International Journal of Science Engineering and Advance Technology, IJSEAT,Vol.3,Issue 7
ISSN 2321-6905
July 2015

International Journal of Science Engineering and Advance Technology

An Ontology- Based Multi-Document Summarization in Apocalypse Management

Smita Bachal
Dept. of Computer Engineering
Dnyanganga College of Engg. and Research
Pune- 411041, India
Email:smitabachal@yahoo.com

Mr. S.M. Sangve
Dept. of Computer Engineering
Dnyanganga College of Engg. and Research
Pune- 411041, India
Email:sunil.sangve@zealeducation.com

Abstract:
With the problem of extended information resources and the remarkable evaluate of data removal, the require of having automated summarization techniques revealed up. As summarization is needed the most at present searching information on the internet, where the user moves for a specific space of passion as per his question, area centered on summaries would provide the best. Ontology based summarization system for is provided. The ontology is a subjective model, which gives the important framework for semantic representation of textual data. In our suggested system implement the hierarchical levels of ontology to even more enhance the high summary and to execute hierarchical text classification in the field of earth quake management. We signify a scientific study of different techniques in which ontology has been applied for summarization practice. Comprehensive experiments on a selection of press launch appropriate to 2011 Sikkim earth quake illustrate that ontology centered multiple documents summarization techniques outperforms other baselines with regards to the conclusion top quality. Also we are designing a Hierarchical clustering algorithm instead of K-means clustering algorithm for better precision. It is found that the greater part of the current techniques often focus on sentence scoring and less attention is given to the appropriate information content in various records.

Keywords *- Earthquake management, multi-document summarization, ontology, generic, query expansion.*

I. INTRODUCTION

It is amazing that stormy weather, quakes, and other natural problems cause incredible physical destruction and loss of life and property way and large. In order to effectively evaluate the pattern of the problems and reduce the major loss for upcoming scenario, effective information collecting methods are essential. The domain professionals anticipate acquiring compacted details about the specific Earth quake occasion information, e.g., the transformative propensity of the problems, the functional status of the public services, and the renovation procedure of the homestead. In few scenario, it is incredibly hard for domain professionals to identify either the most essential data overall or the most appropriate details to a specified subject targeted summarization. So ontology based summary creation i.e. our system is very helpful. The requirement of summarization has as of delayed lengthy because of the development of information on the Internet. With the availability and rate of web, information removal from on the internet information has been rate down. On the other hand, it is not easy for user to personally summarize those huge on the internet information. In scenario like, when a client search down information about earthquake which occurred in Sendai, Japan, the user will most likely get amazing articles recognized with that occasion [1]. The user would without doubt choose summary that could summarize those document. The purpose of multi-document summarization is collecting the resource content into a smaller type defending its information and common importance. The focus on and strategy of

summarization of documents clarify the kind of summary that is designed.

Approach towards summarization can be either extractive or abstractive (Radev et al., 2002). In extractive summarization, critical phrases are identified and straightforwardly extracted from the exclusive papers, i.e. the last summary comprises of exclusive phrases. Then again, in abstractive summarization (Ganesan et al., 2010) the phrases which are chosen from the exclusive evaluation are further managed to rebuild them in previous times connecting them into last summary. This methodology usually contains powerful organic language processing and phrase pressure. By knowing the kind of summary i.e., a sign, useful, extractive and abstractive, we can then implement them to either individual document or several papers [1] [2].

This study concentrates on useful and extractive type several papers summarization. The various features that make several documents summarization rather different from single documents conclusion is that several papers summarization contains papers summarization issue contains several sources of details that overlap and supplement. So the important task are not just identifying and changing repeating crosswise over documents, furthermore ensuring that the last conclusion is both consistent and complete.

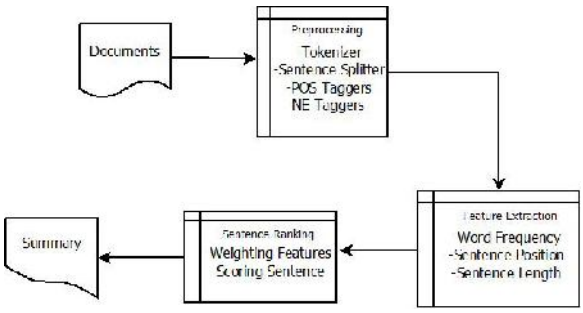


Fig. 1 Feature based summarization

The staying of this research can be classified as: We analyze the four amazing strategies of multiple document summarizations and provide it with related work from literary works. The earnings and demerits regarding these strategies are additionally mentioned. The rest of the study is categorized out as takes after: First we show the evaluation on four multiple documents summarization techniques to be specific the feature based strategy, cluster based strategy, graph centered system and knowledge based system. At that factor we factor the suggested multiple documents summarization strategy; i.e., the element based system. At final we finish with summary.

II. RELATED WORK

A number of research analyses have maintained to multiple documents summarization in the analysis world (Erkan and Radev, 2004a, Wan and, 2008, Haribagiu and Lacatusu, 2010) also reveals unique kinds of strategies and accessible systems for multi documents summarization.

A. Approaches of Multi-document summarization

In this analysis we information our concentrate extremely on four well known strategies to multi documents summarization. Our study will be targeted around the associated with example: For every technique, we will first discuss its main thought. Following that, we will take some analysis from relevant literary works [3] [4] [5].

1) Feature based Method:

Extractive summarization contains identifying the most important phrases from the material and set up them together to create a small summary. Presently acknowledging crucial sentences, feature impacting the significance of phrases are made the decision. Here we show a part of the frequent features that has been regarded for phrase choice.

- Word Frequency:
  The idea of using word frequency is that essential terms seem generally in the documents. The most well-known evaluate generally used to estimate the saying repeat is tf and idf.
- Sentence location:
  Important information in a document is frequently protected by writers at the beginning of the article. Therefore the beginning phrases are required to contain the most crucial material.
- Sentence length:
  Very short phrases are more often than not omitted in summary as they contain less information. Long phrases are furthermore not appropriate to create the summary.

Above fig.1 reveals feature based summarization. In any case not all text features are handled with identical level of importance as a amount of the features have more importance or weight and some have minimum. Hence focus can be given to handling the material functions targeted around their energy. This problem can be getting over by using weight learning strategy. Several scientists have been using

various weight studying systems in their study. Binwahlan et al. (2009) provided a novel text summarization model depending on swarm brainpower method known as Particle Swarm Optimization (PSO) [6].

2) Cluster based method:

The idea of clustering is to cluster similar records into their sessions. The level that multiple reviews are involved, these objects make reference to phrases and the sessions talk with the cluster that a sentence has a position with. Usually, clustering algorithm can be organized as agglomerative or partitioned (Jain et al., 1999).

In agglomerative bunching usually go "bottom up" process, every one phrase is from the get go regarded a different team by its own. Radev et al. led the need of team centroids for their multi report summarizer [7] [8]. Centroids are the top placement tf-idf that talks to the cluster.

These cluster centroids are then applied to identify the phrases in each one cluster that are most like the centroid. Consequently, the summarizer generates phrase which are most appropriate to each one cluster.

3) Graph based method:

The important speculation of graph representation is the organization or linking between items. These relationships are available targeted around their actual relationship. On account of content records, the actual relationship is usually the nearness between objects for this scenario, phrases. As in most wok regarding graph based technique, the most usually used assessment evaluate is the cosine measure. An advantage at that point prevails if the likeness weight is over some predetermined threshold [9]. Once the graph is designed for a located of records, valuable phrases will then be recognized. It requires after the thought that a phrase is regarded important on the off opportunity that it is undoubtedly associated with several other phrases.

4) Knowledge based method:

Most records or content will have its technique recognized with a particular factor or event. These styles or events by and huge fit in with a particular place and every area normally have its own primary learning structure. Therefore, there have been efforts created via experts to utilization the base studying (i.e., ontology) to create summarization outcomes. Certainly, several distinct applications have personalized their model to be ontology motivated (Shareha et al., 2009, Nasir and Noor, 2011).li et al. (2010) designed the Ontology-enriched Multi-Document Summarization (OMS) structure to generate query essential conclusion from a collecting of information [10] [11] [12]. In previous relevant research, Wu and Liu (2003) physically designed a place particular ontology for company information content. A relative believed however with additional ontology peculiarities were suggested by Hennig et al. (2008) for sentence scoring [13] [14].

In [10] A Multi-document Rhetorical Structure (MRS) is suggested for multi-document customized summarization process. This framework can deal with interrelationship between material models at various levels of granularity and can determine at the similar time the occur and modify of unique activities. MRS uncover conventional multi-document representation in combination structure rumours and supplement modify and dispersal information of activities topics which can't be gotten in information combination concept. Determinedly, a activity of figuring's such as building MRS, multi-document information combination centered MRS and explanation period are suggested.

In[15] Enhancement of algorithms for computerized text purchase in enormous text documents sets is a basic evaluation sector of data mining and learning revelation. Most of the text-clustering frameworks were based in the term-based evaluation of partition or similitude, ignoring the dwelling of conditions in records. In this document they show a novel strategy known as Structured Cosine Similarity that apparel papers clustering with a different way of displaying on documents summarize, considering the structure of conditions in documents to be able to enhance the way of talk documents clustering.

III. IMPLEMENTATION DETAIL

In the domain of Earth quake control, over several reports are regularly launched by the local government or regional urgent workplaces through the problems, which cover most activities appropriate to the problems and the time distance will be days to months, based on how severe the problems is. The information will be provided in a structure of newswire, including lot of schedule confirming on several factors of the problems. In such a scenario, it is incredibly hard for domain experts to identify either the most important info overall summarization or the most appropriate details to a specified query. In our system very well implement structure of ontology created by professional for in sentence choice strategy.

A. System Architecture

- First of all, the ontology into multi-document summarization issue in earth quake control domain is applied.
- The chance of using the ontology is researched to make it happen of decreasing information redundancy.
- Also ontology is used in sentence selection process.
- The ordered clustering algorithm is used for better precision.

System mainly contain following components.

- Create ontology: by professional, on earth quake ontology of concepts is designed.
- Input: several documents relevant to earth quake.
- Sentence mapping: Allocate sentence in ontology node.
- General summarization:
  i. Preprocessing: tokenization, stop words eliminating, is arising.

ii. Clustering algorithm: Hierarchical algorithm.
iii. Improved sentence selection: For choice of sentence, hierarchy of ontology is applied; sentence significance is enhancing as per its place. Sentence at reduced stage is more essential than sentence at advanced stage.
iv. Redundancy Reduction: eliminated identical significance information phrases.

- Question Centered Summarization
  i. Query mapping: add query in ontology structure according to ideas consist of in query.
  ii. Finding Similar classes: find relevant sentence category.
- IC based sentence ranking: assign ranking of each sentence.
- Last Summary: choose top obtained sentence.

B. Algorithm

1) k-means Algorithm: Input: n objects (or sentences) and a number k Algorithm

1: Initially randomly select K no of sentences as centroid from input sentences. These sentences represent initial group centroids.
2: Calculate similarity value of all sentences from each cenroid.
3: Assign each sentence to the group that has the max similar centroid.
4: When all sentences have been assigned, recalculate the positions of the K centroids.
5: Repeat Steps until the stopping condition is met i.e. same centroids value meet.

2) Hierarchical Algorithm: Input: no of document
Output: Clusters of similar sentences.

1: Compute the distance matrix between the input data points i.e. sentences.
2: Let each data point be a cluster
3: Repeat
4: Merge the two closest clusters
5: Update the distance matrix
6: Until only a single cluster remains

C. Mathematical Model

- Term Frequency (TF): The term frequency defined as follows:

$$t, f_{i,j} = \frac{n_{i,j}}{\sum_k ni}$$

Where,
- $n_{i,j}$ is the number of occurrences of term ti in sentence sj,
- SUM (n (i,j)) is the sum of number of occurrences of all the terms in sentence sj .

- TFICF

$$\text{icf} = \log \frac{|C|}{C:T:\varepsilon\, e}$$

Where, |C| - is total no of concepts in ontology
$C:T:\varepsilon\, e$ is the number of concepts where term ti appears (that is, ni, j ≠0).
Then TFICF is defined as,

$$\text{TFICFi; j} = t, f_{i,j} * icfi$$

- Similarity between Sentences

$$\text{SentSim (s1, s2)} = \frac{\sum_{i=1}^{n} s1i * s2i}{\sqrt{\sum_{i=1}^{n} s1i * \sum_{i=1}^{n} s2i}}$$

Where,
S1 and S2 is sentence vector of n size
i is sentence number.

- Information Content (IC)

IC of a concept is defined by following way negation of the logarithm of the probability p (a) of encountering a concept a in a given document as IC (a)-log P (a)
The probability of a concept can be calculated as the summation of the each occurrence of all the concepts which are subsumed by it as follows:

$$\text{P (a)} = \sum_{\text{nespecialisation}} \frac{Count(n)}{N}$$

Where,
- Specializations (a) is set of terms subsumed by concept a,
- N is the total number of concepts in the corpus.

- Improving Sentence Ranking:
  Si (w) =Si (w) +x

Where, W - is calculated weight by IC method Si - i th no Sentence

x - is an integer its value depends on ontology hierarchy wise predefined threshold

Set Theory Let system S is represented as:

S = {SM, SR, SS, DC} (1)

- Sentence Mapping
  Consider, SM is a set for the sentence mapping
  SM = {sm1, sm2 …}
  Where, sm1, sm2 ... are the number of sentence mapped on ontology hierarchy
- Sentence Representation
  Let SR is a set for sentence representation
  SR = {sr1, sr2, sr2 …}
  Where, sr1, sr2... are the number of sentence representation used
- Summarization Process
  Let, SS is a set of summarization
  SS = {Gs, Qs}
  Let Gs is a set for generic summarization for sentence selection, redundancy reduction and ranking.
  For this we represent p for this all process.
  Gs = {p1, p2, p3 …pn}
  Where p1, p2, p3... are the number of process Qs is a set for query based summarization
  Qs = {q1, q2, q3…qn}
  Where q1, q2, q3 are the number of queries
- Document Clustering
  Let DC is a set for document clustering
  DC = c1, c2, c3... where,
  C1, c2... are the clusters form after various approaches.

IV. RESULTS AND DISCUSSION

Our system utilization several records (Sikkim Earth quake 2011) of problems information gathered from reports of various news channels and also report launched by National Disaster Response Force. It includes several records in total regarding prior to during and after earthquake approved. For evaluation between k mean and ordered criteria we use entropy and f evaluate principles. The outcome reveals that hierarchical algorithm gives high quality of outcomes equivalent to k-mean but k-mean algorithm works excellent as in evaluation on time base. In order to evaluate produced summaries by distinct techniques, in our system use TFICF based produced conclusion as sources and after it evaluate with our new enhances summary generation system. Result reveals that our system is outshining the current system. System reveals precision and recall is better in contrast to other summary technique.

Below tables and graphs are describes predicted results of our system

Table I. F Measure of K-Mean and Hierarchical

No.of document	K-mean F measure	Hierarchical F measure
D50	0.63	0.94
D100	0.64	0.95
D150	0.62	0.945
D200	0.73	0.946
D250	0.66	0.845
D300	0.83	0.945

In following figure 3 graphs shows that no of documents (x-axis) vs. f measure value (y axis) of k mean and hierarchical.

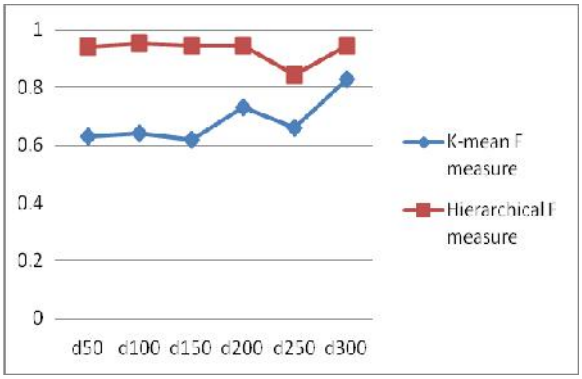


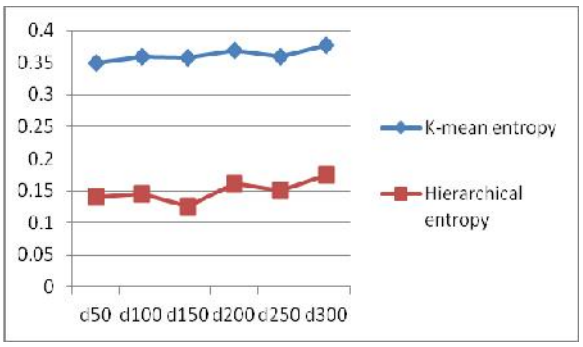
Fig.3: f measure comparison of k-mean and hierarchical

Table II. Entropy of K-mean and Hierarchical

No.of document	K-mean entropy	Hierarchical entropy
D50	0.35	0.141
D100	0.359	0.146
D150	0.358	0.125
D200	0.369	0.162
D250	0.36	0.15
D300	0.377	0.174

In following figure 4 graph shows that x axis represent no of documents vs. y axis represent entropy value of k mean and hierarchical. Entropy increases as the database increase means the quality of cluster decreases as records increases i.e. hierarchical algorithm require less database.

Fig.4: Entropy comparison of k-mean and hierarchical



Comparison between our system and previous system based on recall and precision value.
In following figure 5 graphs describe good recall value of current system. Y axis represents recall value and X axis represent indicate no of documents.

Table III. Recall of system

No.of document	K-mean entropy	Hierarchical entropy
D30	64.17	70.13
D25	62.09	64.05
D23	60.14	97.77
D20	55.12	58.7

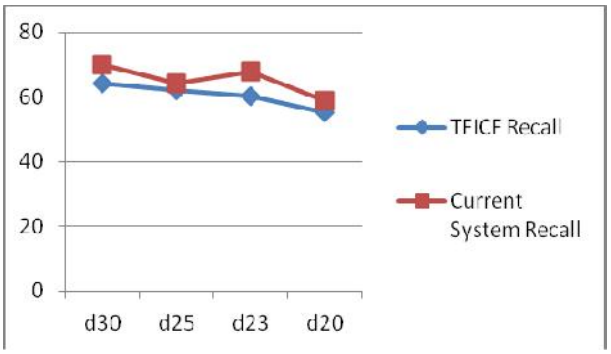


Fig.5: shows recall base comparison between Existing and proposed system

In following figure 6 graph describe existing system give better accuracy as compare with previous system. Y axis indicates precision value and X axis indicates no of documents.

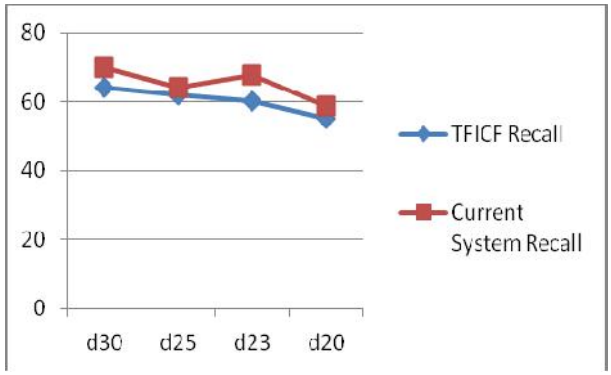


Fig 6: Precision base comparison between Existing and proposed system

Finally according to the outcome obtained when assessing the user satisfaction, summary and abstract generated by our system is give better outcome.

Table IV. Precision of systems

No.of document	TFICF precision	Existing system precision
D30	63.11	65.01
D25	61.96	60.11
D23	59.56	67.56
D20	53.19	54.19

V. CONCLUSION

A scientific research on several techniques that implement the ontology is given to fix various multi-document summarization issues in Earth quake management sector. In this paper, enhanced summary generation strategy is utilized by using ontology structure. For generic summarization, various vector space designs are utilized to signify sentences in the reports gathering, and the practicality of different mixtures of the VSMs is researched. Then the hierarchical algorithm were used to cluster the sentence set and the important sentence close to the centroids of the sentence groups are produced using enhanced sentence selection method.
The last summary was consequently produced by decreasing information redundancy and position sentence is outperforms past outcome of summary. For query focused summarization, we delved into the

impact of query development in summarization works.

ACKNOWLEDGMENT

I would like to thank the scientists as well as publishers for creating their sources available and teachers for their guidance. I'm grateful to the regulators of Savitribai Phule University of Pune and concern associates of cPGCON2015 conference. I'm also grateful to reviewer for their useful recommendations and also thank the college authorities for providing the needed facilities and assistance. Finally, we would like to extend a heartfelt gratitude to friends and family members.

REFERENCES

[1] V. Nastase, "Topic-driven multi-document summarization with encyclopedic
Knowledge and spreading activation", in Proc. EMNLP, 2008, pp. 763-772.

[2] C. Lee, Z. Jian, and L. Huang, "A fuzzy ontology and its application to news summarization," IEEE Trans. Syst., Man, Cybern., B Cybern., vol. 35, no. 5, pp. 859-880,Oct. 2005.

[3] H. Saggion, K. Bontcheva, and H. Cunningham, "Robust generic and query-based summarization," in Proc. ECAL, 2003, pp. 235'238.

[4] F. Wei, W. Li, Q. Lu, and Y. He, "Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization," in Proc. SIGIR, 2008, pp. 283 290.

[5] X. Wan, J. Yang, and J. Xiao, "Manifold-ranking based topic focused multi-document summarization," in Proc. IJCAI, 2007, pp. 2903,2908.

[6] J. Tang, L. Yao, and D. Chen, "Multi-topic based query-oriented summarization," in Proc. SDM, 2009.

[7] A. Highlight and L. Vanderwende, "Exploring content models for multi-document summarization," in Proc. HLT-NAACL, 2009, pp. 362 370.

[8] E. Klien, M. Lutz, and W. Kuhn, "Ontology-based discovery of geographic information services: An application in disaster management, Compute" Environ. Urban Syst., vol. 30, no. 1, pp. 102 123, 2006.

[9] H. Hsu, C. Tsai, M. Chiang, and C. Yang, "Topic generation for web document summarization," in Proc. IEEE SMC, 2008, pp. 3702 3707.

[10] X. Yong-dong, W. Xiao-long, L. Tao, and X. Zhi-ming, "Multi-document summarization based on rhetorical structure: Sentence extraction and evaluation," in Proc. IEEE SMC, 2008, pp. 3034 3039.

[11] D. Wang, T. Li, S. Zhu, and C. Ding, "Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization," in Proc. SIGIR, 2008, pp. 307 314.

[12] G. Erkan and D. Radev, "Lex page rank: Prestige in multi-document text summarization," in Proc. EMNLP, vol. 4, 2004, pp. 365 371.

[13] X.Wan and J. Yang, "Multi-document summarization using cluster based link analysis," in Proc. SIGIR, 2008, pp. 299 306.

[14] D. Radev, H. Jing, M. Sty, and D. Tam, "Centroid-based summarization of multiple documents," Inf. Process. Manage. vol. 40, no. 6, pp. 919 938, 2004.

[15] S. Yuan and J. Sun, "Ontology-based structured cosine similarity in document summarization: With applications to mobile audio-based knowledge management," IEEE Trans. Syst., Man, Cybern. B Cybern. vol. 35, no. 5, pp. 1028-1040, Oct. 2005.

[16] http://en.wikipedia.org/wiki/Cosine-similarity

[17] http://en.wikipedia.org/wiki/2011-Sikkim-earthquake

[18] http://en.wikipedia.org/wiki/Complete-linkage-clustering