Computer and Information Science; Vol. 13, No. 3; 2020
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

Improvements of Automatic Extraction of FA Words Tendency using Non_linear Approach

Talal H. Noor[1], El-Sayed Atlam[1,3] , Ghada Elmarhomy[1], Ahmed Abd Elwahab[4], Rawda Draz[3] & Mahmoud Elmarhoumy[1,2]

[1] College computer science and engineering, Taibah University, Saudi Arabia

[2] Higher Institute Of Computers, Information and Management Technology Of Tanta, Egypt

[3] Faculty of Science, Tanta University, Tanta 31527, Egypt

[4] Faculty of Engineering, system and computer, Al-Azhar University, Egypt

Correspondence: El-Sayed Atlam, College computer science and engineering, Taibah University, Saudi Arabia.

Received: May 15, 2020 Accepted: June 15, 2020 Online Published: June 23, 2020

doi:10.5539/cis.v13n3p66 URL: https://doi.org/10.5539/cis.v13n3p66

Abtract

Field association (FA) terms are used to identify the subject of text (document field) by extracting specific words in that text. In this paper we use FA terms to study the effect of time change on specific terms by calculating the frequency of this terms, which associated with the archive field in a specific period. This paper also introduces a new approach for automatic evaluation of the stabilization classes using non-linear approach. The stabilization classes refer to the changing of FA terms with time in a specific period. The new approach improves the performance of decision tree than linear approach by using non-linear approach. The corpus that used in this approach has number of 1,356 files, and it is about 7.49 MB, after comparing the presented approach with the traditional one, we conclusion that the new approach enhanced the F-measure for increment, steady, decrement classes by 7.7%, 3.1%, 2.2%, sequentially.

Keywords: FA terms, decision tree, time series, non-linear equation, Information Retrieval (IR)

1. Introduction

To evaluate informational technology (IT), there are a huge amount of data that belongs to different of fields. This information can be used in classifying, retrieving information, clustering, and so on. Each field has some words that can be distinguished it, also this words occur repeatedly in that field. Information retrieval used to extract this words but it is a hard process. Information retrieval (IR) is the task, given a set of documents and a user query, of finding the relevant documents to an information need from a set of information resource and retrieving it(Samuel et al.2001). In each period of time there are some words that repeated in each field which can be defined by calculating its frequency. Examples of these words "world cup" is more spread in the period that the competition of the world cup has been played; and "heatstroke" is more spread in summer.

Old Approaches (Atlam et al. 2017(a,b); Atlam et al. 2006; Atlam et al. 2003; Atlam et al. 2002) neglect the time factor when extracting the words. Atlam et al. 2001 considered the spread of words with time change dependent on utilizing watchwords with documents. These keywords are not the most representative of texts and there are no relation between words and fields.

The contribution of this paper is as follows:

- The effect of FA terms frequency within a specific period using non-linear approach.
- The suggested new approach that enhanced the evaluation of the changing classes using the decision tree (DT) algorithm C4.5 (Quinlan 1993; Lima 1996) on FA terms using non-linear approach between parameters.

This paper focus on the nature of change of words with time change, and applying the closest mathematical relationship, that leads to the increasing of DT precision and solve the problem of data scattering. Section 2 in this study discusses related work. Section 3introduces our new upgrade algorithm and methodology. Section 4 presents the experimental evaluation. Section 5 focus in conclusion and future work.

2. Related Work

There are many studies in IR. All of this studies presented different methods, which were useful in classification, clustering, and analyzing documents however they ignore the relation between frequency of words and time changing in given period.

A method for extracting familiar subjects automatically with important keywords in web texts is presented by Morita et al (2012. Morita approach judges word changing with time and their fields in input Texts, and grouped them into two groups that related to the same subjects.

Rokaya et al(2008) and Atlam et al (2018) presented another model of positioning a particular example words called field affiliation terms (FA Terms). The positioning of archives gives an exact orchestrating of results than old strategies. This investigation presents a composition of list items that utilizes relations between FA terms in the question, where the higher co-recurrence of two words implies the closer connection between them.

Samuel et al (2001), presented a strategy for partitioning text into field-reasonable sections, and afterward separating FA terms or expressions from the content by decide how themes develop. In any case, Samuel's methodology couldn't effectively decide the significance of FA terms in a specific period.

Hashim et al (2019), introduces new method to extract, Arabic keywords from corpora based on their recurrences changes in a document over given periods of time using a decision tree. The new approach is applied on new data set field (computer science) which makes it different to traditionally used methods.

In this paper, we focus on the relation between frequencies of FA words with time variation. Moreover, we study the effect of word tendency related to a field using a new model called non-linear.

3. Field Association Terms

This section will introduce more details about the field association terms and their levels using the field tree**.**

3.1 FA Terms

It is normal that anybody can perceive the field of the record when they notice some particular words. These particular words are called field affiliation terms (FA Terms), which characterize as the littlest words that can decide a book field in a diagram named field tree. The tree structure speaks to information in a various leveled structure, and furthermore positions connections between record fields through the field tree (Dozawa 1999; Fukumoto et al. 1996; Ding et al. 2001;Azzopardi 2012). In this paper, a report field speaks to a well known information, which can be effectively utilized in human correspondence. For instance, <MIDICINE/Diseases/Cancer> express the way in tree with super field <MIDICINE> that has subfield < Diseases> and terminal field <Cancer>.

3.2 FA Terms Levels

Since FA terms has different scope to associate with a field. That mean some FA terms can be identify only one field, whereas others may identify 2 or more fields. There are 5 distinct levels defined to put FA term in its correct fields. These levels are defined as follow:

i. Perfect FA terms: words are associated to one subfield (e.g, cancer, flu, etc).

ii. Imperfect FA terms: words are associated to one or more subfield in one super field (e.g, arthritis, asthma, etc).

iii. Super FA terms: words are associated with one super field (e.g, patient, nurse, and hospital).

iv. Various FA terms: words are associated with more than one subfield of more than one super field (e.g, treat, winner, etc).

v. Non FA terms: words don not specify neither subfields nor super fields, and also include stop words (e.g, size, pronouns, etc).

These levels are resolved in the calculation of Atlam et al (2002), the calculation additionally consequently FA terms by utilizing standardization recurrence for each term in the wake of ascertaining it. In this paper, we will consider the calculation which decide the productive FA terms and the impact of time arrangement in word inclination utilizing new methodology (non-straight).

4. The New Methodology

4.1 System Outcome

The outline of the new system is shown in Fig. 1. Assume that $\mathrm{Freq}(\mathrm{FAT}_k^{pi})$ is the frequency of FA term k in a specific period p_i and total_ freq(FATs. p_i) is the sum of frequencies of all FA terms in p_i. To adjust the effect that is happened because of the difference of number of FA terms in each p_i and to determine the change with time correctly. The normalized frequency of FA term $Norm\_Freq_{ki}(FAT_k^{pi})$[4].

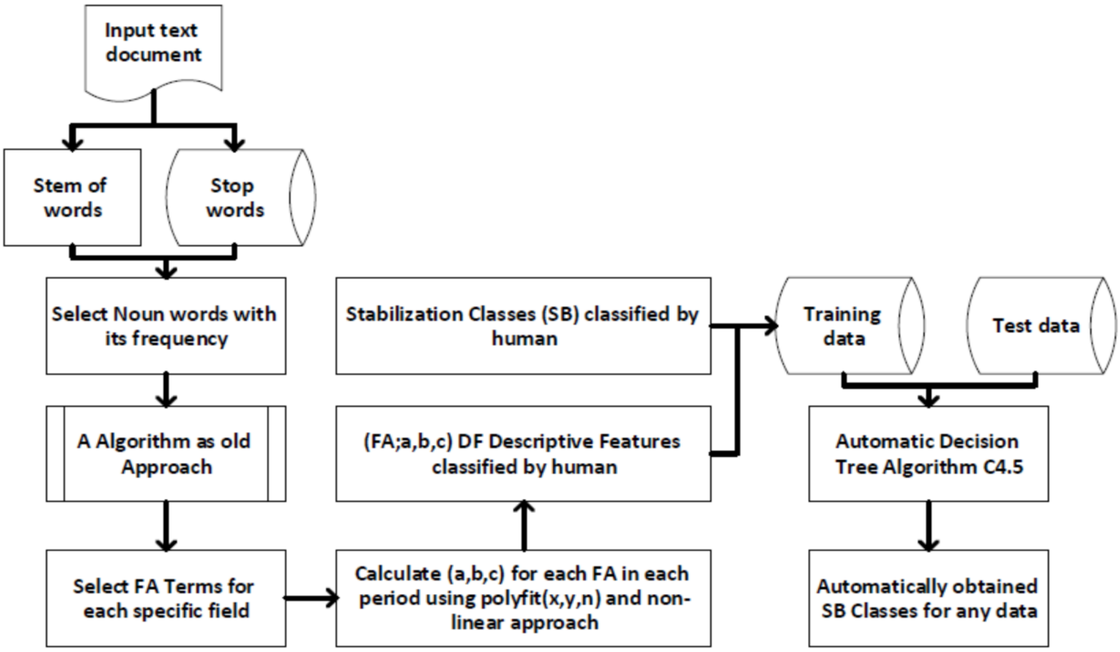
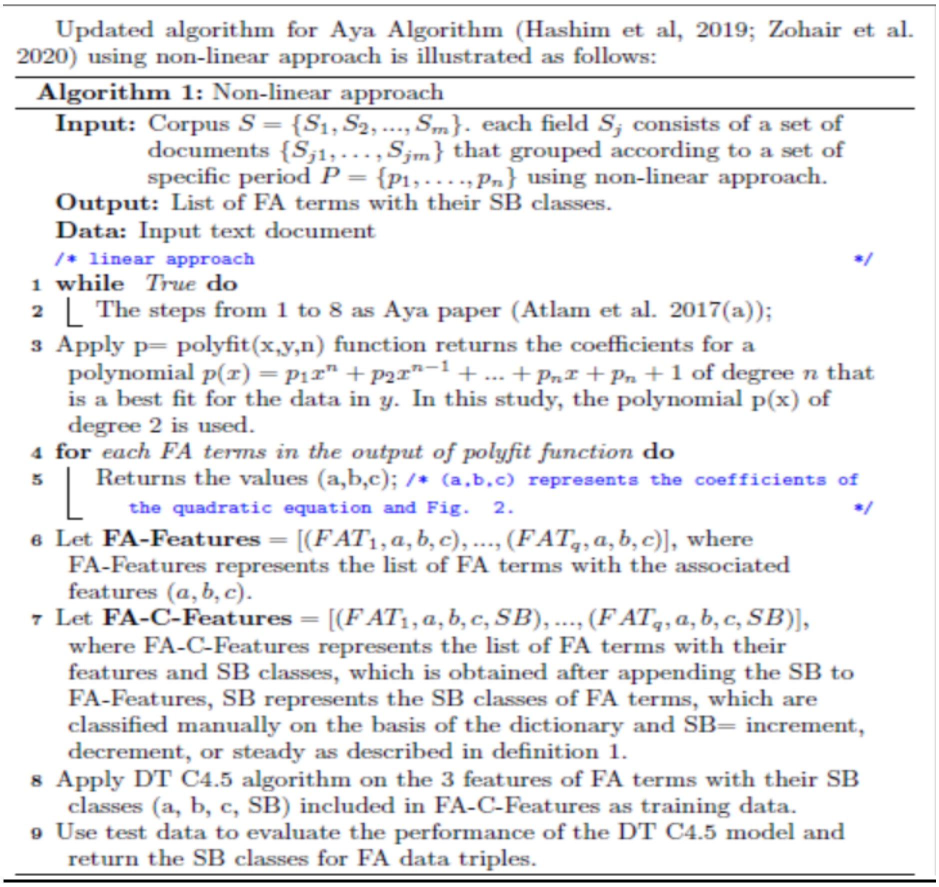

Input text document
Stem of words
Stop words
Select Noun words with its frequency
A Algorithm as old Approach
Select FA Terms for each specific field
Calculate (a,b,c) for each FA in each period using polyfit(x,y,n) and non-linear approach
(FA;a,b,c) DF Descriptive Features classified by human
Stabilization Classes (SB) classified by human
Training data
Test data
Automatic Decision Tree Algorithm C4.5
Automatically obtained SB Classes for any data

Figure 1. System outline

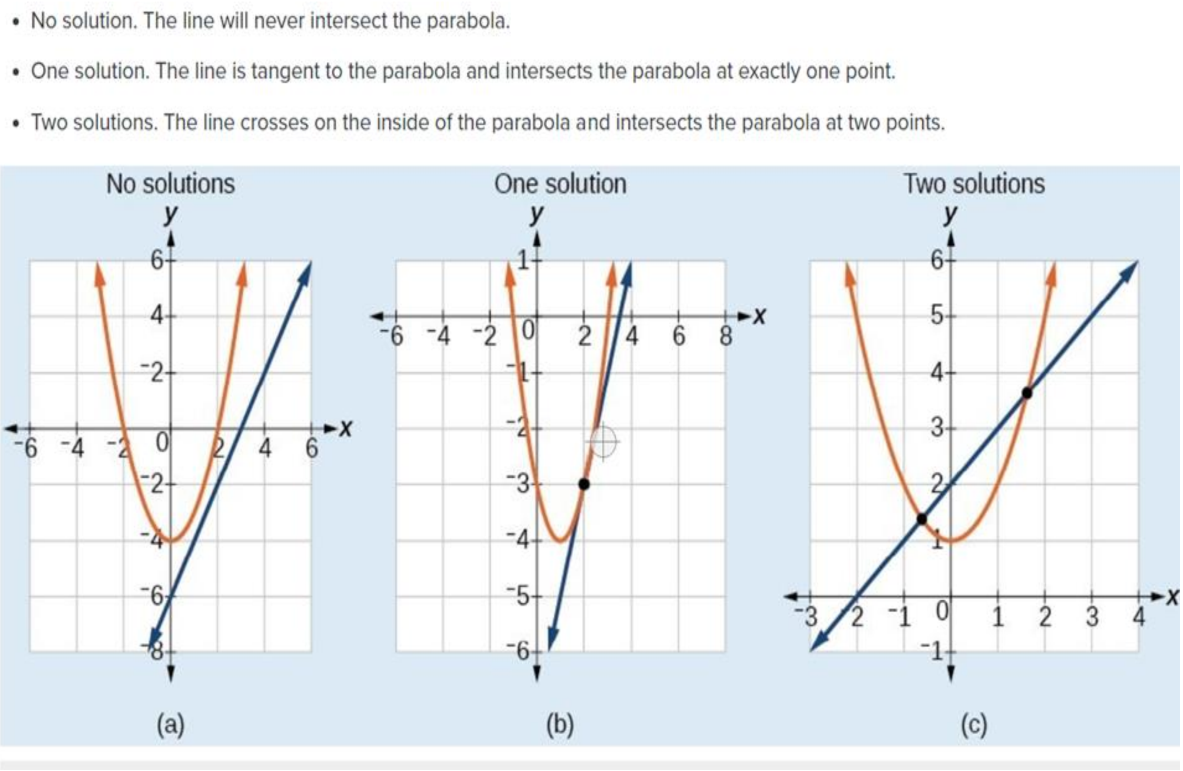
4.2 Definition and Algorithm

Definition 1. Stabilization Classes (SB) include (increment class, decrement class, and steady class)[3].



Updated algorithm for Aya Algorithm (Hashim et al, 2019; Zohair et al. 2020) using non-linear approach is illustrated as follows:

```
Algorithm 1: Non-linear approach
Input: Corpus $S = \{S_1, S_2, ..., S_m\}$. each field $S_j$ consists of a set of
       documents $\{S_{j1}, ..., S_{jm}\}$ that grouped according to a set of
       specific period $P = \{p_1, ...., p_n\}$ using non-linear approach.
Output: List of FA terms with their SB classes.
Data: Input text document
/* linear approach */
while True do
    The steps from 1 to 8 as Aya paper (Atlam et al. 2017(a));
Apply p= polyfit(x,y,n) function returns the coefficients for a
  polynomial $p(x) = p_1x^n + p_2x^{n-1} + ... + p_nx + p_n + 1$ of degree $n$ that
  is a best fit for the data in $y$. In this study, the polynomial p(x) of
  degree 2 is used.
for each FA terms in the output of polyfit function do
    Returns the values (a,b,c); /* (a,b,c) represents the coefficients of
      the quadratic equation and Fig. 2. */
Let FA-Features = $[(FAT_1, a, b, c), ..., (FAT_q, a, b, c)]$, where
  FA-Features represents the list of FA terms with the associated
  features $(a, b, c)$.
Let FA-C-Features = $[(FAT_1, a, b, c, SB), ..., (FAT_q, a, b, c, SB)]$,
  where FA-C-Features represents the list of FA terms with their
  features and SB classes, which is obtained after appending the SB to
  FA-Features, SB represents the SB classes of FA terms, which are
  classified manually on the basis of the dictionary and SB= increment,
  decrement, or steady as described in definition 1.
Apply DT C4.5 algorithm on the 3 features of FA terms with their SB
  classes (a, b, c, SB) included in FA-C-Features as training data.
Use test data to evaluate the performance of the DT C4.5 model and
  return the SB classes for FA data triples.
```



- No solution. The line will never intersect the parabola.
- One solution. The line is tangent to the parabola and intersects the parabola at exactly one point.
- Two solutions. The line crosses on the inside of the parabola and intersects the parabola at two points.

Figure 2. Cases of non-linear points

5. Evalution

5.1 Data Set

The new methodology is prepared utilizing corpus gathered from the web. Especially, information corpus is gathered from Independent News and Medical News Today in different fields as appeared in Fig.3.

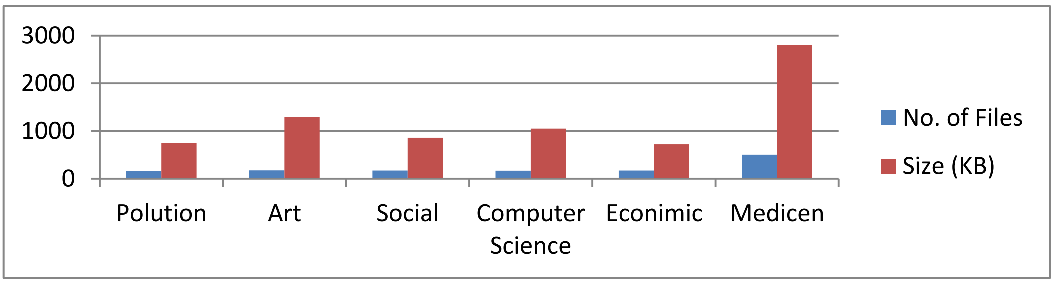


Figure 3. Number and size of Data Set in each Field

The quantity of documents in our corpus is 1,356 records, and it is about 7.49 MB. We utilized our corpus to discover FA terms with their levels. Besides, we have chosen great, semi-great, and super FA terms that are identified with the therapeutic field to examine the impacts of the time change by utilizing frequencies. We have focused on choosing the initial three levels (great, semi-great and super) as they are the best delegate of the report field. The gathered information are isolated into two gatherings; one speaks to the preparation information which thought about the information to the DT (C4.5)-training data and the other one is the test information that are vary from input information totally. The highlights of the two gatherings determined multiple times one by utilizing customary strategy for direct technique and the new methodology by utilizing non-straight technique with the idea that recurrence of FA terms changes by time as Table 1.

Table 1. An Example Of DT Data Using New Approach

A	B	C	DF	Class
0.00035177	-1.4159	1424.7066	D_name	I
0.00021106	-0.84951	854.824	O_name	I
-0.0010948	4.4054	-4431.7997	Di_name	D
0.00028141	-1.1327	1139.7653	m_name	I
-4.9389e^{-06}	0.019733	-19.71	Di_name	C
7.0353e^{-05}	-0.28317	284.9413	T_name	D
0.00014071	-0.56634	569.8827	m_name	C

5.2 Experimental Result

The general rationale of utilizing DT is to make a preparation model which can be utilized to foresee class by taking in choice guidelines induced from earlier information (preparing information). So DT is learned by the preparation information, after that there is an interconnection done between the aftereffect of DT and human results by reliance on the grouping of SB classes of the test information .

Table 2 speaks to the conclusive outcome of DT utilizing the conventional technique. Qualities with super addition letter are speaking to the crossing point between right human choice and right DT choice. This numbers is the quantity of FA terms that are grouped effectively by both Manually and DT framework.

Table 2. DT result using traditional method

	DT Evaluation		
Inc. Class	**Std.Class**	**Des.Class**	
1069^a	182	0	Inc. Class
Human Evaluation			
83	2882^a	0	Std.Claa
155	185	12^a	Des.Class

The crossing point between right human choice and right DT choice

To assess this framework, we use exactness, review, and F_measure rates on each class. This rates are characterized as follows:

$$\text{Precision}=\frac{Correct\ classified\ FA\ terms\ determined\ by\ decision\ tree}{Total\ classified\ FA\ terms\ deremined\ by\ decision\ tree}$$

$$\text{Recall}=\frac{Correct\ classified\ FA\ terms\ determined\ by\ decision\ tree}{Total\ correctly\ classified\ FA\ terms\ deremined\ by\ Human}$$

$$\text{F\_measure}=\frac{2*Precision*Rcall}{Precision+Rcall}$$

Fig. 4 shows the pace of accuracy, review, and F-measure to assess the SB classes that come about because of applying DT on FA terms frequencies in explicit period utilizing new methodology technique.

DT EVALUTION OF THE NEW METHOD

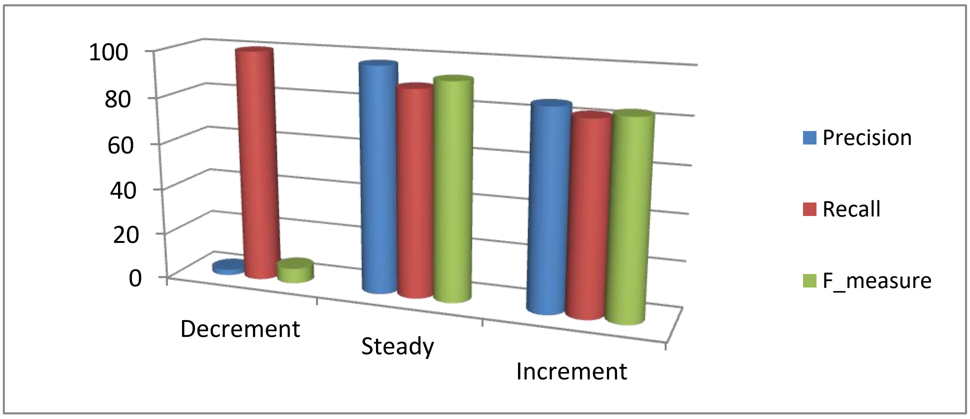


Figure 4. Recall, Precision and F-measure using three stability classes

From Fig. 4, the paces of exactness, review and F-measure show that the precision level of the new framework to characterize FA terms effectively that are evaluated consequently by the DT C4.5 dependent on recurrence change with time.

5.3 Comparison with the Traditional Method

Right now, non-direct methodology is utilized to assess the exactness of the new strategy, the amendment strategies for the accuracy of DT utilizing the straight pattern model of old methods which depended on utilizing basic words and dismissed the significant association among words and fields and furthermore utilizing pattern line to speak to connection between standardization recurrence and time as appeared in Table 3. Table 3 shows the comparison between the rates of Recall, Precision and F-measure for new (non-linear) and linear trend model.

Table 3. Comparison between old and new methods using three stability classes

	Old Method			New method		
Classes	Decrement	Steady	Increment	Decrement	Steady	Increment
Precision	2.3	95.6	78.7	2.4	97.2	85.4
Recall	78.6	84.4	73.3	100	88.7	81.8
F-Measure	4.4	89.7	75.9	6.6	92.8	83.6

From the evaluation results shown in Table 3, it is clear that the rates of Recall, Precision and F-measure for new (non-linear) is increasing by 10% than the rate using the linear trend model.

6. Conclusion

In this paper, a new technique model called non-linear model is introduced to produce automatically SB classes for ordered FA terms. The viability of the new technique (non-linear model) is affirmed by F-measure for as 83.4% for (IC), as 92.8% for (CC), and as 6.6% for (DC), respectively. However, F-measure is 75.9% for (IC), 89.7 for 9CC), and 4.4% for (DC) using old method (linear model). In conclusion, the new methodology upgraded the F-measure for increase, consistent, decrement classes by 7.7%, 3.1%, 2.2%, sequentially. It turns out that the performance is better when using our new approach. Therefore, The new approach improves the performance of decision tree than linear approach by using non-linear approach and other traditional approaches. Future work could focus in applying the new approach for Arabic and other languages.

Refrences

Atlam, E. S., & Elbarbay, O. (2017a). Granular information retrieval using Neighborhood Systems. *Journal of Mathematical Methods in the Applied Sciences*, *41*(15). https://doi.org/10.1002/mma.4610

Atlam, E. S., Doaa, A. S., Fayed, F. M. G. (2018). An Improvement of FA Terms Dictionary using Power Link and Co-Word Analysis. *International Journal of Advanced Computer Science and Applications, 9*(2). https://doi.org/10.14569/IJACSA.2018.090233

Atlam, E. S., Elmarhomy, G., Fuketa, M., Morita, K., & Aoe, J. (2006). Automatic building of new field association word candidates using search engine. *Information Processing and Managements*, *42*(4), 951-962. https://doi.org/10.1016/j.ipm.2005.08.006

Atlam, E. S., Fayed, G., Azaa, T., & Aya, I. (2017b). A new retrieval method based on time series variation using field association terms Journal. *Journal of Mathematical Methods in the Applied Sciences*, *41*(15). https://doi.org/10.1002/mma.4713

Atlam, E. S., Fuketa, M., Morita, K., & Aoe, J. (2003). Documents similarity measurement using field association terms. *Information Processing and Managements*, *39*(6), 809-824. https://doi.org/10.1016/S0306-4573(03)00019-0

Atlam, E. S., Makoto, O., Masami, S., & Aoe, J. (2001). An evaluation method of words tendency depending on time-series variation and its improvements. *Information Processing and Managements, 8*(2), 157-171. https://doi.org/10.1016/S0306-4573(01)00028-0

Atlam, E. S., Morita, K., Fuketa, M., & Aoe, J. (2002). A new method for selecting english field association terms of compound terms and its knowledge representation. *Information Processing and Managements*, *38*(6), 807-821. https://doi.org/10.1016/S0306-4573(01)00062-0

Azzopardi, J., & Staff, C. (2012). Incremental clustering of news reports. *Algorithms, 5*(3), 364-378. https://doi.org/10.3390/a5030364

Ding, Y., Gobinda, G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing and Management, 37,* 817-842. https://doi.org/10.1016/S0306-4573(00)00051-0

Dozawa, T. (1999). Innovative Multi-Information Dictionary. *Imidas'99*. Japan: Annual Series, Zueisha Publication Co.

Fukumoto, F., Suzuki, Y., & Fukumoto, J. (1996). An automatic clustering of articles using dictionary definitions. *Trans Inf Process Soc Japan, 37*(10), 1789-1799. https://doi.org/10.3115/992628.992699

Hashim, H., Atlam, E., Ahmad, R. A., & Malik, A. (In press) An Implementations Method for Arabic Keyword Tendency Using Decision Tree, An International Journal of Computer Applications in Technology (IJCAT), (Accepted Dec., 2019).

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Liman, J. (1996). Cue phrase classification using machine learning. *J Artif Intell Res*., *5*(1), 53-94. https://doi.org/10.1613/jair.327

Morita, K., Atlam, E. S., Fuketa, M., Yuya, I., & Aoe, J. (2012). An automatic extraction method of word tendency judgment for specific subject. *Inf Technol J*., *11*(8), 1007-1015.

Rokaya, M., Atlam, E. S., Fuketa, M., Dorji, T., & Aoe J. (2008). Ranking of field association terms using co-word analysis. *Information Processing and Managements, 44*(2), 738-755. https://doi.org/10.1016/j.ipm.2007.06.001

Sakurai, T., & Utsumi, A (2004). Query-based multi-document Q4 summarization for information retrieval. *In The Proceeding of NTCIR.*

Samuel, S. L., Shishibori, M., Sumitomo, T., & Aoe, J. (2001). Extraction of field-coherent passages. *Information processing and Management*, *38*(2), 173-207. https://doi.org/10.1016/S0306-4573(01)00032-2

Zohair, M. S., Abdallah, A. M., Atlam, E. S., Talal, H. N., Ahmad, R. A., & Ghada, E. (2020). A new Approach of Time Series Variation Based on Power Links and Field Association Words. *Journal of Computer and Communication, 8*, 72-85. https://doi.org/10.4236/jcc.2020.83008

Copyrights