

WEB BASED DOCUMENT RETRIEVAL USING ADVANCED CBRS

**Singaravelan Shanmugasundaram, Anthoni Sahaya Balan,
Murugan Dhanushkodi**

(1) Department of CSE, PSR Engineering College, Sivakasi, India

(2) Department of CSE, Sethu Institute of Technology, Pulloor, India

(3) Department of CSE, Manonmaniam Sundaranar University, Tirunelveli, India

Singaravelan Shanmugasundaram

Department of CSE, PSR Engineering College, Sivakasi, India

singaravelan.msu@gmail.com

Article info

Paper category: Review Paper

Received: 10.5.2017.

Accepted: 8.9.2017.

JEL classification: D8

ABSTRACT

Multi-document summarization is an automatic procedure aimed at extraction of information from multiple texts written about the same topic. Resulting summary report allows individual users, such as professional information consumers, to quickly familiarize themselves with information contained in a large cluster of documents. This proposed work CBRS (Cluster Based Ranking with Significance) summarizes the multi document with semantic meaning of the terms in the documents. Such that it produces a good results while clustering and ranking with retrieving document. As a clustering result to improve or refine the sentence ranking results. The effectiveness of the proposed approach is demonstrated by both the cluster quality analysis and the summarization evaluation conducted on our simulated datasets.

Keywords:

Documentation Summarization, Sentence Clustering, Sentence Ranking

1. INTRODUCTION

The steady and amazing progress of computer hardware technology in the last few years has led to large supplies of powerful and affordable computers, data collection equipment's, and storage media. Due to this progress there is a great encouragement and motivation to the database and information industry to make a huge number of databases and information repositories; which is available for transaction management, information retrieval, and data analysis. Thus, technology advancement has provided a tremendous growth in the volume of the text documents available on the internet, digital libraries and repositories, news sources, company-wide intranets, and digitized personal information such as blog articles and emails. With the increase in the number of electronic documents, it is hard to organize, analyze and present these documents efficiently by putting manual effort. These have brought challenges for the effective and efficient organization of text documents automatically. Data mining is the process of extracting the implicit, previously unknown and potentially useful information from data. Document clustering, subset of data clustering, is the technique of data mining which includes concepts from the fields of information retrieval, natural language processing, and machine learning. Document Clustering is different than document classification. In document classification, the classes (and their properties) are known a priori, and documents are assigned to these classes; whereas, in document clustering, the number, properties, or membership (composition) of classes is not known in advance. Thus, classification is an example of supervised machine learning and clustering that of unsupervised machine learning.

1.1. PROBLEM STATEMENT

The main problems in the existing work is which clusters and ranks according to the corpus or terms in each document, but it doesn't look up the exact meaning of the word and summarize it. This leads to irrelevant results. To overcome the above problem a real similarity measure is needed so find the exact similarity a WordNet tool is applied.

1.2. PROPOSED SYSTEM

The basic idea is as follows. First the documents are clustered into clusters. Then the sentences are ranked within each cluster. After that, a mixture model is used to decompose each sentence into a K-dimensional vector, where each dimension is a component coefficient with respect to a cluster. Each dimension is measured by rank distribution. Sentences then are reassigned to the nearest cluster under the new measure space. As a result, the quality of sentence clustering is improved. In

addition, sentence ranking results can thus be enhanced further by these high quality sentence clusters. In all, instead of combining ranking and clustering in a two stage procedure like the first category, isolation, we propose an approach which can mutually enhance the quality of clustering and ranking. That is, sentence ranking can enhance the performance of sentence clustering and the obtained result of sentence clustering can further enhance the performance of sentence ranking.

The proposed system includes:

- Integrating Clustering and ranking simultaneously terms and sentences
- A Cosine similarity is used to show their relationships.
- A conditional ranking is integratedly used to perform better results.

1.3. WORDNET TOOL

WordNet is a thesaurus for the English language based on psycholinguistics studies and developed at the University of Princeton. It was conceived as a data-processing resource which covers lexico-semantic categories called synsets. The synsets are sets of synonyms which gather lexical items having similar significances, for example the words "a board" and "a plank" grouped in the synset {board, plank}. But "a board" can also indicate a group of people (e.g., a board of directors) and to disambiguate these homonymic significances "a board" will also belong to the synset {board, committee}. The definition of the synsets varies from the very specific one to the very general. The most specific synsets gather a restricted number of lexical significances whereas the most general synsets cover a very broad number of significances.

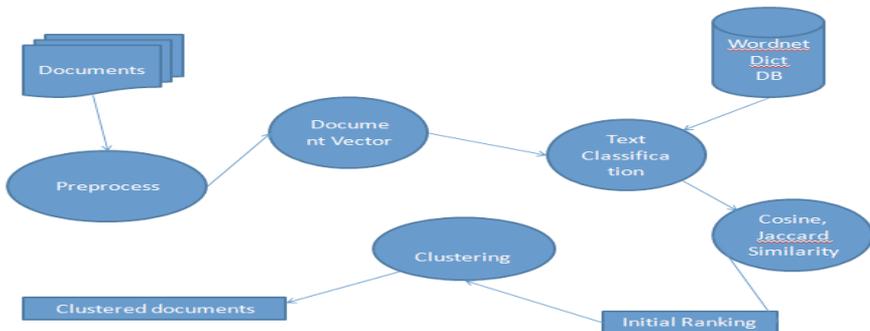
The organization of WordNet through lexical significances instead of using lexemes makes it different from the traditional dictionaries and thesaurus. The other difference which has WordNet compared to the traditional dictionaries is the separation of the data into four data bases associated with the categories of verbs, nouns, adjectives and adverbs. The names are organized in hierarchy, the verbs by relations, the adjectives and the adverbs by N-dimension hyperspaces. The following list enumerates the semantic relations available in WordNet. These relations relate to concepts, but the examples which we give are based on words.

- (1) Synonymy: relation binding two equivalent or close concepts (frail /fragile). It is a symmetrical relation.
- (2) Antonymy: relation binding two opposite concepts (small /large). This relation is symmetrical.
- (3) Hyperonymy: relation binding a concept-1 to a more general concept-2 (tulip /flower).
- (4) Hyponymy: relation binding a concept-1 to a more specific concept-2. It is the reciprocal of hyperonymy. This relation may be useful in information retriev-

- al. Indeed, if all the texts treating of vehicles are sought, it can be interesting to find those which speak about cars or motor bikes.
- (5) Meronymy: relation binding a concept-1 to a concept-2 which is one of its parts (flower/petal), one of its members (forest /tree) or a substance made of (pane/glass).
 - (6) Metonymy: relation binding a concept-1 to a concept-2 of which it is one of the parts. It is the opposite of the meronymy relation.
 - (7) Causality: relation binding a concept-1 to its purpose (to kill /to die).
 - (8) Value: relation binding a concept-1 (adjective) which is a possible state for a concept-2 (poor /financial condition).
 - (9) Has the value: relation binding a concept-1 to its possible values (adjectives) (size /large). It is the opposite of relation value.
 - (10) Similar to: certain adjectival concepts which meaning is close are gathered. A synset is then designated as being central to the regrouping. The relation 'Similar to' binds a peripheral synset with the central synset (moist /wet).
 - (11) Derived from: indicate a morphological derivation between the target concept (adjective) and the concept origin (coldly /cold).

1.4. SYSTEM DESIGN

Figure 1.: Data Flow Diagram



Source: Authors'

1.5. SYSTEM IMPLEMENTATION

The proposed Clustering across Ranking of web documents consists of four main modules. They are:

- Data Preprocessing
- Document Bi-Type Graph
- Ranking
- Similarity Measure

Data Preprocessing: The effectiveness of the proposed approach is demonstrated by both the cluster quality analysis and the summarization evaluation conducted on the DUC 2004-2007 and our simulated datasets. Document pre-processing is a prerequisite for any Natural Language Processing application. It is usually the most time consuming part of the entire process. The various tasks performed during this phase are

Parsing: Parsing of text document involves removing of all the HTML tags. The web pages will contain lot of HTML tags for alignment purpose. They does not provide any useful information for classification. All the text content between the angle braces '<' and '>' are removed in this module. The tag information between them will not be useful for mining purpose. They will occupy more space and it should be removed. This step will reduce lot of processing complexity.

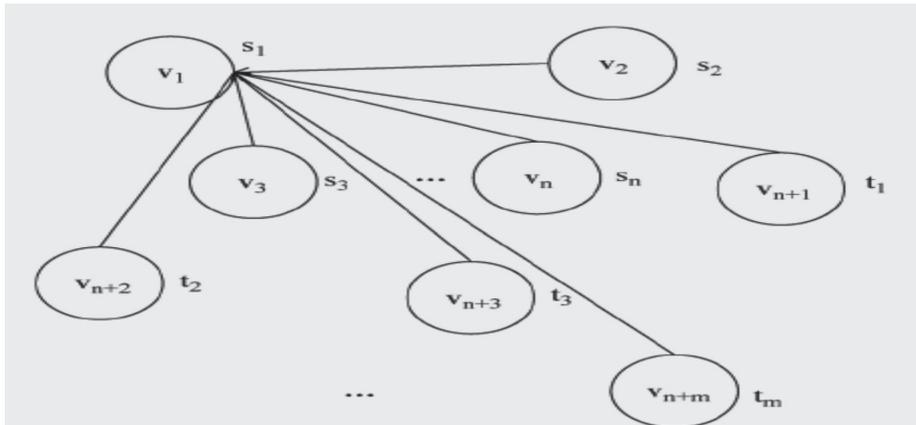
Tokenization: Tokenization is actually an important pre-processing step for any text mining task. Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining. Tokenization usually occurs at the word level. Often a tokenizer relies on simple heuristics.

Stop word Removal : Stop word removal removes the high frequent terms that do not depict the context of any document. These words are considered unnecessary and irrelevant for the process of classification. Words like 'a', 'an', 'the', 'of', 'and', etc. that occur in almost every text are some of the examples for stop words. These words have low discrimination values for the categories. Using a list of almost 500 words, all stop words are removed from the documents.

Stemming : Stemming removes the morphological component from the term, thus reducing the word to the base form. This base form doesn't even need to be a word in the language. It is normally achieved by using rule based approach, usually based on suffix stripping. The stemming algorithm used here is the Porter Stemmer algorithm, which is the standard stemming algorithm for English language. Example: Playing, Plays, Played, Play.

Document Bi-Type Graph: The main contributions of the paper are three-fold: (1) Three different ranking functions are defined in a bi-type document graph constructed from the given document set, namely global, within cluster and conditional rankings, respectively. (2) A reinforcement approach is proposed to tightly integrate ranking and clustering of sentences by exploring term rank distributions over the clusters. (3) Thorough experimental studies are conducted to verify the effectiveness and robustness of the proposed approach. Three different ranking functions are defined in a bi-type document graph constructed from the given document set, namely global, within-cluster and conditional rankings, respectively. In first present the sentence-term bi-type graph model for a set of given documents, based on which the algorithm of reinforced ranking and clustering is developed. Let , $G = \{V, E, W\}$ where V is the set of vertices that consists of the sentence set $S = \{s_1, s_2, \dots, s_n\}$ and the term set $T = \{t_1, t_2, t_3, \dots, t_n\}$, i.e. $= S \cup T$, n is the number of sentences and is the number of terms. Each term vertex is the sentence that is given in the WordNet as the description of the term. It extracts the first sense used from WordNet instead of the word itself. E is the set of edges that connect the vertices. An edge can connect a sentence to a word, a sentence to a sentence, or a word to a word, i.e. the graph G is presented in Fig. below. For ease of illustration, we only demonstrate the edges between v_1 and other vertices. All the documents are represented in the form of a vector called Term.

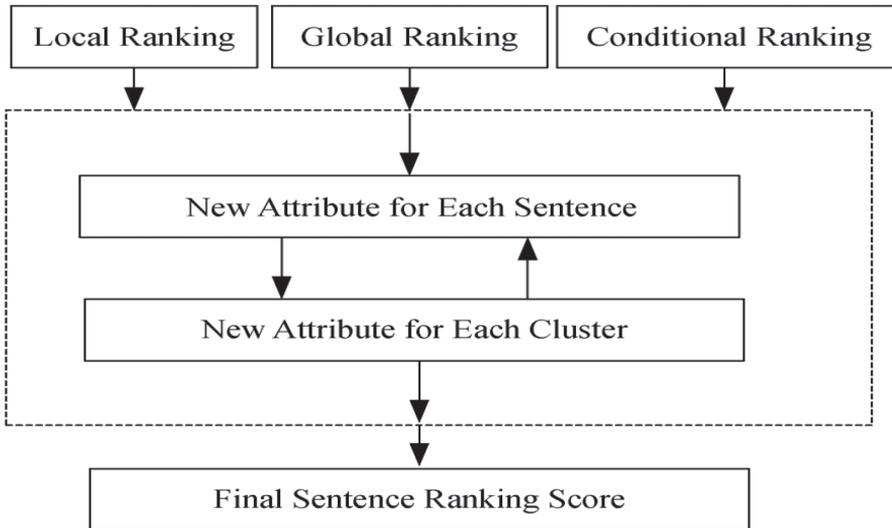
Figure 2.: Bi-Type graph



Source: Authors'

Ranking: Ranking has been done in three types ranking functions

Figure 3.: The sentence ranking process



Source: Authors'

Global Ranking (Without Clustering): A sentence should be ranked higher if it contains highly ranked terms and it is similar to the other highly ranked sentences, while a term should be ranked higher if it appears in highly ranked sentences and it is similar to the other highly ranked terms.

Local Ranking (Within Clusters): We decompose the whole document set into sentences, and obtain K sentence clusters (also known as theme clusters) by certain clustering algorithm. The V theme clusters is denoted as $C = \{C_1, C_2, \dots, C_K\}$ where C_k ($k = 1, 2, 3, \dots, K$) represents a cluster of highly related sentences S_{Ck} , which contains the terms T_{Ck} .

Conditional Ranking (Across Clusters): To facilitate the discovery of rank distributions of terms and sentences over all the theme clusters, we further define two "conditional ranking functions" $r(S|Ck)$ and $r(T|Ck)$. Sentence and term conditional ranks over all the theme clusters and are ready to introduce the reinforcement process. These two rank distributions are necessary for the parameter estimation during the reinforcement process.

Term Ranking: Term ranking is an essential issue in clustering documents. Ranking distinguishing terms higher yields better estimation of similarity between documents and hence higher quality is clustering. Standard frequency based term

ranking methods in Information Retrieval (IR). Term frequency (TF) is the frequency of a term among all the terms in the Web page collection, and calculated as $TF(t) = nt / n$, where nt is the number of occurrences of t in the collection and n is the total number of terms in the collection. Term frequency / inverse document frequency (TF/IDF) is a method to reduce the bias of term frequency by penalizing with the document frequency. It is calculated as $TF/IDF(t) = TF(t) \cdot \log |W| / |D(t)|$ where $D(t)$ is the set of Web pages t appears.

Sentence Ranking: The documents are clustered into k clusters. Then the sentences are ranked within each cluster. Grouping of words or terms and then provide the ranking for sentences.

Similarity measures: The similarity between a sentence and a cluster can be calculated as the cosine similarity between them. Where $WST(i,j)$ is the cosine similarity between the sentence S_i and the term T_j . Thus the value of $WST(i,j)$ is between 0 and 1. If $WST(i,j)$ is near to 1, it means the sentence S_i and the term T_j are semantically similar. If $WST(i,j)$ is near to 0, it means the sentence and the term are semantic different. $WSS(i,j)$ is the cosine similarity between the sentences S_i and S_j . $WTT(i,j)$ is the cosine similarity between the terms T_i and T_j . First we calculate the center of each cluster can thus be calculated accordingly, which is the mean of S_i for all in the same cluster, i.e.,

$$\overrightarrow{\text{Center}}_{C_k} = \frac{\sum_{s_i \in C_k} \overrightarrow{s_i}}{|C_k|},$$

where $|C_k|$ is the size of cluster C_k .

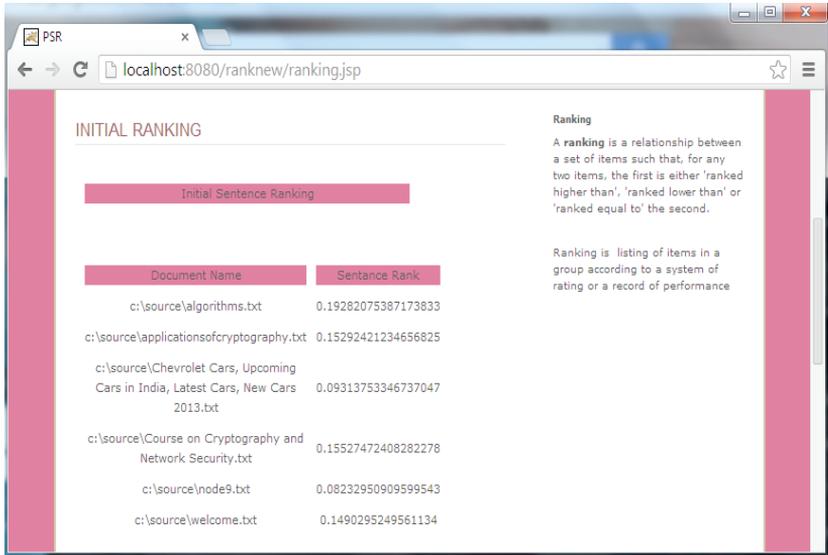
Then the similarity between a sentence and a cluster can be calculated as the cosine similarity between them, i.e.,

$$\text{sim}(s_i, C_k) = \frac{\langle \overrightarrow{s_i}, \overrightarrow{\text{Center}}_{C_k} \rangle}{\sqrt{\|\overrightarrow{s_i}\|^2} \cdot \sqrt{\|\overrightarrow{\text{Center}}_{C_k}\|^2}}.$$

Finally, each sentence is re-assigned to a cluster that is the most similar to the sentence. Based on the updated clusters, within-cluster ranking is updated accordingly, which triggers the next round of clustering refinement.

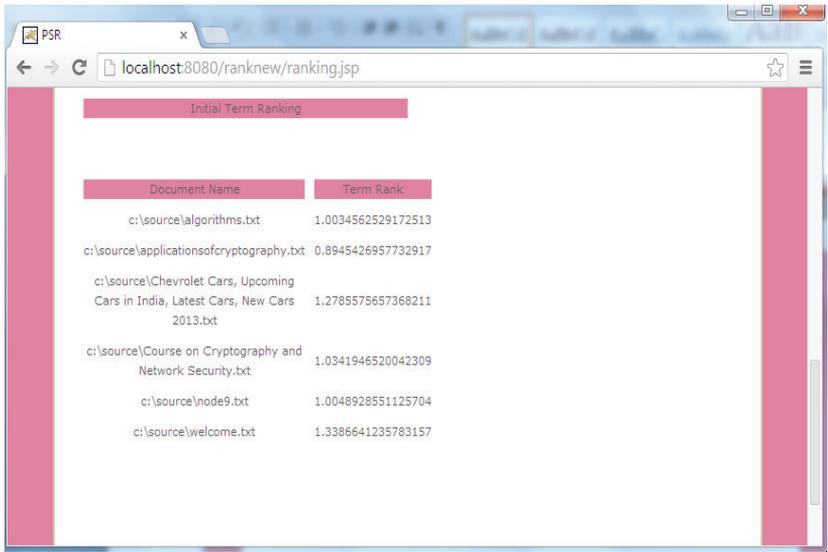
2. EXPERIMENTAL RESULTS

Figure 4.: Initial sentence ranking



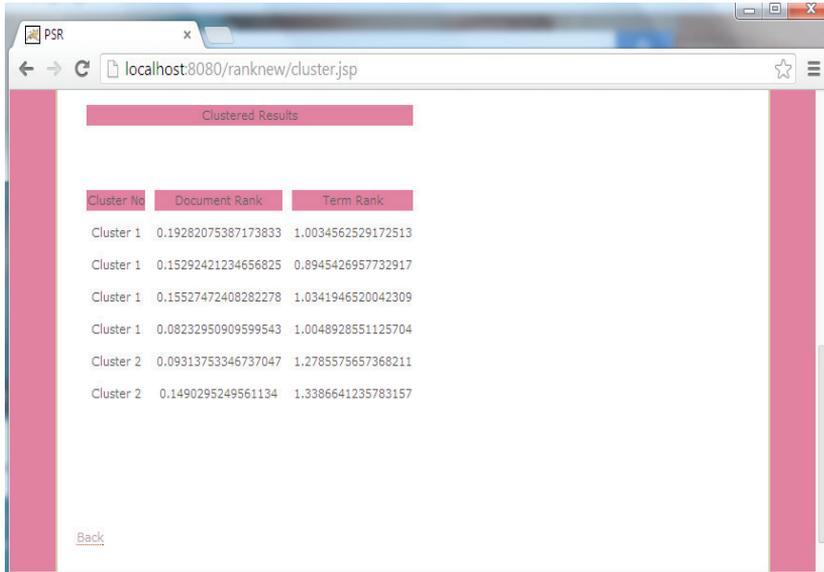
Source: Authors'

Figure 5.: Initial term ranking



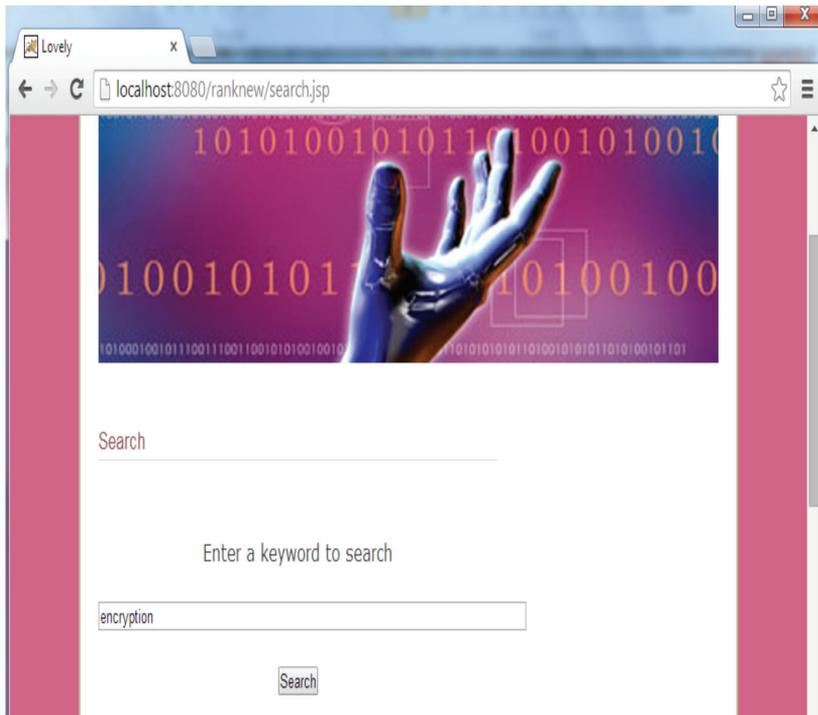
Source: Authors'

Figure 6.: Clustering



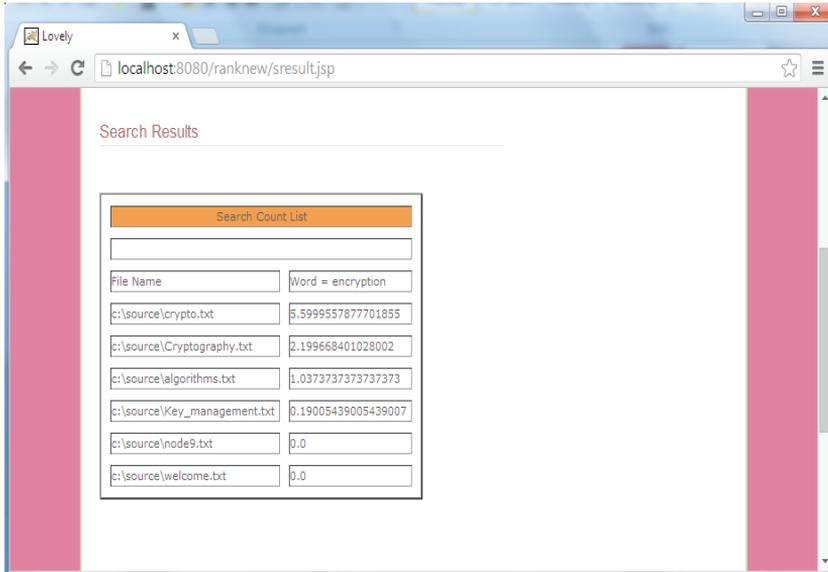
Source: Authors'

Figure 7.: Search engine



Source: Authors'

Figure 8.: Searching results



Source: Authors'

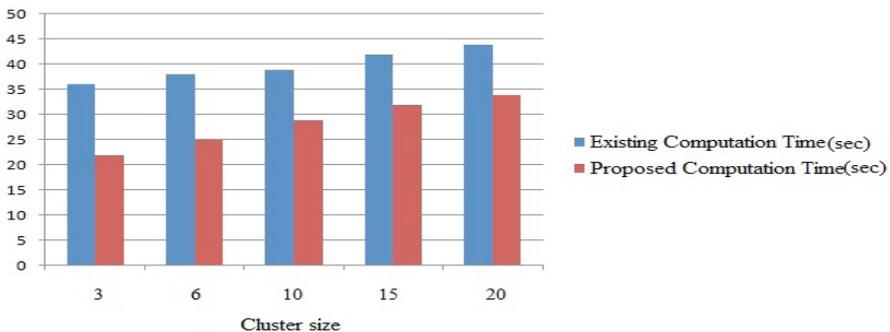
3. PERFORMANCE EVALUATION

Table 1.: Cluster Size and the Computation Time

Clusters size	Existing Time in sec	Proposed Time in sec
3	36	22
6	38	25
10	39	29
15	42	32
20	44	34

Source: Authors'

Figure 9.: Proposed Method Comparison Cluster Size And The Computation Time



Source: Authors'

4. CONCLUSION

This paper, we first define three different ranking functions in a bi-type document graph constructed from the given document set. Based on initial K clusters, ranking is applied separately, which serves as a good measure for each cluster. Sentences then are reassigned to the nearest cluster under the new measure space to improve clustering. As a result, quality of clustering and ranking are mutually enhanced. To further examine how the cluster number influences summarization, we conduct the following additional experiments by varying the cluster number. Given a document set, we let denote the sentence set in the document set, and set in the following way, $K = e * S$. We applied to provide Integrating Clustering and ranking simultaneously terms and sentences and to provide ranking for different word but same meaning and to improve the efficiency of document retrieval.

REFERENCES

- Ahn C.M, Kim D.H, Lee J.H, and Sun P, "Multi-document using weighted similarity between topic and clustering-based non-negative semantic feature," in Proc. APWEB/WAIM Conf., (2007): 60-63
- Barzilay R and Mckeown K.R, "Sentence fusion for multi-document news summarization," *Comput Linguist.*, vol. 31, no. 3, (2005): 297-327
- Cai X.Y, Hong Y, Li W.J, and Ouyang Y, "Simultaneous ranking and clustering of sentences: A reinforcement approach to multi-document summarization," in Proc. 23rd COLING Conf. '10, (2010): 134-142
- Cai X.Y and Li W.J, "A spectral analysis approach to document summarization: Clustering and ranking sentences simultaneously," *Inf. Sci.*, vol. 181, no. 18, (2011): 3816-3827
- Cai X.J and Li W.J, "Mutually reinforced manifold-ranking based relevance propagation model for query-focused multi-document summarization," *IEEE Tran. Audio, Speech, Lang. Process.*, vol. 20, no. 5, (2012): 1597-1607
- Celikyilmaz A and Hakkani-Tur D, "A hybrid hierarchical model for multi-document summarization," in Proc. 48th Annu. Meeting Assoc. Comput. Linguist., (2010): 815-824
- Celikyilmaz A and Hakkani-Tur D, "Discovery of topically coherent sentences for extractive summarization," in Proc. 49th ACL Conf. '11, (2011): 491-499
- Chi Y, Gong Y.H, Li T, Wang D.D, and Zhu S.H, "Integrating clustering and multi-document summarization to improve document understanding," in Proc. 17th CIKM Conf., (2008): 1435-1436
- Filatova E and Hatzivassiloglou V, "Event-based extractive summarization," in Proc. 42nd ACL Conf., (2004): 104-111
- Fisher S and Roark B, "Query-focused summarization by supervised sentence ranking and skewed word distributions," in Proc. DUC'06, 2006.
- Mihalcea R, "Graph-based ranking algorithms for sentence extraction, applied to text summarization," in Proc. 42nd ACL Conf., (2004): 20-24
- Wan X.J and Yang J.W, "Multi-document summarization using cluster-based link analysis," in Proc. 31st SIGIR Conf., (2008): 299-306