ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

Multi-document Summarization Based on Cluster

Khanapure V.M[1], Prof. Chirchi V.R[2]

PG Student, Dept. of CNE, College of Engineering, Ambajogai, Maharashtra, India [1]

Assistant Professor, Dept. of CNE, College of Engineering, Ambajogai, Maharashtra, India [2]

ABSTRACT: A **s**ummary can be loosely defined as a text which is produced from one or more texts, that contain a significant portion of the information in the original text(s), and that is no longer than half of the original text(s).The main goal of a summary is to present the main ideas in a document in less space. Multi-document summarization is the process of producing a single summary of a set of related source documents, is relatively new. For handling multiple input document following are the problems1.Recognizing and coping with redundancy 2.Identifying important differences among document and 3.Covering the informative content as much as possible. In this paper, to address these problems, we propose multidoument summarization based on cluster using sentence-level semantic analysis (SLSS), mixture model and symmetric non-negative matrix factorization (SNMF).

Keywords: Semantic similarity, Symmetric non-negative matrix factorization, Multi-document summarization.

I.INTRODUCTION

Multi-document summarization is the process of generating a generic or topic-focused summary by reducing documents in size while retaining the main characteristics of the original documents. Since one of the problems of data overload is caused by the fact that many documents share the same or similar topics, multi-document summarization has attracted much attention in recent years. With the explosive increase of documents on the Internet, there are various summarization applications are used. For example, the informative snippets generation in web search can assist users in further exploring, and in a question-based system summary is required to provide information asked in the question. Another example is summaries for news groups in news services, which provide users to better understand the news articles in the group.

For handling multiple input document following are the problems1.Recognizing and coping with redundancy 2.Identifying important differences among document and 3.covering the informative content as much as possible. In this paper, to address these problems, we propose multi-document summarization based on cluster using sentence-level semantic analysis (SLSS), mixture model and symmetric non-negative matrix factorization (SNMF).

Since SLSS can better capture the relationships between sentences in a semantic manner, we use it to construct the sentence similarity matrix. Based on the similarity matrix, we perform the proposed mixture language model and SNMF algorithm to cluster the sentences. Finally we select the most informative sentences in each cluster considering both internal and external information.

II.RELATED WORK

Multiple document summarizations have been widely studied recently. The summary can be either generic or query specific. In a generic summary generation, the important sentences from the document are extracted and the sentences so extracted are arranged in the appropriate order. In a query specific summary generation, the sentences are scored based on the query given by the user. The highest scored sentences are extracted and presented to the user as a summary. Following are the two broad level classifications of text summarization techniques.

Extractive summarization and abstractive summarization. Extractive summarization usually ranks the sentences in the documents according to their scores calculated by a set of predefined features, such as term frequency inverse sentence frequency (TF-ISF) [20], sentence or term position [20], and number of keywords. Abstractive summarization involves information fusion, sentence compression and reformulation. In this paper, we study sentence-based extractive summarization. Gong et al. [22] propose a method using latent semantic analysis (LSA) to select highly ranked sentences for summarization. Proposes a maximal marginal relevance (MMR) method to summarize documents based on the cosine similarity between a query and a sentence and also the sentence and previously selected sentences. MMR

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

method tends to remove redundancy, which is controlled by a parameterized model which actually can be automatically learned. Other methods include NMF-based topic specific summarization, CRF-based summarization, and hidden Markov model (HMM) based method. In addition, some graph-ranking based methods are also proposed [22]. Most of these methods ignore the dependency syntax in the sentence level and just focus on the keyword co-occurrence. Thus the hidden relationships between sentences need to be further discovered. The method proposed in group sentences based on the semantic role analysis, however the work does not make full use of clustering algorithms. In our work, we propose a new framework based on sentence-level semantic analysis (SLSS), mixture language model and symmetric non-negative matrix factorization (SNMF). SLSS can better capture the relationships between sentences in a semantic manner, mixture language model is used to measure the similarity between documents and SNMF can factorize the similarity matrix to obtain meaningful groups of sentences.

III.PROPOSED METHOD

3.1 Overview: Figure 1 demonstrates the framework of our proposed approach. Given a set of documents which need to be summarized, first of all, we clean these documents by removing formatting characters. In the similarity matrix construction phase, we decompose the set of documents into sentences, and then parse each sentence into frame(s) using a semantic role parser. Pair wise sentence semantic similarity is calculated based on both the semantic role analysis [11] and word relation discovery using WordNet [20]. Section 3.2 will describe this phase in detail. Once we have the pairwise sentence similarity matrix, we perform the symmetric matrix factorization to group these sentences into clusters in the second phase. Full explanations of the proposed SNMF algorithm will be presented in section 3.3. Finally, in each cluster, we identify the most semantically important sentence using a measure combining the internal information (e.g., the computed similarity between sentences) and the external information (e.g., the given topic information). Section 3.4 will discuss the sentence selection phase in detail. These selected sentences finally form the summary.

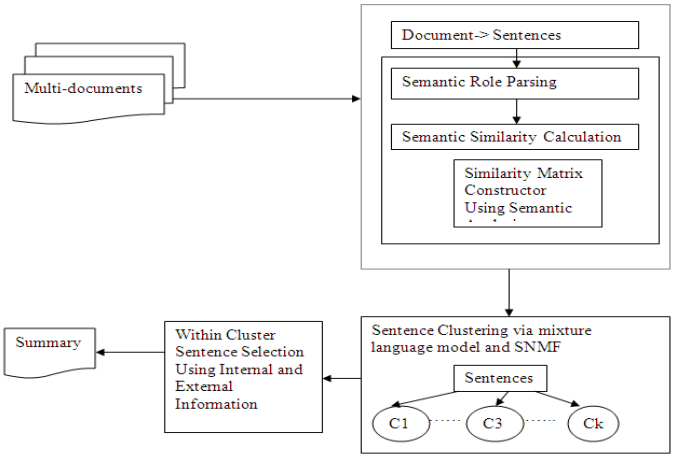

Figure 1: Proposed method overview.

3.2 Semantic Similarity Matrix Construction: After removing stemming and stopping words, we trunk the documents in the same topic into sentences. Simple Word-matching types of similarity such as cosine can not faithfully capture the content similarity. Also the sparseness of words between similar concepts make the similarity metric uneven. Thus, we perform semantic role analysis on sentences and propose a method to calculate the semantic similarity between any pair of sentences.

3.2.1 Sentence-level semantic analysis (SLSS): A semantic role is defined as "a description of the relationship that plays with respect to the verb in the sentence". Each verb in the sentences is labelled with Argument and the verb which is labelled is called "frame". Input to the SLSS algorithm is sentences Si and Sj. Assign labels to each verb in the sentences using Semantic role labler.After assigning label calculate the common semantic roles WordNet. Then to find role similarity between Tm(ri) and Tn(ri) as

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

$$rsim(Tm(ri), Tn(ri)) = \frac{\sum_j tsim(t_{ij}^m, ri)}{|Tn(ri)|} \quad (1)$$

Then, calculate the frame similarity between *fm* and *fn* is

$$fsim(f_{m,} f_n) = \frac{\sum_{i=1}^{k} rsim(T_m(r_i), T_n(ri))}{K} \quad (2)$$

Therefore, the semantic similarity between Si and Sj can be calculated As follows:

$$Sim(S_{i,} S_j) = \max_{f_m \in s_{i,} f_n \in s} fsim(f_{m,} f_n) \quad (3)$$

Where similarity score is between zero and one.

3.3 Mixture Language Model and Symmetric nonnegative matrix factorization: Once we obtain the similarity matrix of the relevant cases, clustering algorithms need to be performed to group these cases into clusters.

3.3.1 Mixture Language Model: Mixture language model [19] is used to measure the similarity between documents while filtering out the general and common information from the request. Mixture model measure is based on a novel view of how relevant documents are generated. We can also view it as a language model with a smoothing algorithm designed specifically for our task.

.

Algorithm Mixture Model( )

1. Input : number of data points n. n*n similarity matrix w
2. Initialization: double r,prob,x,y=0
3. Compute thetaE,thetaT,thetaD

```
if(synsets.length > 0 ||GEWords.contains(alphaStr))
                thetadE.add(alphaStr);
else if(query.contains(alphaStr))
                thetadT.add(alphaStr);
else
                thetadD.add(alphaStr);
```

4. Calculate Probability prob

```
prob = (lmdaE*(tfwiE/tfwjE)) + (lmdaT*(tfwiT/tfwjT)) +
                (lmdaD*(tfwiD/tfwjD));
```

5. Compute relevance r

```
r-=x*Math.log(y/x);
```

6. Output: r

Figure 2:The Mixture Model Algorithm

Figure 2 shows the Mixture language model in which input is sentence similarity matrix w. This document is mixture of three language models: A General English language model thetaE, a user-specific Request Model thetaT, and a document context Model thetaD. Each word wi in the document is generated by each of the three language models with probability lmdaE, lmdaT and lmdaD respectively. Then calculate probability and relevance score so mixture language model is used to measure the similarity between documents. By using mixture model, the effect of the words that occur frequently in the request or in general English on the similarity calculation is naturally reduced.

3.3.2 SNMF: We propose a new multi-document summarization framework based on sentence-level semantic analysis (SLSS) and symmetric non-negative matrix factorization(SNMF). SLSS is able to capture the semantic relationships between sentences and SNMF can divide the sentences into groups for extraction. It has been shown that SNMF is equivalent to kernel K-means clustering and is a special case of trifactor NMF. Another important property is that the

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

simple SNMF is equivalent to the sophisticated normalized cut spectral clustering. Spectral clustering is a principled and effective approach for solving normalized cuts [38]. These results demonstrate the clustering ability of SNMF.

Algorithm SNMF( )

1. Input Sentence pairwise similarity matrix W
2. Initialize H,H=1
3. Compute the norm of Matrix

$$\min_{H\geq 0} J = \| W - HH^T \|^2$$

4. Check the KKT condition

If($(-4WH + 4HH^T H)_{ij} H_{ij} = 0$)

$$H_{ij} \leftarrow H_{ij} - \in_{ij}$$

Else

$$H_{ij} \leftarrow \frac{1}{2}[H_{ij}(1+\frac{(WH)_{ij}}{(HH^T H)_{ij}}]$$

5. Output H

Figure 3: The SNMF Algorithm.

3.4 Within-Cluster Sentence Selection: After grouping the sentences into clusters by the SNMF algorithm, in each cluster, we rank the sentences based on the sentence score calculation . The score of a sentence measures how important a sentence is to be included in the summary.

Algorithm Multidocument_Summarization( )

1. Input : Cluster document
2. Initialize : lmd=0.7
3. Compute f1sim
   F1sim=f1sim+snmf.w[x][y];
   F1sim=f1sim /double(k-1);
4. Compute f2sim
   F2sim=Sim(Si,request)
5. Calculate Score
   Score=(lmd*f1sim)+((1-lmd)*f2sim)

Figure 4: The Multidocument Summarization Algorithm

where F1(Si) measures the average similarity score between sentence Si and all the other sentences in the cluster Ck, and N is the number of sentences in Ck. F2(Si) represents the similarity between sentence Si and the given topic T. λ is the weight parameter, which is set to 0.7 empirically.

3.5 Module :

1. Pre-processing of customer request and past cases and Sentence-level semantic similarity calculation.
2. Top-ranking case clustering using mixture model and SNMF algorithm.
3. Multidocument summarization for each case cluster.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

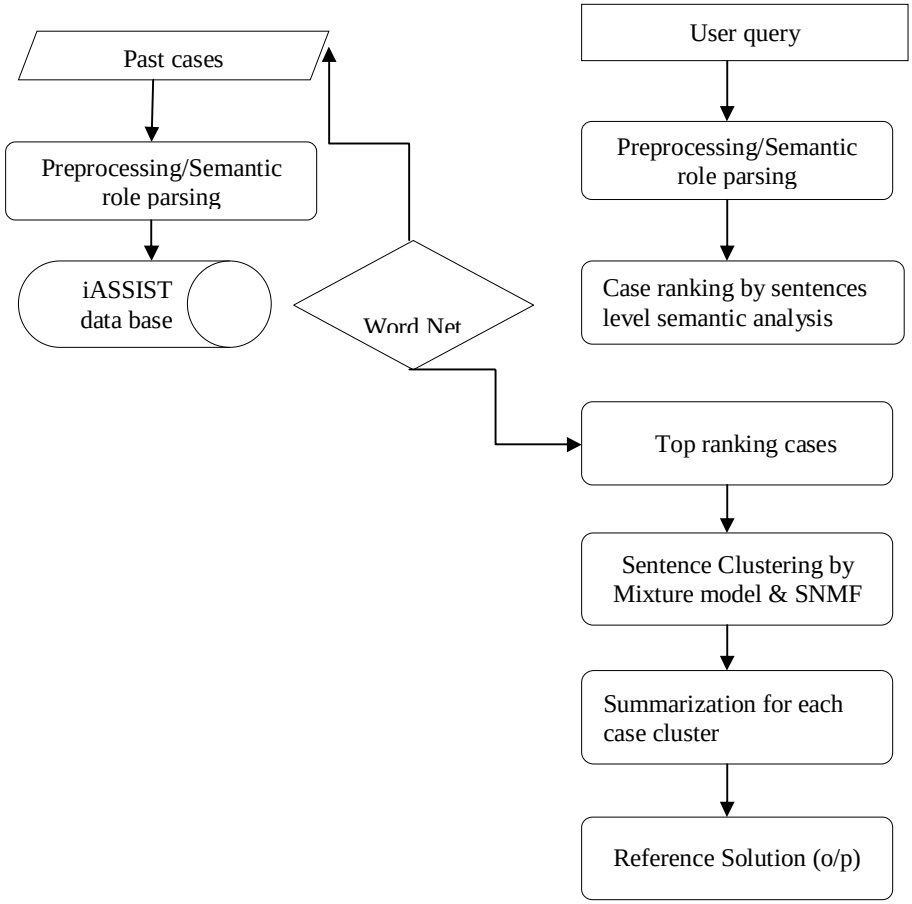

Figure 5: State Flow Diagram of Multi-document summarization.

IV.EXPERIMENTAL RESULT

To improve the usability of the system, we proposed sentence-level semantic analysis approach and SNMF clustering algorithm can be naturally applied to the summarization task to address the aforementioned issues.

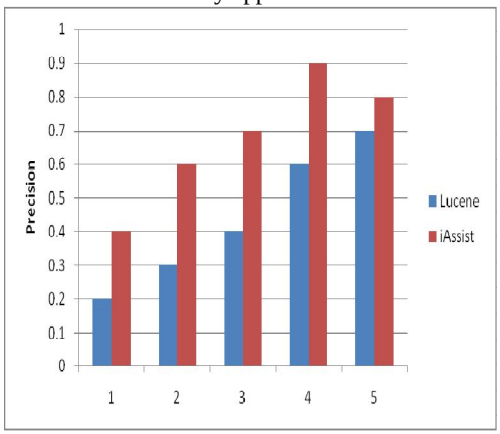

Figure 6: Precision of the retrieved cases.

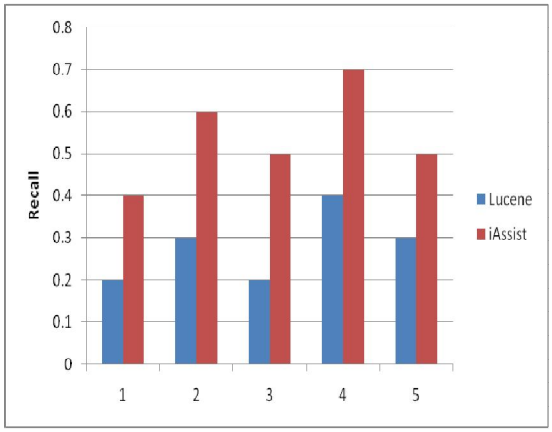

Figure 7: Recall of the retrieved cases.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

4.1 Case comparison: In this set of experiments, we randomly select five questions from different categories and manually label the related cases for each question. Then, we examine the top 10 retrieved cases by keyword-based Lucerne and our proposed system, respectively. Figure 6 and 7 show the average precision of the two methods. In figure 6, the high precision of multidocument summarization demonstrates that the semantic similarity calculation can better capture the meanings of the requests and case documents. In figure 7, we only look at the top 10 retrieved cases while some of the cases may have more than 20 relevant cases, the recall is also reasonable and acceptable.

4.2 Result Analysis: In this section, we compare our proposed request-focused case-ranking results and Apache Lucene, which is one of the most popular keyword-based text-ranking engines.

Example1. Can I update my iPod music collection from more than one computer: The full representation of the abstract arguments of an illustrative example is shown in Table I. Table III shows the top-ranking case samples retrieved by Lucene and Multidocument summarization . For ranking results, we find that Lucene takes the word "iPod", "Computer" as the keyword and return many cases related to them as the search result in list format as shown in figure 8.Obviously they are not what the customer want.

TABLE I
REPRESENTATION OF ARGUMENTS OF AN ILLUSTRATIVE EXAMPLE

Can	-	S-AM-MOD	from	-	B-AM-MNR
I	-	S-A0	more	-	I-AM-MNR
update	-	S-V	than	-	I-AM-MNR
my	-	B-A1	one	-	I-AM-MNR
iPod	-	I-A1	computer	-	E-AM- MNR
music	-	I-A1			
collection	-	E-A1			

Example:
Sentence: Can I update my iPod music collection from more than one computer
Label: Can[S-AM-MOD] I[S-A0] update[S-V] my[B-A1] iPod [I-A1] music[I-A1] collection [E-A1] from[B-AM-MNR] more[I-AM-MNR] than[I-AM-MNR] one[I-AM-MNR] computer[E-AM-MNR]

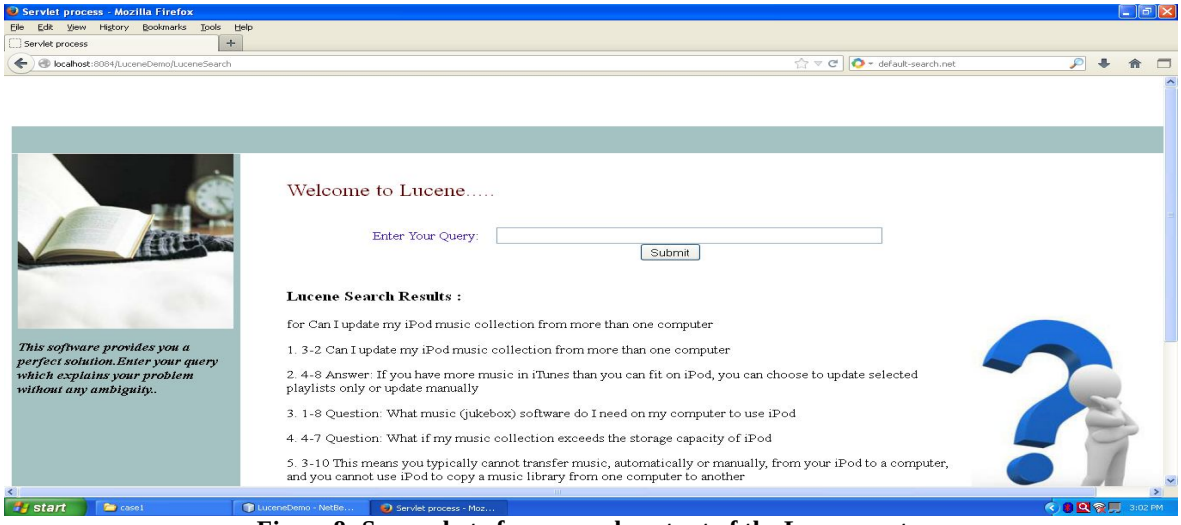

Figure 8: Screenshot of an example output of the Lucene system.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

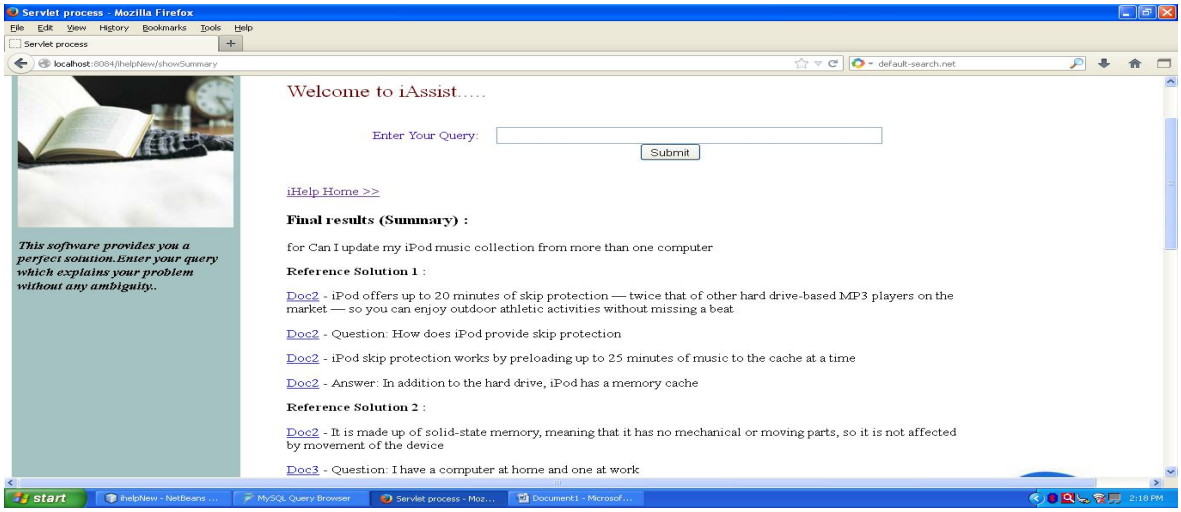

Figure 9 : Screenshot of an example output of the Multi-document summarization

In our proposed system, Multi-document summarization provides the semantic meaning of request. We first calculate sentence-sentence similarities using semantic psychoanalysis and construct the similarity matrix. Then mixture language model and symmetric matrix factorization is used to group sentences into clusters for extraction. Finally, the informative sentences are selected from each group to form the summary.

Table II
Top Ranking Score

Top similarity scores:
0.9, 0.5, 0.3333333333333333, 0.2, 0.2, 0.2, 0.16666666666666666, 0.125, 0.125

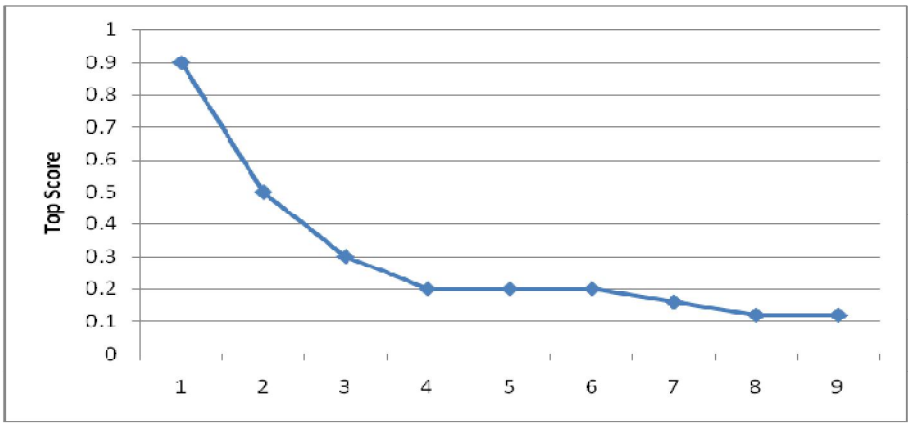

Figure 10: Top ranking Score

The above graph shows the top score of our proposed system which is based on the sentence level semantic analysis. Existing system i.e Lucene system which is based on the keyword matching based ranking scheme for case retrieval and results will be in a list format. Our proposed system we search and rank the existing cases according to their relevance to users' requests in a semantic way i.e. high top score gives the better result.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

TABLE III
TOP RANKING CASE SAMPLES BY Lucene and Multi-document summarization.

Request	Can I update my iPod music collection from more than one computer
Lucene	Top Ranking Cases iPod is compatible with computers running on Mac OS X and PCs running on Windows 2000 or Windows XP
Multidocument summarization	Top Ranking Cases Yes. When you first connect iPod to your computer, iPod find that computer as its "home" computer. Each time you connect, iPod downloads the music library stored on it. This means that you cannot transfer music, automatically or manully, from your iPod to computer, and you cannot use iPod to copy a music library from one computer to another.

V. CONCLUSION

To improve the usability of the system, we perform multidocument summarization to generate a brief summary for each case cluster. In this paper we search and rank the existing cases according to their relevance to users' requests in a semantic way and we provide a better result representation by grouping and summarizing the retrieved past cases to make the system fully functional and usable. The high performance of multidoument summarization based on cluster using sentence-level semantic analysis (SLSS), mixture model and symmetric non-negative matrix factorization (SNMF).

REFERENCES

1. B. K. Giamanco, Customer service: The importance of quality customer service. [Online]. Available: http://www.ustomerservicetrainingcenter.com
2. S. Agrawal, S. Chaudhuri, G. Das, and A. Gionis, "Automate ranking of database query results," in *Proc. CIDR*, 2003, pp. 888–899.
3. D.Wang, S. Zhu, T. Li, and C. Ding, "Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization," in *Proc. SIGIR*, 2008, pp. 307–314.
4. K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is nearest neighbor meaningful?" in *Proc. ICDT*, 1999, pp. 217–235.
5. D. Radev, E. Hovy, and K. McKeown, "Introduction to the special issue on summarization," *Comput. Linguist.*, vol. 28, no. 4, pp. 399–408, Dec. 2002.
6. D. W. Aha, D. Mcsherry, and Q. Yang, "Advances in conversational case-based reasoning," *Knowl. Eng. Rev.*, vol. 20, no. 3, pp. 247–254, Sep. 2005.
7. R. Agrawal, R. Rantzau, and E. Terzi, "Context-sensitive ranking", in *Proc. SIGMOD*, 2006, pp. 383–394.
8. A. Leuski and J. Allan, "Improving interactive retrieval by combining ranked list and clustering," in *Proc. RIAO*, 2000, pp. 665–681.
9. X. Liu, Y. Gong, W. Xu, and S. Zhu, "Document clustering with cluster refinement and model selection capabilities," in *Proc. SIGIR*, 2002, pp. 191–198.
10. R. Collobert and J.Weston, "Fast semantic extraction using a novel neural network architecture," in *Proc. ACL*, 2007, pp. 560–567.
11. M. Palmer, P. Kingsbury, and D. Gildea, "The proposition bank: An annotated corpus of semantic roles," *Comput. Linguist.*, vol. 31, no. 1, pp. 71–106, Mar. 2005.
12. C. Fellbaum, "*WordNet: An Electronic Lexical Database*", in Cambridge, MA: MIT Press, 1998.
13. X. Liu, Y. Gong, W. Xu, and S. Zhu, "Document clustering with cluster refinement and model selection capabilities," in Proc. SIGIR, 2002, pp. 191–198.
14. D. Radev, E. Hovy, and K. McKeown, "Introduction to the special issue on summarization," Comput. Linguist., vol. 28, no. 4, pp. 399–408, Dec. 2002.
15. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 8, Aug. 2000.
16. D.W.Aha, D. Mcsherry, and Q. Yang, "Advances in conversational case-based reasoning," Knowl. Eng. Sep. 2005.
17. D. Bridge, M. H. Goker, L. Mcginty, and B. Smyth, "Case-based recommender systems," *Knowl. Eng. Rev.*, vol. 20, no. 3, pp. 315–320,Sep. 2005
18. Dingding Wang, Tao Li, Shenghuo Zhu, and Yihon Gong,"iHelp: An Intelligent Online Helpdesk System" in IEEE TRANSACTIONS 2011
19. Y. Zhang, J. Callan, and T. Minka, "Novelty and redundancy detection in adaptive filtering," in *Proc. SIGIR*, 2002, pp. 81–88.
20. Y. Gong and X. Liu. "Generic text summarization using relevance measure and latent semantic analysis", in SIGIR 2001.
21. R. Mihalcea and P. Tarau" A language independent algorithm for single and multiple document summarizations", in IJCNLP 2005.