

Clustering of Web Search Results Based on Document Segmentation

Mohammad Hasan Haggag¹, Amal Aboutabl¹ & Najla Mukhtar¹

¹ Faculty of Computers and Information, Helwan University, Cairo, Egypt

Correspondence: Najla Mukhtar, Faculty of Computers and Information, Helwan University, Cairo, Egypt.
E-mail: najlacs2008@yahoo.com; mohamed.haggag@fci.helwan.edu.eg; aaboutabl@helwan.edu.eg

Received: April 26, 2013 Accepted: May 28, 2013 Online Published: June 28, 2013

doi:10.5539/cis.v6n3p89

URL: <http://dx.doi.org/10.5539/cis.v6n3p89>

Abstract

The process of clustering documents in a manner which produces accurate and compact clusters becomes increasingly significant mainly with the vast size of information on the web. This problem becomes even more complicated with the multi-topics nature of documents these days. In this paper, we deal with the problem of clustering documents retrieved by a search engine, where each document deals with multiple topics. Our approach is based on segmenting each document into a number of segments and then clustering segments of all documents using the Lingo algorithm. We evaluate the quality of clusters obtained by clustering full documents directly and by clustering document segments using the distance-based average intra-cluster similarity measure. Our results illustrate that average intra-cluster similarity is increased by approximately 75% as a result of clustering document segments as compared to clustering full documents retrieved by the search engine.

Keywords: Clustering, Lingo algorithm, Document segmentation, Cluster similarity

Abbreviations

VSM: vector space model.

LSI: Latent Semantic Indexing.

SVD: Singular Value Decomposition.

SHOC: Semantic, Hierarchical, Online Clustering.

1. Introduction

Owing to the massive size of information on the web and low accuracy of user queries, finding the right information from the web and the necessity to easily classify them may be difficult if not impossible. One approach that tries to solve this problem is using clustering techniques for grouping similar documents together in order to facilitate presenting results in a more compact form and enable thematic browsing the results set. For instance, when a search engine is given a query, e.g., 'agents', it reacts to a set of search results R_i (where i is the number of search results returned from the query). Each search result is a short description of a web page that matches the query. In practice, a great number of results are returned, and we need a method to group related results and exclude the irrelevant ones (Ueda & Saito, 2006). These irrelevant results come up since the query terms can come out in many different contexts, e.g., travel agents or intelligent agents. Even within a single context there can be multiple sub-topics, e.g., intelligent agent software or intelligent agent conferences (Yan, 2000). As a result, if we set the related results together this will be very useful to the user.

In this paper, every document in the clustering results gained from web search engines is decomposed into segments. Then, the Lingo algorithm is used to cluster all document segments. Clustering quality is evaluated using some distance-based similarity measures such as intra-cluster similarity.

2. Related Work

One of the main studies of clustering criteria was made by Milligan and Cooper (1985), in which they carried out an experimental evaluation of 30 different criteria. Their focus was on stopping rules for hierarchical clustering. They tested these criteria on simulated data sets involving a maximum of (Macskassy, Banerjee, Davison, & Hirsh, 1998) clusters and (Huang, 2008) attributes. These practical differences encouraged us to sensitivity examine

clustering criteria for such large and complex data sets. Our study focuses on distance-based clustering, which depends on a similarity function to compare vectors describing the objects to be clustered. An alternative class of clustering algorithm is known as mixture modelling, where the objects to be clustered and generated from a combination of probability distributions of a known type (Oliver, Baxter, & Wallace, 1996).

Another approach of clustering techniques is called (K-means) and it is broadly used in document clustering. K-means is based on the idea that a core point can stand for a cluster. Particularly, K-means makes use of the notion the notion of a centroid, which is the mean or middle point of a group of points. Note that a centroid almost never corresponds to a real data point (Manning, Raghavan, & Schtze, 2008).

3. Vector Space Model for Document Clustering

The VSM (vector space model) is one of the most widespread models for representing documents to be clustered. Clustering documents using the vector space model is categorized by three characteristics (Frey, 2012):

- Each document is characterized by a vector of word frequencies, where generally occurring words have been excluded using a stop list or heuristic feature selection techniques.
- A distance measure is defined as a function of these document vectors, in order that we can measure the similarity or distance between any pair of documents in the vector space.
- A clustering algorithm utilizes this distance measure to set related documents into clusters.

4. Lingo Algorithm

Lingo algorithm is used to reverse the traditional order of cluster discovery. Initially, there was an attempt to find good, theoretically varied cluster labels and then assign documents to the labels to form groups, rather than computing proximity between documents and then labelling the discovered groups.

In conventional approaches, which determine group labels after discovering the real cluster content, this task proves fairly difficult to be achieved.

Lingo algorithm is used to cluster search results, the input documents are results of a search engine and these documents consist of the title, a short web snippet and URL (Osiński, 2003).

4.1 Stages of Lingo Algorithm

Lingo algorithm can be divided into five stages which are set as follows:

4.1.1 Pre-processing Stage

The pre-processing stage aims at preparing the input data and changes it to a useable form. This consists of text filtering, where HTML tags, character entities, and special characters, excluding sentence boundaries, are removed. Sentence boundaries are essential for the phrase extraction during the second stage (section 4.1.2). In addition, the language of every document is determined before stop words get removed and the text is stemmed. By means of stop words, the identification of the language is carried out. In case of English documents the Porter stemmer algorithm is used, for Polish they had their own simple dictionary stemmer.

4.1.2 Frequent Phrase Extractions Stage

The feature extraction stage aims at discovering phrases and single terms that will potentially be able to explain the verbal meaning behind the LSI (Latent Semantic Indexing). To be regarded as a candidate for a cluster label, a phrase or term must have some characteristics which are as follows (Osiński, 2003):

- Appearing in the input documents at least a specified number of times. This is a result of broadly accepted hypotheses in information recovery that features that reappear regularly in the input documents have the strongest descriptive power. Furthermore, omitting the uncommon words and phrases will significantly increase the time efficiency of the algorithm as a whole.
- Not cross sentence boundaries. Mainly, sentence markers mark a topical change; consequently a phrase extending beyond one sentence is probable to carry little meaning to the user.
- Be a complete phrase. Compared to partial phrases, complete phrases should permit better description of clusters (compare “Senator Hillary” and “Senator Hillary Rodham Clinton”).
- Not begin nor end with a stop word – stripping the phrases of leading and trailing general terms is about to increase their readability as cluster labels.

4.1.3 Cluster Label Induction Stage

In this stage, descriptions of meaningful group are shaped relying on the SVD (Singular Value Decomposition) decomposition of the term-document matrix. There are four steps to this stage; term-document matrix building is built depending on the input group of snippets. In opposition to the SHOC (Semantic, Hierarchical, Online Clustering) approach, in Lingo only single terms are used to construct the A matrix. The cause for this is that it is more natural to use groups of single words instead of groups of phrases to express abstract concepts, which really very often are phrases themselves. Moreover, individual terms permit finer granularity of description, abstract concept discovery in this step the Singular Value Decomposition of the term-document matrix is performed to acquire the orthogonal basis of its column space, phrase matching this step depends on a vital observation that both the abstract concepts and the phrases discovered in the feature extraction stage are expressed in the same space – the column space of the original term-document matrix. Consequently, the classic cosine distance can be used to compute how well a phrase or a single term represents an abstract concept, and label pruning even though the SVD-derived basis of the column space of the term-document matrix is orthogonal, some of the cluster label candidates (verbal representations of the abstract concepts) tend to overlap. The cause for this is that for each abstract concept only the best matching word or phrase is selected and other phrases, which potentially could help express the difference between two abstract concepts, are ignored (Frey, 2012).

4.1.4 Cluster Content Discovery Stage

In this stage, the classic Vector Space Model is used to allocate the input snippets to the cluster labels induced in the previous stage. The assignment process much resembles document recovery based on the VSM model– the only difference is that rather than one query, the input snippets are harmonized against every single cluster label (Osiński, 2003).

4.1.5 Final Cluster Formation Stage

In the final stage of the algorithm, cluster scores are computed and if in the presentation interface the resulting clusters are sorted according to their score, the user will presented with the well-described and relatively large groups in the first place (Osiński, 2003).

5. Clustering Evaluation

The purpose functions in clustering formalize the aim of achieving high intra-cluster similarity (documents within a cluster are similar) and low inter-cluster similarity (documents from different clusters are dissimilar). This is an internal criterion for the clustering quality. But good scores on an internal criterion do not essentially translate into good effectiveness in an application (Raskutti, 1999).

5.1 Computation of Inter-Cluster and Intra-Cluster Similarity

Intra-cluster similarity is used to determine how near the data objects are in a cluster. For a clustering, $\pi(D) = \{C_1, C_2, \dots, C_{|\pi(D)|}\}$, intra-cluster similarity is measured as follows (Raskutti, 1999):

Assume that:

- D: data set,
- C_i : is a cluster (block) of $\pi(D)$, $1 \leq i \leq |\pi(D)|$,
- ϕ : consensus function,
- m: the number of clustering's,
- $votes_{ij}$: the number co-occurrences of i and j in a cluster,
- \forall : universal quantification means for all.

Given a group of clusterings $\Pi = \{\pi_1, \pi_2, \dots, \pi_m\}$, the problem of joining multiple clusterings is defined as finding a new clustering $\pi^* = \{C^*1, C^*2, \dots, C^*|\pi^*|\}$ by using the information provided by Π . A consensus function ϕ , is used for determining quality of the final clustering π^* . See (Strehl & Ghosh, 2003)

$$\forall i(\phi(\pi^*(D)) \geq \phi(\pi_i(D))), 1 \leq i \leq |\Pi| \quad (1)$$

Many consensus functions have been suggested in the literature. In (Strehl & Ghosh, 2003), a suggested consensus function is based on a co-association measure (simultaneous occurrences) stated by:

$$coassoc(i, j) = votes_{ij}/m, \quad (2)$$

$$ICS(\pi(D)) = \sum_{i=1}^{|\pi(D)|} \frac{1}{|C_i|^2} \sum_{d,d' \in C_i} similarity(d, d') \tag{3}$$

For the same clustering, inter-cluster similarity is defined as follows:

$$ECS(\pi(D)) = \sum_{i=1}^{|\pi(D)|} \sum_{j=i+1}^{|\pi(D)|} \frac{1}{|C_i||C_j|} \sum_{d \in C_i, d' \in C_j} similarity(d, d') \tag{4}$$

Finally, the consensus function is

$$\phi(\pi(D)) = k_1 \cdot ICS(\pi(D)) + k_2 \cdot ECS(\pi(D)) \tag{5}$$

Final clustering is expected to have compressed and close clusters; consequently a better clustering formula the greater values 3 offers (from 0.5 to 0.9). In the same way, a better clustering the smaller values 4 supplies., k. K1, and K2 parameters in Formula 5 are user defined values, which satisfy K1 > 0 and K2 < 0.s

Inspection of formula 5 reveals that its time complexity is quadratic concerning the number of objects in the data set, O(|D|²). This difficulty is owing to pair wise similarity calculations performed in formulas 3 and 4.

Example: Let us calculate intra-cluster similarity, ICS $\pi(\pi^*(D))$, of $\pi^*(D)$ of Table 2 using multiple clustering shown in Table1 by using formula 6.

$$\begin{aligned} ICS_{\Pi}(\pi^*(D)) &= \frac{1}{3^2} \left(\binom{2}{2} + \binom{0}{2} + \dots + \binom{1}{2} \right) \\ &+ \frac{1}{2^2} \left(\binom{1}{2} + \binom{0}{2} + \dots + \binom{0}{2} \right) + \dots \\ &+ \frac{1}{1^2} \left(\binom{0}{2} + \binom{0}{2} + \dots + \binom{0}{2} \right) \end{aligned}$$

Table 1. Binary representation of multiple clustering, Π

		d1	d2	d3	d4	d5	d6	d7	d8
Π_1	C_{*11}	1	0	1	0	0	1	0	0
	C_{*12}	0	0	0	1	1	0	0	0
	C_{*13}	0	0	0	0	0	0	0	0
	C_{*14}	0	1	0	0	0	0	1	1
Π_2	C_{*21}	1	1	0	1	0	0	0	0
	C_{*22}	0	0	0	0	0	0	0	1
	C_{*23}	0	0	0	0	1	0	1	0
	C_{*24}	0	0	1	0	0	1	0	0
Π_3	C_{*31}	0	0	0	0	0	0	0	0
	C_{*32}	0	0	1	0	0	1	0	0
	C_{*33}	1	1	0	1	0	0	0	1
	C_{*34}	0	0	0	0	1	0	1	0

Table 2. Binary representation of π^*

		d1	d2	d3	d4	d5	d6	d7	d8
π^*	C_{*1}	1	0	1	0	0	0	1	0
	C_{*2}	0	1	0	0	0	1	0	0
	C_{*3}	0	0	0	1	1	0	0	0
	C_{*4}	0	0	0	0	0	0	0	1

We symbolize each cluster with a bit vector. Existence of an object in a cluster is shown by 1; similarly absence of an object is revealed by 0. Each cluster representation is as long as the size of the database, $|D|$. Three clustering, each having four clusters is shown in Figure 1. An example of a cluster is shown below: C11 cluster has d1, d3, and d6 objects.

$$\begin{array}{c} C_{11} \\ \boxed{1 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 0} \end{array}$$

So as to estimate Formulas 3 and 4, each cluster has to be examined for pair wise objects, and corresponding entries in the co-association matrix has to be updated. For example, C11 cluster increments following object pairs in the co-association matrix: (d1, d1), (d1, d3), (d1, d6), (d3, d3), (d3, d6), and (d6, d6). Because of the pair wise increment nature, this computation has quadratic time complexity.

Let Π be multiple clustering, and π^* be a new final clustering for a data set D . In following we define intra-cluster, and inter-cluster similarities of $\pi^*(D)$. Intra-cluster similarity of $\pi^*(D)$ is shown in Formula 6.

$$ICS_{\Pi}(\pi^*(D)) = \sum_{k=1}^{|\pi^*(D)|} \frac{1}{|C_{*k}|^2} \sum_{i=1}^{|\Pi|} \sum_{j=1}^{|\pi_i|} \binom{|C_{*k} \wedge C_{ij}|}{2} \quad (6)$$

And, formula 7 shows the inter-cluster similarity of $\pi^*(D)$.

$$\begin{aligned} ECS_{\Pi}(\pi^*(D)) = & \\ \sum_{k=1}^{|\pi^*(D)|} \sum_{l=k+1}^{|\pi^*(D)|} \frac{1}{|C_{*k}| |C_{*l}|} \sum_{i=1}^{|\Pi|} \sum_{j=1}^{|\pi_i|} & \left(\binom{|(C_{*k} \vee C_{*l}) \wedge C_{ij}|}{2} \right. \\ & \left. - \binom{|C_{*k} \wedge C_{ij}|}{2} - \binom{|C_{*l} \wedge C_{ij}|}{2} \right) \end{aligned} \quad (7)$$

Finally, consensus function is described as:

$$\phi_{\Pi}(\pi^*(D)) = k_1 \cdot ICS_{\Pi}(\pi^*(D)) + k_2 \cdot ECS_{\Pi}(\pi^*(D)) \quad (8)$$

ICS $\pi(\pi^*(D))$ is computed as follows; every cluster of $\pi^*(D)$ is logically ANDed with every cluster in π in order to find pair wise co-occurrences of the objects in the same cluster. In the same way, when computing ECS $\pi(\pi^*(D))$ every cluster pair in $\pi^*(D)$ is logically Ored to find all pairs of objects; this result is ANDed with every cluster in order to find pair wise co-occurrences of objects. So far we obtained the pair wise co-occurrences of objects in the same clusters and in two different clusters.

By subtracting (last 2 components in Formula 7) pair wise co-occurrences of the objects in the same clusters; we gain pair wise co-occurrences of objects in different clusters as shown in the formula. When calculated the same clustering, π , Formula 6 is equivalent to Formula 3 and Formula 7 is equivalent to Formula 4. Even though equivalent formulas yield same result, it is very significant to note that Formulas 6 and 7 are calculated in cluster level, not in object level for each pair wise object.

6. Experimental Work and Results

In this work, the Lingo algorithm is used for clustering documents (for more details see section 4). Our first experiment performs clustering on the original document set as returned by the search engine. In our second experiment, every document returned by the search engine is segmented and the whole set of document segments are input to our clustering system. Clustering is evaluated for both experiments by using similarity measures. Our proposed model is illustrated in the right part of Figure 1.

The document set used in our experiments consists of 30 documents. Each document in the document set is segmented resulting in a total of 90 segments. We implemented the Lingo algorithm for clustering documents. Intra-cluster similarity is computed for every pair of documents in a cluster in order to measure how near the documents are to each other within a cluster. For both experiments, intra-cluster similarity results are shown in Figure 2. Average intra-cluster similarity is computed for clusters resulting from each experiment. In the case of clustering full documents and segmented documents, the average intra-cluster similarities are 0.425 and 0.743 respectively. Hence, we conclude that the average intra-cluster similarity is increased approximately by 75% as a result of clustering document segments rather than clustering full documents retrieved by the search engine.

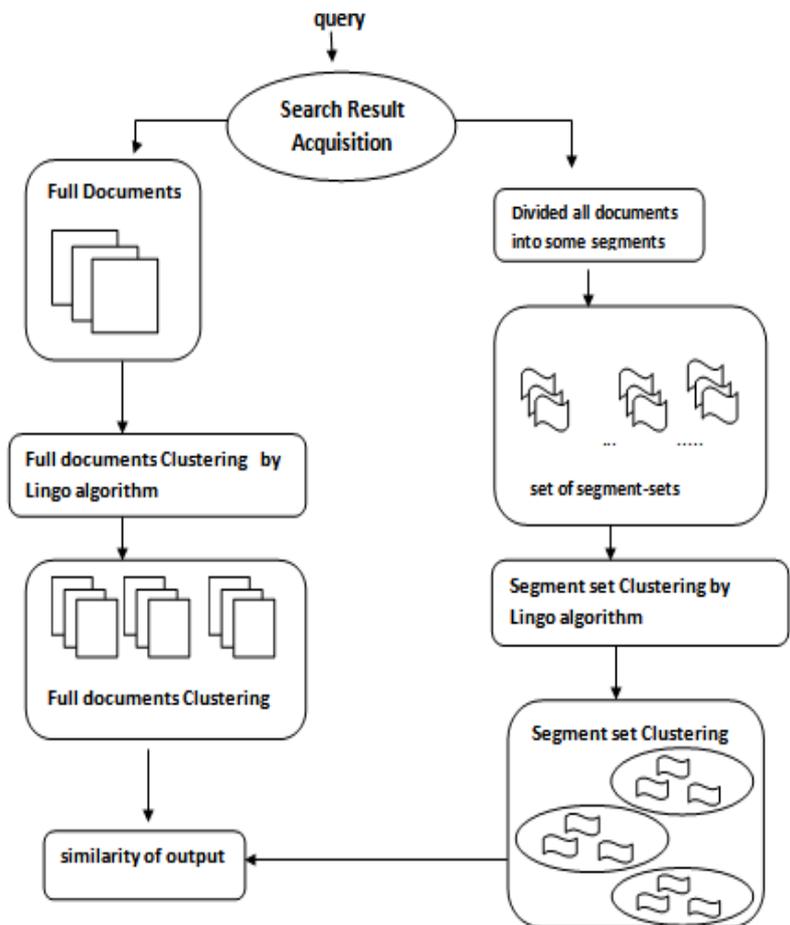


Figure 1. Proposed model

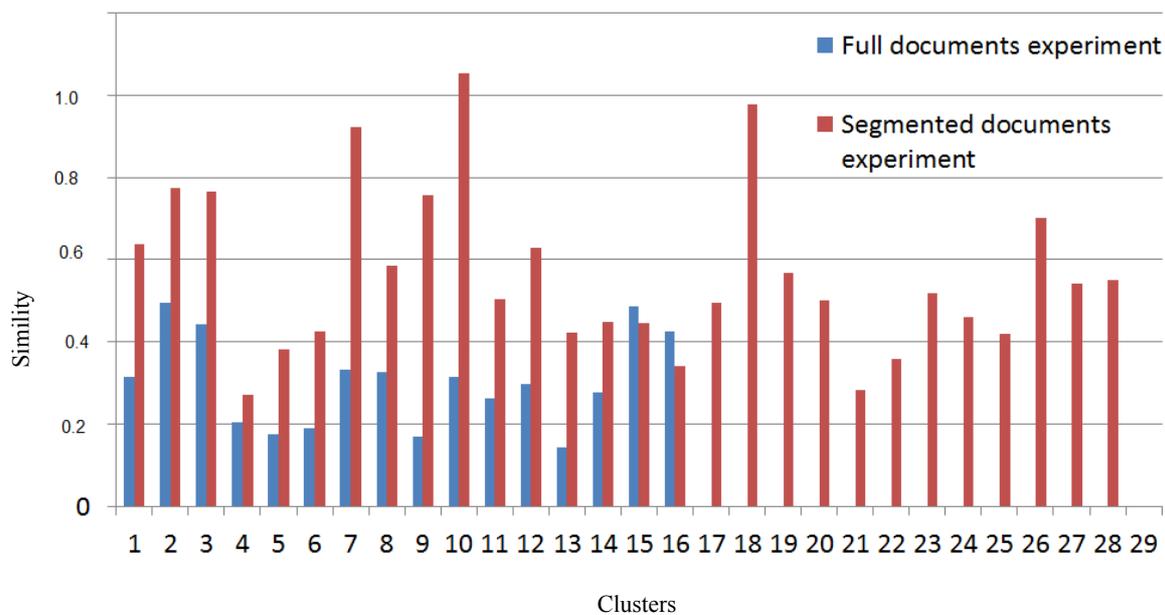


Figure 2. The result of intra-cluster similarity for the full documents experiment (without segmentation) and the segmented documents experiment

7. Conclusion and Future Work

In this paper, we cluster web search results using the lingo algorithm with and without segmentation. We compute the average intra-similarity in each case. An increase of approximately 75% in average intra-similarity is achieved when document segmentation is performed compared to the case of clustering the whole documents.

For future work, we will explore the behaviour of our system with non-document data sets as well as testing the suitability of our criteria as stopping rules for hierarchical clusters. In addition, we plan to extend our criteria to study overlapping clusters.

References

- Frey, S. (2012). *Computing Labels for Scientific Clusters using Lingo Algorithm and Word Net*. Karlsruhe.
- Huang, A. (2008). *Department of Computer Science, Similarity Measures for Text Document Clustering*. Hamilton: The University of Waikato.
- Macskassy, S., Banerjee, A., Davison, B., & Hirsh, H. (1998). Human Performance on Clustering Web Pages: A Preliminary Study. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)* (pp. 264-268).
- Manning, C. D., Raghavan, P., & Schtze, H. (2008). *Introduction to information retrieval*. New York, NY: Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511809071>
- Oliver, J., Baxter, R., & Wallace, C. (1996). Unsupervised Learning Using MML. In *Proceedings of the Thirteenth International Conference on Machine Learning (ICML-96)* (pp. 364-372).
- Osiński, S. (2003). *An Algorithm for Clustering of Web Search Results*. Poland: University of Technology.
- Raskutti, B. (1999). *An Evaluation of Criteria for Measuring the Quality of Clusters*. Telstra Research Laboratories 770 Blackburn Rd, Clayton Victoria 3168, Australia.
- Strehl, A., & Ghosh, J. (2003). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3, 583-617.
- Ueda, N., & Saito, K. (2006). Parametric Mixture Model for Multitopic Text. *Systems and Computers in Japan*, 37(2), 56-66. <http://dx.doi.org/10.1002/scj.20259>
- Yan, W. (2000). Web Mining and Knowledge Discovery of Usage Patterns. *CS 748T Project (Part I)*, February.
- Zhang, D., & Dong, Y. (2004). Semantic, Hierarchical, Online Clustering of Web Search Results. In *6th Asia-Pacific Web Conference on Advanced Web Technologies and Applications (AP Web 2004)* (p. 6978). Hangzhou, China. http://dx.doi.org/10.1007/978-3-540-24655-8_8

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).