Available online at www.ijiere.com

International Journal of Innovative and Emerging Research in Engineering

e-ISSN: 2394 - 3343 p-ISSN: 2394 - 5494

Query focused Multi-Document Summarization in Disaster Management domain using ontology

Pallavi Sonwane, Imran Shaikh
SNDCOE & RC Yeola, Savitribai Phule Pune University,
Maharashtra, India

ABSTRACT:
From last few decades, so many terrific disaster had taken places. After that disaster various media or newspaper gives different news .If anyone wants to get an accurate information about that disaster then he must read all the news of particular disaster. Therefore in this paper I explained a conceptual method called as domain ontology. In this paper, I explore the concept of ontology to summarize all over the information related with a particular disaster.

Keywords: Disaster, Query Focused, Ontology, Generic Summarization, Sentence Mapping.

I. INTRODUCTION

Now a days number of disasters arise like floods, earthquakes, tsunami etc which destroy the life of humen being totally in many manners like natural property and life. to minimize the loss of disaster in future situation it is necessary to analyze the trends, patterns & relationships of the disasters efficiently there are many methods to get the information about the disaster like, to get the information about the disaster like, media or newspaper may be recorded in the form of text documents that are related to the disaster.

If any domain expert wants to get complete information about the related disaster events description then he must read all the information avialable [2]. for e.g. the evolutionary nature of disaster, help status of the public service ,public participation during disaster event, expectation from their government etc.
An example of a disaster news is shown in Table I. As is shown, the six sentences provide a summary on the status of disaster over a week in Jammu & Kashmir.

TABLE I Example of Disaster Information

J & k floods :kins of victims still search for bodies
Village in bandipuraunder water, formers lose livelihood
Sonali, rahul meet j & k flood victims,promise help
J & K Floods:hariyanagovt to send pumps to J & K to drain out water
Gates foundation announces Usd 7 lakh J & K Flood relief
J & K Floods: Death toll climb to 277

Such information can provide domain expert primary information of how the life was affected by the flood, and subsequently, domain experts will contact the respective department and make a set of data that would be helpful if the situation happens again in future.

In the field of disaster management, during the disaster, over thousands of hundreds of reports are often generated by the state government or local emergency offices , which cover most information related to the disaster and the duration will be weeks to months, depending on how intense the disaster is. The data will be represented in a format of different news, containing a lot of routine reporting on multiple issues of the disaster. In such a case, it is very problematic for domain experts to easily find either the most important information overall (generic summarization) or the most relevant information for any query(query /topic-focused summarization)[12][13]. Therefore, multidocument summarization techniques can be used to get helpful information from multiple reports.

II. EXISTING SYSTEM

Most previous work is based on predefined matching rules hand-coded by domain experts or matching rules learned offline by some learning method from a set of training examples. Such approaches work well in a traditional database environment, where all instances of the target databases can be readily accessed, as long as a set of high-quality representative records can be examined by experts or selected for the user to label. Consequently, hand-coding or offline-learning approaches are not appropriate for two reasons. First, the full data set is not available beforehand, and therefore, good representative data for training are hard to obtain. Second, and most importantly, even if good representative data are found and labeled for learning, the rules learned on the representatives of a full data set may not work well on a partial and biased part of that data set. While most previous record matching work is targeted at matching a single type of record. Unfortunately, however, the dependencies among multiple record types are not available for many domains.

III.RELATED WORK

A. GENERIC SUMMARIZATION

To get exact information in less time from a large group of document ,text summarization is useful . It is very difficult for human beings to manually summarize the text of large documents. There is an great amount of text material available on the internet. However, usually the Internet provides more information than is required. Therefore, a big problem is encountered: searching for relevantdocuments through an large number ofdocuments available, and containing a large quantity ofrelevant information. text summarization is used to reduce the source text into a shorter version protecting its information content and overall meaning..

MEAD is a multi-document text summarizer

Proposed by Dragomir R. Radev [5]. The proposed system creates the summary based on cluster centroids.

Information summarization, which is the combination of the rhetorical structure theory and MEAD summarizer proposed by AfnanUllah Khan [3].

Summarization system which is mainly based on sentence-level semantic analysis and non-negative matrix factorization proposed by Dingding Wang [4]. Two techniques for both single and multi document text summarization proposed proposed by Md. Mohsin Ali [5]. The first technique is the combination of The first technique.(cpsl is the Combination of Centroid, Position, and Length Features), second LESM is the combination of LEAD and CPSL[7].

Wepropose generic text summarization methods that create text summaries by ranking and extracting sentences from the original documents. However, most existing methods could not provide conceptualinformation in the sentence level. In most cases, to make summary from the data the conceptual information is required. Some scholars use that concept contained by the sentences to point out multi-document summarization [10].We know that Wikipedia contains too many concepts not relevant to a specific domain, therefore such methods cannot be applied to domain-specific document summarization tasks.

We enhance the generic summarization by using the redundancy based algorithm to save the time for interpreting the text information from the large documents[10].

B. QUERY-FOCUSED SUMMARIZATION

Query focused summarization extracts an important information summary from a set of document collection for query given by users. Exact information related to query is given by the Query-Focused Summarization method.

For example, a user submits a query to a search engine and the search engine usually returns a lot of result documents. To 'click-and-view' each of the returned documents is obviously tedious and infeasible in many cases.

One challenging issue is how to help the user digest the returned documents. Typically, the documents talk about different perspectives of the query. A generalized solution must be created a meaningful summary for each query.

. Supervised method and unsupervised method are the two existing methods. Supervisedmethod limited to predefined domains because it requires training examples. To find 'centered' sentences as the summary, clustering algorithm used by the unsupervised method. The summarization is generally depend on the document collection itself but it does not consider query information. In most of the existing system the topic is related only with the query asked by the user. In this paper, we try to overcome the limitations of the existing methods and study a new setup of the problem of multi-topic based query-oriented summarization

C. QUERY EXPANSION

Query expansion is the technique of segmenting the user's query with additional terms in order to improve search results. Document summarization is also done by using query expansion. Limited effect on web information retrieval performance has shown by expanding with synonyms. However other types of information derived from an ontology are muchmore useful to improve the search results[1].

IV.PROPOSED SYSTEM

A. Ontological Hierarchy: Ontology concept is often used by domain experts in disaster management domain. Such ontology provides the information exists in disaster management, and how such information can be related within a hierarchy and subdivided according to similarities and differences among them[9].

B. **Text Preprocessing and Feature-based estimation:** The following points are considered while preprocessing of the textual data:

Word frequency: The idea of using word frequency is the important words appear many times in the document. The most common measure to calculate the word frequency is TF and IDF.

Title/headline word: Occurrence of similar word in title and sentence indicates that the sentence is highly relevant to the document.

Sentence location: Important information in a document is often covered by writers at the beginning of the article. Thus the beginning sentences are assumedcontaining most important contents.

Sentence length: Very short sentences are usually not included in summary because they express less information. Very long sentences are also not suitable for a summary.

Cue word: There are certain words in a sentence which indicate that the sentence is carrying an important message in the document (e.g., “significantly”, “in conclusion”).

Proper noun: Sentences containing proper noun representing a unique entity suchlike name of a person, organization or place are considered important to the document.

V. SYSTEM ARCHITECTURE

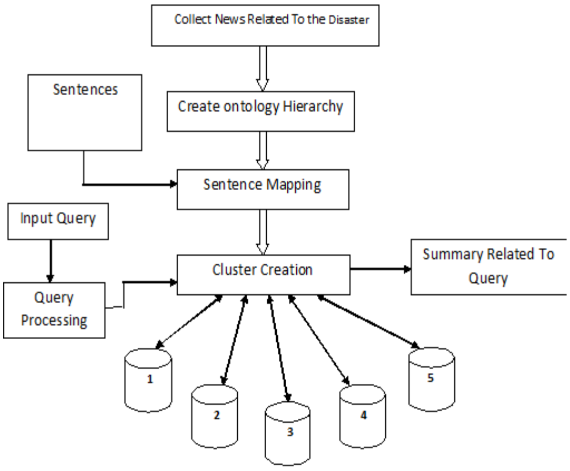

Figure 1. System Architecture

1. Collect news related to the disaster: the firt step is collection of news from different midium and make a document set in the texual format. This process is done manually and the document sets are saved in the particular folder.

2. Create ontology hierarchy: Generally speaking, an ontology is often provided by domain experts in disaster management domain. Such an ontology provides answers for the questions concerning what entities exist in disaster management, and how such entities can be related within a hierarchy and subdivided according to similarities and differences among them.

C. Sentence Mapping: To utilize the ontology for better understanding the documents, we initially decompose the collection of domain-specific documents into sentences, and then map each sentence to the ontology hierarchy. For each concept of the ontology hierarchy, a group of keywords (i.e., nouns) are assigned by the experts for the sake of sentence mapping. The procedure of sentence mapping is executed based on the following criteria.

1) If the sentence is related to only one concept, map this sentence to the corresponding concept.

2) If the sentence is related to two or more concepts, map this sentence to the least common ancestor (LCA) of these concepts. If the LCA is the most general concept of the ontology, then map the sentence to the original specific concepts.

Sentence Representation:

A key question in multi-document summarization using the ontology is how to represent the sentences we have mapped onto the ontology. We examine several ways to model a sentence into a vector, including term frequency (TF) model, term frequency-inverse sentence frequency (TFISF) model , term frequency-inverse concept frequency (TFICF) model, concept hierarchy (CH) model , and the linear combinations of these models. The vector space models mentioned above provide different insights for document summarization.

Term Frequency Model: In this model, each entry of a sentence vector denotes the term weight (normalized term frequency to prevent a bias toward longer sentences and to give a measure of the importance of the term *ti* within the particular sentence *sj*).

TFISF Model: Similar to TFIDF, term frequency, inverse sentence frequency (TFISF) is used to evaluate how important a word is to a sentence in a corpus. The importance increases proportionally to the number of times a word appears in the sentence but is offset by the frequency of the word in the corpus. The inverse sentence frequency is a measure of the general importance of the term (obtained by dividing the total number of sentences by the number of sentences containing the term, and then taking the logarithm of that quotient).

D. Input query: I this step the query is entered by the user and the generated summary is focused on that particular query.

E. Cluster creation: K means algorithim is used to cluster creation.
Algorithmic steps for k-means clustering
Let $X = \{x_1,x_2,x_3,\ldots\ldots..,x_n\}$ be the set of data points and $V = \{v_1,v_2,\ldots\ldots.,v_c\}$ be the set of centers.
1) Randomly select *'c'* cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
4) Recalculate the new cluster center using:

$$\mathrm{Vi} = (\frac{1}{Ci})\sum_{j=1}^{Ci}\mathrm{Xi}$$

where, *'c_i'* represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

VI. MATHEMETICAL MODELING

A. TERM FREQUENCY MODEL:

Each entry of a sentence vector denotes the term weight. The term frequency defined as follows:

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}$$

where $n_{i,j}$ is the number of occurrences of term ti in sentence sj , and the denominator$\sum_k n_{k,j}$ is the sum of number of occurrences of all the terms in sentence sj .

B. TFISF MODEL:

Similar to TFIDF [16], term frequency inverse sentence frequency (TFISF) is used to show how important a word is to a sentence in a group. Its importance increases directly to the number of times a word appears in the sentence but is reduced by the frequency of the word in the group. The inverse sentence frequency is a measure of the general importance of the term (obtained by dividing the total number of sentences by the number of sentences containing the term, and then taking the logarithm of that quotient), which is defined as

$$isf_i = log\frac{|S|}{|s: t_i \in s|}$$

Where mod of $|s|$ is the total no of the sentence in the group and$|s: ti \in s|$is the number of sentences where the term tiappears.

Then TFISF is defined as

$$TFISF_{i,j}=tf_{i,j}*isf_i$$

High term frequency (in the given sentence) and a low sentence frequency of the term in the whole document collection represent by high &low weight respectively to filter out the common terms. . The TFISF value for a term will always be larger than or equal to zero.

C. TFICF MODEL:

Term frequency-inverse concept frequency (TFICF) is used to prove how important a word is to a concept in an ontology sequencing. The inverse concept frequency is computed as

$$icf_i=\log\frac{|C|}{|c:t_i\in c|}$$

Where $|c|$ is the total number of concepts in the ontology $|c: ti \in c|$ is the number of concepts where the term ti appears. If the term is in the group but not appears in any concepts, then the icf value is set to zero. TFICF is defined as

$$TFICF_{i,j}=tf_{i,j}*icf_i$$

D. CH MODEL:

The procedure of sentence modeling by CH is as follows:

1) We number all of the nodes level by level from top to bottom except the root.
2) We represent a sentence as a vector Vc with the size of the total number of concepts in the ontology, denoted as c1, c2, ··· , c| This vector is initialized with zeros. The value of each entry ci is increased by 1 when encountering an existence of the i-th concept node in the sentence.
3) To impose the structure into a vector, all of the ancestor nodes of the 1-valued ci are accounted as existence.
4) s l2-normalized so that ||Vc||2 = 1.

E. LINEAR COMBINATIONS:

The first three vector space models— TF, TFISF, TFICF—are provided based on the term feature in the sentences, whereas CH using the concepts appearing in the whole document collection to represent a sentence. In the following, we try to combine the term-based VSMs and the concept-based VSM together to verify if it is helpful to summarize multiple documents.

$CH+TF: V_{c+tf}=(\lambda_1 V_c, (1-\lambda_1)V_{tf})$;
$CH+TFISF: V_{c+tfisf}=(\lambda_2 V_c, (1-\lambda_2)V_{tfisf})$;
$CH+TFICF: V_{c+tficf}=(\lambda_3 V_c, (1-\lambda_3)V_{tficf})$.

λ1, λ2, and λ3 are the vectors in a combined form respectively.

VII. CONCLUSIONS

In this paper, I use ontology to solve different multi document summarization problems in disaster management. In generic summarization, we are using different vector space models to represent sentences in the document collection by using the combination of different vector space models. By using centroid-based methods we cluster the sentence set & extract the sentences which are close to the centroid of the sentence[6]. By reducing information redundancy and ranking sentences, we get the final summary. We are using query expansion in summarization task for query focused summarization[8][11].The meaningful information is provided by an ontology related with collection of data. we provide information extraction techniques to further improve summarization results.

REFERENCES

[1] L. Zheng, C. Shen, L. Tang, T. Li, S. Luis, S. Chen, and V. Hristidis, “Using data mining techniques to address critical information exchange needs in disaster affected public-private networks,” in Proc. SIGKDD, 2010, pp. 125–134.

[2] X. Wan, J. Yang, and J. Xiao, “Manifold-ranking based topicfocused multi-document summarization,” in Proc. IJCAI, 2007, pp. 2903–2908.

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

[4] G. Erkan and D. Radev, “Lexpagerank: Prestige in multi-document text summarization,” in Proc. EMNLP, vol. 4, 2004, pp. 365–371.

[5] 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

[6] 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.

[7] 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

[8] H. Saggion, K. Bontcheva, and H. Cunningham, “Robust generic and query-based summarisation,” in Proc. ECAL, 2003, pp. 235–238.

[9] E. Klien, M. Lutz, and W. Kuhn, “Ontology-based discovery of geographic information services—An application in disaster management,” Comput., Environ. Urban Syst., vol. 30, no. 1, pp. 102–123, 2006.

[10] 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.

[11] X. Wan and J. Yang, “Multi-document summarization using clusterbased link analysis,” in Proc. SIGIR, 2008, pp. 299–306.

[12] H. Daume and D. Marcu, “Bayesian query-focused summarization,” in ´ Proc. ACL, vol. 44, no. 1.2006, p. 305.

[13] 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.