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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Query Focused Summarization using TF-IDF, K-Mean Clustering and HMM

Sonali Gandhi, Praveen Sharma

M-Tech(Pursuing), Dept. of CSE, NGF college of Engineering, Haryana under the Affiliation of Maharshi Dayanand University at Rohtak, Haryana, India

Assistant Professor, Dept. of CSE, NGF college of Engineering, Haryana under the Affiliation of Maharshi Dayanand University, Rohtak, Haryana - India

ABSTRACT: Under the scheme the proposed approach provide summary using HMM by forming the K-Mean clustering with meaningful words and relationship using TF-IDF giving more information related to document. This will provide better summary as compared to existing algorithms. The proposed approach we have built is a cluster-based summarization system with the knowledge coming from the clustering. The knowledge is composed of not only in recognizing important phrases in the document, but also in recognizing the relationships and the relationship types that exist between them. This extracted knowledge is represented in the form of a hierarchical. Even without the summary, just looking at the nodes and relationships in the thematic graph gives us a rough idea about what the document is taking about. A summary however gives us the actual details. This method makes a lot of sense. Improvements and further experimentation would most definitely make the existing system more reliable than it is now. The proposed approach can be extended with automatic generation of summarized data based on aspect oriented model which will give more efficient result in document summarization and will increase pre-processing speed and accuracy.

KEYWORDS: Term Frequency – Inverse Document Frequency (TF-IDF), Machine Learning (ML), Web Mining, K-Mean Clustering, Hidden Markov Model (HMM), Information Retrieval.

I. INTRODUCTION

Information retrieval (IR) systems provide users with a vast amount of reference material. Along with this tremendous access comes the challenge of effectively presenting a user with relevant information in response to a query. When using an IR engine to search through electronic resources, simple queries often return too many documents and many are not relevant to the intended search. For instance, there are several million documents on the World Wide Web pertaining to "Michael Jordan." Most of these concern the basketball star, so it is difficult to find information about the television personality, the jazz musician, the mathematician, or the many others who share that name. It would be useful to have a system that could overcome this limitation. One approach is to cluster the documents after retrieval and present a synopsis of each cluster so that a user can choose clusters of interest. This is the motivation for our Query, Cluster, Summarize system, which performs the following tasks in response to a query: retrieves relevant documents, separates the retrieved documents into clusters by topic, and creates a summary for each cluster. Our implementation of the Query, Cluster, Summarize system partitions the code into portable modules, making it easy to experiment with different methods for handling the three main tasks listed above. In our current implementation of the Query, Cluster, Summarize system, we use existing software libraries for each task. Throughout this paper, we discuss our choices for each of the modules used, but note that it is possible to exchange individual modules with other existing methods. Previous work on using a combination of clustering and summarization to improve IR is summarized. However existing IR systems employing this combination, Query, Cluster, Summarize system most resembles the NewsInEssence system [25] in that both systems can produce multi document summaries from document sets clustered by topic. However, NewsInEssence is designed for IR from HTML-linked document sets while Query, Cluster, Summarize system has been designed for IR from generic document sets. Another system that leverages clustering and summarization for information organization similarly to Query, Cluster, Summarize is the

DOI: 10.15680/IJIRCCE.2017. 0504155

Columbia Newsblaster system [21]. Newsblaster, like NewsInEssence, is a web-based system which crawls news websites and then clusters and summarizes the news stories, but it does not currently accept queries. Recently, 9 the value of summarization to users in IR has been demonstrated in [20], where a study showed increases in user recall of retrieved information when clustering and summarization were included in the output of the IR system based on HMM, TF-IDF and K-Means Clustering.

II. LITERATURE REVIEW

R. Baeza-Yates, C. Hurtado, and M. Mendoza [6] suggests that, the search engine gives the list of related results. These results are based on the previously searched queries or such technique can be used to tune or redirect the user. In this method the clustering algorithm is used. The clustering is done on the basis of previously fired queries. It clusters the semantically similar queries. It does not only give the clustered data but it also ranks the suggested list of result. The ranking is done on the basis of two conditions, 1. Similarity of queries to the input query 2. Observation that measures the attention of the user attracted towards the result of the query. The combination of both these conditions measures the user interests. In the given algorithm, query session is considered for giving the result. The query session also considers the rank of clicked URL. The relevance ranking is measured by using two components similarity of query and support of query.

Harshada P. Bhambure, MandarMokashi[9] discuses that user search goals for a query by clustering feedback sessions. For that, we use a concept of pseudo document, which is the revised version of feedback session. At the end, we cluster these pseudo-documents to infer user search goals and represent them with some keywords. Since the evaluation of clustering is also an important problem, we used evaluation criterion classified average precision (CAP) to evaluate the performance of the restructured web search results. The clustering is done by bisecting k means where in the existing system it is done by k means clustering. The new algorithm increases the efficiency of result. After the segmented result formation, the result in the every segment is reorganized as per number of clicks of URLs. The link which is clicked more number of times will appear at first location in the segment. This reduces the time requirement for searching.

DasariAmarendra, KavetiKiranKumar[10] suggest that user's information needs due to the use of short queries with uncertain terms. thus to get the best results it is necessary to capture different user search goals. These user goals are nothing but information on different aspects of a query that different users want to obtain. The judgment and analysis of user search goals can be improved by the relevant result obtained from search engine and user's feedback. Here, feedback sessions are used to discover different user search goals based on series of both clicked and unclicked URL's. The pseudo-documents are generated to better represent feedback sessions which can reflect the information need of user. With this the original search results are restructured and to evaluate the performance of restructured search results, classified average precision (CAP) is used. This evaluation is used as feedback to select the optimal user search goals.

BhaveshPandya, CharmiChaniyara, DarshanSanghavi, KrutarthMajithia[11] suggest that ambiguous query, different users may have different search goals when they submit it to a search engine. The inference and analysis of user search goals can be very useful in improving search engine relevance and user experience. In this propose a novel approach to infer user search goals by analysing search engine query logs a framework to discover different user search goals for a query by clustering the proposed feedback sessions. Feedback sessions are constructed from user click through logs and can efficiently reflect the information needs of users a novel approach to generate pseudo-documents to better represent the feedback sessions for clustering. Classified Average Precision (CAP) to evaluate the performance of inferring user search goals. Experimental results are presented using user click through logs from a commercial search engine to validate the effectiveness.

III. PROPOSED WORK & PSEUDO CODE

To achieve the desired results proposed in scheme the following steps are as under:-

Step 1: **Forming K-Means Clustering**

1. Begin with n clusters, each containing one object and we will number the clusters 1 through n.

2. Compute the between-cluster distance D(r, s) as the between-object distance of the two objects in r and s respectively, r, s =1, 2, ..., n. Let the square matrix D = (D(r, s)). If the objects are represented by quantitative vectors we can use Euclidean distance.
3. Next, find the most similar pair of clusters r and s, such that the distance, D(r, s), is minimum among all the pairwise distances.
4. Merge r and s to a new cluster t and compute the between-cluster distance D(t, k) for any existing cluster $k \neq r, s$. Once the distances are obtained, delete the rows and columns corresponding to the old cluster r and s in the D matrix, because r and s do not exist anymore. Then add a new row and column in D corresponding to cluster t.
5. Repeat Step 3 a total of n − 1 times until there is only one cluster left.

Inputs : Data: $X := (x1, x2, . . . , x`) \subset IR^n$, Number of classes: k
Initialization: Choose random centers c1 . . . , ck
Solution: for i = 1, . . . , k do
$Ci = \{x \in X|i = \arg \min 1 \leq j \leq k \ \|\ cj - xk\ 2\}$
for i = 1, . . . , k do
$ci = \arg \min_{z\in R} n\ P\ x \in Ci\ k\ ||z – x||\ k\ 2$

Step 2: **Frequency Generation using TF-IDF**

1. TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)
2. IDF(t) = log_e(Total number of documents / Number of documents with term t in it).
3. Value = TF * IDF

Step 3: **HMM Integration**

```
for i = 1 to m do {Initialization}
$\gamma 0(i) = \pi j$
$\delta 0(i) = 0$
Pi = new node
Pi .i = 0
Pi .j = i
L.Push(Pi)
end for
for j = 1 to n do
for i = 1 to m do
. $\gamma j (i) = \max m\ k=1\ \gamma j-1(k)tk(i)ei(Xj )$
$\delta j (i) = \arg \max m\ k=1\ \gamma j-1(k)tk(i)ei(Xj )$
Ni = new node {Add new leaf nodes}
Ni .i = j
Ni .j = i
Ni .parent = $P\delta j (i)$
L.P ush(Ni)
end for
Compress(L)
if L.root changed then
Partial output(Trace-back($\delta$, (L.root).i, (L.root).j))
P = N
end if
end for
sn = arg maxm k=1 $\gamma n(k)$
Partial output(Trace-back($\delta$, n, sn)) {Trace back the last part}
return
```

IV. SIMULATION RESULTS

For experiment in the scheme 100 XML documents and 10 queries are raised. The experiment was done using c# programming language and NUnit standard package.. To store intermediate and final result Microsoft sql-server is used. Performance Measurement Parameter: In this scheme, two performance parameters are defined to evaluate the proposed approach. These two performance parameters are listed as follows. (Precision): It is the fraction of retrieved documents that are relevant.

Precession=#relevant_items_retrieved /#retrieved_items

Recall(R) is the fraction of relevant documents that are retrieved

Recall=relevant_items_retrieved /relevant_items

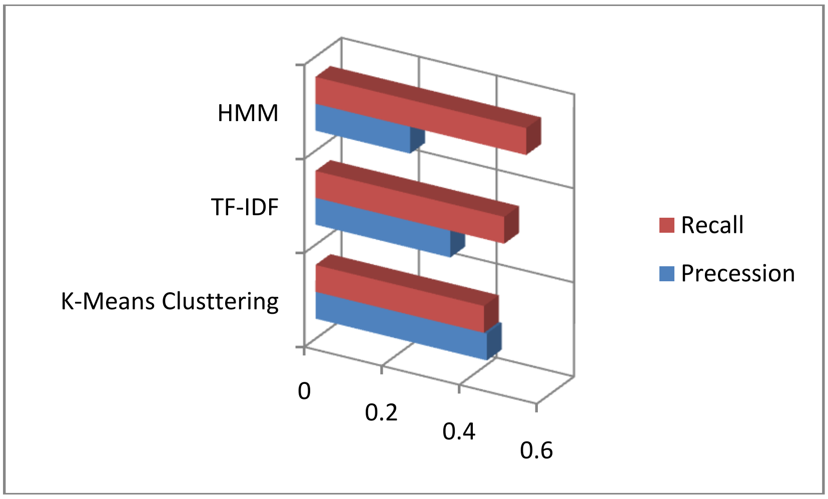
All the stages of the proposed model has been implemented and executed by taking 10 queries and on 200 no of documents. **TABLE I** contains comparison result. The result shows that precision and recall value of the proposed work is increased. This means proposed model is more efficient to retrieve the relevant documents:

Model	Precession	Recall
K-Means Clustering	0.44222	0.434534
TF-IDF	0.34788	0.486777
HMM	0.24435	0.543444

TABLE 1

Graphical Comparison of Result

The graph 1 shows scheme proposed model works in as under with respective results gained using precession recall:



Graph 1: Bar Chart depicting the precession and recall threshold on KMC, TF-IDF & HMM

V. CONCLUSION AND FUTURE WORK

Taxonomically one can distinguish among the following types of summaries: Extractive/non-extractive, generic/query-based, single-document/multi-document, and monolingual/multilingual/cross lingual. Most existing summarizers work in an extractive fashion, selecting portions of the input documents (e.g. sentences) that are believed to be more salient. Non-extractive summarization includes dynamic reformulation of the extracted content, involving a deeper understanding of the input text, and is therefore limited to small domains. Query-based summaries are produced in reference to a user query (e.g. summarize a document about an international summit focusing only on the issues related to the environment) while generic summaries attempt to identify salient information in text without the context of a query which cannot be dealt with. Consequently the query based summarization is best into its aspects as to be dealt with K-Means, TF-IDF and HMM for accurate and effective summarization results. However, for the future work the same can be implemented using hadoop or parallel computing with fully automated scenarios with better performance on the part of the time and speed using map-reduce and pre-processing step can be omitted as such.

REFERENCES

1. Mele, " Web Usage Mining for Enhancing Search –Result Delivery and Helping Users to Find Interesting Web Content," ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '13), pp. 765-769, 2013
2. P. Sudhakar, G. Poonkuzhali, R. Kishor Kumar, "Content Based Ranking for Search Engines," Proc. International Multi Conference of Engineers and Computer Scientists (IMECS 12), 2012.
3. H. Chen and S. Dumais, "Bringing Order to the Web: Automatically Categorizing Search Results," Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI '00), pp. 145-152, 2000.
4 X. Wang and C.-X Zhai, "Learn from Web Search Logs to Organize Search Results," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007.
5. H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma, "Learning to Cluster Web Search Results," Proc. 27th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '04), pp. 210-217, 2004.
6. R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Conf. Current Trends in Database Technology (EDBT '04), pp. 588-596, and 2004.
7. H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, "Context-Aware Query Suggestion by Mining Click Through," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '08), pp. 875883, 2008.
8. Harshada P. Bhambure, MandarMokashi, "Inferring User Search Goals Using Feedback Session" Conf. Research paper, pp.2319-7064 and 2013.
9. DasariAmarendra, KavetiKiran Kumar, "Inferring User Search Goals with Feedback Sessions using K-means clustering algorithm",Volume 2, Issue 11,pp. 780-784,November-2015.
10. BhaveshPandya et al., "A New Algorithm for Inferring User Search Goals with Feedback Sessions ", Int. Journal of Engineering Research and Applications www.ijera.com ISSN: 2248-9622, Vol. 5, Issue 8, (Part - 2), pp.30-33,August 2015.
11. Fang Chen, Kesong Han and Guilin Chen, "An Approach to Sentence Selection Based Text Summarization", In the Proceedings of IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, Volume: l,pp.489-493,2002.
12. Ben Hachey, "Multi-Document Summarization Using Generic Relation Extraction", Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 420-429, 2009.
13. Alok Ranjan Pal, Projjwal Kumar Maiti and Diganta Saha, "An Approach To Automatic Text Summarization Using Simplified Lesk Algorithm And WordNet", International Computer Modelling (IJCTCM)Journal of Control Theory and Vol.3, No.4/5, September 2013
14. A.R.Kulkarni,S.S.Apte, an automatic text summarization using lexical cohesion and correlation of sentences, International Journal of Research in Engineering and Technology
15. A.Kogilavani and Dr.P.Balasubramani, "Clustering and Feature Specific Sentence Extraction Based Summarization of Multiple Documents", International Journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010.
16. J. T. Giles, L. Wo, and M. W. Berry. GTP (General Text Parser) software for text mining. In H. Bozdogan, editor, Statistical Data Mining and Knowledge Discovery, pages 455–471. CRC Press, Boca Raton, 2003.
17. T. G. Kolda and D. P. O'Leary. A semidiscrete matrix decomposition for latent semantic indexing in information retrieval. ACM T. Inform. Syst., 16(4):322–346, 1998.
18. C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization (WAS 2004), Barcelona, Spain, 2004.
19. C.-Y. Lin and E. Hovy. The automatic acquisition of topic signatures for text summarization. In Proc. Document Understanding Conference, 2002.
20. M. J. Ma˜na-L´opez, M. de Beunaga, and J. M. G´omez-Hidalgo. Multidocument summarization: An added value to clustering in interactive retrieval. ACM T. Inform. Syst., 22:215–241, 2004.
21. K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In Proceedings of the Human Language Technology Conference, 2002.
22. A. Mikheev. Tagging sentence boundaries. In Proc. NAACL Conference, pages 264–271, Seattle, WA, 2000. Morgan Kaufmann.
23. P. Over and J. Yen. An introduction to DUC-2004: Intrinsic evaluation of generic news text summarization systems. In Proc. Document UnderstandingConference, 2004.
24. L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. P. IEEE, 77:257–285, 1989.
25. D. R. Radev, S. Blair-Goldensohn, Z. Zhang, and R. S. Raghavan. Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. In Proc. HLT Conference, San Diego, CA, 2001.
26. D. R. Radev, W. Fan, and Z. Zhang. Webinessence: A personalized web-based multi-document summarization and recommendation system. In Proc. NAACL Workshop on Automatic Summarization, Pittsburgh, PA, 2001.
27. G. Salton. Automatic Text Processing – The Transformation, Analysis, and Retrieval of Information by Computer. Addison–Wesley, 1989.