ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Combined Cluster Based Ranking for Web Document Using Semantic Similarity

V.Anthoni sahaya balan[1], S.Singaravelan.M.E. [2]

P.G.Scholar, Department of CSE, PSR Engineering College, Sivakasi-626140, Tamilnadu, India[1]

Assistant Professor, Department of CSE, PSR Engineering College, Sivakasi-626140, Tamilnadu, India[2]

ABSTRACT: Multidocument summarization is a set of documents on the same topic, the output is a paragraph length summary. Since documents often cover a number of topic themes with each theme represented by a cluster of highly related sentences, sentence clustering has been explored in the literature in order to provide more informative summaries. An existing cluster-based summarization approach that directly generates clusters first and with ranking next. Ranking distribution of sentences in each cluster should be quite different from each other, which may serve as features of cluster, we propose an integrated approach that overcomes the drawback that we provide ranking for same meaning of different words. 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

I. INTRODUCTION

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 organizes documents into different groups called as clusters, where the documents in each cluster share some common properties according to defined similarity measure. The fast and high quality document clustering algorithms play an important role in helping users to effectively navigate, summarize, and organize the information. Clustering can produce either disjoint or overlapping partitions. In an overlapping partition, it is possible for a document to appear in multiple clusters whereas in disjoint clustering, each document appears in exactly one cluster.

II. PROBLEM STATEMENT

Cluster based summarization approach directly generates cluster with ranking. Ranking distribution of sentences in each cluster should be quite different from each other. In our work to provide ranking for same meaning of different word by using word net tool. While searching the web document to get better results clustering and ranking very much needed. While performing clustering and ranking one by one it least computational process and time consumption.

III. PROPOSED SYSTEM

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

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 motivation of the approach is that, for each sentence cluster, which forms a topic theme, the rank of terms conditional on this topic theme should be very distinct, and quite different from the rank of terms in other topic themes. Therefore, applying either clustering or ranking over the whole document set often leads to incomplete, or sometimes rather biased, analytical results. For example, ranking sentences over the whole document set without considering which clusters they belong to often leads to insignificant results. Alternatively, clustering sentences in one cluster without distinction is meaningless as well. However, combining both functions together may lead to more comprehensible results.
The three main contributions of the paper are:

- 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.
- A reinforcement approach is proposed to tightly integrate ranking and clustering of sentences by exploring term rank distributions over the clusters.
- Thorough experimental studies are conducted to verify the effectiveness and robustness of the proposed approach.

IV. SYSTEM DESIGN

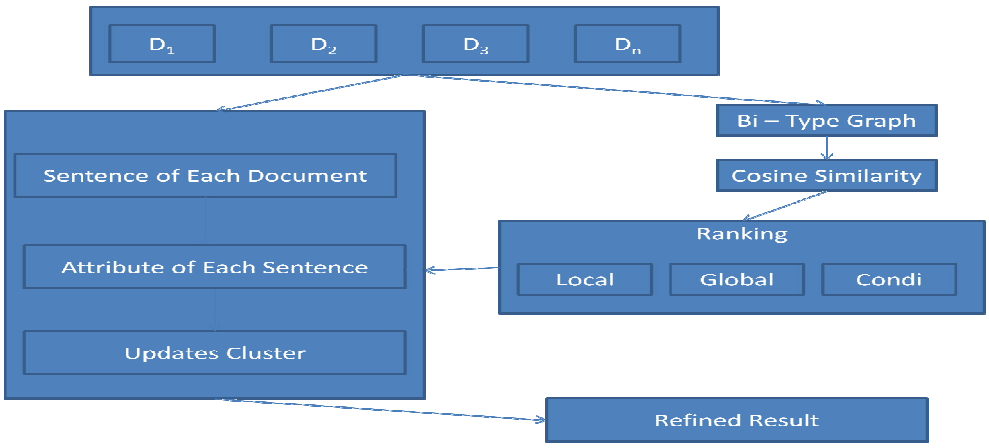

Fig 4.1 Data Flow diagram

V. SYSTEM IMPLEMENTATION

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

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

- Data Preprocessing
- Document Bi-Type Graph
- Ranking
- Sentence Ranking Algorithm

5.1 Data Preprocessing

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, for example:

• 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.

5.2 Document Bi-Type Graph

In this section, we 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={s1,s2,….sn} and the term set T={t1,t2,t3…tn} , i.e.=S U 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.1 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 v1 and other vertices. All the documents are represented in the form of a vector called Term.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

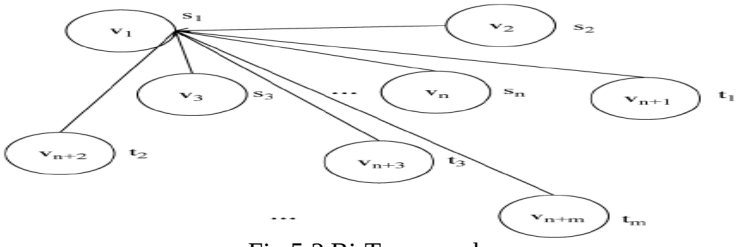

Fig 5.2 Bi-Type graph

5.3 Ranking

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

Frequency-Inverse Document Frequency vector(TF_IDF vector). The Term Frequency and Inverse Document Frequency are calculated as follows:

Term Frequency,

TFdt = freq(d,t)

Inverse Document Frequency,

IDFt = log (|D| / |Dt|)

Where freq (d, t) is the number of occurrences of term t in document d

|D| is the total number of documents

|Dt| is the number of documents containing the term t

Now, the TF_IDF of a term is calculated by,

TF_IDFdt = TFdt X IDFt

The TF_IDF vector of a document will be represented as,

<TF_IDFterm1, TF_IDFterm2, ... , TF_IDFtermn>

And the Rank of the Sentence is defined as

$$r(s_i) = \alpha \cdot \sum_{i=1}^{m} \mathbf{W_{ST}}(i,j) \cdot r(t_j) + (1-\alpha) \sum_{i=1}^{n} \mathbf{W_{SS}}(i,j) \cdot r(s_j), \quad (5.1)$$

And the Rank of the term is defined as

$$r(t_j) = \beta \cdot \sum_{i=1}^{n} \mathbf{W_{TS}}(j,i) \cdot r(s_i) + (1-\beta) \sum_{i=1}^{m} \mathbf{W_{TT}}(i,j) \cdot r(t_i). \quad (5.2)$$

5.4 Sentence Ranking Algorithm

Input: Bi-Type graph

Output: Clusters

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

1. $t \leftarrow 0$;
2. Get the initial partition for S, i.e. C_k^t, $k = 1,2,\ldots K$, calculate cluster centers $\overrightarrow{Center}_{C_k^t}$ accordingly.
3. **For** (t=1; t<IterNum && $\varepsilon > Tre$; t++)
4. Calculate the within-cluster ranking $r_{C_k}(T_{C_k})$, $r_{C_k}(S_{C_k})$ and the conditional ranking $r(s_i \mid C_k)$;
5. Get new attribute $\vec{s}_i$ for each sentence s_i , and new attribute $\overrightarrow{Center}_{C_k^t}$ for each cluster C_k^t;
6. **For** each sentence s_i in S
7. **For** k=1 to K
8. Calculate similarity value $sim(s_i, C_k^t)$
9. **End For**
10. Assign s_i to $C_{k_0}^{t+1}$, $k_0 = \arg\max_k sim(s_i, C_k^t)$
11. **End For**
12. $\delta = \max_k \mid \overrightarrow{Center}_{C_k^{t+1}} - \overrightarrow{Center}_{C_k^t} \mid$
13. $t \leftarrow t+1$
14. **End For**
15. For each sentence s_i in S
16. **For** k=1 to K
17. $$f(s_i) = \sum_{k=1}^{K} \alpha_k \cdot r(s_i \mid C_k)$$
18. **End For**
19. **End For**

VI. EXPERIMENTAL RESULTS

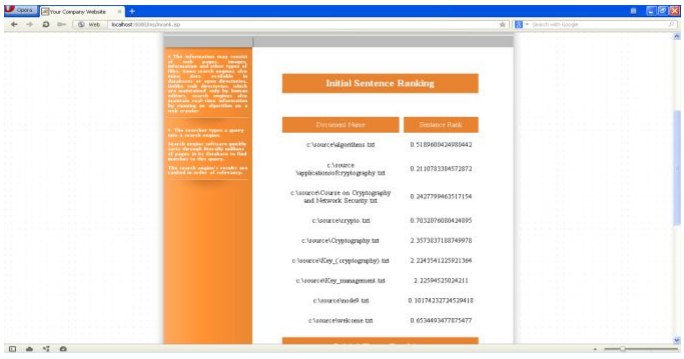

Fig 6.1 INITIAL SENTENCE RANKING

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

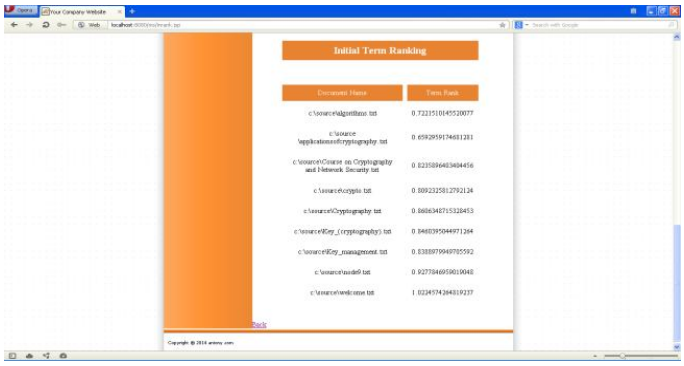

Fig 6.2 INITIAL TERM RANKING

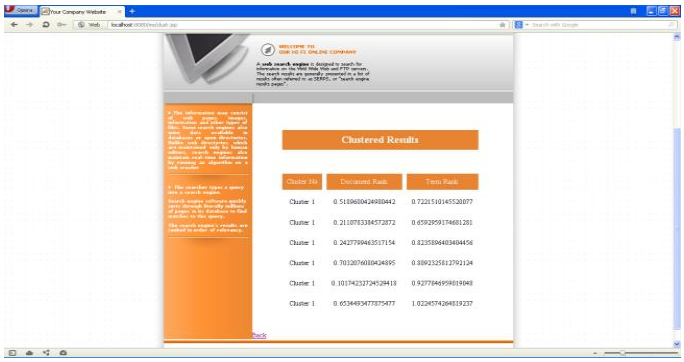

Fig 6.3 CLUSTERING

Table 6.1 Cluster Size and the Computation Time

Clusters	Time
3	0.36
6	0.38
10	0.39
15	0.42
20	0.44

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

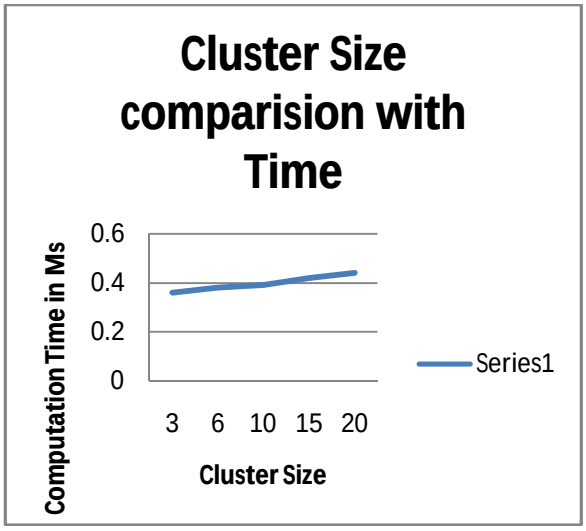

Fig 6.4 CLUSTER SIZE AND THE COMPUTATION TIME

VII. CONCLUSION

In previous experiments, the cluster number is predicted through the eigen values of 1-norm normalized sentence similarity matrix. This number is just the estimated number. The actual number is hard to predict accurately. 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.

$$K = e * S$$

Effectively utilizing multi-faceted associated relationships and distributions of terms and sentences will certainly release the negative impact of the undesired inaccurate clustering results.

REFERENCES

1. L.Antiqueris, O. N. Oliveira, L. F. Costa, and M. G. Nunes, "A complexnetwork approach to text summarization," Inf. Sci., vol. 175, no.5, pp. 297–327, Feb. 2009.
2. R.Barzilay and K. R. Mckeown, "Sentence fusion for multi-document news summarization," Comput Linguist., vol. 31, no. 3, pp. 297–327, 2005.
3. R. Barzilay and L. Lee, "Catching the drift: Probabilistic contentmodels, with applications to generation and summarization," in Proc.HLT-NAACL '04, 2004, pp. 113–120.
4. J. Bilmes, "A Gentle tutorial on the EM algorithm and its applicationto parameter estimation for Gaussian mixture and hiddenMarkov models," Univ. of Berkeley, Berkeley, CA, USA, Tech. Rep.ICSI-TR-97-02, 1997.
5. Xiaoyan Cai and Wenjie Li, "Ranking Through Clustering: An Integrated Approach to Multi-Document Summarization," IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 21, NO. 7, JULY 2013
6. X. Y. Cai, W. J. Li, Y. Ouyang, and Y. Hong, "Simultaneous ranking and clustering of sentences: A reinforcement approach to multi-document summarization," in Proc. 23rd COLING Conf. '10, 2010, pp.134–142
7. Xiaoyan Cai and Wenjie Li "Mutually Reinforced Manifold-Ranking Based Relevance Propagation Model for Query-Focused Multi-Document Summarization," IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 5, JULY 2012 1597.