more important while the security and economic factors becomes more complex. With different objections of planning and operation in power system, the choice of different control variables and constraints may propose different optimization. It needs the help of optimal theory. Especially with extensive search on intelligent heuristic algorithm which applied on power system. It supports a new method to solve the relative question in power system.

Particle Swarm Optimization (PSO) is one of the latest evolutionary optimization techniques developed by Eberhart and Kennedy PSO concept is based on a metaphor of social interaction such as bird flocking. This algorithm provides a robust, parallel processing approach to obtain the global best solution with higher probability. This method is easy to implement and fast to convergence. As its intelligence background, it can apply both to science research and engineering program. It aroused widespread interest as soon as it is proposed, and obtain a large number of achievements in a short time. It is used in many fields such as function optimization, neural network design, pattern recognition, signal processing, and robot technology and so on.

Electric power line overhaul is multi-objective and nonlinear complex optimization. The main task is to find out best overhaul programmed in order to reduce maintenance costs when the power grid happen to be failure. Electric power line overhaul plan is complex optimization for its time-varying, discretion, nonlinear and randomness. So far, this issue has not been satisfactorily resolved.

PSO in the application of power system is on a late start. It shows broad prospects in the field of power system and has begun to attract the attention of the scientist who works on power system. Especially with the establishment and improvement in electricity market, how to take advantages of PSO algorithm to solve the problem of power system will become hot.

2. The principle of PSO

PSO is an evolutionary algorithm which is initialized with a population of random solutions and searches for optima by updating generations. Each individual or potential solution, named particle, flies in the dimensional problem space with a velocity which is dynamically adjusted according to the flying experiences of its own and its colleagues. The particle evaluate their positions relative a global fitness during each generation and share memories of their best positions, and use those memories to adjust their own velocities and subsequent positions.

Suppose that the searching space is *D*-dimensional and *m* particles from the colony. The *i*th particle represents a *D*-dimensional vector $X_i(t) = (x_{i1}(t), x_{i2}(t), ..., x_{id}(t))$ in the searching space, the *i*th particle's "flying" velocity also a

D-dimensional vector, denoted as $V_i(t) = (v_{i1}(t), v_{i2}(t), ..., v_{id}(t))$ .Denote the best position of the *i*th particle as $P_i(t) = (p_{i1}(t), p_{i2}(t), ..., p_{id}(t))$, and best position of the colony as $P_g(t) = (p_{g1}(t), p_{g2}(t), ..., p_{gd}(t))$

The particles velocities and position are updated by the following equations:

$$V_{ik}(t+1) = wv_{ik}(t) + c_1 rand1(p_{ik}(t) - x_{ik}(t)) + c_2 rand2(p_{gk}(t) - x_{ik}(t)) \quad (1)$$

$$X_{ik}(t+1) = x_{ik}(t) + v_{ik}(t+1) \quad (2)$$

w is the weight parameter, it has a global search ability when w is large. It tends to be a local search when *w* is small. Generally, *w* =0.9 in at first and decrease to 0.4 along with the increasing of generation.c_1,c_2 are learning factors, $rand_1$ and $rand_2$ are random numbers between (0,1), i=1,2 , …,m; k = 1,2 , …,d. A particle's velocity in each dimension is clamped to a maximum velocity $V_{i,}$ and is set to a certain fraction of the range of the search space in each dimension.

3. Model of Electric Power Line Overhaul Plan

3.1 Knapsack problem

Knapsack problem is a typical optimization in operations research. It is applied to budgetary controlling, item selecting, materials cutting and goods loading and searched as sub problems. With the development of network, public-key system of knapsack plays an important role in public-key design .Knapsack problem can be described as follows:

$$\max f(x_1, x_2, ..., x_n) = \sum_{j=1}^{n} c_j x_j \qquad \text{j=1, 2… n} \quad (3)$$

$$\sum_{j=1}^{n} a_{ij} x_j \le b_j \qquad \text{i=1, 2… m; } x_j \in \{0,1\} \quad (4)$$

Where n is the number of objects, m is the number of resource, c_j is profit of object *j*, b_i is a budget for resource, a_{ij} is consumes of resource *i*, x_j is a 0/1 decision variables (when the object *j* is chosen, x_j=1, else x_j=0).

KP can be regarded as a resource allocation problem, where we have *m* resources a_{ij} (i=1,2,...,m) and *n* objects j(j=1,2,...,n). Each resource i has a budget b_i, each object j has a profit c_j and consumes a_{ij} of resource i. The problem is to maximize the profit within a limited budget. Knapsack problem can be derived from a series of related optimization problem.

Knapsack problem has wide applications in practice. Much organic combination with simple structure makes complex structure. The way to solve a complex problem can be easy when some simple problem can be solved. In the solution of complex optimization, Knapsack problem is always as a sub problem. It can get a better solution in optimization from the improvement of Knapsack problem.

3.2 Solution knapsack problem based on particle swarm optimization

In 0/1 knapsack problem, the variable *i* is set to be 1 if object *i* is placed in the knapsack. According to genetic algorithm, we can regard $x_1, x_2, ..., x_n$ as binary variables. It can convert to decimal as:

$$y_1 = x_1 \times 2^{n-1} + x_2 \times 2^{n-2} + ... + x_n \quad (5)$$

It can also be pressed as two integers:

$$y_1 = x_1 \times^{n/2-1} + x_2 \times 2^{n/2-2} + ... + x_{n/2} \quad (6)$$

$$y_2 = x_{n/2+1} \times 2^{n/2-1} + x_{n/2+2} \times 2^{n/2+2} + ... + x_n \quad (7)$$

We can decide the number

The solution of knapsack problem mapping from its solution space to some discrete points on one (or more)lines in Euclidean space. Next we can get the solution by using PSO. But we should note that the solution space is not continuous, so we need rounding operation in position computing. Through the test, we can get a better solution using this algorithm.

Generate 100 random numbers in range [0,1], which represent the value of weight. Take an average value when the test runs 10 times.

3.3 The combination of model and the actual

With the constraints in electric power line overhaul and the similarities between workload and working hours, we can apply the model of 0/1 knapsack problem to electric power line overhaul.

In order to reach the optimization objectives, we do some changes in the initial model of knapsack problem:

$$J = \min\left(b_i - \sum_{j=1}^{n} a_j x_j\right) \qquad j = 1,2,...,n \quad (8)$$

$$s.t.\sum a_j x_j \le b_i \qquad j=1,2,...,n \qquad i=1,2,...,m \tag{9}$$

b_i represents workload in ith month, x_j is a 0/1 decision variables(when the task is chosen, x_j=1,else x_j=0). a_j reprents the workload we need to finish ith task.

Formula (9) gives the constraints of electric power line overhaul to make actual workload less than that in plan. Formula (8) is a definition of objection function. J represents a difference between actual workload and planning workload per month.

We should consider the arrangements in actual work condition. For example, we assigned little task in cold winter like January to April. Otherwise we assigned more tasks in May and June.

The correspondence between original and improved model is that: there are three parameters in knapsack problem; they are size of knapsack, weight of object, value of object. In improved model, the weight and value of an object represent the workload. The sizes of knapsack represent the planning workload.

4. Experimental results

Though Matlab is widely used in simulation, the proposed PSO algorithm for electric power line overhaul plan was implemented in Matlab language. In main procedures of knapsack problem, we choose the data in March to June. The actual assignments are 100. We assigned little in March and April and a little more in May and June. The actual for every task is:

w=	3	2	3	5	7	8	6	4	5	3
	4	8	5	1	7	3	9	1	2	8
	4	7	3	2	6	8	1	7	5	6
	10	2	9	8	1	2	3	4	3	1
	6	4	7	5	8	4	3	1	5	4
	2	4	7	5	2	3	6	7	3	4
	4	6	8	3	4	7	5	1	9	3
	2	9	4	7	5	6	1	4	3	7
	2	8	3	4	2	1	6	4	2	3
	1	8	7	4	3	4	6	8	2	4

The second step is to number the task:

order=	1	2	3	4	5	6	7	8	9	10
	11	12	13	14	15	16	17	18	19	20
	21	22	23	24	25	26	27	28	29	30
	31	32	33	34	35	36	37	38	39	40
	41	42	43	44	45	46	47	48	49	50
	51	52	53	54	55	56	57	58	59	60
	61	62	63	64	65	66	67	68	69	70
	71	72	73	74	75	76	77	78	79	80
	81	82	83	84	85	86	87	88	89	90
	91	92	93	94	95	96	97	98	99	100

We can get the information about finished task and unfinished task by numbering the data.

The third step is to put the task which has been numbered into knapsack. Then get the best arrangement. The arrangements of each month are as follows:

m1=	4	7	8	9	14	16	18	19	22	23
	24	33	35	37	38	39	42	45	49	50
	52	60	62	64	65	66	77	79	81	83
	84	88	90	100						
m2=	1	13	20	25	29	30	32	40	51	56
	58	68	74	76	80	82	94	99		

m3=3 10 11 12 15 26 27 31 41 43
44 48 54 55 61 69 71 75 85 86
91 97
m4=2 5 6 17 21 28 34 36 46 47
53 57 63 67 72 73 78 89 92 93
95 98

Then deal with the data by improved knapsack problem model, we get:

The planning workload of March=b1=150

The actual workload of March= 136=49+3+4+3+4+8+5+4+4+4+6+7+5+6+1+3+2+3+4+4+3+4

The planning workload of April=b2=100

The actual workload of April =83=3+5+8+6+5+6+3+2+3+7+8+6+7+8+4+3+3

The planning workload of May=b3=120

The actual workload of May=100=3+4+8+8+1+10+6+7+5+1+5+2+2+9+4+5+3+7+7+2

The planning workload of June=b4=140

The actual workload of June =129=2+7+8+9+4+7+8+4+7+6+8+5+9+4+4+2+8+7+3+8+4+3+2

According to the data above, the workload in each month all satisfy the constraints but the effect are different.

The value of object function in March $J_1 = 150 - 139 = 11$

The value of object function in April $J_2 = 100 - 83 = 17$

The value of object function in May $J_3 = 120 - 100 = 20$

The value of object function in June $J_4 = 150 - 129 = 21$

5. Conclusion

This paper noticed the importance of electric power line overhaul which has some problems in power system, then proposed a knapsack problem model. Use PSO algorithm to solve the problem and simulate with Matlab. The result shows the correctness and feasible to solve the problem using PSO algorithm.

References

Kennedy J, Eberhart R. (1995). Particle swarm optimization. *Proceedings of IEEE Conference on Neural Networks*. Perth, Australia, 4, 1942-1948.

Sensarma P S, Rahmani M. (2002). A comprehensive method for optimal expansion planning using particle swarm optimization. *Proceedings of the IEEE Power Engineering Society Transmission and Distribution Conference*. New York, USA, 1317-1322.

Fukuyama Y. (2002). Modern Heuristic Optimization Techniques with Applications to Power Systems. *IEEE Power Engineering Society*. 45 – 51.

Pirlot P. M. (1996). General local search methods. *European Journal of Operational Research*. 92: 493-511.

Abe S. (1992). Solving inequality constrained combinational optimization problems by the Hopfield neural network. *Neural Network*. 5: 633-670.

FATIH M.A. (2003). Binary particle swarm optimization algorithm for lot sizing problem. *Journal of Economic and Social Research*. 5(2): 1-20.

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

An Efficient Method for Generating Optimal OBDD of Boolean Functions

Ashutosh Kumar Singh (Corresponding Author)

School of Engineering and Science

Curtin University of Technology

Saarwak Campus, CDT, 250, Miri, Malaysia

Tel: 60-85-443-939 E-mail: ashutosh.s@curtin.edu.my

Anand Mohan

Centre for Research in Microprocessor Applications (CRMA)

Department of Electronics Engineering, Institute of Technology

Banaras Hindu University

Varanasi- 221 005, India

Tel: 90-542-257-5272 E-mail: amohan@bhu.ac.in

Abstract

An efficient method of finding optimal (OBDD) of an *n* variable Boolean function is presented that offers a simple and straightforward procedure for optimal OBDD generation along with storage economy. This is achieved by generating *n*! fold tables and applying node reduction rules to each fold table directly instead of generating all *n*! OBDDs of the function.

Keywords: Fold table, Binary decision diagrams, Formal verification

1. Introduction

The pioneering work of Akers (1978) on graphical representation of Boolean functions using Binary Decision Diagrams (BDDs) offered an attractive and convenient technique for simplification and manipulation of complex Boolean functions. Modification of these BDDs was suggested by Bryant (1986) and since then different types of Decision Diagrams (DDs) have been introduced by researchers (see the book written by Sasao T. and Fujita M. 1996 for detail). The graphical representation of Boolean functions using BDDs have been potentially used for simplification of complex functions (Drecher R., Dreshler N. and Gunther W. 2000; Hong Y., et. al 2000; Scholl C. et. al 2000). As a result, BDDs and its variations are being extensively used in logic design, synthesis and testing of digital circuits (Jabir A. M., Pradhan D. K., Singh A. K. Rajaprabhu T. L. 2007; Minato S. 1996; Lai Y. T., Pedram M. and Vrudhula S. B. K. 1996; Gergov J. and Meinel C. 1994; Shen A., Devadas S. and Ghosh A. 1995). However, minimization of number of BDD nodes has been of focal interest in several applications such as formal verification. Ordered Binary Decision Diagram (OBDD) is an important form of decision diagram generated by imposing ordering relation among function variables such that the resulting form is canonical and it provides more compact representation of Boolean functions (Litan L. H. and Molitor P. 2000; Wang Y., Abd-el-Barr M. and McCrosky C. 1997; Wegener I. 1994; Bryant R. E. 1992; Liaw H. T. and Lin C. S. 1992; Friedman S. J. and Supowit K. J. 1990; Bern J., Meinel C. and Slobodova A. 1996). OBDD representation along with the use of the data structures for caching intermediate computations provides a way for the efficient implementation of many Boolean operations. However, one major drawback of OBDD representation is that identification of optimal (minimal node) OBDD requires generating all *n*! OBDDs of *n* variable function. The ordering of the function variables corresponding to optimal OBDD is called optimal ordering and for which many techniques are already reported (Wegener I. 1994; Bryant R. E. 1992; Bern J., Meinel C. and Slobodova A. 1996; Drechslor R., Becker B. and Gockel N. 1996) but they are inefficient in respect of computational complexity and storage requirements.

This paper describes a new efficient method for identification of optimal OBDD of a Boolean function without generating all *n*! OBDDs. The proposed method uses a set of fold table consisting of all *n*! ordering for *n* variable function and directly applying BDD reduction rules to each fold table without really generating decision diagrams for each possible ordering. Further, a new algorithm has been developed for fast generation of fold tables and the table consisting minimum number of nodes is used to generate optimal OBDD. Therefore it offers a simple and computationally efficient procedure for optimal OBDD generation along with storage economy.

2. Ordered Binary Decision Diagram (OBDD)

A Boolean function can be represented using OBDDs by imposing certain ordering relations among function variables where the canonicity of representation allows easy detection of many useful properties such as symmetry and unateness of variables.

Definition_1

Binary Decision Diagrams (BDDs) represents a Boolean functions as a rooted, directed acyclic graph with a vertex set containing two types of vertices, *non-terminal* and *terminal vertices*. A non-terminal vertex *v* has two attributes i.e. (*i*) an argument *index* (*v*) $\in \{x_1,\ldots.., x_n\}$ and (*ii*) two children indicated by *dashed* and *solid* lines for *low* (*v*) and *high* (*v*) respectively. A terminal vertex *v* has an attribute *value* (*v*) $\in \{0, 1\}$ and has no outgoing edge.

An un-simplified BDD is basically a Binary Decision tree which contains 2^{n-1} non-terminal nodes. Considering an example function $f_1 (x_0, x_1, x_2) = \sum(3, 5, 6, 7)$, its BDD is shown in figure 1 (b) which is a direct mapping of the truth table of figure 1(a) in the tree form. In this tree the value of function is determined by tracing a path from the root to a terminal vertex. A BDD representation of an '*n*' variable function will initially have 2^n-1 nodes[4, 12, 15] and the function value in the tree of figure 1(b) is determined by tracing a path from the root to a terminal vertex. The BDD can be further simplified using following node reduction rules (Bryant R. E. 1992; Drechslor R., Becker B. and Gockel N. 1996).

(i) Deletion Rules: If one or more non-terminal nodes are such that their both branches corresponding to "0" & "1" lead to a non-terminal successor node or to a terminal node then that non-terminal node can be deleted as shown in figure 2 (a).

(ii) Merging Rules: If two or more terminal (or non-terminal) nodes of the same label have the same "0" and "1" successors i.e. their left and right sons are equivalent then they can be merged in a single node shown in figure 2 (b).

The application of above reduction rules to the BDD of the function $f (x_0, x_1, x_2) = \sum(3, 5, 6, 7)$ gives simplified BDD given in figure 1 (c).

Definition_2

OBDDs are generated by imposing a total ordering "<" over the set of variables so that for any vertex *u* and either non-terminal child *v*; their respective variables must be ordered *var (u) < var (v)*. The OBDD generated using any ordering arrangement can be reduced to give simplified representation of a Boolean function. The OBDD shown in figure 1 is generated considering variable ordering $x_0 < x_1 < x_2$, however, in principle the variable ordering can be selected arbitrarily. Thus for a three variable function the total number of OBDDs can be 3! but the selection of an appropriate ordering is critical for efficient reduction of OBDD nodes.

Definition_3

The size of the OBDD is defined as the total number of terminal and non-terminal nodes in OBDD, for example, the size of the OBDD shown in figure 1 (c) is 6.

3. Effect of Variable Ordering

The nodal complexity of OBDDs for a given function greatly depends on variable ordering and hence it is possible that different OBDDs of same function can have different number of nodes. The identification of suitable ordering for generating OBDD of a function that has fewer nodes is not very crucial in the case of simple and medium complexity functions, however, for complex functions ($n \geq 5$) variable ordering has dramatic effect on the computational and storage requirements which directly effects the efficiency of the Boolean function manipulation algorithms in generating fewer node OBDD. Most applications requiring OBDD generation choose some ordering of the variables at the beginning and construct all possible OBDDs to identify the optimal OBDD having least number of nodes. This requires more computation as well as storage for *n*! OBDDs of an '*n*' variable function.

The effect of variable ordering on the number of OBDD nodes is demonstrated considering a six variable example function $f = x_0 \cdot x_3 + x_1 \cdot x_4 + x_2 \cdot x_5$ which will have 6! orderings and corresponding number of OBDDs. For simplicity, the OBDDs for only two out of total 6! orderings are shown in figure 3. The significance of variable ordering can be appreciated by observing the difference between the number of OBDD nodes for the two orderings of figure 3 (a) and (b). Although the difference between the numbers of nodes in the two OBDDs is only eight in our example but it may become extremely large for complex functions ($n \geq 5$). Therefore developing an efficient method for identification of appropriate variable ordering to generate *optimal OBDD* is an interesting problem to achieve minimization of storage requirements and reducing computations.

4. Identification of Optimal OBDD

This section describes a new algorithm for generating optimal OBDD of a Boolean function based on pre-calculation of nodes for each possible variable ordering. Some *definitions* that have been used are:

(1) If $I \subseteq \{0, 1, 2, \ldots, n\text{-}1\}$ then ϕ *(I)* is defined as the *set of ordering* on {0,1, ...,n-1}

ϕ *(I)* = $\{\phi : \phi$ is an ordering on $\{0, 1, 2, \ldots, n\text{-}1\}$

For example, an OBDD of a three variable function with ordering $x_1 < x_0 < x_2$ will be represented here as OBDD (1, 0, 2).

(2) A *fold table* symbolized by $TABLE_I$ is constructed corresponding to each of the $n!$ variable orderings in which the entries in each table are made according to the sequence of minterm values of the given Boolean function. The total number of fold tables will be $n!$ and generating all of them becomes a tedious work for large number of variables. For a particular ordering "ϕ" the table is denoted by $TABLE_\phi$.

(3) If $v \in \{0, 1, 2, \ldots,(n\text{-}1)\}$ and ϕ is an ordering on $\{1, 2, \ldots, (n\text{-}1)\}$ $value_v(\phi)$ denotes the number of nodes on label v in the OBDD (ϕ).

Therefore the task of identifying the optimal OBDD can be simplified if an ordering "ϕ" is determined using simplified procedure such that it minimizes $\sum_{v=0}^{n-1} value_v(\phi)$.

4.1 Fold Table Generation

If $f(x_0, x_1, \ldots\ldots, x_{n-1})$ is an "n" variable Boolean function, where $x_i \in \{0, 1\}$ and $i = \{0, 1, \ldots\ldots, (n\text{-}1)\}$ then it shall have 2^n ordered n-tuples. The function f assumes a particular value for each of n-tuples which may be considered as defining n-bit unsigned binary integer having decimal value (d) in the range 0 to 2^n-1. The relation between the unsigned binary integer and its corresponding decimal value can be expressed as:

$$(x_{n-1}, x_{n-2}, \ldots\ldots\ldots, x_1, x_0) \Leftrightarrow \sum_{l=0}^{n-1} 2^i x_i = d \tag{1}$$

If the value of the function y corresponding to decimal value d of an n-tuple is expressed as y_d then $f(x_0, x_1, \ldots\ldots\ldots, x_{n-1}) = y_d$ and the function values $(y_0, y_1, y_2 \ldots\ldots y_{2^n-1})$ define a finite sequence. The fold table generation involves interchanging input variables that changes decimal value of n-tuples reordering of function value and therefore it can be viewed as occurrence of finite sequences. Considering that x_i and x_j are two function variables that are to interchanged to generate new decimal value d' of an n-tuple then

$$d' = \sum_{\substack{l=0 \\ l \neq i,j}}^{n-1} x_l 2^l + x_j 2^i + x_i 2^j \tag{2}$$

Equation (2) can be rewritten as:

$$d' = \sum_{\substack{l=0 \\ l \neq i,j}}^{n-1} x_l 2^l + x_j 2^i + x_i 2^j + x_i 2^i + x_j 2^j - x_i 2^i - x_j 2^j \tag{3}$$

Rearranging equation (3) and putting the value of d from equation (1) we get

$$d' = d + x_j 2^i + x_i 2^j - x_i 2^i - x_j 2^j \tag{4}$$

Therefore equations (1) and (4) can be used to determine all 2^n sequences of entries for all $n!$ tables of the fold table for an n-variable function by changing the positions of two variables at a time.

4.2 Node Reduction using Fold Table

This subsection discusses the method of finding a particular variable ordering for generating optimal OBDD of a given Boolean function by direct application of reduction rules to the fold table instead of generating OBDDs. The proposed method exploits the property that the value of the variables on the first k labels depends only on their ordering and not on the ordering of the remaining (n-k) variables[14] for recording entries in the fold table by considering each k labels ($k \leq n$). The "0" and "1" values of the function are stored as w_0 and w_1 respectively and for each such pair (w_0, w_1) it is determined whether or not a new node is required. This is achieved using following two criterions that are directly related to the *deletion* and *merging* rules:

(i) If $w_0 = w_1$ then do not create a new node since its both branches (0 & 1) point to the same vertex (deletion rule).

(ii) If there are m nodes having $w_0 \neq w_1$ then at the same label and if their left sons and right sons are equivalent then don't create new node since it would be equivalent to m (merging rule).

Otherwise create new node if both the above criterions are violated.

4.3 Algorithm for Optimal OBDD Generation

Optimal OBDD generation algorithm requires following input parameters:

(1) Fold Table *($TABLE_I$)*; which is actually a mapping from $(0, 1)^{n-k}$, where $k = |I|$, $v \in I$ and also satisfying $\phi[k] = v$ to those nodes of OBDD (ϕ) that are either internal nodes labeled with the member of I or terminal nodes ("0" or "1").

(2) $\sum_{v=0}^{n-1} value(\phi)$, which is total number of nodes for each ordering, where $\phi \in \phi(I)$.

The generation of fold table "$TABLE_I$" is achieved using equations (3) and (4) for each ϕ, where $\phi \in \phi(I)$ and computation of total number of nodes for each ordering. This is achieved by considering $TABLE_\phi$ for a particular ordering and storing each k label of $TABLE_\phi$ with paired function value (w_0, w_1). The number of such pairs is given by $2^{\{(n-1)-k\}}$ for $k \le n$ and applying node generation criterion on the paired values (w_0, w_1) to compute total number of nodes. This process is iterated until all orderings have been considered and the specific ordering "ϕ" that gives the minimum value of $\sum_{v=0}^{n-1} value(\phi)$ is identified as the *optimal ordering* corresponding to optimal OBDD of the function. Therefore the algorithm for identification of optimal variable ordering can be given as in figure 4.

The application of the algorithm given in figure 4 can be illustrated considering an example function $f(x_0, x_1, x_2, x_3) = x_1x_3 + x_0x_1'x_2'$ for ordering $TABLE_{(3, 2, 1, 0)}$. Now computation of the "*value*" of each label for the selected ordering can be achieved by assigning k=0, 1, 2 or 3 for generating paired output for each assignment of k as given in the fold table 1. Recalling that total number of (w_0, w_1) pairs would be $2^{\{(n-1)-k\}}$, the number of (w_0, w_1) pairs would be 8, 4, 2 and 1 for k=0, 1, 2 and 3 respectively. Analyzing table 1 (a) it is found that out of total 8 (w_0, w_1) pairs; the first, second, fifth and sixth pairs have $w_0 = w_1$ and thus creation of OBDD nodes is not required for these pairs [criteria 4.2 (*i*)]. The remaining pairs are equivalent and hence they all can be represented using only one node in OBDD [criteria 4.2 (*ii*)]. Similarly out of total four pairs in table 1 (b) the first, second and fourth pairs don't need any node in OBDD representation, however, one node would be necessary for third pair. Finally, there is no scope of node reduction in tables 1 (c) and 1 (d) because their pairs are not covered by either of the node reduction criterions. Therefore they require as many numbers of nodes as the number of pairs and thus the OBDD for ordering $\phi(3, 2, 1, 0)$ will have total "five" nodes as shown in figure 5.

Now if the labels in OBDD are considered such that the first label starts from bottom corresponding to $v = 0$ and the subsequent higher labels are obtained by moving up to root node then it is clear that for the first label of the OBDD "4" nodes are deleted and "3" nodes are merged into one node. Further, "3" nodes are deleted at the second label (for v=1) while no node reduction is possible for third and fourth labels. Similarly the total number of OBDD nodes corresponding to remaining orderings of the function variables can be pre-calculated without really generating the OBDDs. The variable ordering that gives minimum number of OBDD nodes is selected to generate optimal OBDD of the function.

The illustrated method of optimal OBDD generation is equally applicable to all Boolean functions without necessitating actual generation of the OBDDs of a function. Therefore the proposed method offers both computational simplicity and storage economy in generating optimal OBDD which makes it more attractive for optimal OBDD generation of complex functions.

5. Conclusion

This paper described a new computationally efficient method for generating optimal OBDD of complex Boolean functions without generating all n! OBDDs of an 'n' variable function. Our proposed method uses fast generation of *fold table* which is used to compute the total number of OBDD nodes for each possible orderings of function variables. Subsequently, the particular ordering corresponding to minimum number of nodes in OBDD is selected to generate optimal OBDD. The suggested algorithm eliminates the necessity of really generating all OBDDs of the function. Therefore it is computationally efficient as well as economical in storage requirements which make it suitable for manipulation of complex Boolean functions.

References

Abusaleh M. Jabir, Dhiraj K. Pradhan, Ashutosh K. Singh Rajaprabhu T. L. (2007). "A Technique for Representing Multiple Output Binary Functions with Applications to Verification and Simulation" *IEEE transaction on computer*, vol. 56, no. 8, pp. 1133-1145.

P. W. Chandana Prasad, M. Maria Dominic and Ashutosh Kumar Singh (2003). "Improved Variable Ordering for ROBDD's", *Proc. of ICADL*'03, pp. 544-547, December 8-11, Kuala Lumpur, Malaysia.

Drecher R., Drechsler N. and Gunther W. (2000). "Fast Exact Minimization of BDD's," *IEEE Trans. CAD of Integrated Circuits and Systems*, vol. 19, no. 3, pp. 384-389.

Hong Y., Beerel P. A., Burch J. R., and McMillan K. L. (2000). "Sibling-Substitution-Based BDD Minimization Using Don't Cares," *IEEE Trans. CAD of Integrated Circuits and Systems*, vol. 19, no. 1, pp. 44-54.

Scholl C., Moller D., Molitor P. and Drechler R. (2000). "BDD Minimization Using Symmetries," *IEEE Trans. Comput.,* vol. 18, no. 2, pp. 81-99.

Litan L. H. and Molitor P. (2000). "Least Upper Bound for the Size of OBDDs Using Symmetry Properties," *IEEE Trans. Comput.*, vol. 49, no. 4, pp. 360-368.

Wang Y., Abd-el-Barr M. and McCrosky C. (1997). "An Algorithm for Total Symmetric OBDD Detection," *IEEE Trans. Comput.,* vol. 46, no. 6, pp. 731-733.

Bern J., Meinel C. and Slobodova A. (1996). "Global Rebuiliding of OBDD's Avoiding Memory Requirement Maxima", *IEEE Trans Comput.,* vol. 15, no. 1, pp. 131-134.

Minato S. (1996). "Fast Factorization Method for Implicit Cube Set Representation," *IEEE Trans. CAD. of Integrated Circuits and Systems,* vol. 15, no. 4, pp. 377-384.

Lai Y. T., Pedram M. and Vrudhula S. B. K. (1996). "Formal Verification Using Edge-Valued Binary Decision Diagrams," *IEEE Trans. Comput.,* vol. 45, no. 2 pp. 247-255.

Drechslor R., Becker B. and Gockel N. (1996). "Genetic Algorithm for Variable Ordering of OBDDs," *IEE Proc. Comput.-Digit. Tech.,* vol. 143, no. 6, pp. 364-368.

Sasao T. and Fujita M., (1996). "Representations of Discrete Functions", Kluwer Academic Publisher.

Shen A., Devadas S. and Ghosh A. (1995). "Probabilistic Manipulation of Boolean Functions Using Free Boolean Diagrams," *IEEE Trans. CAD. of Integrated Circuits and Systems,* vol. 14, no. 1, pp. 87-95.

Gergov J. and Meinel C. (1994). "Efficient Boolean Manipulation With OBDD's can be Extended to FBDD's," *IEEE Trans. Comput*., vol. 43, no. 10, pp. 1197-1208.

Wegener I. (1994). "The Size of Reduced OBDDs and Optimal Read Once Branching Program for almost all Boolean Functions," *IEEE Trans. Comput.,* vol. 43, no. 11, pp. 1262-1269.

Bryant R. E. (1992). "Symbolic Boolean Manipulation with Ordered Binary-Decision Diagrams," *ACM computing surveys* vol. 24, no. 3, pp. 293-318.

Liaw H. T. and Lin C. S. (1992). "On the OBDD–Representation of General Boolean Functions," *IEEE Trans. Comput.,* vol. 41, no. 6,, pp. 661-664.

Friedman S. J. and Supowit K. J. (1990). "Finding the Optimal Variable Ordering for Binary Decision Diagram," *IEEE Trans. Comput*., vol. 39, no. 5, pp. 710-714.

Bryant R. E. (1986). "Graph Based Algorithm for Boolean Function Manipulation," *IEEE Trans. Comput.,* vol. C-35, no. 8, pp. 677-691.

Akers S. B. (1978). "Binary Decision Diagrams," *IEEE Trans. Comput.,* vol. C-27, no. 6, pp. 509-516.

Table 1. Fold Table of the Function $f(x_0, x_1, x_2, x_3) = x_1x_3 + x_0x_1'x_2'$

(a)

x_0	f	
0	0	w_0
1	0	w_1
0	0	w_0
1	0	w_1
0	0	w_0
1	1	w_1
0	0	w_0
1	1	w_1
0	1	w_0
1	1	w_1
0	0	w_0
1	0	w_1
0	0	w_0
1	1	w_1
0	0	w_0
1	1	w_1

(b)

x_1	f	
0	0	w_0
0	0	
1	0	w_1
1	0	
0	0	w_0
0	1	
1	0	w_1
1	1	
0	1	w_0
0	1	
1	0	w_1
1	0	
0	0	w_0
0	1	
1	0	w_1
1	1	

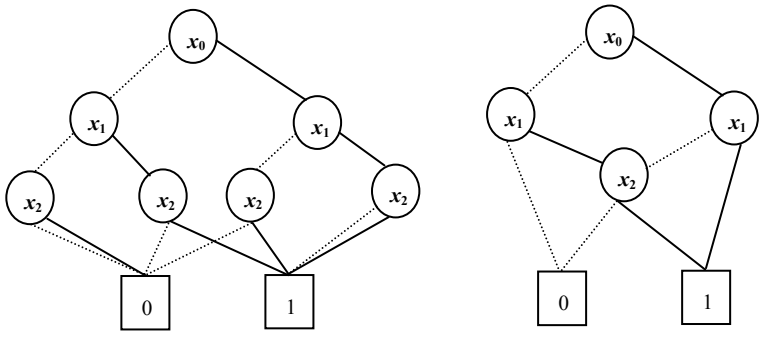
(c)

x_2	f	
0	0	w_0
0	0	
0	0	
0	0	
1	0	w_1
1	1	
1	0	
1	1	
0	1	w_0
0	1	
0	0	
0	0	
1	0	w_1
1	1	
1	0	
1	1	

(d)

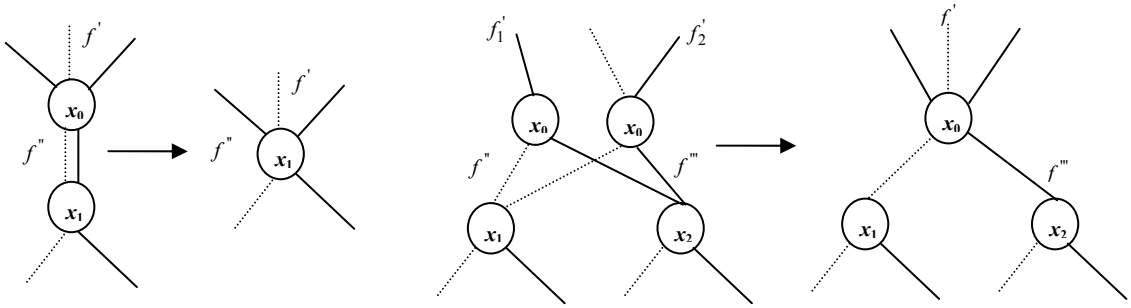
x_3	f	
0	0	w_0
0	0	
0	0	
0	0	
0	0	
0	1	
0	0	
0	1	
1	1	w_1
1	1	
1	0	
1	0	
1	0	
1	1	
1	0	
1	1	

x_0	x_1	x_2	f
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1



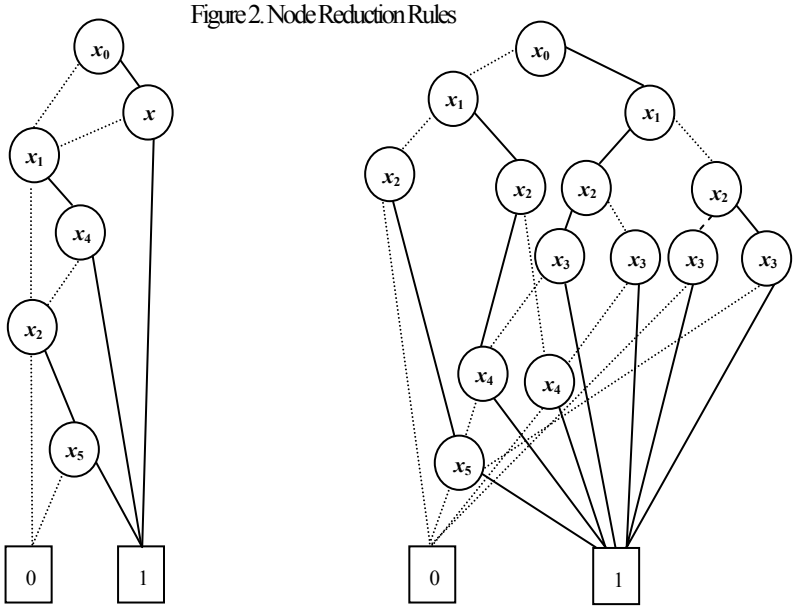
(a) (b) (c)

Figure 1. Truth Table and BDD of $f_1(x_0, x_1, x_2) = \sum(3, 5, 6, 7)$



(a) Deletion Rules (b) Merging Rules

Figure 2. Node Reduction Rules



(a) (b) $x_0 < x_1 < x_2 < x_3 < x_4 < x_5$

Figure 3. Effect of Variable Ordering on OBDD Nodes for $f = (x_0 \cdot x_3 + x_1 \cdot x_4 + x_2 \cdot x_5)$

For any ϕ; $\phi \in \phi(I)$

[Compute $\sum_{v=0}^{n-1} value(\phi)$]

BEGIN

Store first label (k) with paired output values starting from 0

If $w_0 = w_1$

Then count ← count (do not create new node)

Else Begin

If $w_0 \neq w_1$ and their left and right sons are equivalent

Then count ← count (do not create new node)

Except Both case count +1 ← count (create new node)

Store this count value as $\sum_{v=0} value(\phi)$

Store next label with increment k by 1 till $k \leq$ n-1 and repeat above procedure

Store $\sum_{v=0}^{n-1} value(\phi)$

Similarly for each ϕ; $\phi \in \phi(I)$

[Compute $TABLE_{\phi}$, and $\sum_{v=0}^{n-1} value(\phi)$]

Store $\sum_{v=0}^{n-1} value(\phi)$ for all ϕ and selected that ϕ which provides minimum value of $\sum_{v=0}^{n-1} value(\phi)$ that particular ϕ will be *optimal*.

Figure 4. Algorithm for Identification of Optimal Variable Ordering

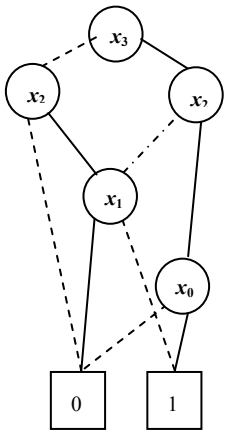


Figure 5. OBDD (3, 2, 1, 0) for $f(x_0, x_1, x_2, x_3) = x_1 x_3 + x_0 x_1' x_2'$

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

The Application of PLC in the Auto-control System of Chemical Makeup Water Treatment of Boiler

Zheng Lai

College of Electronic Science and Engineering

Jilin University

Changchun 130012, China

E-mail: laizheng@email.jlu.edu.cn

Abstract

With the deepening of China modernization course, the application of Programmable Logic Controller (PLC) is used more and more abroad in China. This article adopted the Quantum PLC and iFix industrial configuration software to study the application of PLC in the auto-control system of chemical makeup water treatment of boiler.

Keywords: Chemical water treatment of boiler, PLC, Application

1. Chemical water treatment of boiler

Because the boiler is developed to the trend of large size, high temperature and high pressure, so the requirement of water quality for the boiler makeup water is higher and higher, and that needs to almost get rid of all salts in the water. So anion-cation exchange technology has been developed quickly. For example, use the H cation exchanger to exchange various cations in the water and give out H^+, and use OH cation exchanger to exchange various anions in the water and give out OH^-. Through the exchange treatment of anion-cation exchanger, all salts in the water will be basically ridded. This method is called water chemical salt ridding treatment. When the raw water gets across the cation exchanger, cations such as Ca^{2+}, $M g^{2+}$, K^+ and N^{a+} in the water will be absorbed by the exchanger and commutative H^+ on the exchanger will be replaced to the water, and create corresponding inorganic acid with anions in the water. This sort of cation exchanger is also called as the cation-bed.

When the water with inorganic acid gets across the anion exchanger, anions such as SO_4^{2-}, Cl^- andHCO^{3-} in the water will be absorbed by the exchanger, and the commutative OH^- on the exchanger will replaced to the water and combine with H^+ in the water to compose H_2O. This sort of mixed ion exchanger is also called as mixed-bed.

When the resin in the mixed bed is disabled, the anion resin and cation resin will be separated first and regenerated respectively afterward.

Because the use quantity of boiler makeup water is large and the requirement of water quality is high, so the mixed bed salt ridding treatment is always used in series after 1st class salt ridding system to further purify the water quality.

Anti-infiltration is a sort of new membrane separation technology and a sort of process which depends on the anti-participation membrane to separate the solvent and solute in the liquor under the pressure.

The so-called “infiltration” is a sort of physical phenomenon. When separate two sorts of water with different concentration salts by a piece of half-infiltrative membrane, the water with lower salts will infiltrate to the water with higher salts through the membrane, but the salts will not be infiltrated, so the salt concentrations will be amalgamated until the concentrations are equal. However, this process needs long time, so this process is also called as natural infiltration. If add a pressure on the side with higher salt concentration, it will stop the above infiltration, and this pressure is called as infiltration pressure. If the pressure is increased, it will make the water infiltrate to the opposite direction and leave the salts. Therefore, the anti-infiltration salt ridding treatment is to inflict more pressures than the pressure of natural infiltration to make the infiltration implement to the opposite direction, and press the water molecule in the raw water to the other side and create clean water and achieve the intention getting rid of salts in the water.

2. Control principle and composing configuration of program control system

The whole control system is composed by locale measurement meters (including electric meter, thermal meter and component analysis meter), electromagnetic valve chest and computer system.

2.1 Control principle

The control system adopts the form of PLC controller + CRT upper monitor, i.e. on the one hand, PLC controller can communicate with the upper computer through DH+ network, send the computed and treated transmittal data to display

on the upper computer and accept the control instruction from the upper computer, on the other hand, PLC controller communicate through the remote I/O scanner and the I/O module of remote computer frame, receive signals returned from the locale and send out control instructions, and drive local equipments through control logic. The upper monitor sends out operation instruction, exports various instructions from the upper monitor to the locale equipment through logical operation of PLC controller, and uploads work parameters such as water quality parameter, flux, pressure, temperature, liquid position and operation state of the equipment at the present time to display on the upper monitor.

2.2 Configuration of hardware

The core of PLC is CPU which has functions of memory and logical operation and likes as human brain. It inputs the data collected on the locale to CPU through the input module, transmits the information needed by the control objective through artificial compilatory trapezium logical control figure to the output module by CPU, and automatically control the equipment through the output points of the output module.

The input module is divided into analog input module (AI) and digital input module (DI). The analog signals of 4~20mA such as flux, electric conductivity, acid-base concentration are transmitted to AI. The feedback signals of valves and pumps are generally transmitted to DI. The output module is divided into analog output module (AO) and digital output module (DO). The feed quantity of the feed equipment is adjusted by the points of AO, and the valves and pumps are controlled by the points of DO.

The reconstruction of the trunk network of chemical water treatment control system adopts the star network and Quantum PLC and two sets of iFIX operation stations (one is also the engineer station) which are connected with exchanger to compose TCP/IP Ethernet, and this network is easy to be used, configured and extended. The whole control system includes 3 sets of control computer chests (I/O computer chest) and 1 set of power distribution chest (including 380V/110V AC transformer and dual power supply switch equipment). The control system PLC belongs to Quantum products made by Schneider Company, which controls the automatic regeneration, half-automatic regeneration and manual regeneration of the chemical water treatment system, and the configuration table of I/O point is seen in Table 1.

The concrete listing of PLC configuration includes following contents.

(1) The CPU of PLC adopts the Quantum53414A of Schneider Company, and its processor is 32 bits and its EMS memory is 4MB.

(2) DI adopts 32 points module 140DDB5300, 24VDC and 4 groups of insulation.

(3) DO adopts 32 points voltage output module 140DDO353000, 24VDC, and the capacity of every channel is 0.5A.

(4) AI adopts 16 points module 140ACI0400 and the signals of 4~20mA.

(5) AO adopts 8 points module 140ACO13000 and 4~20mA.

(6) Configure with a piece of TCP/IP Ethernet communication port module 140NOE77100.

(7) This PLC is the newest product, and the module of CPU has main frequency of 96MHz and user logic memory of 1M.

The monitor computer adopts DELL P4/3.0G CPU, 512M EMS memory, 21 inch high differentiation color LCD which can offer clear, high contrast, high definition figures and tables that are easy to be read by operators (definition 1600× 1200 image), and multimedia anti-magnetic sound box. And the exchanger adopts the products of Hirschmann.

2.3 Configuration of software

The program software adopts Concept V2.5 which supports five program languages of IEC1131-3 including standard sequence flow chart, function chart, trapezium chart, structured text and instruction list. Because of the characters of chemical water treatment technology, most controls are sequence controls, and adopt convenient and applied SFC sequence flow chart.

The configuration software adopts iFIX V3.5, and communicated with Quantum PLC by the driver of MBE. At the same time, combining with PLC control program, it can realize following functions.

(1) Display the operation of system equipment and display the parameters of operation.

(2) Set up program control step sequence.

(3) Display the fault and warning of system equipment.

(4) Inquire operation report forms, operation records, historical curves and accident records.

2.4 Electromagnetic valve chest

In the actual operation and regeneration processes of the program control system, except for necessary interlock control, the control systems of the activated carbon filter, cation bed, middle water tank, salt ridding water tank, anion bed,

acid-base measurement system are respectively a relative independent system, so one electromagnetic valve control chest is equipped around every bed body, and the operation of knobs on the control panel can operate various valves of the bed body. When the automatic position is selected, the instruction of PLC will control the valves through the switch valve in the electromagnetic valve chest, and the program control system is equipped 7 spot electromagnetic valve chests.

2.5 Locale measurement meters

To ensure the operation and regeneration of water treatment, the program control system equips corresponding meters to the real-time measure various main parameters and the configuration of meters are seen in Table 2.

3. Conclusions

According to the actual statistics, the technology of water treatment directly decides the normal operation of industrial equipment, and many tube explosion accident of industrial boiler were induced by bad water quality. With the enhancement of industrial automatization level and the extensive application of computer in the operation management domain, the computer monitor technology is highly regarded in the domain of water treatment. The water treatment monitor system integrates inspection online, fault diagnosis, computer management, assistant decision and other technologies, which largely enhances the automatization degree of water treatment, and reduces faults of manual factor, and further increase the water saving efficiency.

Using the technology principle of water treatment and the data collection of sensor, the application of PLC in the auto-control system of chemical makeup water treatment of boiler not only can contribute the management of industrial water making, but can properly adjust the technical parameters according to the water quality to make system pressure stable, equipment operation stable, fault ratio largely reduced, and water making cost reduced to large extents. Through this system is still in the perfection stage, but the actual application has important meanings to enhance the reliability of the operation, and this system has high economic benefits and social benefits, large research and development potentials and application foreground.

With the deepening of China modernization course, the application of Programmable Logic Controller (PLC) is used more and more abroad in China. China should expedite the research and development of PLC products and the cultivation of talents to make various industries can expertly grasp the technology of PLC and drive the development of China economy.

References

Chang, Dounan. (2002). *Principle, Application and Experiment of PLC.* Beijing: China Machine Press.

Chinese Society of Power Engineering. (2002). *Technical Manual of Fuel-Burning Power Plant Equipments.* Beijing: China Machine Press.

Liao, Changchu. (2002). *Programming and Application of PLC.* Beijing: China Machine Press.

Li, Peiyuan. (2000). *Water Treatment and Water Quality Control for Thermal Power Plant.* Beijing: China Power Press.

Liu, Zhiyuan. (1999). *PLC and Its Application in the Power Plant.* Beijing: China Power Press.

Song, Deyu. (2005). *Design Technology of PLC Principle and Application System.* Beijing: Metallurgical Industry Press.

Tian, Ruiting. (1994). *Application Technology of PLC.* Beijing: China Machine Press.

Wang, Tingyou. (2005). *Principle and Application of PLC.* Beijing: National Defense Industry Press.

Wang, Weibing. (2002). *The Principle and Application of PLC.* Beijing: China Machine Press.

Wang, Zhaoyi. (1993). *Tutorial of PLC.* Beijing: China Machine Press.

Yu, Qingguang. (2001). Applications of Programmable Logic Controller (PLC) in Large Scale Thermal Power Plant. *Chinese Journal of Scientific Instrument.* No. 22(3).

Zhou, Weidong. (1999). Auto-control Technique for Chemical Feed Water Treatment of Boilers. *Hubei Chemical Industry.* No. 16(6).

Table 1. Configuration table of I/O points

	DI	DO	AI	AO	Total
Actual use	370	299	51	8	728
Actual configuration	416	352	64	12	848
Provided quantity of module	13	11	4	3	31
Point quantity of every module	32	32	16	4	84

Table 2. Configuration of measurement meters

Meter name and type	Factory	Quantity	Purpose
EJA pressure flux transducers	Chengdu Henghe Chuanyi Electromechanical Equipment Co., Ltd	24	Be used to monitor input/output fluxes and backwash fluxes of various bed bodies and regeneration systems
Static pressure level meter- YU 21	Wuhan Guangming Meter Factory	3	Be used to monitor liquid positions of salt ridding water tank and middle water tank
Electric conductivity meter- 9782C	HNOEYWELL	4	To monitor water qualities of various bed bodies and take them as the proof to judge whether the bed body is disabled
Micro natrium meter- 079721-2	HNOEYWELL	1	To monitor the operation of cation bed and take it as the proof to judge whether the cation bed is disabled
PH meter- 9782PH/ORP	HNOEYWELL	3	To monitor the water quality of salt ridding water and neutralization pool waste water treatment
Acid-base meter- 9782	HNOEYWELL	4	To monitor the concentration of regeneration liquid when the cation bed and anion bed are regenerated
Magnetic reversible level meter- UHZ257-57/C	Shanghai Yuanwang	10	Be used to monitor the liquid position of various liquor tanks of mordant acid-base measurement system

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

Seat Booking System for a Cineplex

J. Condell, J. McDevitt, D. McGilloway, J. McGlinchey & G. Galway

School of Computing and Intelligent Systems

Faculty of Computing and Engineering

University of Ulster at Magee

Northland Road, Londonderry, N.I., U.K.

Tel: 44-028-7137-5024 E-mail: j.condell@ulster.ac.uk

Abstract

This paper describes a cinema seat booking system and its design. The idea behind the system is to allow public kiosks to be installed in local shopping areas, where members of the general public could pre-book cinema seats for the film of their choice. A discussion follows on the design of a prototype for the system. Requirements specification is discussed and a design and task analysis is carried out. Evaluation is very important in computer interface development and improvement. Here analytical, observational and usability evaluations are carried out. The system created is found to be visually pleasing, user-focused and fully functional.

Keywords: Evaluation, Task analysis, User interface

1. Introduction

Major corporations are increasingly looking for ways they can reduce their costs. Cinema corporations are no exception. This paper describes a cinema seat booking system designed as a kiosk which customers can use at local shopping areas prior to going to the cinema to pre-book cinema seats for the film of their choice. This could reduce costs allowing for less staffing costs. A discussion follows on the design of a prototype for the system. `

Traditionally the focus has been on technical functionality of systems and interfaces. In recent years the 'HCI approach' has been more prevalent [Norman, 1992; Baecker and Buxton, 1987]. The evaluation carried out in this paper on the pilot system concentrates on the needs and goals of the users of such a system. As designers we must work to a set of principles, which must be interpreted within the context of the task in hand, which may apply constraints (rules of style). We will concentrate on looking at certain generally accepted guidelines for the design of 'usable' human-computer interfaces [Brooks, 1988; Smith and Mosier, 1986; Shneiderman, 1992]. There are eight general guidelines relating to HCI design [Preece et al, 1994]. These are consistency (internal, external and real-world); visual clarity; compatibility with expectations; flexibility and control; explicit structure; continuous and informative feedback; error prevention and correction; user documentation and support.

In Section 2 a requirements specification is shown for the pilot system. Section 3 discusses a design and task analysis. The database design is briefly shown in Section 4 while Section 5 includes detailed descriptions and screenshots of the actual interface and system. Evaluation is known to be an important part of design and the user-centred design process. Analytical and observational evaluations are detailed in Section 6 which includes expert and usability evaluations. Section 7 concludes the paper.

2. Requirements Specification

Requirements gathering or analysis is the process of finding out what a client or customer requires from a system. The requirements specification for this study consisted of the following tasks:

1. Add details of new films.
2. Search for a particular actor.
3. Search for films of a particular type.
4. Keep a database of existing customers.
5. Update movie schedule on database.
6. Record booking of seats and reduce seats accordingly.
7. Provide a list of upcoming movies.
8. Provide pictures or clips from movies.

9. Provide actor information.
10. Process credit-card details.
11. Pick-a-seat.

3. Design and Task Analysis

Task analysis is concerned with what people do to get things done. Tasks in this respect are meaningful for the user in that the users believe it necessary and desirable to undertake tasks. The term task here refers to an intentional, purposeful level of description. These simple tasks are also referred to as actions [Payne and Green, 1989]. Various tasks or actions have to be performed on this pilot system. Figure 1 illustrates the process flow diagram for this pilot kiosk booking system, showing the tasks that a customer may carry out and correspondingly how the system will respond.

Here we provide an analysis of the task designs.

3.1 Booking a film

If a kiosk user wants to book a film they are presented with the start screen which is displaying a picture of the four films that are currently being shown in the cinema. To continue they must touch the screen.

3.1.1 Displaying the film list by date

After touching the screen, the user is presented with a screen showing the four films again, from this screen they can take an option to display information about the film or by choosing an alternative option they can display a list of dates and times that the film will be shown.

3.1.2 Selecting a film

After the user has taken the option to display the list of dates and times that the film will be shown they must select the actual date and time that they wish to view the selected film. They do this by pointing to the required row displayed on the screen. They must confirm their decision by pointing to the `book seats' button that is displayed on the screen.

3.2 Recording seats

3.2.1 Displaying the seating arrangement and selecting seats

The user will then be shown a screen with a seating plan of the cinema. While this screen is loading the system will check the database and display which seats are available or unavailable by using different colours (dark grey for unavailable white for available). The user then points to an available seat that they want to book. The seat changes colour to show that they have chosen the seat. If they wish to change their mind they point to the seat again and it becomes available. After choosing the seat / seats that they want they either choose `cancel' to go back to the previous screen or `accept' to continue to the next screen.

3.2.2 Assigning seat types

After the user has chosen the seat / seats that they want they must decide on the category and quantity of seats for each category. The three categories are adult, concession and child. If the film is rated as any thing other than U or PG the child category will not be available. Also children cannot be unaccompanied so they must reserve a seat for an adult. If they have not done this they can go back to the seating arrangement and select another seat for the adult. When all seats have been assigned to the relevant categories then the accept button is displayed. The user must touch this button to continue otherwise `cancel' to exit.

3.2.3 Displaying ticket charges

After selecting the accept button the user is shown the total cost of the tickets and the total cost of each category of ticket. If the price is acceptable they touch the `accept' button to continue otherwise the `cancel' button to exit. After touching the `accept' button the user is asked for their credit card details.

3.3 Accepting payment

3.3.1 Requesting credit card payment and processing credit card details

The user is asked to place the credit card in the machine and touch the enter button. At this point the details are read from the card and a new display is shown prompting the user to enter their pin number. After entering their pin number the user can select `clear pin' if they have made a mistake or `enter' to continue to process the payment. The payment is processed via the bank.

3.3.2 Displaying payment confirmation

A message is shown to confirm that the payment has been processed and that the tickets will be printed out. After a short period of time the screen will return to the start screen.

3.3.3 Printing tickets

The tickets are printed out.

4. Database Design

Figure 2 shows the details of the design of the tables included in the database for this pilot system. The system was coded using Visual Basic.

5. Customer System Interface

There now follows a series of screenshots of the actual interface itself.

5.1 Welcome screen

Figure 3 is the welcome screen, which greets the user. The user can enter the program by touching anywhere on the screen.

5.2 Select Film screen

The "select film" screen in Figure 4 contains a lot of information for the user. By pressing the buttons at the left side of the screen, the user can access information from the database showing the dates and times for all the films. By pressing on the movie posters at the top of the screen, the user can access a pop-up screen containing information about the film. The user can then select a show to attend by touching the "showno" they wish to attend. Once this has been pressed the book film button appears. This only appears now so that the user cannot try to book a film without having selected one. The user can now either book the chosen film by pressing the `book film' button or they can press the `cancel' button and return to the welcome screen.

5.3 Show Storyline screen

Figure 5 is a screenshot of the storyline screen, which provides the user with information about the film.

5.4 Choose Seat Form

This screen shown in Figure 6 allows the user to pick the exact seat position they wish to purchase for the film of their choice. The user simply clicks on the desired seat position and this is then stored in the database so that the same seat cannot be double booked. Available seats will be coloured white while seats which are already booked and therefore unavailable will show up as grey. A colour key is provided to show the user which seats they can choose. Once they have chosen seats these turn green showing that they have been selected. Also the number of seats chosen shows up at the bottom of the screen as well as the `id' of the last seat chosen. If the user is unhappy with their selection they can deselect the chosen seat by pressing it again and the colour of the seat will return to white showing that it available to be chosen again. Once seats have been chosen the accept button becomes visible. This prevents the user from pressing this button without having chosen any seats. At this stage the user can accept the chosen seats by pressing the accept button which will bring them on to the next screen or they can cancel and return to the previous screen by pressing the cancel button.

5.5 Help screen for Choose Seat

There is a help screen (see Figure 7) which further explains the booking process to the user.

5.6 Seat Breakdown screen

This screen (Figure 8) allows the user to enter the number of each type of ticket which is required. They can enter adult, child or concession. If the movie is for over 18`s, the child option is not available. The number of unaccounted seats shows the number of seats which have been selected on the seat layout screen. The user cannot continue until the correct number of tickets has been selected from the choice of adult, child and concession. As each ticket is chosen the number shown in the "number of unaccounted seats" section at the top decreases by one. It will decrease until it reaches 0, which means that the correct number of tickets has been chosen. As a form of validation, the accept button only appears once the correct number of tickets have been chosen. At this point the user can accept the tickets chosen by pressing the chosen button or can cancel using the cancel button, thereby returning to the welcome screen.

5.7 Seat Breakdown Help screen

Figure 9 shows this screen which gives the user detailed instructions on using the seat breakdown screen.

5.8 Confirmation screen

Figure 10 shows this screen which allows the user to view and confirm their order before they have to enter their credit card details. They can see the film details and the amount they will be charged. They can cancel at this stage without the transaction being written to the database and the screen will return to the opening welcome screen. By pressing the accept button the user is confirming that the details are correct and they wish to continue.

5.9 Card Swipe screen

This screen (see Figure 11) simulates the user swiping their credit card in a provided slot. Once they have swiped the card they must press the enter button to continue.

5.10 Card Swipe Help screen

Figure 12 shows the user provided with more detailed information on swiping their card.

5.11 Enter pin screen

This screen shown in Figure 13 allows the user to enter their credit card pin number. Once they have entered the four pin numbers the enter button appears and they can continue to the next screen. If they make a mistake at any stage they can delete the numbers entered by pressing the `Clear Pin' button and starting again. The enter button is not visible until the user has entered four pin numbers. This validation is to ensure that they cannot continue without entering the correct number of characters. Also the pin numbers disappear once the four entries have been made so that the user cannot try and enter more than four characters. The entries show up in the textbox as "*" characters so that no-one else can see the pin number which has been entered.

5.12 Enter pin Help screen

This screen provides the user with more detailed help on entering the credit card pin number in Figure 14.

5.13 Transaction Complete screen

This screen shows (Figure 15) the user that the transaction has been successfully completed and the tickets are printing out to the tray.

5.14 Film Update Password screen

From the welcome screen it is also possible for staff to access the film data entry screen (Figure 16). By pressing the button at the bottom right of the screen, staff can bring up a password screen where they must enter the correct password in order to access the entry screen. If the incorrect password is entered the password screen disappears. Once the correct password has been entered the staff member gains access to the film update screen.

5.15 Film Update screen

This screen (Figure 17) is used by a staff member to enter new films to be shown or to update current details. The user can enter the film title, certificate, description and associated picture or movie poster. Once the update button has been pressed the details are written to the database. The user closes the screen by pressing the `exit' button.

6. Interface Evaluation

Evaluation is concerned with gathering information about the usability of a design and interface by a specified group of users for a particular activity within a specified environment [Preece, 1994]. These evaluations are used to inform design by providing answers to questions. An analytical evaluation is carried out as well as an observational evaluation. Feedback is given from these evaluations and improvements suggested.

6.1 Analytical Evaluation

The Cineplex Kiosk prototype is a Cinema seat booking system that can be installed in local shopping areas. It is a Kiosk that makes use of Touch Screen technology to allow members of the public to book seats for a film of their choice in cinemas of a local cinema corporation. The system also allows for cinema management to change the listings as new films are released or to update the existing selection.

The design of the kiosk system from the users' perspective has two main focus points: the provision of information of films to be shown and the booking of seats for a specific showing.

Consistency is evident. As the user progresses through the various screens there are icons that the user touches to evoke a function which are located in the same area of each screen. The icons are supplemented by a written description.

The user accesses the system expecting to find information about films showing in local cinemas giving details about the title of the film, start time, duration and certificate. A list of currently showing films are displayed, the user selects a film by touching the name and the show times are displayed. If the times do not suit the user they can go back to select another film and view the times of its showings. What the user expects to see when navigating these screens is what they get. When a choice is made the user will expect to pay and the screen for inputting credit card details appears. A simple numeric keypad allows for the inputting of credit card numbers and a drop down list of dates for the expiry date.

The manager who adds and removes the films as they are distributed or come to the end of their time is presented with a form which he fills out with the details of the new releases and inserts them into the database, where they can be edited or deleted. It is password protected so is only accessible to those who know the password. Either user finds a system that is compatible with their expectations.

In terms of flexibility and control the public user has very little control over what can be done with/to the system. Information about films can be viewed and when a choice is made payment will be processed and the tickets printed. The information required for payment is validated and corrections can be made prior to committal. If the validation of credit card details is not met the user is so informed. This is what is required for public kiosks since user modification of the system is not desired.

The simplicity and clarity of the prototype is by choice and evident in the design. An interface or system should be visually and conceptually clear [Microsoft, 1992]. Only the required information is available and is presented in such a way as to fulfil the user's expectations and requirements. It is a case of 'just the facts' presented in an aesthetic and entertaining fashion.

The interface is structured logically, sequentially and is rationally laid out. It is uncluttered and keeps the user's attention as the information is absorbed at which point the user can move on or move back a screen. When processing is occurring (as in the validating of credit card details) the user is advised that something is happening.

Entered data appears on-screen as it is being input. Screen updates are instantaneous and where a response is delayed (as in the validation of credit card details or in seat selection) the user is informed. In the seat selection section of the system users are shown a colour coded key that informs them which seats are available and which are not and when a seat is selected the colour changes to indicate the choice. A box also displays the seat number which will be printed on the ticket when purchased.

The amount of typing required is kept to a minimum and the touch screen technology is maximised together with drop down lists. The facility to exit from any screen is also incorporated into the design. These functions virtually eliminate the possibility of user error.

This system is designed to be used by the general public. Some, but not all, of the public are technology literate and have no fear about using computerised systems while others would be nervous at first. The use of touch screen technology can bring computer systems to the masses, there is no skill required to use it. It is something they can walk away from if they are uncomfortable with it and people in general are tactile in nature. They like to handle things like clothing and fruit, furniture and cds. This Cineplex kiosk system does not require grammar or spelling ability just to read the number of their credit card.

There is no ambiguity present on the screen because the layout is logical as is the sequence of events. Titles are included in each screen so that the user knows where they are at any point. Only information relevant and pertinent to the user is presented. The selections available are limited and clear. Everything the user needs to correctly use the system is obvious.

6.2 Observational Evaluation

Here we carry out an expert evaluation as well as a usability evaluation [Preece, 1992]. Feedback is given and improvements suggested.

6.2.1 Expert Evaluation

An Information Technology expert was asked to evaluate the system. This was his response:

"While using the Cineplex kiosk, I found the screen very visually pleasing, but I found I had to read the instructions to learn how to go through each of the screens, although the instructions were clear they could have probably been written more clearly. The second time through, I found it a lot easier to use. The user input is very minimal and does not allow for many mistakes. Selecting the seats during the booking is a useful feature to have and I found it simple to select the seats I require. Overall I was impressed with the system."

6.2.2 Usability Evaluation

It is always necessary in designing interfaces to think of improving the usability of the system for the user as much as possible. A generic usability evaluation booklet [Patterson, 1997] was used to evaluate the system in terms of usability. This gave a lot of very useful feedback and information on the system.

The evaluation was based upon a set of software ergonomics criteria, which a well-designed user interface should aim to meet. The booklet comprises twelve sections which are not in any order of importance within the booklet. Each of these sections is based on a different criterion, or "goal", which a well-designed interface should aim to meet. Each of the sections has a number of checklist questions based upon the specific criterion.

The evaluator must select the grade which best describes their answer to the question with respect to the application. The grading system is: "Always", "Most of the Time", "Some of the Time", and "Never". Specific comments relating to specific checklist questions can also be added. At the end of each section there is a rating scale ranging from "Very Satisfactory" to "Very Unsatisfactory". The evaluator must tick the box which best describes the way they feel about the user interface in terms of the issues in that section.

The first ten criteria are as follows: Visual Clarity, Auditory Presentation, Consistency, Compatibility, Informative Feedback, Explicitness, Appropriate Functionality, Flexibility & Control, Error Prevention & Correction, and User Guidance & Support.

Two further sections are included, which relate to the general usability of the application.

The first of these is *System Usability Problems.* The questions in this section ask about the specific problems which were experienced when carrying out the evaluation tasks. The grading system here used is: "No Problems", "Minor Problems" and "Major Problems". The final section included is *General Questions on System Usability.* This section asks for views on the interface which has been evaluated.

This method of evaluation through the usability booklet was used on this booking system. The system was graded very well and overall it was classed as a well designed interface. Very useful feedback was provided. Some very minor problems were reported which were investigated.

6.2.3 Feedback on Evaluation

During the evaluation of this interface a number of bugs were found and improvements had to be made to the system which include:-

1. Some screen help buttons were missing or did not work.
2. The font on some screens was not consistent with the scheme; some of the text boxes were too small.
3. The icons had a graphic which made the text difficult to read, the graphic was removed.
4. Screens / values not refreshing / updating after loading more than once.
5. No password error message for wrong passwords on the film admin section.

A number of possible upgrades were apparent which include:-

1. The final kiosk software would not have the admin button on the start screen. This is for evaluation purposes only. Also the `minimize', `restore' and `close' buttons would be hidden. The windows taskbar would have to be hidden.
2. Replace `swipe card' remarks with `insert card' with the development of chip and pin credit cards. This will also require additional upgrades to tell the user to remove their card.
3. What happens if the print option fails? What happens if the printer runs out of ink? How does the user get their tickets? Would it be better that the system does not print ticket but rather when the person arrives at the cinema there is another machine into which they put their credit card and it reads the credit card and prints out the tickets that are charged to that credit card?
4. A screen could show general information at the start, such as price, opening times, facilities, map of the location of the cinema and cinema telephone number.
5. On U & PG films do not display the child option until at least one adult has been assigned versus the current system to show a message box.

7. Conclusion

This paper describes a pilot study to design a kiosk system for customers of a local cinema corporation to use in order to pre-book and pay for tickets in local shopping areas. The system has been described and evaluated. Feedback has been discussed and the system is currently being improved on accordingly.

References

Baecker, R, Buxton, W. (1987). *Readings in Human-Computer: A multi-disciplinary approach.* Morgan Kaufmann.

Brooks, F.P. (1988). "Grasping reality through illusion: Interactive graphics serving science." In Proceedings of Human Factors in Computing Systems CHI'88 Conference. ACM Press.

Microsoft. (1992). *The Windows interface: An application design guide.* Microsoft Press.

Norman, D.A. (1992). *Turn signals are the facial expressions of automobiles*. Addison-Wesley.

Patterson, G. (1997). "A new perspective on HCI Education, Research and Practice for User Interface Design." In Proceedings of 6th International Conference on Man-Machine Interactive Intelligent Systems in Business, Montpellier, pp1-3.

Payne, S, Green, T.R.G. (1989). Task-action grammar – The model and its development. In Task analysis for Human-Computer Interaction. Ellis Horwood.

Preece, J. (1992). *A guide to usability: Human factors in computing*. Addison-Wesley.

Preece, J, Rogers, Y, Sharp, H, Benyon, D, Holland, S, Carey, T. (1994). *Human-computer interaction.* Addison-Wesley.

Shneiderman, B. (1992). *Designing the user interface: Strategies for effective human-computer interaction.* Addison-Wesley.

Smith, S.L., Mosier, J.N. (1986). *Guidelines for designing user interface software.* MITRE Corporation.

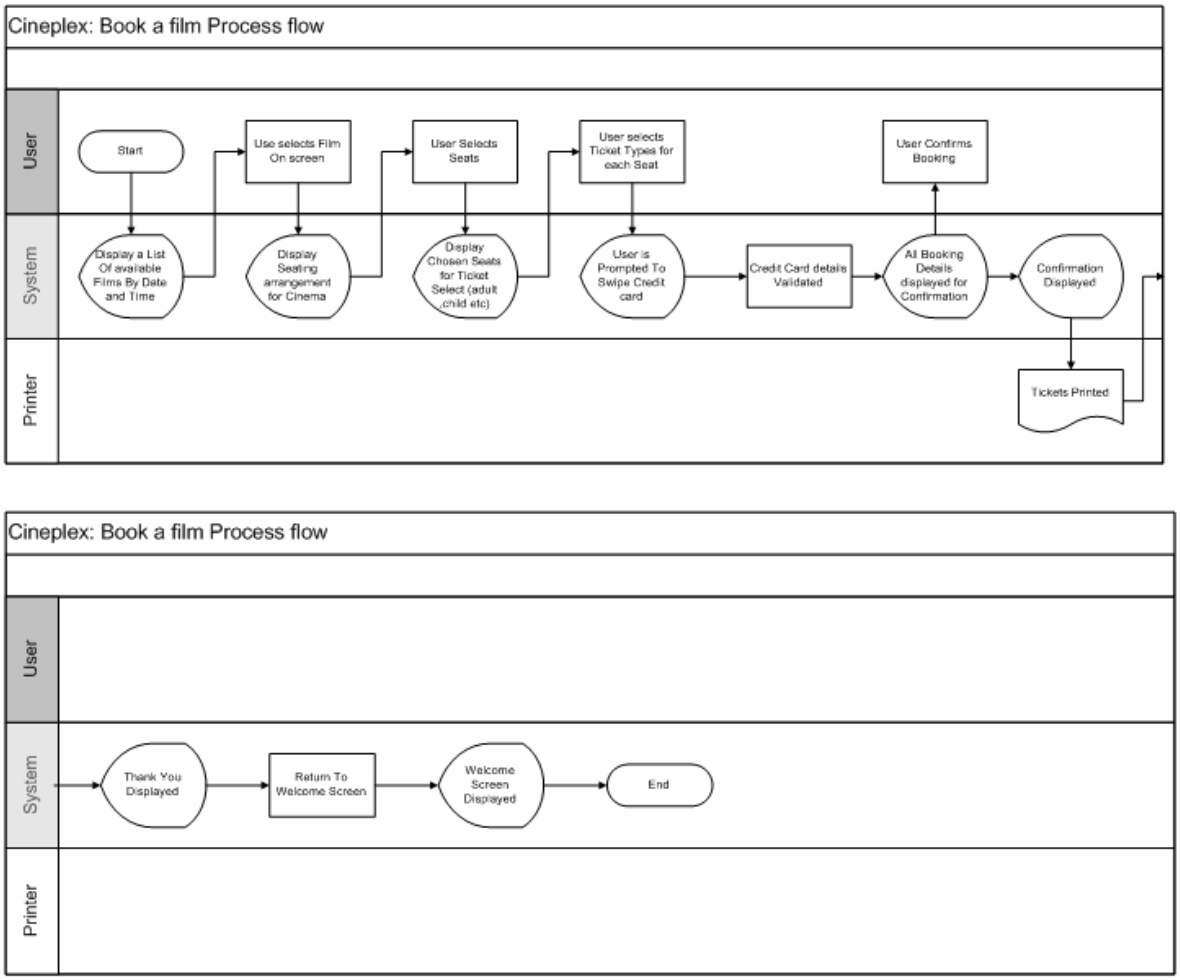


Figure 1. Process flow Diagram for Kiosk System

Tables	Description
Booking **BookingRef** ShowNo CardNo	Main booking table to record a booking by individuals. Recorded By show No, given a unique Booking ref no, and records Credit Card Number.
BookingDetails **SeatNo** **BookingRef**	Records the seats for each booking as there are usually multiple seats chosen for each booking. This table is queried for the seating availability.
Films **FilmNo** FilmTitle Cert Description Picture	Store all film information (updated by the Cineplex Staff).
Showings **ShowNo** FilmNo FilmDate FilmTime ScreenNo	This table shows the schedule for the films to be shown.

Figure 2. Database tables used in design

Figure 3. Welcome screen

Figure 4. Select Film screen

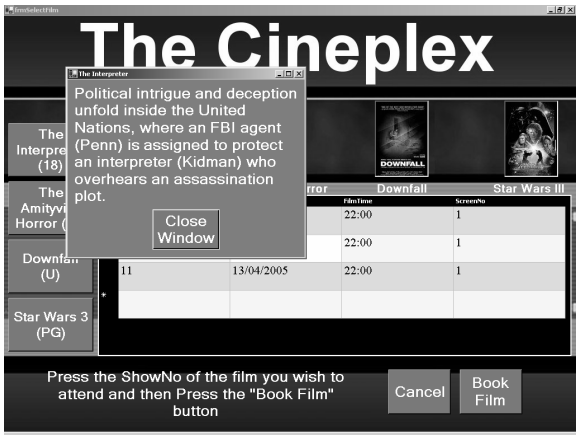


Figure 5. Show Storyline screen

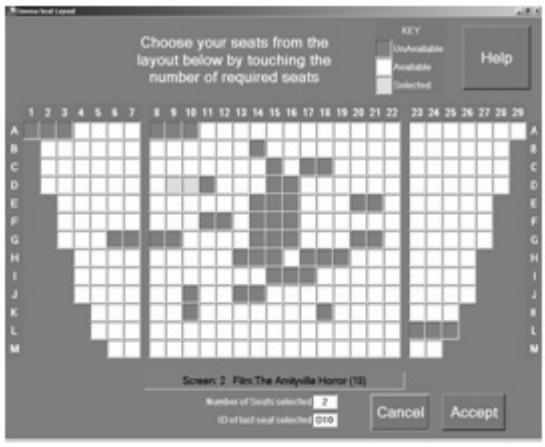


Figure 6. Choose Seat Form

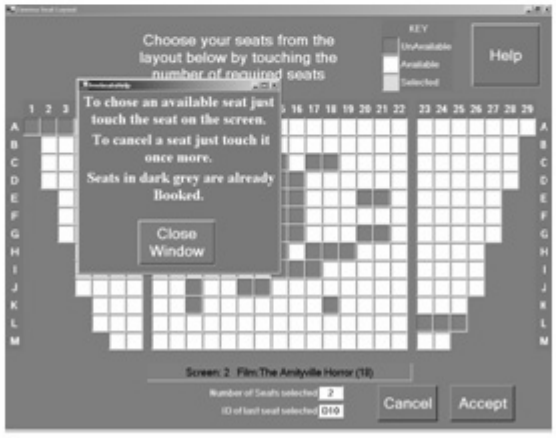


Figure 7. Help screen for Choose Seat

Figure 8. Seat Breakdown screen

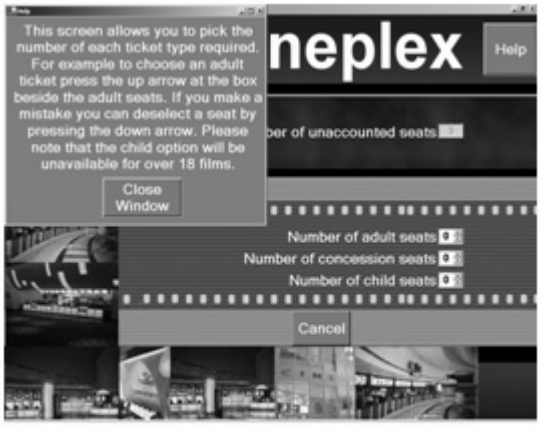


Figure 9. Seat Breakdown Help screen

Figure 10. Confirmation screen

Figure 11. Card swipe screen

Figure 12. Card swipe Help screen

Figure 13. Enter pin screen

Figure 14. Enter pin Help screen

Figure 15. Transaction Complete screen

Figure 16. Film Update Password screen

Figure 17. Film Update screen

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

The Research on the Mode of Making Use of the Microcomputer Circularly

Qinghai Bai

College of Mathematics and Computer Science

Inner Mongolia University for Nationalities

Tongliao 028043, China

Tel: 86-475-239-5155 E-mail:baiqh68@163.com

Chunsheng Zhang

College of Mathematics and Computer Science

Inner Mongolia University for Nationalities

Tongliao 028043, China

Tel: 86-475-831-0225 E-mail: zhangcs_817@sina.com

Yan Li

College of Mathematics and Computer Science

Inner Mongolia University for Nationalities

Tongliao 028043, China

Tel: 86-475-831-3300

Yufeng Bai

College of Mathematics and Computer Science

Inner Mongolia University for Nationalities

Tongliao 028043, China

Tel: 86-475-831-3300

Abstract

The eliminated speed of the microcomputer has become aggravated year by year, it brings huge of economic loss and causes very bad result of the environment pollution, much attention have already been paid to this problem internationally. For this, they put forward the mode of making use of the microcomputer circularly, point out that the monopoly of the operating system has a negative influence to the microcomputer, and analyze the real need of the enterprise for the microcomputer hardware. Meanwhile, they give some effective plans to resolve it. They offer the grid technique that will postpone the speed of the microcomputer's elimination with the mode of making use of the microcomputer circularly, the technique of the balance of the network resource and so on.

Keywords: Microcomputer, Eliminate, Make use of the mode circularly, Operate system

1. Introduction

1.1 The International Attention Paid to The Reuse of the Eliminated Computers

Initiated by UNESCO ,a meeting called Cooperating to Improve the Reuse of the Informative Technological Instruments was held in Paris in 2003(Informative enterprisers from all over the world get together in Paris to explore the reuse of the eliminated computers,2003) . The meeting means to have knowledge of the businesses tentative plan for the reuse of the eliminated informative instruments, explores and makes the specific measures to reuse effectively these eliminated instruments all over the world.

At present, with the informative technology being replaced at a greatly accelerated speed, some instruments originally designed and planned to be used for three years are eliminated after being put into use for less than one year even half a year. According to the survey of UNESCO, seven million computers will be eliminated within three years in the world first one hundred power companies and thirty million computers will be eliminated in the world first one thousand power companies, which will cause a great loss for human being.

During the meeting, the attendants' discussion centers on the possibility of reuse of the eliminated informative instruments in the different social, economic and cultural backgrounds. And they also put forward a series of suggestions and measures about the reuse of the eliminated informative instruments.

Thus it can be seen that the reuse of the eliminated computers is a tough problem to be resolved urgently by human being.

1.2 Recycling Economic Mode and Computers' Circularly being Used Mode

The eliminated computers contain several kinds of poisonous substances suck as lead , tin , mercury , chromium , dalvor and plastic, If dealt with improperly or discarded, these substances will pollute soil, water, animals and plants , and eventually do a lot of harm to human's health . So they are a serious threat to humans' lives.

Recently, a new economic concept called Recycling Economy has been put forward internationally(Wang, Chunmei& Cheng , Aijun,2005). It changes the traditional economic mode that is "Resources-Products-Junk" into the new mode that is "Resources-Products-Regenerated Resources". The new economic mode features low consumption, less pollution and high profits, which can gain the greatest environmental benefits at the price of the least resources and environmental pollution.

As far as computer products are concerned, the recycling economic mode means classifying the eliminated computers. These that can be reused will be continuously put into use after being transformed and reassembled. Those that are actually eliminated will be reclaimed to extract mineral resources.mode of making use of the microcomputer circulary at presented in Figure 1.

1.3 The Present Condition of Our Country's Dealing with the Eliminated Computers

China is still at the elementary phase in preventing and controlling pollution caused by computer junks(http://www.people.com.cn/electric/990910/x1050.htm). The collective managerial system for electron junks hasn't been set up. Only few computers have been reclaimed by firms and large numbers of computer junks are either buried as common rubbish or reclaimed for dissembling in less qualified workshops, which causes severe potential pollution for the environment.

Another, the research on computers being reused has been seldom reported. Thus, it can be concluded that enough emphasis hasn't been put on the reuse of the eliminated computers in our society.

2. Are the Eliminated Computers Really Put Out of Use?

2.1 The Role of Operating System

What causes the computers to be eliminated at such a high speed? Is it because of the lack of computer hardware resources? What role does the operating system play in computers being eliminated so quickly?

This problem can be analyzed from the development history of microcomputers and operating system(http://www.cxsyzx.com/readnews.asp?newsID=172,2004;
http://yjs.bit.edu.cn/blog/user1/83/archives/2005/672.html, 2005).

The development of microcomputers goes through four stages: From 1971 to 1973 the computers were at 4 bits or 8 bits computer times; From 1974 to 1978,there appeared 8 bits computers; From 1978 to 1981,there existed 16 bits computers; hereafter in 1985, the computers began to go into 32 bits times. From the development of microcomputers, we can see that the developing speed is very high and the working speed and the volumes of the computers are on increase continuously.

Let's deal with the development of the operating system. The operating system called MS-DOS1.0 was designed for IBM by Microsoft in 1979 which established Microsoft as the monopolizer of the microcomputer operating system. Later in 1978 and 1995 DOS1.1-7.0 were developed successively. In 1995 the operating system Windows95 was developed by Microsoft and then Windows97, Windows me, Windows 2000, Windows XP and Windows 2003 became popularized one after another. All of these operating systems were produced with the renovation and upgrading of CPU. This operating system has the characteristics of graphics Interface, convenient operation and multitasks, which make it very popular among the consumers. So this operating system almost monopolizes the whole market of microcomputer operating system.

Among the too many characteristics of Windows operating system, the graphics interface makes it look very pleasing to the eye and makes it easily operated. But it is exactly because of this advantage that makes the Windows operating

system itself larger and larger, which makes the most parts of computer resources occupied by this beautiful picture and this in fact is a kind of waste.

Also, multitasks, another important characteristic of Windows enables us to do several tasks and this brings a lot of convenience for the consumers. However, multitasks haven't been used completely in the average businesses and it is also unnecessary for the average businesses to use them completely. Single task is enough for the average businesses. So multitasks can be regarded as another kind of waste as far as the average businesses are concerned.

Therefore, the operating system in fact adds fuel to the flame in eliminating computers.

2.2 The Businesses' Real Need for Microcomputer Hardware Resources.

On one hand, microcomputers are being eliminated at a very high speed; On the other hand , with the time passing and the development of the businesses, is it necessary for the firms to have a higher and higher demand for the hardware?

Take a supermarket as an example, with the time passing, the nature of its business hasn't been changed fundermentally. The increase of its business quantity and statistics hasn't made it necessary for the supermarket to demand for the hardware. And this can be solved by adding the configuration of the server. Meanwhile, it isn't necessary for the operating system to have complicated graphics interface and multitasks system. So high level microcomputers and too many kinds of operating system haven't brought big profits to the supermarket. In fact hardware configuration for payee are inferior to the configuration for currently popular computers. The monitor for the payee is usually black and white.

3. The Solution to the Reuse of the Microcomputers

3.1 Take Operating System Itself as the Point of Departure

3.1.1 Use the Earlier Operating System

In views of different businesses' demands, earlier operating system can be suitable for the applying system with small business quantity and fixed business mode. With the help of corresponding programming appliances, the software for this type of computers can be developed to decrease the demands for the hardware.

For example, by applying VFP3.0 in Windows 3.1 I develop a paying system for a restaurant. This paying system has also a graphics interface and can be operated effectively in microcomputer 386. In the same way, another system is developed by applying VFP6.0 in Windows 98. However, this system can be operated beyond computer 586.

Although these two systems occupy different spaces of memory, they are the same in functions and are both convenient to be operated by the users.

3.1.2 Perfect an Operating System or Transform an Operating System

Computers used in the businesses rely mainly on database. And according to this, the operating system with simple interface, occupying little memory and single task can be developed by perfecting or transforming the existing computers . However, this operating system must be strengthened in the management of database and supportive of the work of applying programme in database and network environment.

Windows 3.X is a multitasking windows operating environment with graphics interface which is developed by Mircosoft based on DOS. The driver supporting C/S system, for example ODBC, can be developed based on Windows 3.X. , which makes it possible to operate C/S programme under Windows 3.X. in computer 386 and computer 486 . Meanwhile, an operating system, such as Linux system, can be transformed to enable it to have high ability to manage database and lower the demands for hardware. This transformed operating system can be used in low level computers to support C/S system.

3.1.3 Take Full Use of Grid Technology

Grid is a recently developed computing technology. Following a group of open standards and agreements, the organizations can visit statistics and store medium and computing resources from other organizations by internet and inner nets. Eventually all the resources including network, data, computing resources and etc, scattered in different environments can be integrated into a perfect computing environment

Internationally, opening the source code and cooperating openly are main ways employed in the research on grid, which provides us with a means . If we transform the source code of the grid technology, apply grid technology to low level computers, to use the low level computer resources in a distributive and balanced way and integrate the resources to make the computing ability and hardware resources increase unlimitedly in theory, the microcomputers will not be eliminated unless the hardwares are damaged naturally .

3.2 Take the Balance of Network Resources as the Point of Departure

At present, the database system exists basically with the form of C/S. C/S system can be divided into double layers structure including the consumer layer and server layer and multilayers structure with business regularity layer as the third one. Because these layers are logic and not physical, they don't rely on the physical structure. Therefore, we can

design reasonably the hardware resources and locations occupied by these layers to take use of low level microcomputers from the point of the network resources balance.

3.2.1 The Theory for Intelligent Client

With the theory for intelligent client , your applying programme can use the hardware in a more effective and better way by using the local resources as much as possible and assembling the local resources with your own intelligent client applying programme. Most of the programme logics are located in the server, including the third layer and beyond the third layer. The server can adopt stored procedure, trigger,regularity and etc to reduce the stream of the network to keep the client programme logic at the lowest level. By combining with the lower level operating system, such as operating system Windows 95,most of work can be taken by the server and in this way low level microcomputers can be used in client .

3.2.2 Use the Terminal Technology of Windows

Having realized the microcomputers are being eliminated at a very high speed, Microsoft finishes the terminal server technology in Windows 2000 , and perfects it in Windows 2003(http://www.chinapohelp.com/aubt.htm). And this itself is the regression to master computer applying mode. Client is only responsible for the input and output of the data and it doesn't do any computation. All the programmes are to be operated in server. By using the balance theory of the network resources, this exactly means the Windows software supported by this technology can be applied at a higher speed to the such low level computers as model 386, model 486 and model 586. In this way ,the speed at which computers are eliminated can be postponed.

3.3 Start with the System Structure

From the point of the system structure, the speed of working out a problem can be quickened by improving the ratio of the hardware functions and the stored quantity needed can also be decreased .However, the cost of the hardware will be improved. The ratio of utilization of the hardware and flexibility and suitability of the computer system will also be decreased(Li, Xuegan&Su, Dongzhuang,1998).This theory presupposes the constant demands of the computer system. If this theory is applied to low level computers, the latent power can be tapped.

For the specific cases, some fixed functions such as multimedia, network and business regularity can be carried out by the hardware. Strictly speaking, the hardware here can be called firmware that is a card. At present, the cost of computer firmware is so low that it can be use for transforming the low level computers to decrease the stored quantity of the computer and improve the speed of computer work. And this is a feasible way for transforming a low level computer.

4. Conclusion

In accordance with the huge economic loss and the threat to the environment caused by the quickly eliminated computers, this article analyzes the reasons why the computers are being eliminated at such a high speed, points out the negative effect of the operating system, analyzes the real needs of the enterprises for the computer hardwares, puts forward the mode of making use of the computers circularly and meanwhile provides the specific solution plans for the problem. Of course, the reuse of the eliminated computers is a knotty problem faced with human being, and it is being paid much attention internationally. However, this problem will exist for a long period of time . All human beings, even several generations need to make constant efforts to work out it .

References

Computer Junks: A Threat to Twenty-first Century. [EB/OL]. http://www.people.com.cn/electric/990910/x1050. htm

Informative enterprisers from all over the world get together in Paris to explore the reuse of the eliminated computers(2003). [EB/OL]. http://www.china.org.cn/Chinese/2003/mar/294091.htm.

Introduction to Windows Terminal Technology. [EB/OL]. http://www.chinapohelp.com/aubt.htm

Jia, Yueqin, Zhang, Ning & Song, Xiaohong. (2005). The Discussion on Grid Safety Techniques. *Microcomputer Information,* 21(3), pp. 199-200.

Li, Xuegan & Su, Dongzhuang. (1998). Computer System Structure Publish by Xi'an Electronic Science and Technology University.

Operate system taction for Windows system. (2005). [EB/OL].

http://yjs.bit.edu.cn/blog/user1/83/archives/2005/672.html. 2005.8.15

The Developing History of Microcomputers.(2004). [EB/OL]. http://www.cxsyzx.com/readnews.asp?newsID=172.

Wang, Chunmei & Cheng,Aijun. (2005). Recycling Economy and Economical Society. *Outlook on Global Science and Technology and Economy.* No. 10.

Wei, Hongbo. (2003). Analysis and Techniques on Network Safety of Distributive System. *Microcomputer Information*, No.11.

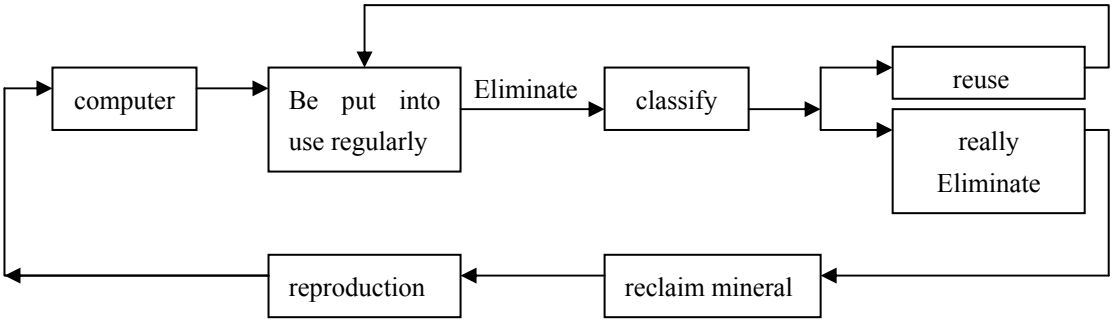


Figure 1. the Mode of Making Use of the Microcomputer Circularly

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

Identification of Sensitive Items in Privacy Preserving - Association Rule Mining

Dr. K. Duraiswamy

K.S. R. College of Technology

Tiruchengode- 637-209, Tamil Nadu, India

E-mail: kduraiswamy@yahoo.co.in

N. Maheswari (Corresponding Author)

P.G. Department of Computer Science

Kongu Arts and Science College

Erode-638-107, Tamil Nadu, India

E-mail: mahii_14@yahoo.com

Abstract

The concept of Privacy-Preserving has recently been proposed in response to the concerns of preserving personal or sensible information derived from data mining algorithms. For example, through data mining, sensible information such as private information or patterns may be inferred from non-sensible information or unclassified data. As large repositories of data contain confidential rules that must be protected before published, association rule hiding becomes one of important privacy preserving data mining problems. There have been two types of privacy concerning data mining. Output privacy tries to hide the mining results by minimally altering the data. Input privacy tries to manipulate the data so that the mining result is not affected or minimally affected. For some applications certain sensitive predictive rules are hidden that contain given sensitive items. To identify the sensitive items an algorithm SENSIDENT is proposed. The results of the work have been given.

Keywords: Data Mining, Privacy Preserving, Association Rules, Sensitive Items, Minimum Support, Minimum confidence

1. Introduction

In recent years, data mining or knowledge discovery in databases has developed into an important technology of identifying patterns and trends from large quantities of data. Successful applications of data mining have been demonstrated in marketing, business, medical analysis, product control, engineering design, bioinformatics and scientific exploration, among others. The current status in data mining research reveals that one of the current technical challenges is the development of techniques that incorporate security and privacy issues. While all of the applications of data mining can benefit commercial, social and human activities, there is also a negative side to this technology: the threat to data privacy. The main reason is that the increasingly popular use of data mining tools has triggered great opportunities in several application areas, which also requires special attention regarding privacy protection. The concept of privacy preserving data mining has recently been proposed in response to the concerns of preserving privacy information from data mining algorithms (Agrawal, Srikant, 2000). There have been two types of privacy concerning data mining. The first type of privacy, called output privacy, is that the data is minimally altered so that the mining result will preserve certain privacy (Dasseni, Verykios, Elmagarmid, Bertino, 2001, Oliveira, Zaiane, 2003 a, Oliveira, Zaiane, 2003 b). The second type of privacy, input privacy, is that the data is manipulated so that the mining result is not affected or minimally affected (Evfimievski, 2002).

For example, through data mining, one is able to infer sensitive information, including personal information, or even patterns from non-sensitive information or unclassified data. As a motivating example of privacy issue in data mining discussed in (Clifton, Marks, 1996). Suppose we (as purchasing directors of BigMart, a large supermarket chain) are negotiating a deal with the Dedtrees paper company. They offer to us a reduced price if we agree to give them access to our database of customer purchases. We accept this deal and Dedtrees starts mining our data. By using association rule mining tool, they find that people who purchase skim milk also purchase Green paper. Dedtrees now runs a coupon

marketing campaign "50 cents off skim milk when you buy Dedtrees products," cutting heavily into the sales of Green paper, who increase prices to us based on the lower sales. When we next go to negotiate with Dedtrees, we find that with reduced competition, they are unwilling to offer us as low a price, and we start to lose business to our competitors. This example indicates the need to prevent disclosure not only of confidential personal information from summarized or aggregated data, but also to prevent data mining techniques from discovering sensitive knowledge which is not even known to the database owners. So the highest sales product can be easily identified using this rule generation. This makes the supplier of the product to demand or hike the product rate. Such products are considered to be sensitive. In the previous works (Dasseni, Verykios, Elmagarmid, Bertino, 2001, Oliveira, Zaiane, 2003 a) of association rule hiding the sensitive items are manually selected by the user. In the proposed work, the selection of sensitive items would require data mining process to be executed first. Based on the discovered rules sensitive items are selected. While generating the association rules the sensitive items has to be hidden. To identify the sensitive items an algorithm SENSIDENT is proposed.

The rest of the paper is organized as follows. Section 2 gives the view of the previous works. Section 3 presents the statement of the problem. Section 4 presents the proposed algorithm for selecting sensitive items. Section 5 shows the example of the proposed algorithm. Section 6 shows the experimental results of the performance of the algorithm. Concluding remarks and future work are described in Section 7.

2. Related Work

In (Dasseni, Verykios, Elmagarmid, Bertino, 2001) the authors selected the rules manually, in order to hide the sensitive knowledge by reducing the support and confidence of the rules. In (Oliveira, Zaiane, 2003 a), in order to protect the sensitive knowledge in association rule mining, the sensitive rules are selected. In (Oliveira, Zaiane, 2003 b) authors introduced a heuristic approach to hide restrictive association rules that requires the victim item. The victim item has to be identified manually and it has to be removed from its transactions. In (Wenliang Du, Zhijun Zhan, 2003), the authors selected the sensitive attributes and protect the attributes by randomization process.

Instead of selecting the sensitive rules and sensitive data manually, the proposed algorithm helps to identify the sensitive data.

3. Problem Statement

The problem of mining association rules was introduced in (Agrawal, Imielinski, Swami, 1993). Let $I = \{i1, i2, \ldots, im\}$ be a set of literals, called items. Given a set of transactions D, where each transaction T in D *is* a set of items such that $T \preceq I$, an association rule is an expression $X => Y$ where $X \preceq I$, $Y \preceq I$, and $X \cap Y = \Phi$. The X and Y are called respectively the body (left hand side) and head (right hand side) of the rule. An example of such a rule is that 90% of customers buy hamburgers also buy Coke. The 90% here is called the confidence of the rule, which means that 90% of transaction that contains X (hamburgers) also contains Y (Coke). The confidence is calculated as $|X \cup Y| / |X|$, where $|X|$ is the number of transactions containing X and $|X \cup Y|$ is the number of transactions containing both X and Y. The notation U here is not the set union operator. The support of the rule is the percentage of transactions that contain both X and Y, which is calculated as $|X \cup Y| / N$, where N is the number of transactions in D. In other words, the confidence of a rule measures the degree of the correlation between item sets, while the support of a rule measures the significance of the item sets. A typical association rule-mining algorithm first finds all the sets of items that appear frequently enough to be considered significant and then it derives from them the association rules that are strong enough to be considered interesting. The problem of mining association rules is to find all rules that are greater than the user-specified minimum support and minimum confidence

However, the objective of privacy preserving data mining is to hide certain sensitive information so that they cannot be discovered through data mining techniques (Oliveira, Zaiane, 2003 a, Oliveira, Zaiane, 2003 b). In this work, sensitive information is identified using the proposed algorithm.

4. Proposed Algorithm

In order to select the sensitive item(s) the frequent item sets and the association rules are generated based on the given support and confidence values (Agrawal, Imielinski, Swami, 1993). By selecting the consequent of the rule, the sensitive item can be identified. One or more items can be selected as sensitive items. The algorithm SENSIDENT is described below:

Input:

(1) input database D

(2) minimum support

(3) minimum confidence

Output: sensitive item(s) to be hidden.

Algorithm:

1. Find the frequent item sets and generate the association rules from D using minimum support and minimum confidence.
2. Identify the rules with single antecedent and consequent.
   (i.e.) x -> y
3. Sort the rules in descending order based on the confidence values.
4. Select the rules with the highest confidence.
5. Count the frequency of the consequent (y) in the rules.
6. Sort the count in descending order.
7. Choose the highest two different counts c1, c2
8. If c1-c2 < threshold value
   If more than one consequent with the same count
   Select the corresponding consequents
   else
   Select the corresponding consequents
   else
   If one (or) more consequent with the same count c1
   Select the consequent(s)
9. Selected consequent(s) will be the sensitive item(s).

The threshold value is the user specified value, the minimum difference between the two counts. The selected sensitive items are further used to hide sensitive rules.

The algorithm SENSIDENT first tries to find the frequent item sets and the association rules using the Apriori algorithm (Agrawal, Imielinski, Swami, 1993). Rules are sorted and the frequency of the consequent items are considered and compared with the threshold value. The selected consequent items are finally considered as sensitive items.

5. Example

This section shows the example to demonstrate the proposed algorithm to select the sensitive items .

TID	Items
T1	ABC
T2	ABC
T3	ABC
T4	AB
T5	A
T6	AC

Frequent item sets are generated with minimum support 0.50. The following are the frequent item sets and its support value A (1.0), B (0.6), C (0.6), AB (0.6), AC (0.6), BC (0.5), ABC (0.5). Association rules with minimum confidence 0.75 are generated.

The rules and the corresponding confidence values are as follows:

C->B0.75, B->C0.75, C->A1.0, B->A1.0, BC->A1.0, AC->B0.75, AB->C0.75, C->AB0.75, and B -> AC0.75.

The rules with highest confidence are

C->A 1.0

B->A 1.0

From the above rules the item A is selected as the sensitive item with the threshold value 2.

6. Experimental Results

6.1 Methodology

In order to better understand the characteristics of the proposed algorithm numerically, a series of experiments is performed to measure various characteristics. The experiments are conducted on a PC, with Pentium IV Processor with 512 MB of RAM running on Windows XP Operating System. To measure the effectiveness, the dataset used are mushroom dataset from UCI machine learning repository (www.ics.uci.edu/mlearn), the two synthetic datasets are csc and c20d10.The mushroom dataset contains 128 different items, with 8124 transactions. The c20d10 dataset has 2000 transactions and 386 items. The csc contains 298 transactions and 88 items.

The Apriori algorithm is used to generate the frequent item sets and the association rules. The minimum support and minimum confidence, which are used to generate the frequent item sets and the rules, are given by the user. From that the rules with single antecedent and consequent are identified. Sorting is performed to find the highest confidence. The two consequents with highest counts are compared with the user threshold value. The threshold value is given to select the sensitive items with minimum difference in their counts.

6.2 Performance Evaluation

For each dataset, the frequent item sets and the association rules are generated with minimum support 0.40 and minimum confidence 0.50. Figure 1 shows the time effects of the various sizes of the data sets, in frequent item set generation. The mushroom dataset produces the frequent item sets in 14,843 ms, c20d10 in 3969 ms and csc in 79ms. The time can be varied with various minimum support and minimum confidence values. Figure 2 shows the selection of the rules with single antecedent and consequent. The mushroom data set is used for this effect. Under the minimum supports 0.4, 0.5 and 0.6 with the minimum confidence 0.75 the rules generated are 31, 19 and 7. The count of rule rapidly decreases with the increase in minimum support value. Figure 3 shows the count of sensitive items with various threshold value. The mushroom data set is used for this effect. With the threshold value 1, 5 and 10, under the minimum support value 0.5 and the minimum confidence value 0.75 the sensitive items identified are 1, 1 and 3. The counts of sensitive items are increased for higher values of threshold value.

7. Conclusion and Future work

In this work it is discussed about the objective of privacy preserving data mining, to hide certain sensitive information so that they cannot be discovered through data mining techniques. In the previous works for association rule hiding the sensitive information is selected manually.

The proposed algorithm is used to identify the sensitive information using the set of association rules. Results illustrating the algorithm are given. In future the sensitive items can be identified using the classification, clustering and regression techniques. Also the proposed work can be integrated with association rule hiding.

References

Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in large databases. *In: Proceedings of ACM SIGMOD International Conference on Management of Data*, Washington DC

Agrawal R, Srikant R (2000) Privacy preserving data mining. *In ACMSIGMOD Conference on Management Of Data*, Dallas, Texas, pp 439–4501.

Clifton C, Marks D (1996) Security and privacy implications of data mining. *In: SIGMOD Workshop on Research Issues on Data Mining and knowledge Discovery.*

Dasseni E, Verykios V, Elmagarmid A, Bertino E (2001) Hiding association rules by using confidence and support. *In: Proceedings of 4th Information Hiding Workshop*, Pittsburgh, PA, pp 369– 383

Evfimievski A (2002) Randomization in privacy preserving data mining. *SIGKDD Explorations* 4(2), Issue 2:43–48.

Oliveira S, Zaiane O (2003 a) Algorithms for balancing privacy and knowledge discovery in association rule

mining. *In: Proceedings of7th International Database Engineering and Applications Symposium (IDEAS03)*, Hong Kong

Oliveira S, Zaiane O (2003 b) Protecting sensitive knowledge by data sanitization. *In: Proceedings of IEEE International Conference on Data Mining*.

Wenliang Du , Zhijun Zhan(2003)Using Randomized Response Techniques for Privacy Preserving Data mining. *SIGKDD'03*

www.ics.uci.edu/mlearn

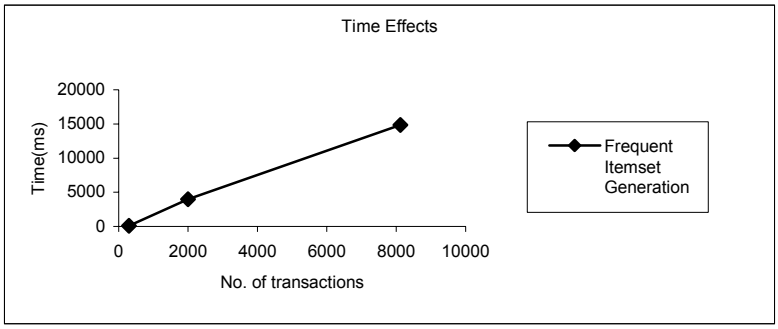


Figure 1. Time effects

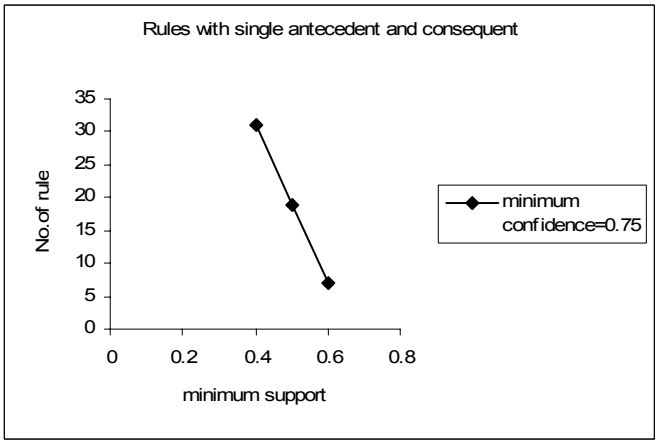


Figure 2. Rules with single antecedent and consequent

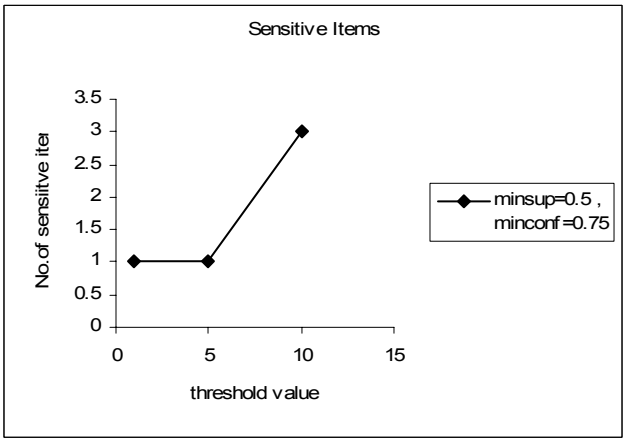


Figure 3. Sensitive items

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

Improved Gabor Filtering Application in the Identification of Handwriting

Yongping Liu (Corresponding author)

Henan Institute of Engineering

Zhengzhou 451191, China

E-mail: yongpingliu@sina.com

Xiaobo Guo

Henan Institute of Engineering

Zhengzhou 451191, China

E-mail: dvwt@163.com

Abstract

In this paper, handwriting image will be regarded as a texture image,The textural features of it were extracted by Improved multi-channel 2-D Gabor filtering,and was added in the features database as the basis for the identification of handwriting after processed.This is a content independent method,with a broad applicability.the speed and accuracy of identification were increased after optimizing the parameters of Gabor filtering, the author also made a lot of experiments by the platform of vc++6.0, it proved the effectiveness of the algorithm.

Keywords: Handwriting identification, Multi-channel Gabor wavelet transform**,** Texture analysis, Feature extraction

1. Introduction

Handwriting identification is a very effective method in distinguishing among identities, it has many significant advantages compared to other means of identification, such as handwriting has the peculiarity of uniqueness, stability, acquisition and non-invasive(Yang,Zi-Hua,2004,pp.67-79). The extraction of different handwriting features is the core of handwriting identification system, According to the object and the characteristics extracted,it can be divided into two methods(text-independent and text- dependent), most of the current means need to rely on the letter being identified(the most typical is signature verification),this method lacks broad applicability. The handwriting identification is intended to get a person's writing style, without too much concern to the specific content of the writing.this paper presents an improved two-dimensional multi-channel Gabor transform the text of an independent handwriting identification methods,the paper put forward a text-independent handwriting identification method based on the improved two-dimensional multi-channel Gabor transform. this method exacted the texture characteristic of handwriting image in different frequencies and directions fastly and accurately by means of Gabor filters.

2. The preprocessing of handwriting image

Because handwriting image itself has many noise information, in order to obtain the unified texture image and make preparations for characteristics extraction,image preprocessing must be carried out, the main pretreatment steps are as follows:

(1) Remove the background grid of the paper and gray the handwriting image. First,according to octree structure color quantization algorithm, convert the handwriting bitmap's color depth to 8-Bit bitmap to facilitate computer processing. In order to eliminate the impact of paper type, background colors, stains, grid lines, as well as other Mixed colors on the handwriting, the author designed a Color-screen obtaining method to take the color of handwriting text(indicated with RGB), a threshold is intercalated to reset the colors which have a large difference between the text color to background color. The ink color of handwriting text has nothing to do with with the handwriting written style,the image can be grayed to reduce the complexity of the system according to the weighted average method(He,Bin,2002).

(2) Remove the noise and binarize the image.Eliminate the random noise introducted when the image was scanned according to filtering algorithm,then binarize the handwriting image to make it only contain 0 and 255 gray levels.

(3)Character Segmentation and normalized treatment. Project the handwriting image in the horizontal direction,the

trough between two adjacent peaks in the projection curve corresponding to the gap between two lines,the distance between the two trough corresponding to the character height of a line. Every character's width and the space between characters can be obtained by vertical projection on each line. Set a threshold set according to the regional average gray to remove too sparse (random graffiti) and too crowded (stains) handwriting. Calculate the size of each text and the scaling factor in horizontal and vertical direction scaling comparing to standard size,then can normalize the character facilitately and remove spaces between words and gap between lines simultaneously.

(4)The text block splice. After the above pretreatment steps,the author splice the normalized hangwriting image into blocks,each block has the size of 256*256,we can splice the charcaters to obtain a unified texture image in the circumstances that Handwriting image itself contains only a few characters(For example, signature, criminal password,ect)

3. Extract handwriting features by Gabor filters.

Texture analysis is an important image analysis method in the frequency domain.Texture is the element gradation distributed orderliness of the image,it manifests the shape and reciprocity of texture element.The handwriting image may regard as to constitute by a group of texture unit,statistics of the whole image texture unit can reflect the writer's handwriting characteristic. The Gabor transformation used in the texture analysis overcomes the time-frequency localization contradictory flaw of the Fourier transformation, It has been recognized as one of the best signal expressive methods in the correspondence and the signal processing fields, particularly in image expression,and applied to many image processing areas(Shen,Cong,2002,pp.20-25).

The Gabor filter's form commonly used in the extraction of image texture features is(Wang,Yun-Hong,2001,pp.229-234):

$$\begin{cases} h_e(x,y) = g(x,y,\sigma)\cos\left[2\pi f(x\cos\theta + y\sin\theta)\right] \\ h_o(x,y) = g(x,y,\sigma)\sin\left[2\pi f(x\cos\theta + y\sin\theta)\right] \end{cases}$$

h_e and h_o denote odd and even Gabor filters, f,θ,σ are three important parameters of Gabor filter: Spatial frequency, phase and space constant. $g(x,y,\sigma)$ specifies the Guass function:

$$g(x,y,\sigma) = \frac{1}{2\pi\sigma^2}\exp\left[-\frac{x^2+y^2}{2\sigma^2}\right]$$

This Gabor filter's frequency form is:

$$\begin{cases} H_e(u,v) = \dfrac{H_1(u,v) + H_2(u,v)}{2} \\ H_o(u,v) = \dfrac{H_1(u,v) - H_2(u,v)}{2j} \end{cases}$$

$j = \sqrt{-1}$ is the Imaginary unit. And

$$\begin{cases} H_1(u,v) = \exp\left\{-2\pi^2\sigma^2\left[(u - f\cos\theta)^2 + (v - f\sin\theta)^2\right]\right\} \\ H_2(u,v) = \exp\left\{-2\pi^2\sigma^2\left[(u + f\cos\theta)^2 + (v + f\sin\theta)^2\right]\right\} \end{cases}$$

The purpose of Gabor transform is analysing the handwriting image on the multi-scale and in multi directions according to the multi-resolution characteristic of two-dimensional Gabor wavelet.Extrac a number of filter coefficients from many sub-planes as the statistical characteristics.Choose a group of different filter parameter, then the output image on different channel can be obtained,which is equivalent to the original image in different expansions on the basis function(Liu.Cheng-Lin,1997,pp.56-63).

Supposes the texture image is described by $f(x,y)$,and it's Fourier transform by $F(u,v)$,the filter by $h(x,y)$, we record the filter output image's duplicate response output as $q(x,y)$,that is: $q(x,y) = q_r(x,y) + jq_i(x,y)$,then

$$\begin{cases} q_r(x,y) = real\{q(x,y)\} = h_e(x,y) \otimes f(x,y) = FFT^{-1}\left[F(u,v) \bullet H_e(u,v)\right] \\ q_i(x,y) = imag\{q(x,y)\} = h_o(x,y) \otimes f(x,y) = FFT^{-1}\left[F(u,v) \bullet H_o(u,v)\right] \end{cases}$$

Define power matrix:

$$q(x,y) = \sqrt{q_e^{\,2}(x,y) + q_o^{\,2}(x,y)}$$

Take the average value and the variance of the power matrix from each channel as the handwriting texture features.

4. The choice of filter parameters

Gabor filters mainly decided by three parameters:the direction θ, the radial center frequency f and constant space σ, it is a band-pass filter centered by (f,θ),the filter's wavelength decided by f ,the Gabor kernel function's direction decided by θ,and σ is Gaussian envelope standard deviation in x and y directions, its value and the filter center frequency is in inversely proportion,assign $\sigma = 2\pi / f$ in this paper. A group of Gabor filters can be obtained through the selecting different parameters,then gain a group of non-orthogonal basis, the information of frequency-domain in different frequency and phase can be obtained by expanding the signal under these basis.

As Gabor filter is conjugate symmetry in frequency domain, so choose the direction parameters variable in 0^o~180^o,four phase interval parameters θ were chosen In the literature [4],they were: $0, \pi/4, \pi/2, 3\pi/4$,the filter's center frequency and the extraction of the characteristics relate to the scale of the texture. For a image size of $N \times N$, the experienced choice of center frequency is $f \le N/2$, The lower center frequency,the bigger the scale of texture analysis,the weaker reflection of handwriting characteristic change.

The unified handwriting texture image is obtained after pretreatment,choose the Gabor filter's center frequency f as 4,8,16,32,64,128 for each phase θ,then we have 24 Gabor filtering channels for 4 direction angles, each filter has the respective choice of frequency and direction, after the handwriting image filtrated,24 group of output coefficients can be obtained, take the power matrix's mean (M) and variance (S) as a characteristic value from each channel output,then a 48-dimensional feature vector is gained.

$$M = \sum_{x=0}^{255} \sum_{y=0}^{255} q(x,y) \Big/ 256^2$$

$$S = \sqrt{\sum_{x=0}^{255} \sum_{y=0}^{255} [q(x,y) - M]^2 / 256^2}$$

5. The improvement of multi-channel Gabor filter

The non-orthogonality of Gabor filter mains there having redundant information after images filtered,thus affect the speed and discrimination capabilities, This article proposed the following selection strategy of the best multi-channel Gabor filter group based on the above experiments,and experiment proved the feasibility of the method.

Suppose there have H sample images belong to C categories, record each sample image as $\{I_t\}, t = 1, 2 \cdots H$,then the Gabor features of each sample image can be manifested as $f_t^{(i)} = \left\{ M_{tj}^{(i)}, j = \{1, 2\mathrm{L}\ 24\}, i = \{1, 2\mathrm{L}\ C\} \right\}$,the category of the sample image is denoted by i ,the channel of the filter is denoted by j , t is the mark of sample, $t \in \{1, 2\mathrm{L}\ H\}$,for the output of all filtes, assume that only preserve the j-dimensional features,that is $f_t^{(i)} = M_{tj}^{(i)}, j \in \{1, 2\mathrm{L}\ 24\}$, then calculate the dispersion matrix S_W in one kind and dispersion matrix S_B between kinds for each category(Bian.Zhao-Qi,2002). The formula is:

$$S_W^{(j)} = \sum_{i=1}^{C} E\left[(M_{tj}^{(i)} - m_j^{(i)})(M_{tj}^{(i)} - m_j^{(i)})^T \right]$$

$$S_B^{(j)} = \sum_{i=1}^{C} E\left[(m_j^{(i)} - m_j)(m_j^{(i)} - m_j)^T \right]$$

$m_j^{(i)} = \overline{M_{tj}^{(i)}}$ shows the mean of jth eigenvalue for all the samples belonging to one category, m_j is the jth eigenvalue's average value of all samples.

Use the criterion:

$$R_j = (|S_B^{(j)} + S_W^{(j)}|) \Big/ |S_W^{(j)}|$$

to judge each eigenvalue's classification ability for different kind of handwriting image,choose the channels corresponding to Several largest values of R_j to filter the handwriting image.

After choosing the most superior Gabor filter group,carry on bandpass filter to the handwriting image with them,select the mean and variance of the output image to idenfify the handwriting.By such optimized processing, both the Gabor filter's channel quantity and the comparing time of characteristic are reduced,it also enhanced the reliability of the identification. 25 individuals each 10 handwriting samples were collected to carry on the massive experiments in this article, and a group composed of the best 12 multi-channel Gabor filters were decided,The parameters of these filters are shown in table 1.

6.The classification of handwriting image and experimental analysis

For simplicity,use the weighted Euclidean distance classifier to classify, compare the unknown handwriting's eigenvector with the sample's that has been trained,only when its eigenvector has the minimum weighted Euclidean distance(WED) between category k,the input handwriting is classified to category k.the WED can be calculated with the formula:

$$WED = \sum_{1}^{N} \frac{(f_i - f_i^{(k)})^2}{(\delta_i^{(k)})^2}$$

f_i denotes the ith character of the unknown samples, $f_i^{(k)}, \delta_i^{(k)}$ denote the mean and the variance of the ith character of category k,and N is the total number of characters for each sample.

The author sort the checked materials according to the distance between the characteristic of each category in the storehouse in the experiment,choose the writers whose handwriting image is near to the samples as the candidate,12individuals each 5 handwriting samples were collected in an experiment,four for training and one for identifing. Because the algorithm calculate the overall image texture features,the character in the handwriting texture picture can be completely different.Take a 256*256 image block with unify texture from each handwriting image,then get the image's Gabor filtering characteristics of each category through the best multi-channel Gabor filter,compute the distance between the checked image and each classified texture image in database,and sort according to the result,then observe the position of one person's handwriting image.

The author encode the checked materials and the corresponding sample,for example,the first person's handwriting is classfied to Y1,the second person's handwriting is classfied to Y2,the first person's checked material is named J1,the second person's checked material is named J2,and so on.gain the most likely three categories according to the distance.The table 2 gives the relevant data of handwriting identification respectively based on the multi-channel Gabor filter and the best multi-channel Gabor filter.

For 12 candidates' checked materials and classfied samples match,when adopt multi-channel Gabor filtering algorithm,there are seven cheeked samples arrange first(accounts for 58.3%) and ten arrange first two(accounts for 83.3%),when adopt the best multi-channel Gabor filtering algorithm proposed in this article,the result is ten(accounts for 83.3%) and eleven(accounts for 91.7%) for the same materials,and the time the latter used in feature extraction and classification is about half of the former, it shows that the improved feature extraction algorithm is effective. The method is also of significant reference meaning for other image processing question,such as face recognition,vehicle license plate recognition,ect.

References

Bian, Zhaoqi., & Zhang, Xuegong. (2002). *Pattern recognition.* (2rd ed.). Beijing: TSINGHUA UNIVERSITY PRESS.

He, Bin., & Ma, Tianyu. (Dec., 2002). *Digital image processing in Visual C++*. (2rd ed.). Beijing: POSTS & TELECOM PRESS.

Liu, Chenglin., Liu, Yingjian.,&Dai, Ruwei. (1997). *WRITER IDENTIFICATION BY MULTICHANNEL DECOMPOSITION AND MATCHING*. Beijing: BActa Automatica Sinica. 1997. 23(1). 56-63.

Shen, Cong. (2002). *Handwriting Identification Based on Improved Multi-channel Gabor Wavelet*. Beijing: Beijing University of Technology.

Wang, Yunhong., Tan, tieniu., & Zhu, yong. (2001). *WRITER IDENTIFICATION BASED ON TEXTURE ANALYSIS.* Beijing: BActa Automatica Sinica. 2001. 27(2). 229-234.

Yang, Zihua., &Wu, Min. (2004). *Handwriting Identification System Based on Texture Analysis,* Journal of Hunan Institute of Engineering(Natural Science Edition). 2004, 14(2): 67-69.

Table1. the parameters of the best muti-channel Gabor filters

Center frequency	4	8	16	32	64
Direction angle	135°	90° 135°	45° 135°	0° 45° 90° 135°	45° 90° 135°

Table 2. Comparison of different Gabor filtering result

Multi-channel Gabor filter		Best multi-channel Gabor filter	
Code of checked material	Ranking results of the corresponding category	Code of checked material	Ranking results of the corresponding category
J1	2	J1	1
J2	1	J2	1
J3	1	J3	1
J4	3	J4	2
J5	2	J5	1
J6	1	J6	1
J7	6	J7	3
J8	1	J8	1
J9	1	J9	1
J10	1	J10	1
J11	1	J11	1
J12	2	J12	1

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

High Availability with Diagonal Replication in 2D Mesh (DR2M) Protocol for Grid Environment

Rohaya Latip (Corresponding author)

Faculty of Computer Science and Information Technology

Universiti Putra Malaysia

Tel: 60-3-8946-6536 E-mail: rohaya@fsktm.upm.edu.my

Hamidah Ibrahim, Mohamed Othman, Md. Nasir Sulaiman, Azizol Abdullah

Faculty of Computer Science and Information Technology

Universiti Putra Malaysia

Tel: 60-3-8946-6510 E-mail: hamidah, mothman, nasir, azizol@fsktm.upm.edu.my

The research has been supported by Malaysian Ministry of Science, Technology and Innovation (MOSTI) under the Fundamental Grant No: 02-01-07-269FR.

Abstract

Replication is a useful technique for distributed database systems and has been implemented in EU data grid and HEP in CERN for handling huge data access. Replica selection in their prototypes still can be enhanced to provide high availability, fault tolerant and low in communication cost. This paper introduces a new replica control protocol, named Diagonal Replication in 2D Mesh (DR2M) for grid environment and compares its performance with the previous protocols. The performance in this paper is data availability for read and write operation, which are compared to the Read-One Write-All (ROWA), Voting (VT), Tree Quorum (TQ), Grid Configuration (GC), Three Dimensional Grid Structure (TDGS), and Diagonal Replication on Grid (DRG). This paper discusses the protocol of replicating data for grid environment, putting the protocol in a logical 2D mesh structure by employing the quorums and voting techniques. The data file is copied in a selected replica from the diagonal sites in each quorum. The selection of a replica depends on the diagonal location of the structured 2D mesh network where the middle replica is selected because it is the shortest path to get a copy of the data from most of the direction in the quorum. The algorithm in this paper also calculates the optimized number of nodes to be grouped in each quorum and how many quorums are needed for the number of nodes, *N* in a network. DR2M protocol also ensures that the data for read and write operations are consistent, by ensuring the quorum must not have a nonempty intersection quorum. To evaluate the DR2M protocol, we developed a simulation model in Java. Our results prove that our protocol improves the performance of data availability compared to the previous data replication protocol, namely Read-One Write-All (ROWA), Voting (VT), Tree Quorum (TQ), Grid Configuration (GC), Three Dimensional Grid Structure (TDGS), and Diagonal Replication on Grid (DRG).

Keywords: Data replication, Grid, Data management, Availability, 2D Mesh protocol

1. Introduction

A grid is a distributed network computing system, a virtual computer formed by a networked set of heterogeneous machines that agree to share their local resources with each other. A grid is a very large scale, generalized distributed network computing system that can scale to internet size environment with machines distributed across multiple organizations and administrative domains (Krauter, 2002; Foster, 2002). Ensuring efficient access to such a large network and widely distributed data is a challenge to those who design, maintain and manage the grid network. The availability of a data in a large network and replicating data at a minimum communication cost are also some of the issues (Ranganathan, 2001; Lamehamedi, 2002; Lamehamedi, 2003; Lamehamedi, 2005). EU data grid and HEP in CERN used "Reptor" as the prototype to manage replica in grid (Guy, 1997; Kunzst, 2003). Figure 1 shows the main component of the Replica Management System, "Reptor" was implemented for EU Data Grid. In our work, we investigate on the replica selection for optimizing and improving data accessing by using replica control protocol in distributed database to the grid environment.

Distributed computing manages thousands of computer systems and it has a limited memory and processing power. On the other hand, grid computing has some extra characteristics. It is concerned to efficient utilization of a pool of

heterogeneous systems with optimal workload management utilizing an enterprise's entire computational resources (servers, networks, storage, and information) acting together to create one or more large pools of computing resources. There is no limitation of users or originations in grid computing. Even though minimum number of nodes for grid is one but for DR2M protocol the best minimum number of nodes should be more than five to suite the large network size such as grid environment.

Quorums improved the performance of fault tolerant and availability of data (Mat Deris, 2003; Mat Deris, 2004; Yu, 1997). Quorums reduce the number of copies for reading or writing data. To implement quorum, a protocol must satisfy two constraints which are total of quorum for read, q_r and write quorum, q_w must be larger than the total number of votes, v assigned to the copies of the data object and the quorum for write, q_w is larger than $v/2$ (Mat Deris, 2001). For voting approach, every copy of replicated data object is assigned a certain number of votes and a transaction has to collect a read quorum of r votes to read a data object, and a write quorum of w votes to write a data object. To address the availability, DR2M replicates data on the middle node of a quorum of read or write in the logical structured of 2D mesh topology network. The term replica means the selected nodes that have the copy of the data file. Java is used to run this replication protocol.

The paper is organized as follow: in Section 2, DR2M protocol and its algorithm are introduced. In Section 3, we present the previous replica control protocols. This section includes the formulation for read/write availability for the previous protocols in distributed database and grid computing. Section 4 describes the simulation framework and Section 5 discusses the simulation results. Brief conclusions and future works are discussed in Section 6.

2. Diagonal Replication in 2D Mesh (DR2M) Protocol

In DR2M protocol, all nodes are logically organized into two dimensional Mesh structure. We assume that the replica copies are in the form of text files and all replicas are operational meaning that the copies at all replicas are always available. The data are replicated to only one node of the diagonal site which is the middle node of the diagonal site in each quorum.

This protocol uses quorum to arrange nodes in cluster. Quorum is grouping the nodes or databases into small cluster to manage the replica for read or write operations. Figure 2 illustrates how the quorums for network size of 81 nodes are grouped by nodes of 5 x 5 in each quorum. Nodes which are formed in a quorum intersect with other quorums. This is to ensure that each quorum can communicate or read other data from other nodes which is in another quorum.

The number of nodes grouped in quorum, R must be odd so that only one middle node from the diagonal sites can be selected such as the black circle in Figure 2, which reduces the communication cost. Example, $s(3,3)$ in Figure 2 is selected to have the copy of data.

2.1 The correctness

This section shows that DR2M protocol is accessing the updated and consistent data. From the definition 2.1 and proof shows the quorums intersect with each other and follow the two conditions of making sure that the data are consistent.

Definition 2.1: Assume that a database system consists of n x n nodes that are logically organized in the form of two dimensional grid structure. All sites are labeled $s(i,j)$, $1 \le i \le n$, $1 \le j \le n$. $D(s)$, is the diagonal sites, $D(s) = s(i,j)$, where $i = j = 1, 2, \ldots, n$ for each quorum.

For example, Figure 2 has 81 nodes where the size of the network is 9 x 9 nodes. For $q1$, the diagonal site $D(s)$ is $\{s(1,1), s(2,2), s(3,3), s(4,4), s(5,5)\}$ and the middle node $s(3,3)$ has the copy of the data file. Figure 2 has four quorums where each quorum actually overlaps with each other. Example node e in $q1$ is actually node a in $q2$ and node a in $q3$ is actually node e in $q4$.

Since the data file is replicated only on one node in each quorum, thus it minimizes the number of database update operations. The selected node of data file is assigned with vote one and the rest of the nodes will have vote zero. A vote assignment on grid, B, is a function such that,

$$B(s(i,j)) \in \{0, 1\},\ 1 \le i \le n,\ 1 \le j \le n$$

where $B(s(i,j))$ is the vote assigned to site $s(i,j)$. This assignment is treated as an allocation of replicated copies and a vote assigned to the site results in a copy allocated at the selected node. That is,

1 vote ≡ 1 copy.

Let $L_B = \sum B(s(i,j)), \qquad s(i,j) \in D(s)$

where L_B is the total number of votes assigned to the selected node as a primary replica in each quorum. Thus $L_B = 1$ in each quorum.

Definition 2.2: Let q_r and q_w denote the read quorum and write quorum respectively. To ensure that read operation always gets the updated data, $q_r + q_w$ must be greater than the total numbers of copies (votes) assigned to all sites. To

make sure the consistency is obtained, the following conditions must be fulfilled (Mat Deris, 2004).

i. $1 \le q_r \le L_B$, $1 \le q_w \le L_B$

ii. $q_r + q_w = L_B + 1$.

These two conditions ensure that there is a nonempty intersection of copies between read and write quorum. Thus, these conditions ensure that a read operation has the most recently updated copy of the replicated data. Let *S(B)* be the set of sites at which replicated copies are stored corresponding to assignment *B*, then,

$S(B) = \{s(i,j) | B(s(i,j)) = 1, 1 \le i \le n, 1 \le j \le n\}$

From Figure 2,

$q_r = \{s(3,3)$ from $q1, s(3,3)$ from $q2, s(3,3)$ from $q3, s(3,3)$ from $q4\}$

$L_B = 4$ for the whole network

and in DR2M protocol $q_r = q_w$ then

$q_w = \{s(3,3)$ from $q1, s(3,3)$ from $q2, s(3,3)$ from $q3, s(3,3)$ from $q4\}$

$L_B = 4$ for the whole network.

Definition 2.3: For a quorum q, a quorum group is any subset of $S(B)$ where the size is greater than or equal to q. The collection of quorum group is defined as the quorum set.

Let $Q(B,q)$ be the quorum set with respect to assignment B and quorum q, then

$Q(B,q) = \{G | G \subseteq S(B)$ and $|G| \ge q\}$

For example, from Figure 2, let site $s(3,3)$ be the primary database of the master data file m. The diagonal site are $s(1,1)$, $s(2,2)$, $s(3,3)$, $s(4,4)$, and $s(5,5)$. Consider an assignment B for the data file m, such that

$$B(s(1,1)) = B(s(2,2)) = B(s(3,3)) = B(s(4,4)) = B(s(5,5)) = 1$$

and $L_B = B(s(3,3))$. Therefore, $S(B) = \{s(3,3)\}$.

For simplicity, a read quorum for data file m, is equal to write quorum. The quorum sets for read and write operations are $Q(B,q1)$, $Q(B,q2)$, $Q(B,q3)$ and $Q(B,q4)$ as in Figure 2, where $Q(B,q1) = \{s(3,3)\}$ in $q1$, $Q(B,q2) = \{s(3,3)\}$ in $q2$, $Q(B,q3) = \{s(3,3)\}$ in $q3$, and $Q(B,q4) = \{s(3,3)\}$ in $q4$.

Therefore, the number of replicated data file m is equal to 4.

Theorem 2.1. The DR2M protocol is accessing consistent data.

Proof. The theorem holds on condition that the DR2M protocol satisfies the quorum intersection properties for the read-write properties and write-write properties by Definition 2.2.

The availability of a read ($A_{DR2M,R}$) and write ($A_{DR2M,W}$) operation is calculated as in Eq. (1) and Eq. (2) respectively, where n is the grid column or row size, example n is 7, for 7 x 7 nodes of grid network size and p is the probability of data available which is between 0 to 1, qr and qw are the number of quorums for read and write operations respectively. Thus, the formulation for read availability for DR2M, $A_{DR2M,R}$ is as given in Eq. (1)

$$A_{DR2M,R} = \sum_{i=qr}^{n} \binom{n}{i} (p^i (1-p)^{n-i}) \qquad (1)$$

and the formulation for write availability, $A_{DR2M,W}$ is as given in Eq. (2)

$$A_{DR2M,W} = \sum_{i=qw}^{n} \binom{n}{i} (p^i (1-p)^{n-i}) \qquad (2)$$

As the network size grows bigger, the more quorums are identified. The number of columns and rows in each quorum must be odd, to get the middle node. If the input of nodes, n is not odd then the simulator will create virtual empty nodes to the last row and column of the 2D mesh. By selecting only one middle node in each quorum has minimized the communication cost. The algorithm for this DR2M protocol is shown in Figure 3. It illustrates how the algorithm was designed and implemented.

3. Related Work

There are few protocols to replicate a data in distributed database and grid computing as discussed in the following subsections:

3.1 Read-One Write-All (ROWA) Protocol

ROWA is a simple and straightforward protocol (Ozsu, 1996). It requires all copies of all logical data items that are updated by a transaction be accessible for the transaction to terminate. Failure of one site may block a transaction and

reduce database availability.

In ROWA, a read operation needs only one copy, while a write operation needs to access the n number of copies. Therefore, the availability for a read operation can be represented as one out of n (Jain, 1991), and for a write operation as n out of n. Thus, the formulation for read availability for ROWA, $A_{ROWA,R}$ is as given in Eq. (3)

$$A_{ROWA,R} = \sum_{i=1}^{n} \binom{n}{i} p^i (1 - p)^{n-i} \quad (3)$$

and the formulation for write availability $A_{ROWA,W}$ is as given in Eq. (4)

$$A_{ROWA,W} = \sum_{i=n}^{n} \binom{n}{i} p^i (1 - p)^{n-i} \quad (4)$$

where p is the probability that a copy is accessible and p is from 0.1 to 0.9.

ROWA reduces the availability of the database in case of failure since the transaction may not complete unless it reflects the effects of the write operation on all copies. Therefore, there have been a number of algorithms that have attempted to maintain mutual consistency without employing the ROWA protocol (Ozsu, 1999).

3.2 Voting (VT) Protocol

In VT approach, every copy of replicated data object is assigned to a certain number of votes and a transaction has to collect a read quorum of r votes to read a data object, and a write quorum of w votes to write the data object. Quorum must satisfy two constraints which are $r + w$ must be larger than the total number of votes, v assigned to the copies of the data object and $w > v/2$, where the total of write quorum of w votes must be larger than half of the total number of votes, v.

For n replicas, VT protocol allows n choices for the read and the write quorum. It starts from (read 1, write n) to (read n, write 1). To avoid the read availability becomes expensive, the read quorum k is selected where k is smaller than the majority quorum.

From Mat Deris (2001), the formulation for read availability in VT, $A_{VT,R}$ is as given in Eq. (5), where n is the total number of nodes that has the votes or sometime it is called replica

$$A_{VT,R} = \sum_{i=k}^{n} \binom{n}{i} p^i (1 - p)^{n-i}, k \geq 1 \quad (5)$$

and the corresponding formulation for write availability in VT, $A_{VT,W}$ is as in Eq. (6), where k is the number of votes for read or write quorum

$$A_{VT,W} = \sum_{i=n+1-k}^{n} \binom{n}{i} p^i (1 - p)^{n-i} \quad (6)$$

This protocol is popular and easy to implement but writing an object is fairly expensive (Mat Deris, 2001), where write quorum (w) of copies must be larger than the majority votes (v), $w > v/2$.

3.3 Tree Quorum (TQ) Protocol

TQ protocol proposed by (Agrawal, 1992; Agrawal, 1990) is a logical tree structured over a network. Figure 4 illustrates the tree quorum structured. One advantage of this protocol is that a read operation may access only one copy and the number of copies to access write operation is always less than a majority of quorum. The size of quorum may increase to maximum of $n/2 + 1$ site when failures occur where n is the number of copies. Write operations in the TQ must access a write quorum, which is formed from the root, a majority of its children and a majority of their children and so forth until the leaves of the tree are reached. The size of a write quorum is fixed but the members can be different. The root or majority of the children of the root can form the read operations.

Let $A_{TQ,h,R}$ and $A_{TQ,h,W}$ be the availability of the read and the write operations with the height, h of tree, respectively. D denotes the degree of sites in the tree and M is the majority of D. Under this protocol, as given by Chung (1990), the availability formulation of a read operation for a tree of height h can be represented as in Eq. (7)

$$A_{TQ,h,R} = p + (1 - p) \sum_{i=M}^{D} \binom{D}{i} A_{TQ^i,h-1,R} \left(1 - A_{TQ,h-1,R}\right)^{D-i} \quad (7)$$

and the availability formulation of a write operation for a tree of height h is as given in Eq. (8)

$$A_{TQ,h,W} = p \sum_{i=M}^{D} \binom{D}{i} A_{TQ^i,h-1,W} \left(1 - A_{TQ,h-1,W}\right)^{D-i} \quad (8)$$

where p is the probability that a copy is available. If the height of a tree is one, the read operation, R_0 and write operation, W_0 is equal to p.

Tree quorum (TQ) uses quorum that is obtained from a logical tree structure imposed on data copies. However, if more than a majority of the copies in any level of the tree become unavailable write operation cannot be performed (Agrawal, 1992; Maekawa, 1992).

3.4 Grid Configuration (GC) Protocol

GC protocol proposed by (Maekawa, 1992) has n copies of data objects that are logically organized in the form of $\sqrt{n}$ x $\sqrt{n}$ grid as shown in Figure 5. In Figure 5, the number of nodes is 25 in 5 x 5 grid network where n is five. The figure shows three grey circles, which represent nodes that are down or not active. The nodes which are downed can be placed logically anywhere in the grid structure.

Read operations on the data item are executed by acquiring a read quorum that consists of data copy from each column in the grid, while write operations are executed by acquiring a write quorum that consists of all copies in one column and a copy from each remaining column.

From (Agrawal, 1996), the formulation of read availability in the GC protocol, $A_{GC,R}$ is as given in Eq. (9)

$$A_{GC,R} = \left(\sum_{i=1}^{\sqrt{n}} \binom{\sqrt{n}}{i} p^i (1-p)^{\sqrt{n}-i} \right)^{\sqrt{n}} \quad (9)$$

where p is the probability that a copy is accessible and p is from 0.1 to 0.9. While, the formulation of write availability in the GC, $A_{GC,W}$ is as given in Eq. (10).

$$A_{GC,W} = \left[1-(1-p)^{\sqrt{n}}\right]^{\sqrt{n}} - \left[\left(1-(1-p)^{\sqrt{n}}\right) - p^{\sqrt{n}}\right]^{\sqrt{n}} \quad (10)$$

This protocol requires a bigger number of read and write quorum for the read operations to be executed. Read quorum must be at every column and for write operations to be executed, write quorum must exist at one of the entire column and exist at least once at other columns. Thus, this decreases the data availability. It is also vulnerable to the failure of entire column or row in the grid (Agrawal, 1996).

3.5 Three Dimension Grid Structure (TDGS) Protocol

TDGS protocol replicated its data in logical box shape structure with four planes (Mat Deris, 2001; Mat Deris, 2007). Figure 6 illustrates eight copies of data object.

The read operations in TDGS are executed by acquiring a read quorum that consists of any hypotenuse copies. For the example shown in Figure 6, hypotenuse copies are {A, H}, {B, G}, {C, F}, and {D, E}. Read operations are executable by these pairs of hypotenuse copies.

Write operations are executable from any planes that consist of hypotenuse copy. Planes in Figure 6 are {H, A, B, C, D}, {C, F, E, G, H), and etc. Example, to execute read operation, copies from {A, H} must be accessible and to execute write operation, copies from {H, A, B, C, D} must also be accessible.

In TDGS protocol, read quorum can be constructed from four hypotenuse copies. From (Mat Deris, 2001), the formulation of read availability, $A_{TDGS,R}$ is as given in Eq. (11), where p is the probability that a copy is accessible and p is from 0.1 to 0.9.

$$A_{TDGS,R} = 1(1-p^2)^4 \quad (11)$$

whereas, formulation of write availability, $A_{TDGS,W}$ is as given in Eq. (12),

$$A_{TDGS,W} = 1-(1-\beta)^4 \quad (12)$$

where $\beta = p\,\varphi + p\,\phi - p^2(\varphi * \phi)$ (Mat Deris, 2001),

$\varphi = p^4(1+p-p^2)$ and $\phi = p^4(2-p^2)$

Read operations are executed by acquiring a read quorum, which must come in pairs at the vertices of the box shape structure. If one of the copies of each pairs is unavailable, thus that hypotenuse copies are not accessible. Therefore, write operations are not executable at the write quorum.

Read and write quorum must intersect otherwise the hypotenuse of read quorum is not accessible and the write quorum is also not accessible to update the latest data. This affects the consistency of the data.

3.6 Diagonal Replication on Grid (DRG) Protocol

Diagonal Replication on Grid (DRG) proposed by (Mat Deris, 2004), is a protocol which is logically organized in a two dimensional grid structure. For example, if a DRG consists of twenty-five sites, the network is logically formed into 5 x

5 grids as shown in Figure 5. Each site has a master data file. A site is either operational or failed and the state (operational or failed) of each site is statistically independent to the others. When a site is operational, the copy at the site is available; otherwise it is unavailable.

Sites can be down or not active. In Figure 5, sites 23, 24, and 25 are not active or fail. Sets of diagonal sites in the grid are selected such as set {1, 7, 13, 19, 25}. After the diagonal sites are identified, the primary copy of the data is placed on the replica which is distributed diagonally (Mat Deris, 2004). For example, diagonal set $D^2(s)$ is {s(2), s(8), **s**(14), s(20), s(21)} and $D^3(s)$ is {s(3), s(9), s(15), s(16), s(22)}.

Each site has the same copies of replica. Figure 7 illustrates the diagonal sets of D^2(s) which is grey in color and $D^3(s)$ which is in doted circles.

The total number of the diagonal nodes is presented as d and L_v is the total number of votes on grid. Let $S(B)$ be the set of sites at which the replicated copies are stored corresponding to B. Whereas i and j are the number of columns and rows respectively.

$$S(B) = \{S(i, j) \mid B(S(i, j)) = 1,\ 1 \le i \le n\ ,\ 1 \le j \le n\}$$

$$L_v = \sum_{s(i,j) \in D(s)} B(S(i, j)) = d$$

In estimating the availability for DRG, all copies are assumed to have the same availability p. Let Av(t) be the read/write availability of protocol t. If the probability that an arriving operation of read and write for data file x are f and (1-f), respectively, then the read/write availability can be defined as,

$$Av(t) = fAv(t_{read}) + (1\text{-}f)Av(t_{write})$$

For any assignment B and quorum q for the data file x, Eq. (13) defines $\varphi(B_x, q)$ to be the probability that at least q sites in $\Omega(B_x)$ are available, then

$$\varphi(B_x, q) = Pr\ \{\text{at least } q \text{ sites in } \Omega(B_x) \text{ are available}\}$$

$$= \sum_{G \in Q(B_x, q)} \left(\prod_{j \in G} P_j \prod_{j \in S(B_x) - G} (1 - P_j) \right) \qquad (13)$$

In Eq. (14) the availability of read and write operations for the data file x, are $\varphi(B_x, r)$ and $\varphi(B_x, w)$, respectively.

$$\text{Av(DRG)} = f\ \varphi(B_x, r) + (1\text{-}f)\ \varphi(B_x, w) \qquad (14)$$

4. Results and Discussion

In this section, DR2M protocol is compared to the results of read and write availability of the previous protocols, namely: ROWA, VT, TQ, GC, TDGS, and DRG. Figure 8 shows the results of read availability in a 81 nodes size of network. ROWA protocol has the highest read availability about average of 11.883% for probability of data accessing 0.1 to 0.9 even when the number of nodes increases. This is because only one replica is accessed by a read operation for any n nodes of network size but ROWA has the lowest write availability. Figure 8 illustrates the write availability for 81 numbers of nodes, where the probability is from 0.1 to 0.9.

The result shown in Figure 9 proves that the DR2M protocol has average of 28.708% higher for write availability for all probabilities of data accessing. This is due to the fact that replicas are selected from the middle location of the diagonal site in each quorum.

5. Conclusions

In this paper, a new protocol, called Diagonal Replication in 2D Mesh (DR2M) protocol has been proposed to manage the data replication in a large network size such as in distributed system and especially in grid environment. DR2M protocol, selects one replica in a diagonal site of a quorum in a 2D mesh logical structure. The number of nodes in each quorum is odd so it is easy to select only one node from the diagonal site in each quorum. The analysis of the DR2M protocol was presented in terms of read and write availability. The results demonstrate that the DR2M protocol provides a convenient approach for write operations. This is due to the minimum number of quorum size required. DR2M has overcome the read and write availability issues of TDGS which is the latest replica control protocol.

References

Krauter, K., et al. (2002). A taxanomy and survey of grid resource management systems for distributed computing. *International Journal of Software Practice and Experience,* 32(2), 135-164.

Foster, I., et al. (2002). Grid services for distributed system integration. *Computer,* 35(6), 37-46.

Ranganathan, K., & Foster, I. (2001). Identifying dynamic replication strategies for a high performance data grid.

Proceedings of International Workshop on Grid Computing, Denver.

Lamehamedi, H., Szymanski, B., Shentu, Z., & Deelman, E. (2002). Data replication strategies in grid environment. *Proceedings of ICAP'03,* Beijing, China, IEEE Computer Science Press, Los Alamitos, CA, 378-383.

Lamehamedi, H., Shentu, Z., & Szymanski, B. (2003). Simulation of dynamic data replication strategies in data grids. *Proceedings of the 17th International Symposium on Parallel and Distributed Processing,* 22-26.

Lamehamedi, H. (2005). *Decentralized data management framework for data grids*. New York: Rensselaer Polytechnic Institute Troy, (Ph.D. thesis).

European Datagrid Project. Advanced replica management with reptor. [Online] Available: http://www.eu-datagrid.org

Guy, L., et al. (1997). *Replica management in data grids.* Presented at Global Grid Forum 5.

Kunzst, P., et al. (2003). Advanced replica management with reptor. *5th International Conference on Parallel Processing and Applied Mathematics*, Czestochowa, Poland.

Mat Deris, M., et al. (2003). Binary vote assignment on grid for efficient access of replicated data. *International Journal of Computer Mathematics*, Taylor and Francis*,* 1489-1498.

Mat Deris, M., Abawajy, J. H., & Suzuri, H. M. (2004). An efficient replicated data access approach for large scale distributed systems. *IEEE/ACM Conf. On Cluster Computing and Grid (CCGRID2004)*, Chicago, USA.

Yu, T. W., & Her, K. C. (1997). A new quorum based replica control protocol*. Proceedings of the 1997 Pacific Rim International Symposium on Fault-Tolerant Systems,* 116-121.

Mat Deris, M. (2001). *Efficient access of replication data in distributed database systems*. Universiti Putra Malaysia. (Ph.D. thesis).

Mat Deris, M., et al. (2004). High system availability using neighbor replication on grid. *2nd Workshop on Hardware/Software Support for High Performance Scientific & Engineering Computing, Special Issue in IEICE Trans*. On Information and System Society, E87-D (7), 1813-1819.

Ozsu, MT., & Valduriez, P. (1996). Distributed and parallel database systems. *ACM Computing Surveys*, 28(1).

Jain, R. (1991). *The art of computer systems performance analysis.* Wiley.

Ozsu, M.T. (1999). *Principles of distributed database systems*. New Jersey Prentice Hall Second Edition.

Agrawal, D., & El Abbadi, A. (1992). The generalized tree quorum protocol: an efficient approach for managing replicated data. *ACM Transactions Database System,* 17(4), 689-717.

Agrawal, D., & El Abbadi, A. (1990). The tree quorum protocol: an efficient approach for managing replicated data. *Proceeding 16th International Conference on Very Large Databases,* 243-254.

Chung, SM. (1990). Enhanced tree quorum algorithm for replicated distributed databases. *International Conference on Database,* Tokyo, 83-89.

Maekawa, M. (1992). A √n algorithm for mutual exclusion in decentralized systems. *ACM Transactions Computer System,* 3(2), 145-159.

Agrawal, D., & El Abbadi, A. (1996). Using reconfiguration for efficient management of replicated data. *IEEE Transactions on Knowledge and Data Engineering,* 8(5), 786-801.

Mat Deris, M., H. Abawajy, J., & Mamat, A. (2007). An efficient replicated data access approach for large-scale distributed systems. *Future Generation Computer Systems,* 24, 1-9.

Mat Deris, M., et al. (2004). Diagonal replication on grid for efficient access of data in distributed database systems. *ICCS 2004, LNCS 3038,* 379-387.

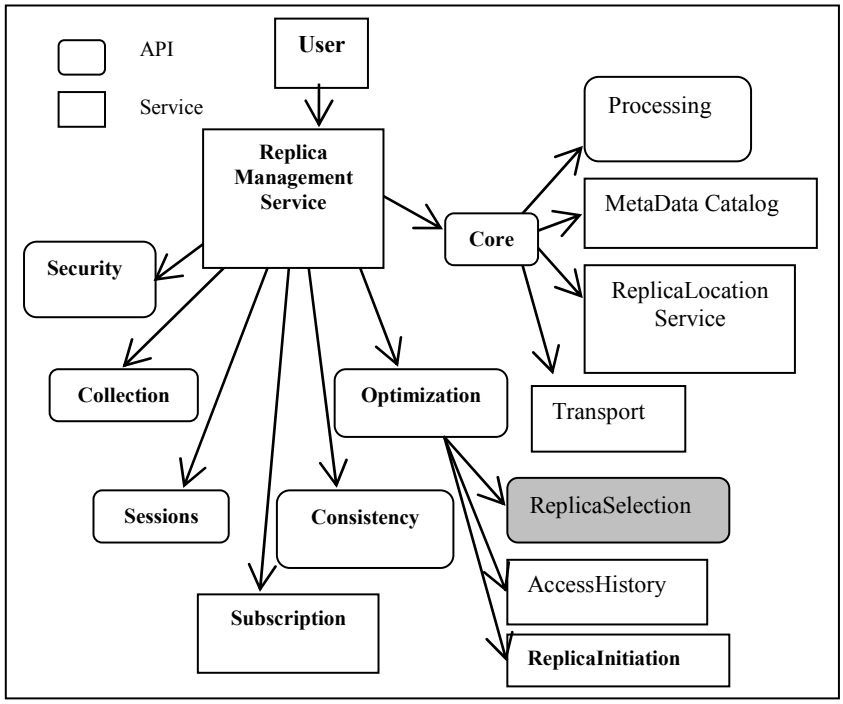


Figure 1. The main component in Replica Management System

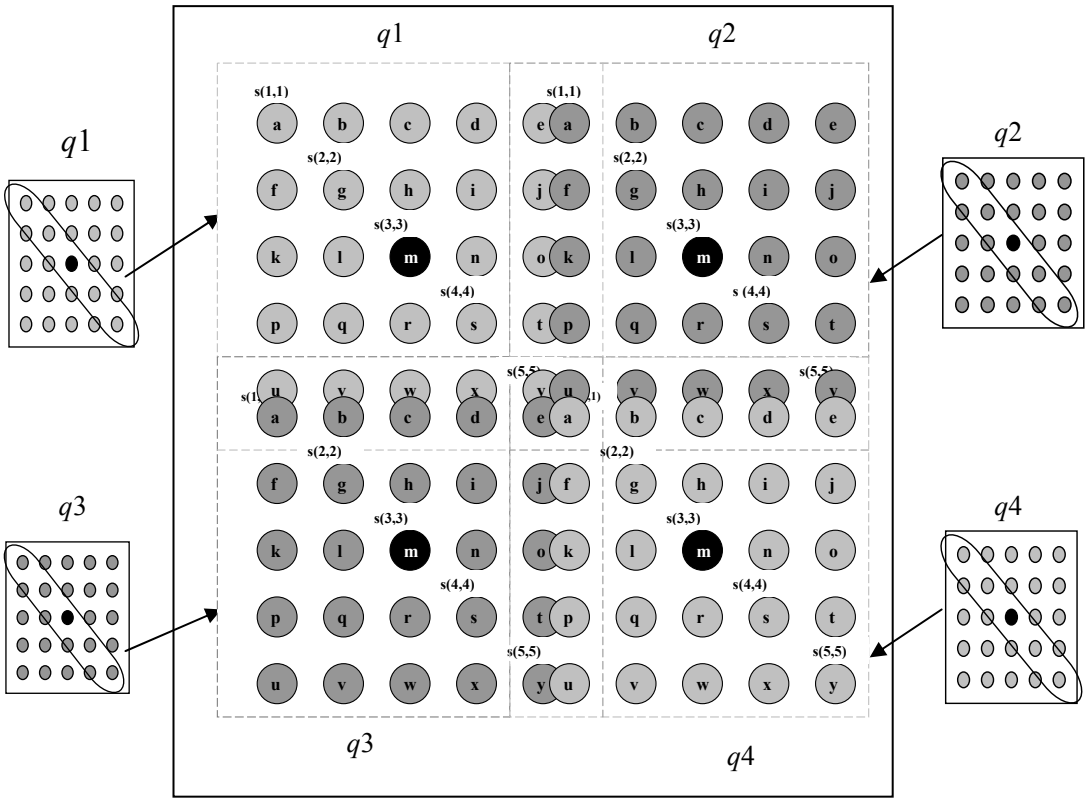


Figure 2. A grid organization with 81 nodes, each of the nodes has *a* data file *a*, *b*,…, and *y* respectively.

Main

Input number of row or column, N

n = next odd number after $\sqrt{N}$

If n is an odd integer then

Find the number of quorum, Q

$q = \left\lfloor \sqrt{n} - \frac{n}{10} \right\rfloor$

$Q = q^2$

Find the number of nodes X in each quorum,

$X = n/q$

Get the next odd integer after X

Select the middle replica, R

Copy the data file at R

Else add one virtual column and row, $Col_{new} + Row_{New}$

$N = Col_{new} * Row_{New}$ then

Return N to Main.

Figure 3. Algorithm of the data replication in DR2M protocol

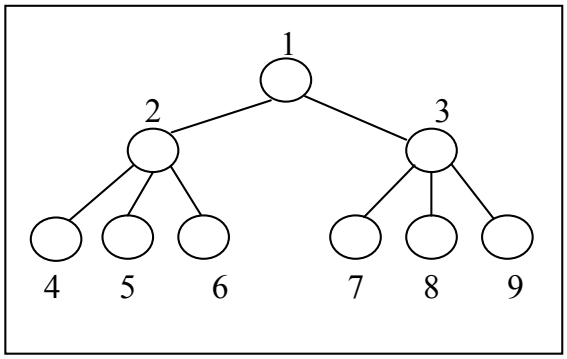


Figure 4. A tree organization of 9 copies of data object.

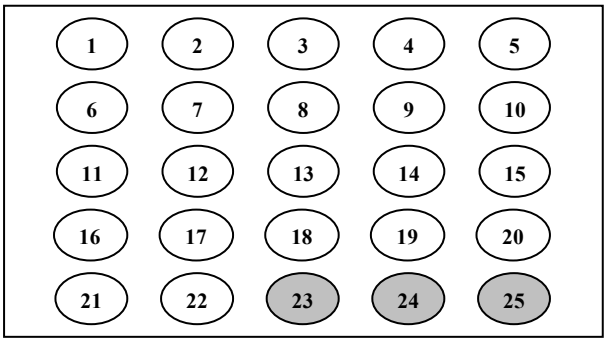


Figure 5. 25 copies of sites in DRG

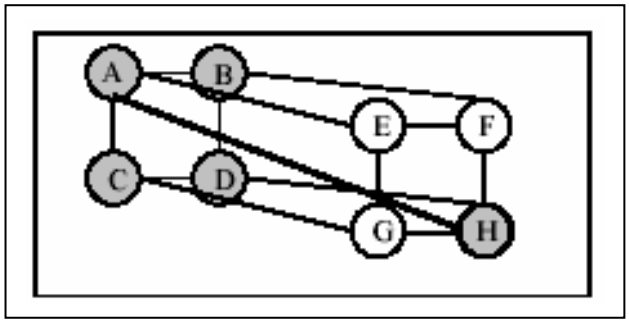


Figure 6. Eight copies of data object

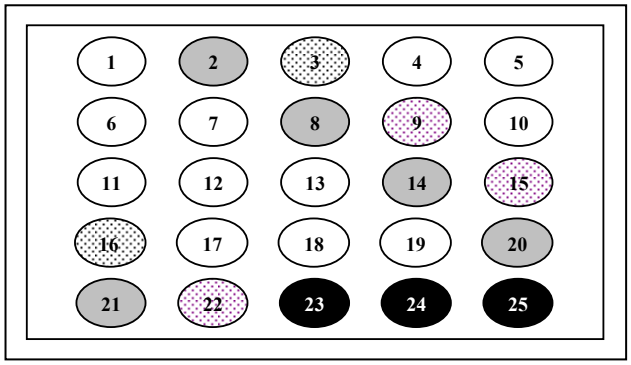


Figure 7. Diagonal sets of $D^2(s)$ and $D^3(s)$

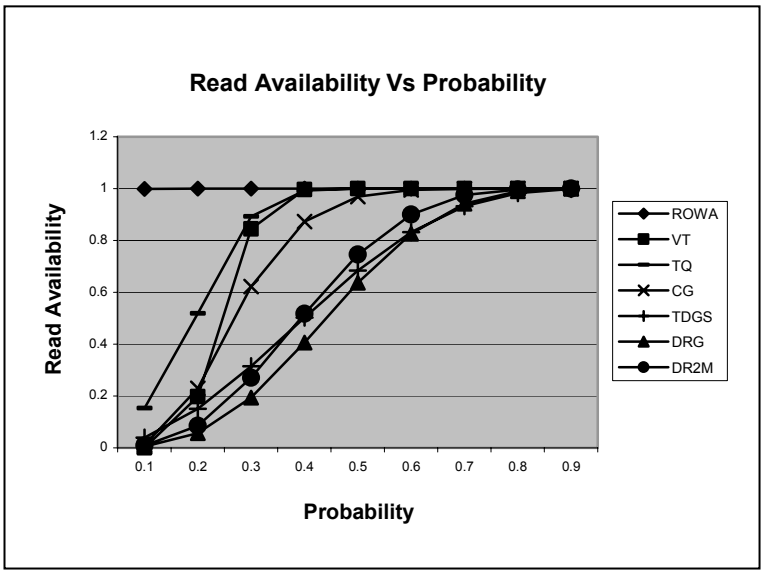


Figure 8. Results for read availability

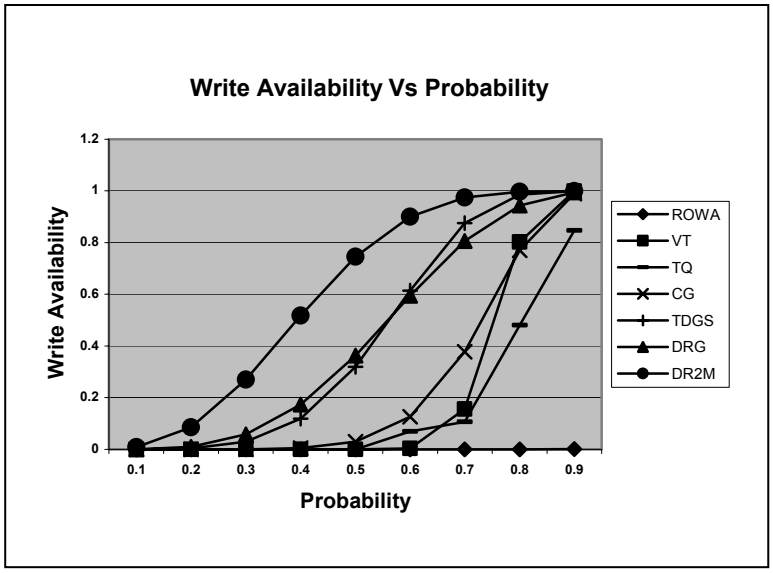


Figure 9. Results for write availability

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

A New Mixing Programming Method

Yuan Cui

Department of Computer Science

Chengdu Medical College

Chengdu 610083, China

Tel: 86-028-8818-0917 E-mail: bubblecui@163.com

Abstract

This paper is aimed at the realization of merging programming of MATLAB and VB. We mainly discuss how to realize merging programming by MATLAB Automation Server technology, and demonstrate how to incorporate visual programming interface of VB with the powerful function of MATLAB in numerical calculation and graphic display.

Keywords: Matlab, VB, Program, OLE

1. Introduction

MATLAB, developed by Mathworks co., is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation. Because of powerful matrix computation, it is called Matrix Lab. Matlab Language is similar with natural language. The usage of Matlab is extremely convenient. It also have plenty functions that can be called easy. But, it also has some shortcomings. Because the MATLAB language is a kind of explanation execution script language, it is very slow regarding the loops sentence execution. Under the same condition, compared to some high level languages such as Vc and VB, It executes loops sentences at the slow speed. The graphic user interface (GUI) of The MATLAB is not very friendly, and the parameter input and the output is not convenient. VB, as a kind of high level computer language, executes loops sentences more quickly, and its GUI is user-friendly. By incorporating visual programming interface of VB with the powerful function of MATLAB in numerical calculation and graphic display, we can take their advantages and avoid their weaknesses.

2. Methodology

There are two methods to merging programming of Matlab and vb.

2.1 Transfer the Matlab functions into dynamic link library (DLL). Matlab provides a conversion tool, which can transform its own functions into DLL files by VC.The high level languages such as VB, VC and VB can call DLLs very easily.

2.2 By some communication Technology between two different tools such as (Dynamic Data Exchange) DDE, Object Linking and embedding (OLE) automation and ActiveX controls. In this method, Matlab is taken as a object server and client can control Matlab by the associated methods of Matlab objects.

Among the above mentioned 2 methods, Among above mentioned 2 methods, the process of the first method is more complicated, require the operator higher computer level, and it is also time-consuming.

Compared to method one, the second is easy to carry out and have a lot of advantages. Therefore, in this paper, we mainly discuss method two. One interested in the other method can further read the reference.

MATLAB can not only be the automation controller, can also be the automation server. In order to realize the call the function of MATLAB, one can take Matlab as the automation server, naturally, VB as the automation controller.

Before calling one must at first know the name of the Matlab ActiveX objects in the system registry, namely, ProgID.Generally, the name is MATLAB. Application or MATLAB.Application.Single.

The former means to automate MATLAB the server be a share of server, other procedures can adjust to use;

The former means that the Matlab automation server is a shared server and other program can call its service. While, the latter means that Matlab is taken as a exclusive server and, if other program need to call its service, a new Matlab automation server have to run. The concrete realizations of the hybrid programming is illustrated as the following,

(1) Run the Matlab automation server

In VB integrated development environment, the sentences to run Matlab automation server is as the following,

```
Dim matlab as object
Set Matlab=createobject("matlab.application")
//Had better minimize the automation server after it run.
MATLAB. MinimiZeCommandWindow().
```

(2) Input and output of data

In Matlab, there are two methods to input and output data, PutFullMatrix and GetFullMatrix. The definition of GetFullMatrix in Matlab is as the following,[3,4]

```
void GetFullMatrix(
        BSTR Name,
  BSTR Workspace,
SAFEARRAY(double) pr,
    SAFEARRAY(double) pi);
```

This function can transfer one or two dimensional arrays in Matlab workspace into VB program. According to its definition, VB can call this function of Matlab. For example, one need to transfer matrix A in default Matlab workspace into a 4X4 matrix in VB.

The definition of PutFullMatrix in Matlab follows,

```
void PutFullMatrix(
BSTR Name,
BSTR Workspace,
SAFEARRAY(double) pr,
SAFEARRAY(double) pi);
```

Function PutFullMatrix can transfer the arrays variables in VB into Matlab matrix. For instance, the following sentence will load complex number with real part called Mreal and imaginary part Mimage into matrix A in Matlab workspace.

an instance that demonstrates how to use these two functions as the following,

```
Dim MatLab As Object
Dim XReal(5, 5) As Double
Dim XImag(5, 5) As Double
Dim ZReal(5, 5) As Double
Dim ZImag(5, 5) As Double
Dim i, j As Integer
For i = 0 To 4
    For j = 0 To 4
    XReal(i, j)    = Rnd * 6
    XImag(i, j) = 0
    Next j
Next i
Set Matlab = CreateObject("matlab.application")
Call MatLab.PutFullMatrix("A", "base", XReal, XImag)
Call MatLab.GetFullMatrix("A", "base", ZReal, ZImag)
```

(3)Execute the Matlab sentences.

While the automation server is running, by calling the methods 'execute'of Matlab automation server, one can execute any sentences in Matlab. For instance, running the following codes in VB,

```
x = 0:pi/100:8*pi;
```

```
y = sin(x)+cos(2*x);
plot(x,y);
title('Graph of X and Y');
xlabel('X - Axis');
ylabel('Y - Axis');
//at first save the above codes into the LINES property of component MEMO1 and then run the following code
FOR loop= 0 TO Memo.Lines.Count-1
        MATLAB.Execute(Memo.Lines[loop]);// Run Matlab sentences line by line
    next
```

(4) Display the figures,created in Matlab, in VB

Matlab have powerful ability in scientific drawing. VB can use Matlab to draw figures and then display it in VB proram[5]. The codes is like,

```
MATLAB.execute('figure(gcf)');//display the current figures created by Matlab
MATLABn.execute('saveas(gcf,"c:\temp.bmp")');// save the figures into temporary files
Image1.Picture.LoadFromFile('c:\temp.bmp');
// display the figures in IMAGE Component of VB. see the fig.1.
MATLAB.execute('close'); // close the display windows created by Matlab.
```

(5) Close Matlab automation server.

Having finished all calculations, in order to save system resources and memory, Matlab automation server should be closed. The method of closing is easy. Just one line, MATLAB. quit;

Through the above five steps, we realized VB and Matlab mixed programming. This approach is simple and effective, suitable for use in a variety of applications.

3. Conclusions and Discussions

Matlab have poor performance when implementing cyclic sentences and its GUI is unfriendly. all these shortcomings of Matlab is just the advantages of VB. But the VB's ability of drawings is worse than Matlab. Mixing program will take good use of their advantages and make software development become more concise and efficient. However, we should note that the codes created by this method can run in computers with installing Matlab, which, to some extent, limits the portability of the system. the methods of removing this restrictions is to compile the concerned Matlab functions in to DLL(dynamic link library) files by VC++ and then call them in VB.

References

Zhu, Xiaosong & Guo, Xiaoli. (2003). Discussions on vb and Matlab mixing program. *eletronic technology allications.* 2003, 029(009). 18-19. (in Chinese).

Deborah's rudimentary home page. http://www.djpate.freeserve.co.uk/MATLAB. htm.

Liu, Zhijian. (2000). The users's guide of Matlab application interface, *Science Press*. (In Chinese).

Xu, Wenshang, (2003). Matlab and vb based data exchange and integration development. *industry control computer*. 2003, 16(11). (in Chinese).

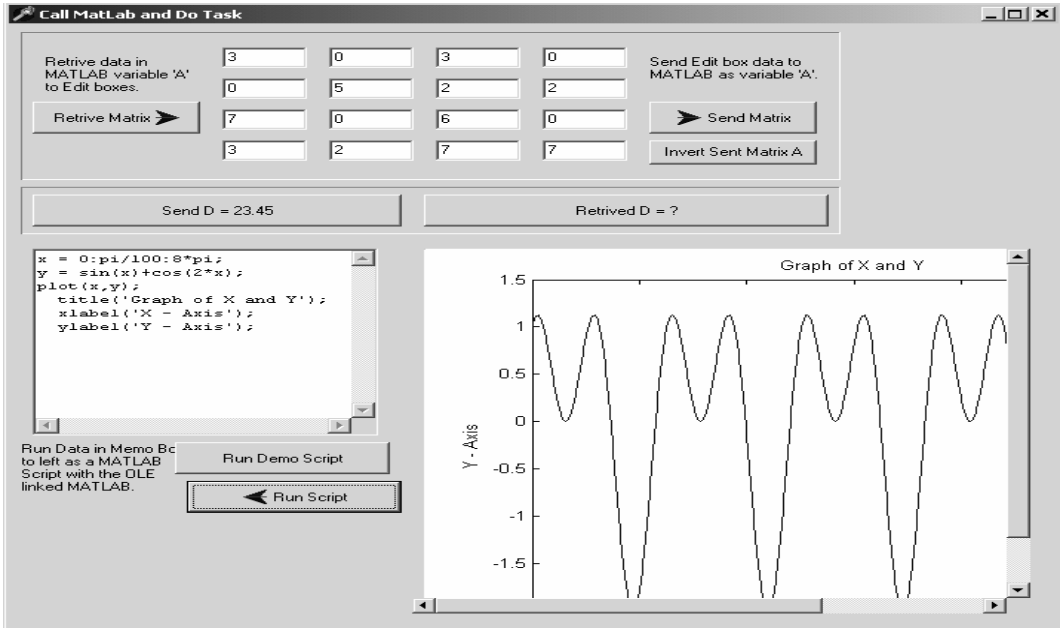


Figure 1. The results(the left is the code and the right is the results.)

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

A Study of Taxpayers' Intention in Using E-Filing System: A Case in Labuan F.T s

Azleen Ilias

School of International Business and Finance Labuan

Universiti Malaysia Sabah

Labuan International Campus 87000, F.T Labuan, Malaysia

Tel: 60-87-460-517 E-mail: neelza80@yahoo.co.uk

Norazah Mohd Suki

Faculty of Industrial Management

Universiti Industri Selangor (UNISEL)

Bestari Jaya Campus

Jalan Timur Tambahan 45600 Berjuntai Bestari

Selangor Darul Ehsan, Malaysia

Tel: 60-33-280-6068 E-mail: azahsuki@yahoo.com

Mohd Rushdan Yasoa'

Rahida Abdul Rahman

School of International Business and Finance Labuan

Universiti Malaysia Sabah

Labuan International Campus 87000, F.T Labuan, Malaysia

Tel: 60-87-460-714 E-mail: rahida70@yahoo.com

Abstract

This study used the Technology Acceptance Model (TAM) to examine taxpayers' intention in using e-Filing system. Data are collected from three higher learning institutions' staffs particularly in Labuan F.T. The purpose of this study is to determine the relationship between taxpayers' intention to use e-Filing towards attitude, perceived usefulness, perceived ease of use, information system quality, information quality and perceived credibility of the system. Further, this study also examines critical determinant (attitude, perceived usefulness, perceived ease of use, information system quality, information quality and perceived credibility of the system) in TAM that influence most taxpayers' intention. This study has replicated few methods from previous studies. i.e (Davis., et al. (1989), DeLone and Mclean (1992), Wang.,Y.S. (2002) and Chang., I.C., et.al.(2005)). This study is analyzed with reliability analysis, correlation analysis and Standardized Regression Weight (using Structural Equation Modelling). The result confirms a strong relationship between TAM determinants and taxpayers' intention. Consequently, this paper will assist Inland Revenue Board (IRB) to improve their Internet e-Filing system. This in turn, will be useful for them to educate Malaysian taxpayers in order to improve their attitude using e-Filing as their mode to file tax return undoubtedly.

Keywords: Technology acceptance model (TAM), E-Filing, Structural equation modeling, Tax, Labuan

1. Introduction

1.1 Introduction

TAM was developed to explain computer usage. The goal of TAM is to provide an explanation of the determinants of computer acceptance that is capable of explaining user behavior across a broad range of end user computing technologies and user populations, while at the same time being both economical and theoretically justified (Davis, 1989).

1.2 Malaysian E-Filing

Previously, the tax return is manually processed. At present, Inland Revenue Board (IRB) shift to a new paradigm towards e-Filing due to Self Assessment System (SAS) and will focus more on audit field. IRB also decided to aim for paperless. E-Filing process is more convenience, fast, accurate and secured in terms of payments. There are four steps in e-Filing. First, taxpayers need to enroll and verify a digital signature and MyKad into the reader. Then, they are required to enter the gross earnings, relief and deductions before the system compute automatically. After that, IRB will receive the tax form electronically and email verification of tax form return will be sent to taxpayers. The most important aspect in e-Filing system is security. As stressed out by IRB CEO, the e-Filing system is secure and difficult to get into anyone's personal tax file because they need to enter PIN (personal identification number) and a password (The Star Online, April 27th, 2007). According to IRB, about 657,000 taxpayers filed their Self-Assessment System (SAS) for year 2006 through the e-Filing system compared to year 2005 which were only 189,048 taxpayers (Berita Harian, June 26th, 2007). Thus, the number of taxpayers used this e-Filing system as a method of filing tax return increased by 247%.

1.3 Problem Statement

It seems clear that there is lack of study that has been conducted in the area of TAM among e-Filing taxpayers', particularly in Labuan F.T. Thus, the study aims to contribute to the knowledge of information technology. Due to the problem on using e-Filing, taxpayers' were uncomfortable with e-Filing as they were unfamiliar with electronic transactions and some said they were not computer savvy (The Star, May 4th, 2006). In addition, most of taxpayers' were very concerned if IRB directly changes the whole manual tax return process to e-Filing because of their inability to use Internet and less computer skill. Bedsides, slow response to e-Filing was mainly because of people's habit of doing their assessment at last minute. Some of them are difficult to accept a new technology since they are very concern about the security.

1.4 Significant of the Study

Through this study, the main reason of taxpayers' intention in using e-Filing will be explored. In addition, this study will assist Lembaga Hasil Dalam Negeri (IRB) to improve e-Filing performance according to TAM determinants and in line with the Government's Information Technology Policy. Based on the result of this study, it is expected there are more trainings and seminars might be conducted in order to improve e-Filing usage and compliance. However, this system will be fully implemented after it is widely accepted (Rahimah Abdullah, 2006). Besides, the IRB chief executive officer and director-general have stated that the process of upgrading e-Filing system services for taxpayers will improve the level of taxpayers' compliance. Consequently, the intention of taxpayers will be improved positively and in turn it will increase their compliance toward the IRB.

1.5 Objective of the Study

The objective of this study is twofold:

To determine the relationship between taxpayers' intention to use e-Filing towards attitude, perceived usefulness, perceived ease of use, information system quality, information quality and perceived credibility of the system.

To examine critical determinant (attitude, perceived usefulness, perceived ease of use, information system quality, information quality and perceived credibility of the system) in TAM that contributes most to influence taxpayers' intention.

The remainder of this paper is organized as follows. A review of related literature on technology acceptance model and research questions is discussed. Next, the methodology employed in this study, research instruments used and data analysis method involved are described. Finally, the empirical results and discussion of the study are drawn

2. Literature Review

2.1 Technology Acceptance Model

The TAM adopts the theory of reasoned act (TRA) model to explore the IT acceptance. TAM and TRA, both of which have strong behavioral elements, assume that when someone forms an intention to act, they will be free to act without limitation (Davis. et al., 1989). In addition, Davis et al., (1989) also stated that TAM indicates both perceived usefulness (PU) and perceived ease of use (PEOU) as key independent variables that determine or influence potential users' attitudes toward IT intention of use.

This study also used DeLone and McLean model of information system success (2003) consists of information system quality (ISQ) and information quality (IQ). Another new dimension is perceived credibility of a computer system developed by Wang (2002) and Chang et al. (2005).

Previously, Wang (2002); Chang et al. (2005); Hung. et al., (2006); and Fu et al., (2006) applied TAM in their study on tax filing methods especially in Taiwan. However, most of researchers construct hypothesized affect the use of Internet

tax filing indirectly through their affect on perceived usefulness (PU), perceived ease of use (PEOU), information system quality (ISQ), information quality (IQ), and perceived credibility (PC) toward attitudes of using (ATT) and behavior intention (BI).

However, there has been little study on TAM and DeLone and McLean model particularly in e-Filing system. Nevertheless, Lai et al. (2004) examined the level of technology readiness of Malaysian tax practitioners and their usage intention towards an e-Filing system. They found a significant positive relationship between the level of technology readiness and the usage intention towards the e-Filing system. Besides, Hanudin et al. (2006) has applied TAM particularly on the intention to use the SMS as a mode for banking transactions. According to this study, perceived expressiveness, perceived usefulness and perceived ease of use are important determinants of intention to use SMS banking among male respondents.

2.2 Technology Acceptance Model (TAM) Determinants

Attitude

Attitude is defined in terms of individual preferences and interests regarding the use of Internet tax-filing system. The measurement is adapted from Davis et al. (1989). Attitude is one of important determinant in increase the level of behavior intention among taxpayers. This can be supported by Chang et al. (2005) that stated as attitude has a significant impact on behavior intention (BI) of using the system.

Perceived Usefulness

Perceived usefulness is defined as the degree of taxpayers' believes from using Internet tax-filing system that would enhance their job performance and the measurement adapted from Davis (1989). In addition, Davis (1989) has stated that perceived usefulness was found to have a strong influence on people's intentions. However, Chang et al.(2005) study has found that perceived usefulness has no direct impact on behavior intention but has significant on attitude, which consequently has an impact on behavior intention of using the system.

Perceived Ease of Use

Perceived ease of use was defined as the degree to which a user aspects the use of Internet tax-filing system to be free of effort and was measured by Davis (1989). In Davis (1989), perceived ease of use which test to had a smaller but significant affect that subsided over time. According to Chang et al., (2005), perceived ease of use also found to have a significant impact on attitude, thus affects behavior intentions.

Information System Quality

According to DeLone and McLean (2003), information system quality is associated with the issue of whether the technical components of delivered is provide the quality of information and service required by stakeholders. Besides, information system quality was defined by the degree to which the technical components of Internet tax-filing provide the quality information and service required by users (Chang et al., 2005).

Information Quality

Based on Chang et al. (2005), information quality has been defined by the degree to which users are provided with quality information regarding their needs. Information quality also represents the users' perception of the output quality generated by an information system and includes such issues as the relevance, timeliness and accuracy (DeLone and McLean, 2003).

Perceived Credibility

Perceived credibility is defined as the extent of users' confidence in the Internet tax-filing system's ability to protect the user's personal information and security. This measurement was adapted from Wang (2002). According to Chang et al. (2005), a credible website needs to safeguard personal information from unauthorized access or disclosure, accidental loss and alteration or destruction. In Lai et al. (2004) study, some of the respondents specifically expressed that they would only use the e-Filing system if the IRB could assure them that the e-Filing system were safe and secure, and if the usability and reliability of the e-Filing system were fully tested and well documented.

2.3 Research Questions

RQ1: How strong the six TAM determinants influence taxpayers' intention?

RQ2: What are critical determinants in TAM those contribute most to influence taxpayers' intention?

3. Research Methodology

3.1 Study Sampling Procedure

The population of this study covered e-Filing users from three higher learning institutions in Labuan F.T. Before data collections were carried out, phone calls were made to each institution in order to ensure total number of academic and

administration staffs. Respondents consist of experience and non-experience individual taxpayers from Universiti Malaysia Sabah (UMS), Institut Latihan Perindustrian (ILP) and Pusat Matrikulasi Labuan (PML). Questionnaires were distributed during January until April 2007. This is due to the fact that individual taxpayers need to submit their e-Filing return before April 30th 2007. We have distributed 80 questionnaires for every institution and the total samples are 240 respondents. Finally, only 100 respondents completed and returned the questionnaires, which represents about 42.0% response rate.

3.2 Instrumentation

The questionnaire has two sections namely TAM determinants and demographic section. The instrument of this study is based on Chang et. al (2005). It presents a new set of instrument with some modification according to seven variables using seven-point Likert scale. The variables consist of behavior intention, attitude, perceived usefulness, and perceived ease of use (Davis, 1989); information system quality and information quality (DeLone and McLean, 1992) and perceived credibility (Wang, 2002). Demographic section consists of gender, education level, job, time of computer using, experience of handling e-Filing and experience of learning e-Filing.

4. Result

4.1 Data Analysis and Results

Based on Table 1 below, a total respondent of gender was fairly accounted (50%:50%). Most of the respondents are degree holder (67%); followed by certificate holder (14%), lower than STPM (8%), STPM (6%) and diploma (5%). Besides, 63% of the respondents are academician and 13% are non-academician. It is followed by support staff (24%). In addition, about 35% of respondents claimed that they used computer more than 28 hours per week. Most of respondents do not have an experience in handling and learning e-Filing system. About 33% of them have experience in handling e-Filing system and 36% have learning experience through IRB courses and seminar.

Table 2 presents the Reliability Analysis, Cronbach's Alpha reliability coefficients of the six determinants and behaviour intention (dependent variable). All the determinants were all above 0.7. It seems that this study provides quite reliable instruments because the score is higher than Chang (2005). For example, the behaviour intention is 0.96 as compared to 0.94; attitude = 0.96 (0.94); and perceived credibility = 0.80 (0.70). Perceived credibility is a new determinant and presents more reliable compared to previous study. Reliability less than 0.6 is considered poor, those in the 0.7 ranges, acceptable, and those 0.8 good (Sekaran, 2000). It is of evidence that the Cronbach's alpha value for the seven factors in this study ranged from 0.8 to 0.97. Therefore, the internal consistency reliability of the measures used in this study can be considered to be good.

4.2 Correlation Analysis

A correlation analysis in Table 3 indicates that all the six TAM determinants (i.e. attitude, perceived usefulness, perceived ease of use, information system quality, information quality and perceived credibility of the system) are positively correlated. Each construct shares greater variance with its own block of measures than with the other constructs representing a different block of measures.

4.3 Structural Equation Modeling

In order to answer the second objective of this study which is to examine the critical determinant (attitude, perceived usefulness, perceived ease of use, information system quality, information quality and perceived credibility of the system) in TAM that contributes most to influence taxpayers' intention tu use e-Filling system, advanced statistical technique know as Structural Equation Modeling (SEM) was utilized.

SEM is a versatile statistical technique that is particularly useful for analyzing nonexperimental data (Byrne, 2001). It has become an increasingly popular data-analytic technique in psychology, counseling, and rehabilitation. Quintana and Maxwell (1999) highlighted several applications of SEM to research, including the use of SEM for testing for mediational relationships, interaction effects, and mean differences; for confirmatory factor analysis and multiple sample analysis; for longitudinal designs; and for handling missing data. Recent innovations have allowed SEM to become a broad data-analytic framework with flexible and unique capabilities. Furthermore, SEM involves an analysis of carefully defined a priori hypotheses about the relationships among both measured and latent variables. It is imperative for researchers to become familiar with this data-analytic technique so that they can use this technique in their research endeavors. It is equally important for practitioners to become familiar with SEM to make judicious assessments of published studies.

4.4 Model-estimation

Analysis of Moment Structure (AMOS) Version 5 was used to estimate the model using SEM with observed variables. Recognition of the reliability of AMOS computations has been established by its increasing use in published studies in reputable journals over the last few years (e.g. Zuroff et al., 1999). Prior to model estimation, each of the multi-item constructs were transformed into totalled scores using equally weighted scales developed from the results of the CFA.

This path analytic procedure was used due to the complexity and difficulty of using a full structural equation model. For a similar use of this technique, see Li and Calantone (1998, p. 88) and the references cited by these authors to justify this approach.

4.5 Model Testing Results

The structural model was assessed by using established measures and evaluative criteria for model fit. Several goodness-of-fit indices are commonly used to evaluate how well the structural model fits the data. The chi square goodness-of-fit test is one of the most commonly used indices. In SEM, a nonsignificant chi square value is an indication that the hypothesized model has a good fit with the data. The problem with using chi square, however, is that it is hypersensitive to sample size (Ullman, 2001). Because SEM is grounded in large-sample theory, finding well-fitted hypothesized models, where the chi square value approximates the degrees of freedom, has proven unrealistic, leading SEM methodologists to develop additional practical or ad hoc indices of fit.

One approach is to divide the chi square (χ^2) value by the degrees of freedom. According to Carmines and Mclver (1981), χ^2/df ratios in the range of 2:1 or 3:1 indicated an acceptable fit between the hypothetical model and the sample data. The most popular alternative measures of fit for SEM analysis, however, are the goodness-of-fit index (GFI), the normed fit index (NFI), the comparative fit index (CFI), and the root mean square error of approximation (RMSEA). The GFI, NFI, and CFI all have values ranging from 0 to 1; a good fit is indicated by values greater than .90 for GFI and NFI and .95 and greater for CFI. For RMSEA, a value of 0 is interpreted as an exact fit; values less than .05 are a close fit, values between .05 and .08 are a fair fit, values between .08 and .10 are a mediocre fit, and values more than. 10 are a poor fit. Regarding the precision of the RMSEA estimates, AMOS reports a 90% confidence interval around the RMSEA value. MacCallum, Browne, and Sugawara (1996) indicated that a small RMSEA and a very narrow confidence interval suggest good precision of the RMSEA value in reflecting model fit in the population. Finally, Martens (2005) indicated that chi square/df, GFI, and NFI tend to be substantially affected by sample size and number of indicators per factor and do not generalize well across samples. Marten (2005) recommended using CFI and RMSEA as the primary goodness-of-fit indexes.

The results suggest that the data fit the current conceptual model well, with a χ^2 of 9.158 (df =6, p =0.165), χ^2/df =1.526, GFI =0.975, CFI =0.996, NFI =0.987, and RMSEA =0.073. Moreover, the squared multiple correlation for the predictors of INT is 0.701, which shows that the variables included in the model explain 70.1 per cent of the variance in the outcome variable. In other words, the error variance of INT is approximately 29.9 percent of the variance of INT itself.

4.6 Hypotheses Testing

The current study proposed to test six hypotheses in identifying the critical determinant in TAM that contribute most to influence taxpayers' intention. Details of the hypotheses are stated below:

H1: Attitude of the taxpayers has significant influence on taxpayers' intention to use e-Filing

H2: Perceived usefulness of the system has significant influence on taxpayers' intention to use e-Filing

H3: Perceived ease of use of the system has significant influence on taxpayers' intention to use e-Filing

H4: Information system quality has significant influence on taxpayers' intention to use e-Filing

H5: Information quality of the system has significant influence on taxpayers' intention to use e-Filing

H6: Perceived credibility of the system has significant influence on taxpayers' intention to use e-Filing

The results of the hypotheses testing are accessible in Table 4. It can be clearly seen in the table that from a total of six hypotheses stated above, significant relationship was found in **Hypothesis 1** between the influence of attitude of the taxpayers on their intention to use e-Filing (β=0.605, C.R=5.955, p=0.000). By comparing the standardized path coefficient of this construct among the other proposed hypotheses, attitude construct lead the list. Therefore, it is certainly become the most critical determinant in TAM that contributes most to influence Malaysian taxpayers' intention to use e-Filing. This infers that taxpayers have positive and strong attitude and interest to use e-Filing. The upside to e-Filing is that it's greatly reducing the volume of paperwork. New technology allows a paperless filing system and provides taxpayers with an electronic version for their files as well. A paperless process saves paper, toner, and file storage costs. Another benefit of e-Filing was that it enabled e-filers to be more productive, presumably because it saves on paperwork costs, makes it easier to correct errors, and is quicker than filing on paper. Further, e-Filing reduced number of errors, attributable to the one-time entry of figures and the checks performed by preparation software.

Next, **Hypothesis 2** avers that perceived usefulness of the system has significant influence on taxpayers' intention to use e-Filing. Table 4 and Figure 1 illustrate that the effect of perceived usefulness of the system on taxpayers' intention to use e-Filing was not recognized (β=0.026, C.R=0.223, p=0.823). Malaysian taxpayers' have less intention to use

e-Filing because of the downside they perceived to obtain such as using an electronic technology that they may not thoroughly understand can be daunting. At the same time, the idea of a paper-based tax return over which they have more control is more comfortable. They still prefer to go to the basics where traditional way to fill up form using pen and paper.

Further, null hypothesis of **Hypothesis 3** asserts that the extent of perceived ease of use of the system has significant influence on taxpayers' intention to use e-Filing. The comparison of the effect as reflected in Table 4 evinced that perceived ease of use of the system (β=0.059, C.R=0.502, p=0.616) had insignificant relationships on taxpayers' intention to use e-Filing. Thus, Hypothesis 3 of perceived ease of use of the system has significant influence on taxpayers' intention to use e-Filing was not established. The present study provides evidence of less effect of perceived ease of use of the system in relation to Malaysian taxpayers' intention to use e-Filing. There are Malaysians taxpayers who are still reluctant to use e-Filing to reveal tax transactions though they can e-file any time of the day or night, and both complex and simple returns can be filed electronically. Most probably, some taxpayers feel that this e-Filing is very complex and they need to take a longer time to accustom with the system.

Meanwhile, **Hypothesis 4** states that information system quality has significant influence on taxpayers' intention to use e-Filing. Inspection of the structural loading of the standardised path coefficients results as exemplified in Table 4 showed insignificant relationship between information system quality and taxpayers' intention to use e-Filing (β=0.189, C.R=1.139, p=0.255). Thus, Hypothesis 4 of information system quality has significant influence on taxpayers' intention to use e-Filing was not acknowledged. However, it should be noted that information system quality is an important construct to influence Malaysian taxpayers to use e-Filing as it has significant influence with perceived usefulness, perceived ease of use, attitude of taxpayers, information quality and perceived credibility of the system with p-values of 0.000 (see Table 4). It is clearly understood that e-Filing can be done at any time from a personal computer. The taxpayer able to get his/her refund quicker or schedule payment to meet the deadline. Indeed, Malaysian taxpayers found that the data transmitted nearly instantaneously to the clearinghouse and then to the system, nothing needs to be scanned computers because the data is transmitted directly into them. However, there still a bigger drawback to doing own taxes is that taxpayer might make errors and/or overlook legitimate deductions. Taxpayers also face risk making mathematical errors. It is advisable to check arithmetic carefully. Even if use a computer, inputting errors occur, thus check each field carefully to make sure figures correspond to withholding and expense records. Occasionally, glitches also show up even in reputable tax programs.

The hypothesis about information quality of the system has significant influence on taxpayers' intention to use e-Filing (**Hypothesis 5**) was not confirmed (β=-0.026, C.R=-0.237, p=0.812). Certainly, barriers to adoption of e-Filing perceived by Malaysian taxpayers includes a continued preference by taxpayers and certain segments of tax practitioners for paper filing, lack of awareness of e-Filing and how to do it and concern about privacy, security and the role of third parties in the process.

Hypothesis 6, the final hypothesis tested in this study, hypothesized that perceived credibility of the system has significant influence on taxpayers' intention to use e-Filing. Surprisingly, the result inferred that the p-value is greater than 0.05 (β=0.033, C.R=0.426, p=0.670). Thus, the proposed relationship is not significant. Malaysian taxpayers perceived that it takes a lot of time to test software compatibility on network computer systems, develop training programs, and convince users that they should use a new system when their old system works just fine and they have yet to see any benefit in changing. Overall, they found that using the electronic filing is yet to be compatible with their lifestyle.

5. Conclusion

5.1 Concluding Remarks

This study is assured to have strong reliable determinant to assess taxpayers' intention in using e-Filing as a medium to file their tax return. This is based on the result of Cronbach Alpha that performed by reliability analysis for six independent variables and one dependent variable. The empirical results of our study can provide support for Davis (1989), DeLone and McLean (1992), Wang (2002), and Chang (2005) models.

Furthermore, this study proves strong and positive relationship between TAM determinants with taxpayers' behaviour intention. The intention of taxpayers seems to be influenced by attitude (0.823), perceived usefulness (0.724), perceived ease of use (0.691), information system quality (0.751), information quality (0.639) and perceived credibility (0.545). From the result, attitude seems to play strong role to influence taxpayers' intention to file tax return in future. This study suggests that taxpayers' attitude can be changed based on their first experience of handling e-Filing system. Furthermore, taxpayers' experience is influenced by their perceived usefulness, perceived ease of use, information system quality, information quality and perceived credibility to file tax return.

5.2 Implication and Future Research

Some of these barriers can be addressed through education and marketing of the advantages of e-Filing, such as faster refunds, electronic receipts that offer proof of filing, convenience, accuracy and reduced likelihood of receiving a notice from the system provider. The result of this study is able to assist Inland Revenue Board (IRB) to improve their e-Filing system and implement e-Filing seminar or on-hand course for taxpayers. This in turn, will also increase taxpayers' compliance toward their responsibility as a Malaysian resident. However, Inland Revenue Board (IRB) need to implement useful activities and programmes to educate taxpayers to use e-Filing. In that case, taxpayers will believe this medium is more useful compared to fill in the form and submit to IRB office manually before the due date. They also might realize e-Filing will process their work faster and more easier as compared their experience before. Once taxpayers' attitude change, the five determinants of TAM will also influence taxpayers' intention. In addition, the intention of taxpayers to file tax return in future might also influence the level of tax compliance.

This study is an exploratory study of TAM particularly in e-Filing system. However, it is quite difficult to be generalised due to small sample sizes, which are only 100 public servants at Labuan. In future, the sample of the study should consider both public and private sectors' staffs in Malaysia. This is because the result can be generalized to Malaysian taxpayers as a whole.

References

Byrne, B. M. (2001). *Structural equation modeling with AMOS*. Basic concepts, applications, and programming. Mahwah, NJ: Erlbaum.

Chang, I.C., Li., Y.C., Hung., W.F., & Hwang., H.G. (2005). An empirical study on the impact of quality antecedents on tax payers' acceptance of Internet tax-filing systems. *Government Information Quarterly 22*, pp. 389-410.

Carmines, E.G. and McIver, J.P. (1981), "*Analysing models with unobserved variables", in Bohrnstedt, G.W. and Borgatta, E.F.* (Eds), Social Measurement: Current Issues, Sage, Beverly Hills, CA.

Davis, F.D.(1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), pp. 319-339.

DeLone, W.H.,& McLean, E.R.(1992). Information systems success: The quest for the dependent variable. Information Systems Research, 3(1), pp. 60-95.

Fu., J.R., Farn., C.K., & Chao., W.P. (2006). Acceptance of electronic tax filing: A study of taxpayer intentions. *Information &Management* 43, pp. 109-126.

Hanudin., A., Zulkifli., M.M., Rizal., M.A.H., & Suddin., L. Explaining intention to use SMS Banking among Bank Islam Malaysia Berhad (BIMB) Consumers: Is Gender a Good Indicator? *Proceeding of International Business Borneo Conference 2006*, pp. 92-101.

Hung., S.Y., Chang., C.M., &Yu., T.J. (2006). Determinants of user acceptance of the e-Government services: The case of online tax filing and payment system. *Government Information Quarterly* 23, pp. 97-122.

Lai., M.L., Siti Normala., S.O., & Kameel., A.M. (2004) Towards an electronic filing system: A Malaysian survey. *eJournal of Tax Research*, pp. 1-15.

Li, T. and Calantone, R.J. (1998), "The impact of market knowledge competence on new product advantage: conceptualization and empirical examination", *Journal of Marketing*, Vol. 62, October, pp. 13-29.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). *Power analysis and determination of sample size for covariance structure modeling.* Psychological Methods, 1, 130-149.

Martens, M. P. (2005). *The use of structural equation modeling in counseling psychology research.* Counseling Psychologist, 33, 269-298.

Quintana, S. M., & Maxwell, S. E. (1999). *Implications of recent developments in structural equation modeling for counseling psychology*. Counseling Psychologist, 27, 485-527.

Ullman, J. B. (2001). *Structural equation modeling*. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using mulnVariate statistics* (4th ed., pp. 653-771). Boston, MA: Allyn & Bacon.

Wang, Y.S.(2002). The adoption of electronic tax filing systems: An empirical study. *Government Information Quarterly*, 20 (4), pp. 333-352.

Zuroff, D.C., Blatt, S.J., Sanislow, C.A. III, Bondi, C.M. and Pilkonis, P.A. (1999), "Vulnerability to depression: reexamining state dependence and relative stability", *Journal of Abnormal Psychology*, Vol. 108 No. 1, pp. 76-89.

(26 th June 2007). 657,000 guna e-Filing. *Berita Harian.*

(10th August 2007). New IRB chief on mission to improve tax compliance. *New Strait Times.*

(18th April 2007). New tax assessment system to be in place next year. *New Strait Times*.

(27th April 2006). Good response to E-Filing. *The Star Online*.

(4th May 2006). Taxpayers need to switch to e-Filing next year, say IRB. *The Star*.

(3rd May 2006). Manual Income Tax Return Are Acceptable Next Year. *Bernama*.

Table 1. Demographic Data

	Items	Frequency	Percentage
Gender	Female	50	50%
	Male	50	50%
Education	Certificate	14	14%
	Diploma	5	5%
	Degree	67	67%
	STPM	6	6%
	Lower than STPM	8	8%
Job	Professional (Academic)	63	63%
	Professional (Non-academic)	13	13%
	Support Staff	24	24%
Time of using computer per week	< 14 hours	34	34%
	3 ours	31	31%
	>28 hours	35	35%
E-Filing handling experience	Yes	33	33%
	No	67	67%
E-Filing learning experience	Yes	36	36%
	No	64	64%

Table 2. Reliability Analysis

Variable	Chang study	Current study
Behavior Intention	0.94	0.96
Attitude	0.94	0.96
Perceived Usefulness	0.98	0.96
Perceived Ease of Use	0.97	0.94
Information System Quality	0.92	0.95
Information Quality	0.97	0.97
Perceived Credibility	0.70	0.80

Table 3. Correlation between Constructs

	isq	pu	pc	iq	peou	ATT	INT
isq	1.000						
pu	.841	1.000					
pc	.667	.657	1.000				
iq	.856	.720	.648	1.000			
peou	.877	.799	.585	.751	1.000		
ATT	.807	.807	.580	.691	.735	1.000	
INT	.751	.724	.545	.639	.691	.823	1.000

(Note: isq = information system quality; pu = perceived usefulness; pc = perceived credibility; iq = information quality; peou = perceived ease of use; ATT = attitude; INT = intention)

Table 4. Standardised Regression Weights Of The Structural Model

Structural Path			Standardised Path Coefficient	S.E.	C.R.	P	Hypothesis Testing
pu	<---	isq	0.841	0.059	15.457	0.000	Supported
peou	<---	isq	0.877	0.050	18.191	0.000	Supported
ATT	<---	isq	0.437	0.109	4.356	0.000	Supported
iq	<---	isq	0.856	0.051	16.496	0.000	Supported
pc	<---	isq	0.667	0.077	8.912	0.000	Supported
ATT	<---	pu	0.439	0.101	4.377	0.000	Supported
INT	<---	ATT	0.605	0.104	5.955	0.000	Supported
INT	<---	pu	0.026	0.121	0.223	0.823	Not Supported
INT	<---	peou	0.059	0.127	0.502	0.616	Not Supported
INT	<---	isq	0.189	0.186	1.139	0.255	Not Supported
INT	<---	iq	-0.026	0.124	-0.237	0.812	Not Supported
INT	<---	pc	0.033	0.084	0.426	0.670	Not Supported

(Note: isq = information system quality; pu = perceived usefulness; pc = perceived credibility; iq = information quality; peou = perceived ease of use; ATT = attitude; INT = intention)

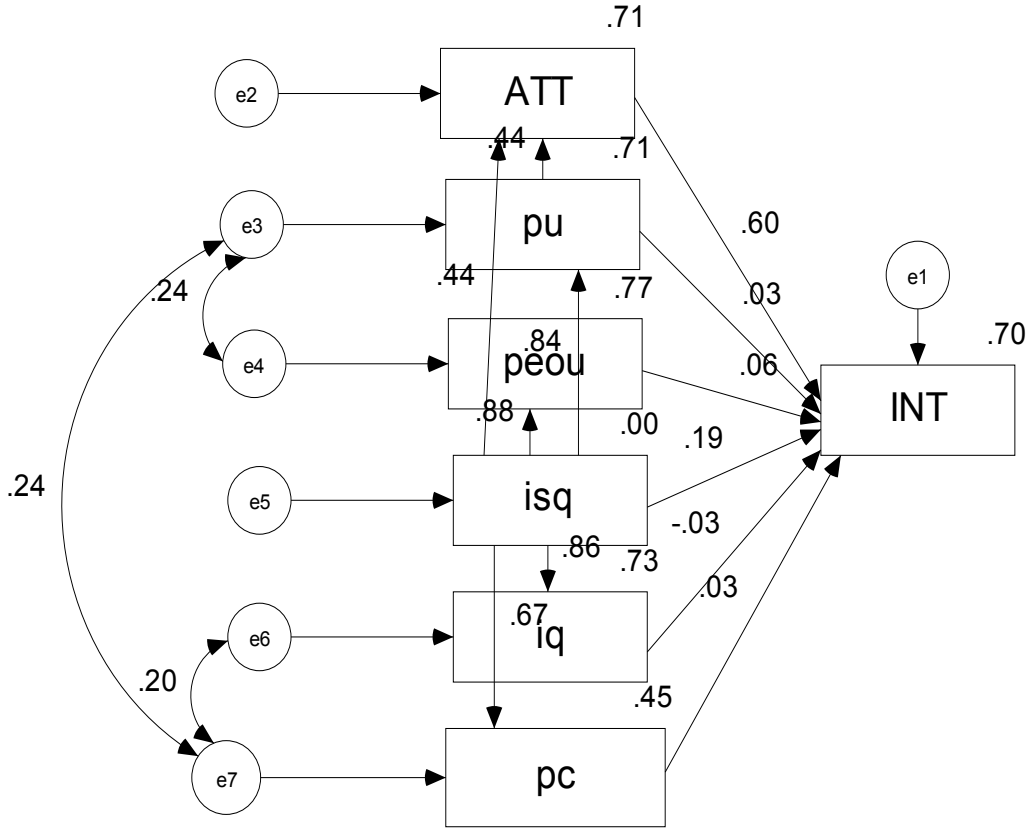


Figure 1. Structural Model

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

Fast Feature Value Searching for Face Detection

Yunyang Yan

Department of Computer Engineering

Huaiyin Institute of Technology

Huai'an 223001, China

E-mail: areyyyke@163.com

Zhibo Guo

School of Information Engineering

Yangzhou University

Yangzhou 225009, China

E-mail: zhibo_guo@163.com

Jingyu Yang

School of Computer Science and Technology

Nanjing University of Science and Technology

Nanjing 210094, China

E-mail: jingyuyang@mail.njust.edu.cn

This research is supported in part by the NSFC under grant 60632050, the High School Technology Fund of Jiangsu province under grant 06KJD520024, and Technology Fund of Huai'an under grant HAG05053 and HAG07063.

Abstract

It would cost much and much time in face detector training using AdaBoost algorithm. An improved face detection algorithm called Rank-AdaBoost based on feature-value-division and Dual-AdaBoost based on dual-threshold are proposed to accelerate the training and improve detection performance. Using the improved AdaBoost, the feature values with respect to each Haar-like feature are rearrange to a definite number of ranks.The number of ranks is much less than that of the training samples, so that the test time on each training samples is saved corresponding to the original AdaBoost algorithm. Inheriting cascaded frame is also proposed here. Experimental results on MIT-CBCL face & nonface training data set illustrate that the improved algorithm could make training process convergence quickly and the training time is only one of 50 like before. Experimental results on MIT+CMU face set also show that the detection speed and accuracy are both better than the original method.

Keywords: Rank- AdaBoost, Feature value division, Dual-AdaBoost, Face detection, Hnheriting cascade

1. Introduction

For its interesting applications, automatic face detection has received considerable attention among researchers in many fields, such as content-based image retrieval, video coding, video conference, crowd surveillance, and intelligent human– computer interface.

Many methods have been proposed to detect faces in a gray image or a color image, such as Template Matching, Mosaic Image, Geometrical Face Model, Difference Pictures, Snake, Deformable templates, Statistical Skin Color Models et al. Now the methods based on statistical learning algorithms have attracted more attention, including PCA, Artificial Neural Networks and Support Vector Machines, Bayesian Discriminant Features, etc. These methods show good performance in the detection precision, but their detection speed needs to be increased(Liang Luhong,2002,pp.449-458).

The first real-time face detector was proposed by Viola and Jones (Viola and Jones, 2001,2004,pp.137-154). They described a face detection framework that is capable of processing images extremely rapidly while achieving high detection rates. There are three key contributions of this detection framework. The first is the introduction of a new

image representation called an "Integral Image" which allows the features used by the detector to be computed very quickly. The second is a simple and efficient classifier which is built using the AdaBoost learning algorithm to select a small number of critical visual features from a very large set of potential features. The third contribution is a method for combining classifiers in a "cascade" which allows background regions of the image to be quickly discarded and focus on promising face-like regions. Simple Haar-like features are extracted and used as weak classifiers. The Viola and Jones frontal face detector runs at about 15 frames per second on 320 by 240 image. As in the work of Rowley (Rowley et al,1998,pp.22-38) and Schneiderman(Schneiderman,2000,pp.746-751), Viola and Jones(Viola and Jones, 2003) built a multi-view face detector with AdaBoost to handle profile views and rotated faces.

However, one problem with these approaches is that there are too many Haar-Like features in a single face image. Another difficulty is that a great deal of non-face training samples are used to reach good performance. The big set of training samples not only slow down the training, but also increase the number of weak classifiers greatly in cascaded detector. So AdaBoost on every round needs to search a large pool of candidate weak classifiers and the computation is very complex.

Attempting to get more efficient detector, improved AdaBoost algorithm called Rank-AdaBoost and Dual-AdaBoost based on feature-value-division are proposed to speed-up the training and detection performance. Firstly, for one Haar−Like feature, distribution of feature values of all face samples is divided into definite ranks. A small quantity of value is used as possible threshold in training instead of all feature values of face samples. Then, the approach of fast dual-threshold finding is developed in the enhanced AdaBoost algorithm, which makes the training process faster and the detection accuracy higher. In the training process of cascaded detector, the formers classifiers are transferred to the later, so that more non-face would be ignored. Both computational speed and the performance are improved obviously by this approach simultaneously.

Experimental results on MIT-CBCL face set and MIT+CMU face set show that our method yields higher classification performance than Viola and Jones' both on training speed and detection accuracy.

2. AdaBoost using Haar-Like feature

2.1 Haar-Like features and integral image

Haar-Like features are a kind of simple rectangle features proposed by Viola et al. as shown in Figure 1. The squares represent a face image. The value of each Haar-Like feature is computed as a difference of the sum over the white and black regions. It describes the local gray feature in the image. Using parity and threshold on this value, a class is predicted.

Viola et al. use three kinds of features. They differ by their division of two, three or four rectangular areas. Rotating these three types could easily generate other kinds of features. Every feature is characterized by its position in the face frame, pre-specified size and type. Given that the base resolution of the classifier is 24 by 24 pixels, the exhaustive set of rectangle filters is quite large, over 100,000, which is roughly $O(N^4)$ where $N=24$ (i.e. the number of possible locations times the number of possible sizes). The actual number is smaller since filters must fit within the classification window.

Computation of rectangle filters can be accelerated using an intermediate image representation called the integral image. Using this representation any rectangle filter, at any scale or location, can be evaluated in constant time. Integral image at location (x, y) is computed as the sum of the pixel values above and to the left of (x,y). A original gray image $\boldsymbol{I}$ and its integral image $\boldsymbol{II}$ is described as follows:

$$\boldsymbol{II}(x, y) = \sum_{i,j=1}^{x,y} \boldsymbol{I}(i, j)$$

The integral image can be computed in one pass over the original image by using the following pair of recurrences:

$$\boldsymbol{S}(x, y) = \boldsymbol{S}(x, y-1) + \boldsymbol{I}(x, y)$$

$$\boldsymbol{II}(x, y) = \boldsymbol{II}(x-1, y) + \boldsymbol{S}(x, y)$$

(where $\boldsymbol{S}(x, y)$ is the cumulative row sum, $\boldsymbol{S}(x,-1) = 0$, and $\boldsymbol{II}(-1, y) = 0$). Using the integral image, the value of a Haar-Like feature can be computed by plus or minus using the integral image. Any rectangular sum can be calculated in four array references. Clearly the difference between two rectangular sums can be calculated in eight references. Since the two-rectangle features defined above involve adjacent rectangular sums they can be computed in six array references, and eight and nine references in the cases of three and four-rectangle features respectively.

In Figure 2, (d) is an integral image corresponding to the left image (a) and the feature (b). (c) shows the feature on the image. Simple $p4+p1-p2-p3-(p6+p3-p4-p5)$ could give the value of the feature. Much computing time is saved.

2.2 AdaBoost algorithm

In its original form, the AdaBoost learning algorithm is used to boost the classification performance of a simple learning

algorithm (e.g., it might be used to boost the performance of a simple perceptron). It does this by combining a collection of weak classification functions to form a stronger classifier. In the language of boosting the simple learning algorithm is called a weak learner. So, for example, the perceptron learning algorithm searches over the set of possible perceptrons and returns the perceptron with the lowest classification error. The learner is called weak because we do not expect even the best classification function to classify the training data well (i.e. for a given problem the best perceptron may only classify the training data correctly 51% of the time). In order for the weak learner to be boosted, it is called upon to solve a sequence of learning problems. After the first round of learning, the examples are re-weighted in order to emphasize those who were incorrectly classified by the previous weak classifier. The final strong classifier takes the form of a perceptron, a weighted combination of weak classifiers followed by a threshold.

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$$

where $f(x)$ is the final strong classifier, ($h_1(x), h_2(x), \ldots\ldots, h_t(x)$) is the serials of weak classifiers. The $h_t(x)$ can be thought of as one feature with a threshold. The original form of the AdaBoost named as Init- AdaBoost is shown as:

1) Given example images: $(x_1, y_1), \ldots, (x_L, y_L)$, where $y_i \in \{1, 0\}$ indicates positive or negative examples; $g_j(x_i)$ is the jth

$$w_{1,i} = \begin{cases} 0.5/m & i \le m, \\ 0.5/n & otherwise \end{cases}$$

Haar-Like feature of ith example x_i.

2) Initialize weights

Where m, n are the number of positive or negative examples respectively. $L=m+n$.

3) For $t = 1 \ldots T$

a. Normalize the weights

$$w_{t,i} = w_{t,i} / \sum_{j=1}^{L} w_{t,j}$$

b. For each feature j, train a weak classifier h_j, and evaluate its error ε_j with respect to w_t, $\varepsilon_j = \sum_{i=1}^{L} w_{t.i} \left| h_j(x_i) - y_i \right|$,

$$h_j(x) = \begin{cases} 1 & p_j g_j(x) < p_j \theta_j \\ 0 & otherwise \end{cases}$$

Where $p_j \in \{1, -1\}$ is a parity bit and θ_j is a threshold.

c. Choose the classifier h_t with the lowest error ε_t

d. Update the weights $w_{t+1,i} = w_{t,i} \beta_t^{1-ei}$, where e_i=0 if example x_i is classified correctly, e_i=1 otherwise, and $\beta_t = \varepsilon_t/(1-\varepsilon_t)$.

4) Final classifier:

$$H(x) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_t h_t(x) \ge 0.5 \sum_{t=1}^{T} \alpha_t \\ 0 & otherwise \end{cases}$$

where $\alpha_t = \log(1/\beta_t)$.

2.3. Cascaded detector

In an image, most sub-images are non-face instances. In order to improve computational efficiency greatly and also reduce the false positive rate, a sequence of gradually more complex classifiers called a cascade is built.

An input window is evaluated on the first classifier of the cascade and if that classifier returns false then computation on that window ends and the detector returns false. If the classifier returns true, then the window is passed to the next classifier in the cascade. The next classifier evaluates the window in the same way. If the window passes through every classifier with all returning true, then the detector returns true for that window. The more a window looks like a face, the more classifiers are evaluated on it and the longer it takes to classify that window. Since most windows in an image do not look like faces, most are quickly discarded as non-faces. Figure 3 describes the cascade.

By using cascaded detector, it is possible to use smaller and efficient classifiers to reject many negative examples at early stage while detecting almost all the positive instances. Classifiers used at successive stages to examine difficult cases become more and more complex.

Since easily recognizable non-face images are classified in the early stages. Subsequent classifiers are trained only on examples that pass through all the previous classifiers. That is to say, classifiers of the later stages of the cascaded detector can be trained rapidly only on the harder, but smaller, part of the non-face training set. Therefore this detection approach would save the computational cost and maintain the performance simultaneously.

Stages in cascade are constructed by training classifiers using AdaBoost. Cascaded detector is trained as follows:

1) Input: Allowed false positive rate f, and detection rate d per layer; target overall false positive rate F_{target}. P denotes set of positive examples, N denotes set of negative examples, n_i is the number of weak classifiers in the ith layer classifier.

2) $F_0 = 1$, $D_0 = 1$, $i=0$

3) While $F_i > F_{target}$

a. i++, $n_i=0$, $F_i=F_{i-1}$

b. While $F_i > f \times F_{i-1}$

- n_i++.
- Use P and N to train a ith layer classifier with n_i weak features using AdaBoost.
- Evaluate current cascaded classifier on validation set to determine F_i and D_i.
- Decrease threshold for the ith classifier until the current cascaded classifier has a detection rate D_i of at least $d \times D_{i-1}$, evaluate F_i.

c. If $F_i > F_{target}$ then evaluate the current cascaded classifier on the set of non-face images, and put any false detection into the N.

3. Proposed Improved AdaBoost

3.1 Problem of time cost

In AdaBoost algorithm, the step 3)b time costs expensively, because all simple classifier $h_j(j=1:k)$ is desired to compute where k is the number of the Haar-Like features and the k is a very large number. Moreover each h_j is obtained by exhaustive searching every samples so that it would take much time to only get a weak classifier $h_t(x)$.

If the *Onetime* is the time needed to get only one simple classifier, and the training time to get one weak classifier is *Traintime*, then

$$Traintime = k^* Onetime.$$

The step is done repeatedly until the face detection rate is satisfied. If the number of weak classifiers is T, then it would cost *Alltime* to obtain the final strong classifier.

$$Alltime = Traintime * T = k^* Onetime * T$$

If there are 600 weak classifiers needed, the mean search time of each h_j is 0.1s, only 24,000 features is used to train, then the consumed time would be $600 \times 24{,}000 \times 0.1 = 1{,}440{,}000$s(i.e. 400 hours, about 16 days). It is too long. So it is necessary to save the time for computing h_j.

3.2 Division of the feature values

According to Init-AdaBoost, for one Haar-like feature, the feature value of each face sample is used as possible threshold. Under a given parity, the possible threshold is compared with that of all training samples. Consequently the false detection rate is calculated. Do it again under another parity. The threshold and the parity with this Haar-like feature could be determined by the face sample which causes the minimum false detection rate, and the feature value of this face sample is used as the threshold for the Haar-like feature.The cost time is $O(m \times n)$ where m ,n is the number of face samples and non-face samples. In face detection training, m ,n are all very large, generally thousands even million. So the time corresponding to $O(m \times n)$ is very expensive.

By the experiment results, we find that false detection occurred frequently when there are some noise on the image and we also find that there are little difference between the feature values of some face samples, the ability of these feature values to discriminate face sample and non-face sample is similar. So to reduce the number of possible threshold, we could get the maximum and the minimum feature values of all training face samples with one Haar-like feature, and obtain r ranks of the feature value from minimum to maximum feature value. That is:

$$\Delta j = (\max(g_j(x_i)) - \min(g_j(x_i)))/r, \quad i=1 \ldots m$$

Where $g_j(x_i)$ is the jth Haar-Like feature of ith example x_i.

Then using each Th_k as a possible threshold to find the weak classifier and its threshold with the parity. Th_k is computed by $Th_k = \min(g_j(x_i)) + k \times \Delta j \quad (k=1 \ldots\ldots r)$.

Now the cost time would be $O(r\times(m+n))$ instead of $O(m\times n)$. After do many experiments,we find that the detection performance could be the same as Init-AdaBoost when r is less then 100. If let r=100, then the value of $r\times(m+n)$ would be smaller and smaller than $m\times n$ because m, n are usually thousands even millions. So the time used to train would be falled rapidly and could be denoted as $O(m+n)$.

Furthermore, let W_k is the sum of weight of every samples under threshold Th_k, G_k is the sum of weight of the samples whose feature value is more than Th_k, but less than Th_{k+1}, then the sum of weight of every samples under threshold Th_{k+1} is W_{k+1}.

$$W_{k+1}= W_k +G_k.$$

That is to say only those samples are tested whose feature value is more than Th_k, and it is not necessary to test all samples to calculate W_{k+1} when W_k has been known. So the time for training could be also saved.

The improved AdaBoost based on the division of the feature values is called Rank- AdaBoost.

3.3 Finding of dual-threshold

According to experimental results, the local feature distributes regularity corresponding to a weak classifier. The typical local features of face and non-face are shown in Figure 4.

In Figure 4, vertical axis y says the ratio of face or non-face in total samples, the horizontal axis x is values of feature. Threshold used to indicate face or non-face could be obtained rapidly. As Figure 4 shows, face is higher than non-face from $\theta^{(1)}$ to $\theta^{(2)}$, so $\theta^{(1)}$and $\theta^{(2)}$ are used as dual-threshold. The specific method to find threshold is described as follows.

1) For each x, compute *face(x)- nonface(x)*.

2) Choose x' with the maximum: x'=argmax (*face (x)- nonface (x)*).

3) From x' to left or right to find the crossing point which cause *face(x)-nonface(x)<0*. If no crossing point is found, then the boundary point is selected as crossing point. So we could get two crossing points and use them as dual-threshold. AdaBoost based dual-threshold is called Dual- AdaBoost.

When determining a feature as a current weak classifier or not, the dual-threshold of the features may be adjusted to ensure weak classifier h_t to meet the demand of detection. So this method can greatly accelerate each weak classifier search. If there were 24,000 features, the search time would be 1/50 of that of exhaustive search.

4. Proposed Inheriting cascaded detector

In cascaded detector, fewer and fewer sub-window images need to be detected by layer classifier at later stages, a little error would make overall detection performance decline. So, a sequence of gradually more complex and more powerful classifiers are trained to increase classification performance with examples that have pass through all the previous classifiers. During cascaded detector training, the predecessor is used as a part of its successor, that is to say each layer is considered not only as an independent node of the cascaded classifier but also as a component of its successor. So the later classifier includes more classification features.

$$f_k(x) = f_{k-1}(x) + \sum_{i=1}^{T} \alpha_t h_t(x)$$

where $f_k(x)$ and $f_{k-1}(x)$ is a strong classification function on the kth and k-1th layer respectively. It is called inheriting cascaded detector. Based on this algorithm, the overall performance of the cascaded detector is enhanced. Moreover, the threshold of classifier on different layers is adjusted to separate the training samples of face and non-face as far as possible, so that more non-face would be ignored. Both computational speed and the performance are improved obviously by this detection approach simultaneously.

As an example, we need to train strong classifiers on the kth layer. $f_k(x)$ and $H_k(x)$ is a strong classification function and strong classifier respectively. A total of L samples consist of m positive examples (face) and n negative examples (non-face). Positive examples arrange in a sequence before n negative examples.

1) Given example images: (x_1, y_1), ..., (x_L, y_L), where $y_i \in \{1, -1\}$ represents positive or negative examples; $g_j(x_i)$ is the jth Walsh feature of ith example x_i. $L=m+n$.

2) Using the last weights resulted from the strong classifiers training on (k-1)th layer as $w_{1,i}$. If k=1 then initialize

$$w_{1,i} = \begin{cases} 0.5/m & i \le m, \\ 0.5/n & otherwise \end{cases}$$

weights

3) Search $\theta_j^{(1)}$ and $\theta_j^{(2)}$ of each local feature based on its distribution in all face and non-face examples. Use $\theta_j^{(1)}$ and θ_j

$^{(2)}$ as dual-threshold .

4) For $t = 1, ..., T$

a. Normalize the weights

$$w_{t,i} = w_{t,i} / \sum_{j=1}^{L} w_{t,j}$$

b. For each feature j, get a weak classifier h_j with $\theta_j^{(1)}$ and $\theta_j^{(2)}$ by the method discussed above, and evaluate its error ε_j, $\varepsilon_j = \sum_{i=1}^{L} w_{t,i} h_j(x_i) \operatorname{sgn}(y_i)$. Choose a weak classifier h_t with the lowest error ε_t from all these weak classifiers, then calculate coefficient

$\alpha_t^{(0)} = (\ln((1-\varepsilon_t)/\varepsilon_t))/2$.

c. Get a strong classifier function

$$f_k(x) = f_{k-1}(x) + \sum_{i=1}^{t-1} \alpha_i h_i(x) + \alpha_t^{(0)} h_t(x)$$

The corresponding strong classifier is:

$$H_k(x) = \begin{cases} 1 & f_k(x) \geq 0, \\ -1 & otherwise \end{cases}$$

d. Using the thresholds to test on positive samples and to make $f_k(x)$ achieve the default requirements $D_k \geq d \times D_{k-1}$.

e. Use $f_k(x)$ to test on negative samples, if $F_i \leq f \times F_{i-1}$ then exit the iteration.

f. Update the weights $w_{t+1,i} = w_{t,i} e^{-\alpha_t h_t}$, where $\alpha_t = (\ln((1-\varepsilon_t)/\varepsilon_t))/2$.

5) Get a strong classifier function

$$f_k(x) = f_{k-1}(x) + \sum_{i=1}^{T} \alpha_t h_t(x)$$

the corresponding strong classifier is:

$$H_k(x) = \begin{cases} 1 & f_k(x) \geq \beta_k, \\ -1 & otherwise \end{cases}$$

where $\beta_k = \min_{i=1,m}(f_k(x_i))$.

5. Experimental results

5.1 Experiments on MIT-CBCL data set

The publicly available MIT-CBCL face database is used to evaluate the performance of the proposed face detection system. The original MIT-CBCL training set contains 2,429 face images and 4,548 non-face images in 19 ×19 pixels grayscale PGM format. The training faces are only roughly aligned, i.e., they were cropped manually around each face just above the eyebrows and about half-way between the mouth and the chin. The data set willserve our purpose of comparing our detection system with their original system, which we shall train using the same training set. Some samples are shown as Figure 5.

The training data set we used is the subset of MIT-CBCL and consist of 1,429 face images and 3,548 non-face images. The test set consists of 1,000 face images and 1,000 non-face images left. The compuer we used is with P4/2.4GHz CPU, 1 GB memory. The results are shown as Table 1.

According to Table 1, the Detection rate and False positive under Rank-AdaBoost or Dual-AdaBoost are similar with Init-AdaBoost, but the training time is obviously different. The training time used in Dual-AdaBoost is only 1/50 of that of Init-AdaBoost. The detection time also fall about half because of less classifiers used. Moreover,the robustness and the ability of generalization become better and better.

5.2 Experiments on a Real-World Test Set

5.2.1 Training data sets selection

Face samples must be selected carefully with variability in face orientation (up-right, rotated), pose (frontal, profile), facial expression, occlusion, and lighting conditions. Moreover some unimportant features of the face should be removed, such as hair and or so.

Non-face samples could be selected randomly. Any sub-window of an image containing no face can be used as a non-face sample. Almost arbitrary large training set can be easily constructed using these non-face samples.

In our experiment, the training data were collected from various sources. Face images were taken from the MIT-CBCL face training dataset, FERET face dataset, NJUST603 and web. The dataset contains face images of variable quality, different facial expressions and taken under wide range of lightning conditions. The dataset contained 8145 face samples including rotated versions of some faces.

Non-face images were collected from the web and MIT-CBCL non-face dataset. Images of diverse scenes were included. The dataset contained images such as animals, plants, countryside, man-made objects, etc. Some non-face samples were selected by randomly picking sub-windows from hundreds of images that did not contain face. More than 2,180 thousand non-face samples were used. Hundreds of these non-face images is very similar to face. Each image of all samples was cropped into size of 19×19. Figure 6 shows some face samples. Figure 7 shows some non-face samples.

In the experiment, 9954 total Haar-like features were selected and used as weak classifiers. The smallest rectangle filter was defined as 4×4, the biggest was 16×16 on the window of 19×19 image.

5.2.2 Experimental results

Given that non-face samples discarded on each layer was more than 40%, at the same time the face samples detection rate was more than 99.99% . We get a cascaded detector using the proposed improved AdaBoost and inheriting cascade model. The face cascaded detector using Rank-AdaBoost has 35 layers of classifiers. The face cascaded detector using Dual-AdaBoost has 21 layers of classifiers.

We tested our system on the MIT+CMU frontal face test set(Rowley et al,1998, pp.22–38). MIT+CMU dataset consists of 130 images with 507 total frontal faces. Every image was resized by 0.85 each iteration during test. The the performance of our detection system are shown in Table 2.

Experimental results on MIT+CMU face set shows that our method provides higher classification performance than Viola and Jones' method both on training speed and detection accuracy.

6. Conclusions and Future Work

In this paper, we presented a speed-up technique to train a face detecotor using AdaBoost by improveing the threshold searching method and new inheriting cascaded frame. Our proposed face detection system incorporating the technique reduces the number of subwindows that need preprocessing and verification. The proposed system is much faster than the original AdaBoost-based detection systems in training speed and is also higher in testing accuracy. It is suitable for realtime applications. Further, the system performs well for frontal faces in gray scale images with variation in scale and position.

A larger training set would be essential for the detector to be of practical use. In particular, the number of non-face images would have to be drastically increased in order to decrease false positives. Moreover, as mentioned earlier, using a larger number of Haar-like features would also improve the accuracy. Implementing and inproving the cascade is required in order to achieve the ultimate aim of our work, i.e., to improve the accuracy of the detector while maintaining real-time detection speed.

References

H.Rowley, S.Balujaand T.Kanade.(1998). Neural Network-based Face Detection. *IEEE Pattern Analysis and Machine Intelligence*, 20, 22–38.

H.Rowley, S.Baluja and T.Kanade. (1998). Rotation Invariant Neural Network-based Face Detection. In *Proc.* of IEEE Conference on *Computer Vision and Pattern Recognition* [C], Australia, 38-44.

Liang Luhong, Ai Hai-zou and Xu Guang-you. (2002). A Survey of Human Face Detection. *Chinese Journal of Computer*, 25, 449~458.

P. Viola and M. Jones. (2001). Rapid object detection using a boosted cascade of simple features [A]. In *Proc. of* IEEE Conference on *Computer Vision and Pattern Recognition* [C], USA, 511~518.

P. Viola and M.Jones. (2004). Robust real-time face detection. *International journal of Computer Vision*, 57, 137~154

P. Viola and M. Jones. (2003). Fast Multi-view Face Detection. Shown as a demo at the IEEE Conference on *Computer*

Vision and Pattern Recognition [C], USA.

Schneiderman, H. and Kanade, T. (2000). A Statistical Method for 3D Object Detection Applied to Faces and Cars. In *Proc.* of IEEE Conference on *Computer Vision and Pattern Recognition*, USA, 746-751.

Table 1. Comparison o f the training and detection

	Init-AdaBoost	Rank-AdaBoost	Dual-AdaBoost
Time for getting a Simple classifier (s)	0.1192	0.0071	0.0023
Sum of weak classifier	96	98	56
Detection rate (%)	97%	96.4%	98.6%
False positive	1	2	0
Detection time (s)	1.766	1.781	0.953

Table 2. Results on the MIT+CMU test set

	Viola's Detector			Rank-AdaBoost			Dual-AdaBoost		
False Positive	50	78	167	48	82	169	51	69	138
Detection Rate(%)	91.4	92.1	93.9	91.2	92.3	94.0	91.8	92.6	94.2

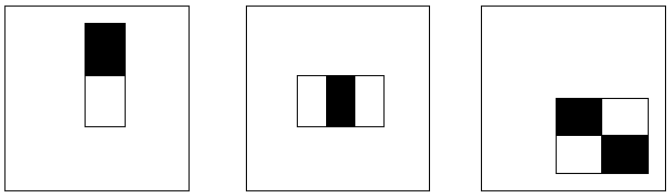


Figure 1. Examples of the Viola and Jones features

(a) (b) (c) (d)

Figure 2. Features extraction using integral image

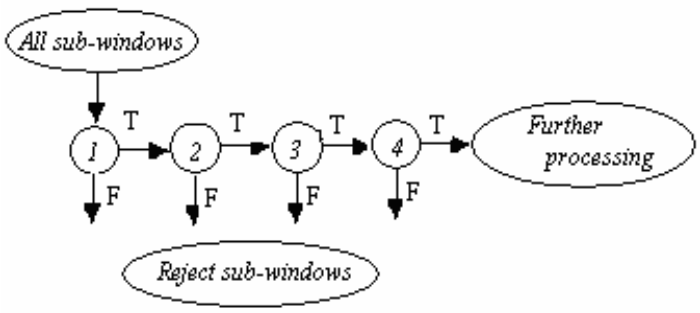


Figure 3. Schematic depiction of a cascaded detector

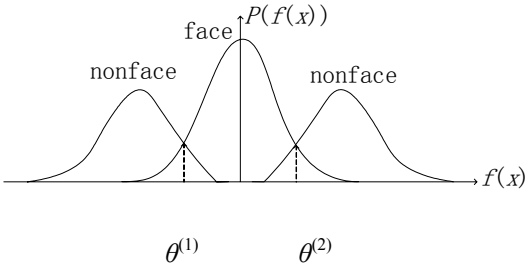


Figure 4. Distribution of typical local features

(a) Some face examples (b) Some nonface examples

Figure 5. Some training examples

Figure 6. Some faces from the training set

Figure 7. Some non-faces from the training set

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

Geospatial Information Technology for Conservation of Coastal Forest and Mangroves Environment in Malaysia

Kamaruzaman Jusoff

Forest Geospatial Information & Survey Laboratory

Department of Forest Production, Lebuh Silikon, Faculty of Forestry

Universiti Putra Malaysia

43400 UPM, Serdang, Selangor, Malaysia

Tel: 60-3-8946-7176 E-mail: kamaruz@putra.upm.edu.my

The r*esearch is financed by Universiti Putra Malaysia (Sponsoring information)*

Abstract

Mangrove forests are one of the most productive and bio-diverse wetlands environments on earth. Yet, these unique coastal tropical forests environment are among the most threatened habitats in the world. Growing in the intertidal areas and estuary mouths between land and sea, mangroves provide critical habitat for a diverse marine and terrestrial flora and fauna. The important need of living being is opportunity to continue their life in sustainable environment and suitable conditions. Potential stand is the place that obtains the possibility of germination and establishment of a plant species according to their physical, chemical, biological demands. In many cases are seen that because of unsuitable selection of site and species, afforestation and reforestation projects after spending time, cost and labor are forced to failure. Therefore, it is an obligation by the relevant authorities, especially Forestry Department to ensure that the rate of seedlings survival in the afforestation and reforestation activities is successfully monitored, mapped and quantified. One of the most efficient techniques available is the use of Geospatial Information Technology consisting of Geographical Information Systems (GIS), Global Positioning System (GPS) and remote sensing (RS). Using this technology and integrating the different thematically maps that shows environmental conditions of specific region, suitable and potential positioning of different species for plantation and rehabilitation programs could be well determined and monitored. For mapping and detection of individual mangrove species for reforestation and afforestation purposes, mathematical functions such as Boolean logic, fuzzy logic, and neural network can be easily applied. It is expected that suitable species-site matching for reforestation and afforestation of mangroves could be implemented with such geospatial tools.

Keywords: GIS, RS, Fuzzy, Boolean, Mangrove, Reforestation, Afforestation

1. Introduction

Mangrove forest is one of the most valuable coastal resources, important for its multiple economic, ecological, and scientific and culture resources for present and future generations. It is an important component of Malaysia's coastal zone ecosystem. In addition, mangrove forests are utilized as a source of fuel, wood and pole production. In recent decades, the coastal zone has been subjected to the effects of a growing population and economic pressure. In order to use coastal resources on a sustainable basis, a proper management planning process is necessary. Mangrove forest assessment and monitoring has been conducted continuously in Malaysia.

The integration of remote sensing and GIS for mangrove forest management is considered as an important tool for the development of effective plans by natural resource managers and planners. Successful application of GIS tools and concepts to the coastal zone is one of the great challenges still facing developers and users of the technology (Hussin et. al., 2007) GIS and remote sensing methods have been used as successful planning and management tools to reforestation of mangroves that have been destroyed in some parts of the world (Ratnasermpong , 2004). After is earned the different data layer according to demand of species, these layers were divided according of suitable or unsuitable for species growth. At the end were determined the region that is suitable for growth and establishment of species considered to relative factors with integrate of data layer, and so were earned the map of suitable site for forestation with foresaid species. In land classification, the final map is one of the main outputs but the information about utilization types and management specifications is also important. After the land suitability evaluation has been done, the decision about the use to be selected depends on physical and economical factors; in this way, perhaps the most suitable use will not be chosen due to its viability.

The main objective of this study is to demonstrate an effective approach for sustainable site selection with mangrove reforestation by using remote sensing and GIS technology. Preparing data layer in GIS is flexible and depends on used layers. In some layers such as topography map, basin boundary, ways network, slow changes are expected using them, but in some layers the change is fast such as cloud cover, snow cover, soil moisture condition, therefore we can use satellite images which are up to date and result in higher accuracy and exactitude in projects. Another advantage is that the GIS users have found that they can get the most of the input data with RS. Also digital satellite data are directly transmittable to GIS. Use of both GIS and RS techniques not only progress the geographical information but also give the ability to users to get economical data. Use of these two technologies are an emergency instruments in natural resources and agriculture that give clear and controlled points versus unclear points and out of control to programmers and managers. But how does the fuzzy logic works in a GIS? The processing is the same as classification of satellite image, but its module and boundary are normally determined by the fuzzy classification user (Juan Francisco, 2007).

2. Methods and Materials

2.1 Determining of Potential Site Selection using Boolean Method

The traditional concept of modelling employs a Boolean approach: the value is true or false. This approach tends to represent reality in a discrete way. But what can be found in the nature is that few elements are discrete, they are rather continuous. In the real world, some objects are quite differentiated from others and their boundaries are quire evident: a river crossing through a valley is quite distinguishable from its surroundings when in full discharge, an area covered by a lake is distinct from the land areas surrounding it; but soil and vegetation and other patterns in nature change transitionally: the limit between two types of soil or vegetation is, in most of the cases, not so clearly defined. Fuzzy modelling appears as an alternative to deal with these continuous or uncertain environments. While in Boolean logic a value is true of false, with fuzzy logic the value could be partially false of partially true which allows for a representation more according to the reality.In this method only were used numbers 0, 1. The pixels which are suitable for growing of species according ecological condition are assigned code1 and the pixels that are not suitable for are assigned code 0. Then all made layers that are consist of 0, 1 will algebra multiply together to district out put points in process of multiply layer. That is enough one of the corresponding pixels would has code 0, so related point that have assigned code 0 in output part will omit from stand. But if all corresponding points that have code 1, are defined with code 1 and be belonging to talented growing for specific species. In Boolean logic, the finality is considered not both sides.

2.2 Determine Potential Site Selection in Fuzzy Method

When done limitations for a topic be great on the earth, whereas we can not get suitable boundary by Boolean logic, should forgo some ideal condition and adjust, in this condition we can use Fuzzy logic, whereas also are classified between two numbers, i.e 0 and 1 and will determine Fuzzy degree membership. In this study, some ecological factors were defined, border domain that observed in before tables separately. In this method also, layers have been identified and classified based on the ecological demands of species and determined domain in Fuzzy logic. Then did action of integration with operator ADN and selection of minimum numeric quantity in each pixels, in integration layers, some layers are naturally Boolean and can not classified them with Fuzzy method, therefore some layers with Boolean classification in integration with Fuzzy layers were used.

2.3 Fuzzy Modelling and its Applications for Coastal Environment Suitability

Fuzzy logic was initially developed by Lotfi Zadeh (Iranian scientist) in 1965 as a generalization of classic logic. He defined a fuzzy set as “a class of objects with a continuum of grades of memberships”; being the membership a function that assigns to each object a grade ranging between zero and one, the higher the grade of membership the closest the class value to one. Traditionally thematic maps are represented with discrete attributes based on Boolean memberships, such as polygons, lines and points. These types of entities have a value or do not have it; an intermediate option is not possible. With fuzzy theory, the spatial entities are associated with membership grades that indicate to which extent the entities belong to a class. Figure 1 presents a representation of traditional Boolean sets and fuzzy sets: while with Boolean logic the boundary between sets is clearly defined (A and B), with fuzzy logic there is a transition zone where each set has less membership grade in relation to the other. In fuzzy theory, the map for A shows membership values closer to 1 when the set falls within A category, while the values are close to 0 when they are far from the category; the same applies for category B.

<Figure 1. Representation of crisp and fuzzy sets>

Fuzzy logic is also a generalization of Boolean logic that instead of using the binary TRUE and FALSE values applies “soft” variables such as deep, moderately deep, steep, etc. These variables are defined in an interval ranging between 0 and 1 allowing a continuous range of values.

Qualitative parameters obtained through interviews to stakeholders and quantitative parameters obtained through field measurements and recorded by FAO in the ECOCROP database (http://ecocrop.fao.org/) were considered to evaluate

the suitability. The quantitative parameters include soil fertility, drainage, texture, depth and pH. The purpose of their study was to asses the performance of Boolean classification methods such as FAO framework for land evaluation versus a fuzzy classification methodology. In their study it was found that the assignment of suitability orders with the Boolean theory (that is, S1 for suitable, S2 for less suitable, N for non-suitable and so on) restricts the results for available land for a potential use: large areas of the study areas were classified with the same rating while for the fuzzy classification a higher variation of suitability were found. These results are a consequence of the matching between suitability-class requirements and land characteristics, where the land is a member of the suitability class or is not and no intermediate values are possible. In this study it is concluded that fuzzy processing allows obtaining information about the degree of land suitability class, which is relevant for land use planers to know how highly or moderately suitable is the land for a crop (Rodrigo and Emmanuel, 2005).

Fuzzy Sets are sets (or classes) without sharp boundaries; that is, the transition between membership and non membership of a location in the set is gradual. A Fuzzy Set is characterized by a fuzzy membership grade (also called a possibility) that ranges from 0.0 to 1.0, indicating a continuous increase from non membership to complete membership. Four fuzzy set membership functions are provided in IDRISI for Windows: Sigmoidal, J-Shaped, Linear and User-defined.

Sigmoidal: The Sigmoidal ("s-shaped") Membership function is perhaps the most commonly used function in Fuzzy Set theory. It is produced using a cosine function. In use, FUZZY requires the positions (along the X axis) of 4 points governing the shape of the curve. These are indicated in the figure below as points a, b, c and d and represent the inflection points of the curve as follows:

a = membership rises above 0

b = membership becomes 1

c = membership falls below 1

d = membership becomes 0

In the fuzzy methodology the same parameters have been considered for fertility, but without taking into account pH, which may have strong fluctuations within the same soil unit. The calculation of the fuzzy memberships for the different factors influencing fertility was evaluated using a linear function as given in Figure 2.

<Figure 2. Linear or asymmetrical triangular membership function. Where x is the input data and, *a* and *c* are the limit values according to related Tables>

For depth an asymmetrical second grade function has been employed:

This function was tested successfully for soil depth. In the equation, *a* is a parameter that controls the shape of the function and the position of the cross-over points; the expression (x-c)2 controls the dispersion

<Figure 3. Membership function for asymmetrical second grade function (adapted from Burrough, 1989)>

A combination of symmetrical Gaussian functions was employed to assess the membership functions for depth. In this way the overlapping nature of soil depth can be assessed.

<Figure 4. Gaussian membership functions for fuzzy subsets of soil depth.>

For slope, an S membership function was employed. The limits a and g corresponded to the limit conditions of steep slopes and flat terrain respectively. This function gives better results when compared to other membership functions, and for this reason has been used four fuzzy sets with the linguistic labels{P, L, M, H}, which stand for *poor*, *low*, *moderate*, and *high*, respectively Fuzzy membership classes.

<Figure 5. S membership function >

3. Results and Discussion

Boolean logic is appropriate to potential site selection and can justice with high convenience. Fuzzy logic also is appropriate to potential site selection because district the points between 0 and 1. Closer number to 1 is more successful. This research could be used as a model for site selection in agriculture, natural resources and biological environment.

As a main advantage of the Boolean theory, it is possible to control and trace easily which factors are affecting the suitability of a plot, while with the fuzzy model it is necessary to review the interaction between membership functions and weights, which is not a straightforward process. Fuzzy theory allows intermediate possibilities of suitability beyond the traditional classes given by the Boolean methods, but on the other hand it can over estimate the potential of a land (Yanar and Akyure, 2007). Oppositely, the Boolean theory can underestimate the real potential of a plot. In this sense, perhaps the land evaluator has to try with both theories and check with information on the field which one agrees better with the reality. Traditional methodologies which rely on Boolean logic require high accuracy and data detail that is

difficult –if not impossible- to find in reality. Fuzzy logic can cope with low detail levels and allows for more flexibility in the suitability classification.

For selecting the region that is suitable for forestation of Mangrove species, Boolean model consider environment and parameters 0 and 1 (black and white) while Fuzzy logic has ability to consider gray, we should pay attention that the most of the environmental parameters are accordance with fuzzy logic. For species forestation the costs also is significant, in Boolean the risks and failure of project is less than the region that the Fuzzy has determined, because in fuzzy has determined the more extensive region that exist the possibility of the forestation than the Boolean one. The figure 6 is example which has compared the suitable region for forestation with Boolean logic and Fuzzy logic methods. The major results are the compilation of relevant thematic databases, assessment of forest land use and forest distribution in 1973, 1987, 1993 and 1998, as well as change in land use and land cover between 1987-1993 and 1993-1998 and development of a proposed forest land use plan. Remote sensing appears to be a significant tool for assessment and monitoring of coastal zone resources, especially mangrove forest.

<Figure 6. Sample of Boolean and fuzzy logic comparison; Boolean method have done in left one and Fuzzy logic in right one>

In addition, planning and management of forest land use is easily and effectively conducted using GIS. However, the integration of remote sensing and GIS for the development of mangrove forest management plans by natural resource managers and planners is necessary. The result of factor maps overlay is multiplied by the result of limitation maps overlay could result to selection of suitable locations for mangroves. The mapping of land use/land cover is essential for natural resource survey. In this context we can say that GIS is a special purpose digital database in which a common spatial coordinate system is the primary means of reference. Comprehensive GIS requires a means of (a) data input, from maps, aerial photograph, satellites, surveys and other sources, (b) data storage, retrieval and query, (b) data transformation, analysis and modelling, including spatial statistics, and (d) data reporting such as maps, reports and plans. GIS is a very powerful integration tool between the various data source.

It is found that the spatial distribution of the problems and their effects are very typical. This shows the role of spatial analysis and the importance of the GIS functionality. A GIS can provide better information to support this type of complex decision-making. With the rapid advancements taking place in computer hardware and GIS software, more complex models are being developed. These models help researchers and planners to simplify complex systems and to develop theory to understand the process at work better. Present analytical functions and conventional cartographic modelling techniques in GIS are based on Boolean logic, which implicitly assumes that objects in a spatial database and their attributes, can be uniquely defined. In the land evaluation process, with Boolean classification, all land units with values that exceed the given threshold may be defined as the class or set of acceptable land units which are to be rejected. These uncertain boundary definitions create some problems with loss of information. The deficiencies of the traditional Boolean logic for the design of spatial databases have been recognized in recent years. As an alternative to Boolean logic, Zadeh's fuzzy set theory has been proposed as a new logical foundation for GIS design.

The mangroves are considered to be highly biologically productive estuarine system and serve as nursery, feeding and breeding ground for many kinds live being. They also help to restrict cyclone damage as a biological shield against the cyclones and prevent soil erosion as well. Innumerable species of fish use them as nesting and feeding grounds. As the land runoff carrying coastal pollutants end up in the seas, the mangroves serve as sink or trap for the coastal pollutants. The mangroves are food source for many Phytoplankton's feed on the allochthonous detritus produced by the mangroves Understanding the status of mangroves including degradation areas and its causes is important decision making and also for creating public awareness to conserve these important coastal resources.

4. Conclusion and Recommendation

Map of land use planning is more accurate and totally conformable with applicative criteria future planning. Different factors in the field such as climate, soil, physiographic are highly interactive with each others and should be incorporated in future studies.

References

Hussin, Yousif, A., Mahfud M. Zuhair, and Michael W. (2007). Monitoring Mangrove Forests using Remote Sensing and GIS, GISdevelopment.net

Juan Francisco, S.M. (2007), Applicability of knowledge-based and fuzzy theory-oriented approaches to land suitability for upland rice and rubber, as compared to the farmers' perception. A case study of Lao PDR, Master of Science theses in Geo-information Science and Earth Observation, ITC.

Ratnasermpong, S.(2004). Coastal Zone Environment Management With Emphasis On Mangrove Ecosystem, A Case Study Of Ao-Sawi Thung Khla, Chumphon, Thailand, GISdevelopment.net. 2004.

Rodrigo, S. S, and Emmanuel J. C. (2005). Fuzzy modeling of farmers' knowledge for land suitability classification, *Agricultural Systems* 83: 49–75.

Yanar, T.A., and Akyüre, Z. (2007). Artificial Neural Networks as a Tool for Site Selection within GIS, *Geodetic and Geographic Information Technologies, Natural and Applied Sciences*, 06531 Ankara, Turkey.

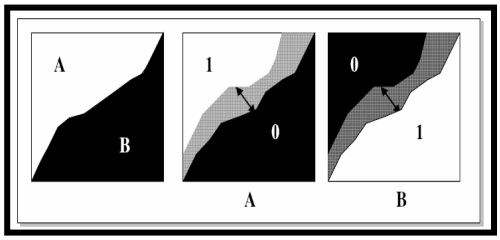


Figure 1. Representation of crisp and fuzzy sets

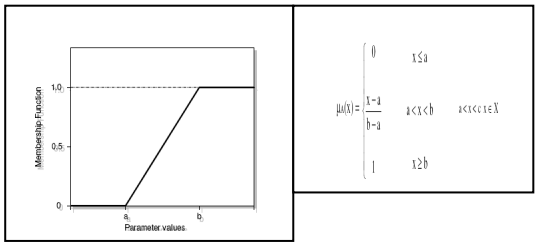


Figure 2. Linear or asymmetrical triangular membership function.

Where x is the input data and, *a* and *c* are the limit values according to related Tables

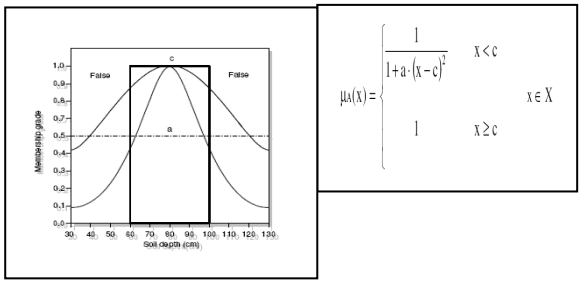


Figure 3. Membership function for asymmetrical second grade function.

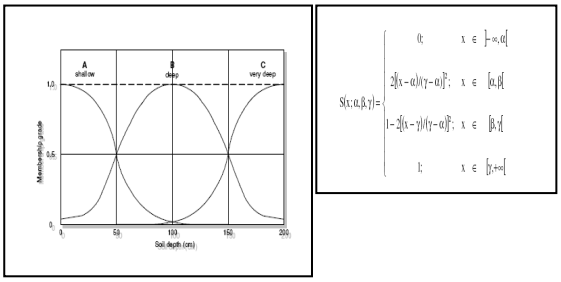


Figure 4. Gaussian membership functions for fuzzy subsets of soil depth.

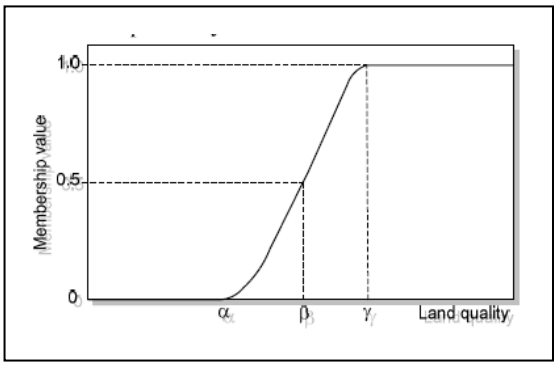


Figure 5. S membership function

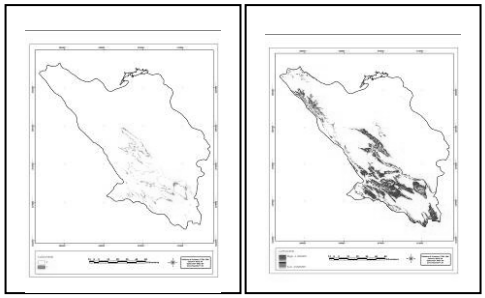


Figure 6. Sample of Boolean and fuzzy logic comparison;
Boolean method have done in left one and Fuzzy logic in right one

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 2

May 2008

The Application of AHP in Electric Resource Evaluation

Chunlan Qiu & Yonglin Xiao

Library of Jiangxi Science and Technology University

Ganzhou 341000, China

E-mail:jxust@163.com

Abstract

This article utilizes the analytic hierarchy process (AHP) method to study and establish the hierarchy model and its evaluation system for the electric resource evaluation.

Keywords: AHP, Electric resource, Evaluation system

1. Introduction

In recent years, with the continual increase of electric resource, the purchasing of electric resource has becoming the important part of the construction of library literature resource. The outlay used in electric resource in many libraries has exceeded 25%, so whether for purchasing new electric resource or continuing order and maintaining present electric resource, a new problem occurs, i.e. how to select and evaluate the electric resource, and how to enhance benefits and make the construction of electric resource more reasonable and scientific through the evaluation of electric resource?

However, present scholars' researches only focused on the evaluation of internet information resource or single database, and there are few researches to study the integrated evaluation to the collection of electric resource, and the evaluation and system are not perfect, the present evaluation methods include qualitative method and quantitative method. In this article, we will adopt the AHP method which combines qualitative analysis with quantitative analysis, and try to establish the electric resource evaluation system that can be applied in the pre-evaluation before purchasing and the after-evaluation in the purchasing.

The AHP method (Cai, 2005, p.58-63) was put forward by US famous operational research expert T. L Saaty in 1970s, which tried to simulate three human basic characters (i.e. decomposing, estimation and integration) to deal with complex problems through analytic hierarchy, quantitative analysis and standardization, and added statistical test in the whole process. It adapts to solve those decision problems that have complex structure and many decision rules and are difficult to be quantified.

The basic approach of AHP method include following steps. (1) Establish the concept of the complex problem and find out main factors involved in the study objective. (2) Analyze the association and subjection relationships among factors and establish orderly ladder hierarchy model. (3) Compare both relative essentialities of various factors on the same layer to the certain rule on the upper layer, and establish the evaluation matrix. (4) Compute the relative weight of the compared factor to the rule on the upper layer according to the evaluation matrix and implement the coherence test. (5) Compute the integrated weight of various layers to the total objective of the system and implement total compositor of the layers.

2. Establishment of electric resource evaluation system

2.1 Establishing hierarchy model

Base on many evaluation indexes of electric resource, we build a ladder hierarchy model (Xiang, 2004, p.26-29 & Wang, 2005, p.67-70 & Xiao, 2002, p.35-42) (seen in Figure 1, the 3rd index are not be concretely explained, which are seen in Table 4).

2.2 Constructing comparison evaluation matrix

We adopt the 1-9 standard degree method (seen in Table 1) to evaluate the relative essentialities of the indexes, and evaluate the proportion degree of the relative essentiality through both comparison among them.

For example, for electric resource, we think the content of database is comparatively more important than the searches system and function and endows it 3 points, and it is more important than the uses and endows it 4 points, and it is little more important than the cost accounting and endows it 2 points, and it is more important than the manufacturer than manufacturer service and endows it 4 points, and in this way, we can establish A-B evaluation matrix (seen in Table 2).

Analogously, we can establish evaluation matrixes B_1-C, B_2-D, B_3-E, B_4-F, B_5-G, C_1-P, C_2-P, C_3-P, C_4-P, C_5-P, C_6-P, C_7-P,

D_1-P, D_2-P, D_3-P, D_4-P, E_1-P, E_2-P, E_3-P, E_4-P, F_1-P, F_2-P, F_3-P, F_4-P, F_5-P, G_1-P, G_2-P, G_3-P, G_4-P, G_5-P.

2.3 Computing the weights W_I of various indexes

The information base of AHP is the evaluation matrix, and it utilizes the compositor principle to obtain the matrix compositor vector and compute the weight coefficients of various indexes. The computation approach (Li, 2004, p.75-78) includes following steps.

(1) Compute the product M_i of factors on every raw of the evaluation matrix B: $M_i=\prod_{j=1}^{n} b_j$, j=1, 2… n.

(2) Compute the root of n of M_i on every raw: $w_i=\sqrt[n]{M_i}$, i=1, 2… n, and n is the order number of the matrix in the equation.

(3) Implement normalized processing to $(w_1, w_2 \ldots wn)^T$, and make $W_I=\frac{w_i}{\sum_{j=1}^{n} w_j}$, so $W_I=[W1, W2, \ldots Wn]^T$ are the eigenvectors, i.e. the weighted coefficients of various indexes.

The concrete evaluations of various layer index weights are seen in Table 4.

2.4 Implementing coherence testing to various evaluation matrix

Because of the complexity of things and human difference of objective evaluation, every evaluation can not achieve completely identical, and to ensure the rationality of the conclusion of AHP method, we need to implement coherence test to various evaluation matrix, so we introduce the negative square values of other latent roots except for the maximum latent root of the evaluation matrix in AHP method and take them as the deviation coherence index of the matrix departures, i.e. use $CI=\frac{\lambda_{\max}-n}{n-1}$ to test the coherence of the evaluation thinking. The maximum latent root $\lambda_{\max}=\sum_{i=1}^{n}\frac{(AW)_i}{nW_i}$.

To test whether different evaluation matrixes have satisfactory coherence, we must introduce the average random coherence index RI value of the evaluation matrix, and RI values of 1-9 order evaluation matrixes can be seen in Table 3.

To the 1st and 2nd evaluation matrixes, they always have satisfactory coherence, but the order number exceeds 2, the ratio of the coherence index CI with the some order random coherence index RI is called the random coherence ratio CR, and when $CR=\frac{CI}{RI}<0.10$, we think this evaluation matrix has satisfactory coherence, i.e. the thinking on various layers is coherent, the conclusion obtained by the AHP is coherent, or else, the evaluation matrix should be adjusted to make it possess satisfactory coherence.

The maximum latent roots of the evaluation matrixes are obtained by the computer program, and according to them we can obtain the coherence index CI and random coherence ratio CR, and then we implement the coherence test, and the results are complete coherence or satisfactory coherence.

2.5 Integrated weight

Though above approach, we can only obtain the weighted vectors of a group of factor to the certain factor on its upper layer. To obtain the relative weights of various factors to the total objective, especially to obtain the compositor weights of various indexes on the lowest layer to the objective, i.e. “integrated weights”, we need superincumbent computation and integrate weights under the single rule, and obtain the relative weight of every evaluation objective in the layer objective to the total objective and implement total evaluation coherence test layer by layer. Relative to the total objective, the integrated weights of various indexes can be denoted as $W=a_i a_{ij} a_{ijk}$, where a_i, a_{ij} and a_{ijk} respectively are 1st, 2nd and 3rd class index weight. Then we implement total compositor to the relative weights.

So we can obtain a clear evaluation index system of electric resource (seen in Table 4).

Though confirming the evaluation grading (seen Table 5), we can evaluate various indexes of the electric resource evaluation system, and the method is to use the weighted adding method, multiply the evaluation value of every evaluation index with the corresponding weight of this index, obtain the weighted evaluation value of the index, add these weighted evaluation values and obtain the total evaluation valves of the evaluation objective. The formula is $S=\sum W_i P_i$ (here, W_i is the integrated weight of the i’th index, P_i is the evaluation value of evaluated object on the i’th index, and i is the sequence number of the concrete index on the lowest layer in the evaluation model).

If the quantity of electric resources participated in the evaluation has Ql, Q2, Q3…Qn, we should adopt the AHP method. We respective establish evaluation matrixes aiming at 63 evaluation indexes, and obtain the compositor vectors

of 63 evaluation matrixes. Multiply every compositor vector with weighted coefficient of corresponding index and add them, we can obtain the total compositor vector of Ql, Q2, Q3…Qn, and its result is also the compositor of the electric resources Ql, Q2, Q3…Qn.

3. Conclusions

The character of AHP is to combine qualitative analysis with quantitative analysis, which has high validity, reliability, conciseness and extensive applicability. But AHP still has limits, and its result only aims at the evaluation index in the rule layer, so the confirmation of evaluation index largely influences the system evaluation, in addition, human subjective evaluation has certain influence to the evaluation results of the system, so this method usually is combined with Delphi method to confirm the values of various indexes.

The evaluation of electric resources by AHP compensates human limit of subjective blur ability, which quantifies decision-makers' experiences and judgments, compares relative factors layer by layer, tests the rationality of comparison result layer by layer, avoids the subjective random of simple evaluation to make the result more exact and make the evaluation decision possess more objectivities and persuasions.

References

Cai, Haipeng & Yang, Kunyu. (2005). Determining the Evaluation Criterion Weight of Digital Campus Based on AHP. *Journal of Changsha Aeronautical Vocational and Technical College.* No. 5(2). p. 58-63.

Li, Chaokui & Tao, Weiguo. (2004). The Application of Analysis Hierarchy Process on the Evaluation of the Navigation System for Network Information Resources. *Journal of the Library Science of Sichuan.* No. 3. p. 75-78.

Wang, Hongfei. (2005). The Purchasing and Evaluation on Electronic Analysis Resources. *Researches in Library Science.* No. 4. p. 67-70.

Xiang, Yingming, Tan, Yiman & Linhuan. (2004). Research on Evaluation Method and Mathematical Model for Electronic Resources. *Library Journal.* No. 1. p. 26-29.

Xiaolong & Zhang, Yuhong. (2002). On the Development of the Electronic Resources Evaluation System. *Journal of Academic Libraries.* No. 3. p. 35-42.

Table 1. 1-9 standard degree method

Comparison between index A and index B	Extremely important	Very important	Important	Little important	equal	Little unequal	unimportant	Very unimportant	Extremely unimportant
Evaluation value of index A	9	7	5	3	1	1/3	1/5	1/7	1/9
Remark	Taking 8, 6, 2, 1/2, 1/4, 1/6, 1/8 as the middle values of above evaluations								

Table 2. Evaluation matrix

A	B_1	B_2	B_3	B_4	B_5
B_1	1	3	4	2	4
B_2	1/3	1	3	2	3
B_3	1/4	1/3	1	1/3	1
B_4	1/2	1/2	3	1	3
B_5	1/4	1/3	1	1/3	1

Table 3. RI values of evaluation matrix

Order number	1	2	3	4	5	6	7	8	9
RI	0.00	0.00	0.58	0.90	1.12	1.24	1.32	1.41	1.45

Table 4. Evaluation index system table of electric resource

A electric resource	1st index (a_i)	2nd index (a_{ij})	3rd index (a_{ijk})	Integrated weight W
	Contents of database B_1 (0.41)	Embodied content C_1 (0.34)	Knowledge range P_1 (0.33)	0.0460
			Magazine sorts P_2 (0.17)	0.0236
			Article periodical proportion P_3 (0.17)	0.0236
			Nuclear periodical proportion P_4 (0.33)	0.0460
		Fixed number of year of data C_2 (0.09)	Database time limit P_5 (1.00)	0.0369
		Data type C_3 (0.23)	Article database P_6 (0.50)	0.0472
			Tabloid database P_7 (0.25)	0.0236
			Reality database P_8 (0.25)	0.0236
		Data repetition ratio C_4 (0.05)	≥30% P_9 (0.10)	0.0020
			10%~30% P_{10} (0.25)	0.0051
			≤10% P_{11} (0.65)	0.0133
		Data update/lag C_5 (0.15)	Daily update P_{12} (0.65)	0.0400
			Weekly update P_{13} (0.25)	0.0154
			Monthly update P_{14} (0.10)	0.0062
		Data source C_6 (0.09)	Data credibility P_{15} (0.50)	0.0184
			Information provider's reputation P_{16} (0.50)	0.0184
		Information resource organization C_7 (0.05)	Classifying according to topic and subject P_{17} (0.40)	0.0082
			Rationality of classification P_{18} (0.40)	0.0082
			Frame structure P_{19} (0.20)	0.0041
	Searches system and functions B_2 (0.24)	Searches function D_1 (0.35)	Browse searches P_{20} (0.10)	0.0084
			Simple searches P_{21} (0.16)	0.0134
			Second time searches P_{22} (0.28)	0.0235
			Checkable field P_{23} (0.46)	0.0386
		Searches technology D_2 (0.35)	Boolean searches P_{24} (0.20)	0.0168
			Truncate searches P_{25} (0.20)	0.0168
			Quotation searches P_{26} (0.20)	0.0168
			Clustering searches P_{27} (0.20)	0.0168
			Position logic P_{28} (0.10)	0.0084
			Weight searches P_{29} (0.10)	0.0084
		Searches results D_3 (0.19)	Completely exact P_{30} (0.43)	0.0196
			Output format P_{31} (0.27)	0.0123
			Hyperlink P_{32} (0.10)	0.0046
			Compositor mode P_{33} (0.10)	0.0046
			Marker P_{34} (0.10)	0.0046
		User service D_4 (0.11)	Help document P_{35} (0.37)	0.0100
			User training P_{36} (0.24)	0.0063
			Word and table tool P_{37} (0.15)	0.0040
			Adjustment of searches interface P_{38} (0.08)	0.0021
			Searches history record P_{39} (0.08)	0.0021
			Literature transfer service P_{40} (0.08)	0.0021
	Uses B_3 (0.08)	Entry time E_1 (0.12)	Database opened time P_{41} (1.00)	0.0096

		Searches time E_2 (0.23)	Searches mode used time P_{42} (1.00)	0.0184
		Download tabloid/entire article E_3 (0.23)	Article number downloaded to the client computer P_{43} (1.00)	0.0184
		User evaluation E_4 (0.42)	Easily using character P_{44} (0.33)	0.0111
			Practicability P_{45} (0.67)	0.0225
	Cost accounting B_4 (0.19)	Data base price F_1 (0.48)	Unit price of database P_{46} (1.00)	0.0912
		Searches cost F_2 (0.12)	Cost of database searches P_4 (1.00)	0.0228
		Entry cost F_3 (0.12)	Cost of database entry P_{48} (1.00)	0.0228
		Investment of hardware and software F_4 (0.21)	Cost to purchasing equipment and software P_{49} (1.00)	0.0399
		System maintenance F_5 (0.07)	Professional maintenance P_{50} (1.00)	0.0133
	Manufacturer services B_5 (0.08)	Data transfer mode G_1 (0.31)	International network P_{51} (0.16)	0.0040
			Special line mode P_{52} (0.30)	0.0074
			Local mode P_{53} (0.54)	0.0134
		Probation time G_2 (0.10)	Below 6 months P_{54} (0.54)	0.0043
			3~6 months P_{55} (0.30)	0.0024
			Below 3 months P_{56} (0.16)	0.0013
		Visiting mode G_3 (0.18)	Password entry P_{57} (0.40)	0.0058
			IP limit P_{58} (0.40)	0.0058
			Concurrent user P_{59} (0.20)	0.0029
		Technology support G_4 (0.31)	User service P_{60} (0.67)	0.0166
			Visiting purview P_{61} (0.33)	0.0081
		Relative document G_5 (0.10)	MARC record provision P_{62} (0.67)	0.0054
			ISSN, web address information provision P_{63} (0.33)	0.0026

Notice: The numbers in the bracket are the weights relative to superior indexes.

Table 5. Evaluation grading

Grading	good	comparatively good	general	comparatively bad	bad
Value interval	10-8	8-6	6-4	4-2	2-0

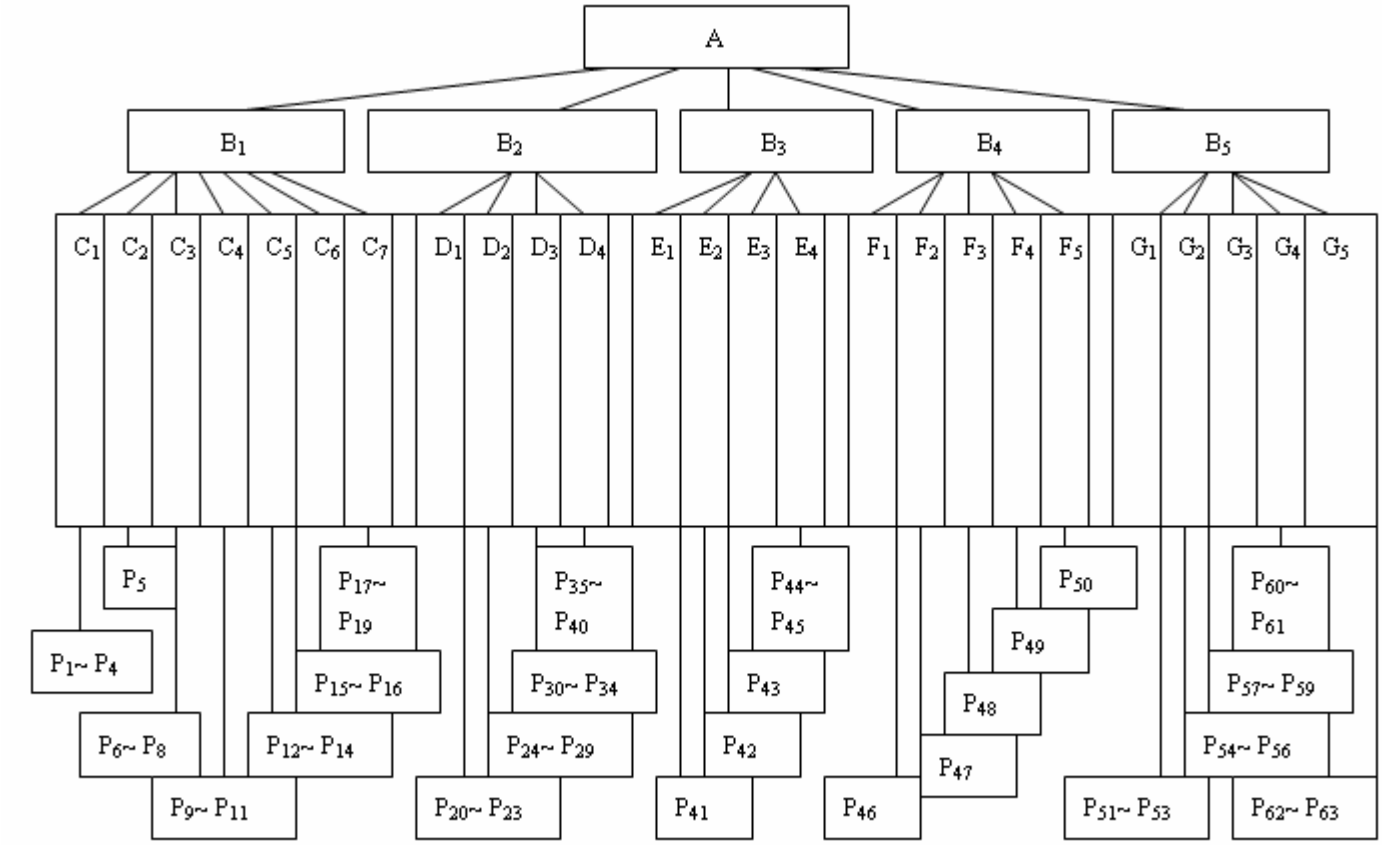


Figure 1. Ladder Hierarchy Model

Notice: A- the evaluation of electric resource, B_1- the contents of database, B_2- searches system and function, B_3- uses, B_4- cost accounting, B_5- manufacturer service, C_1- embodied content, C_2- fixed number of year of data, C_3- data type, C_4- data repetition ratio, C_5- data update/lag, C_6- data source, C_7- information resource organization, D_1- searches function, D_2- searches technology, D_3- searches results, D_4- user service, E_1- entry time, E_2- searches time, E_3- download tabloid/entire article, E_4- user evaluation, F_1- data base price, F_2- searches cost, F_3- entry cost, F_4- investment of hardware and software, F_5- system maintenance, G_1- data transfer mode, G_2- probation time, G_3- visiting mode, G_4- technology support, G_5- relative document.

A journal archived in Library and Archives Canada
A journal indexed in CANADIANA (The National Bibliography)
A journal indexed in AMICUS
A leading journal in computer and information science

Computer and Information Science

Quarterly

Publisher Canadian Center of Science and Education
Address 4915 Bathurst St. Unit # 209-309, Toronto, ON. M2R 1X9
Telephone 1-416-208-4027
Fax 1-416-208-4028
E-mail CIS@ccsenet.org
Website www.ccsenet.org
Printer William Printing Inc.
Price CAD.$ 20.00

ISSN 1913-8989