www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 11 Nov 2015, Page No. 14980-14982

Ontology-Based Multi-Document Summarization.

Bhakti Mehta ,Varsha Marathe ,Priyanka Padvi and Manjusha Shewale
MITCOE,,Savitribai Phule Pune University,Pune, Maharashtra,India.

Bhakti Mehta- B.E. in Computer Engineering from MIT College of Engineering,Savitribai Phule Pune University,Pune,Maharashtra,India.
bhaktimehta0909@gmail.com
Varsha Marathe- B.E. in Computer Engineering from MIT College of Engineering,Savitribai Phule Pune University,Pune,Maharashtra,India.
varshamarathe7979@gmail.com
Priyanka Padvi- B.E. in Computer Engineering from MIT College of Engineering,Savitribai Phule Pune University,Pune,Maharashtra,India.
priyu.padvi@gmail.com
Manjusha Shewale- B.E. in Computer Engineering from MIT College of Engineering,Savitribai Phule Pune University,Pune,Maharashtra,India.
manjusha.shewalen24@gmail.com

Abstract: **Ontology is defined as conceptual representation data and relationships between this data.In this paper, we propose to use this ontology for summarization of multiple documents related to a specific domain. We explore various techniques that can be used for summarization. We then focus upon a particular approach for summarization of documents belonging to a particular domain.The domain that we have considered is disaster management.As an example we will be taking the earthquake that took place in Nepal in April 2015.Using this example we will demonstrate summarization of multiple documents.**

Keywords: Ontology ,Nepal earthquake ,multiple documents, summarization.

1 .INTRODUCTION

As we all are aware, natural calamities like earthquakes,floods etc.cause lot of destruction of property and life. It is very important to analyze the consequences of the disasters and minimize the destruction caused and the loss to handle the future situations.Information gathering plays a very important role here. effective. More specifically, the news reports regarding the disaster can be in the form of text documents.It is a lot helpful for domain experts if the detailed information about the disaster is available in condensed format. For e.g., the worst case scenario that could occur due to the disaster, the the status of the rescue plan implemented, and the measures taken to control the consequences of the disaster.The following scenario gives the information about the aspects of disaster managemenent that are important to analyse by the domain experts.

Scenario: An earthquake took place in Nepal in April 2015.Due to this earthquake the habitats in Nepal was destroyed and devastated .Domain experts need to analyse the status of rehabilitation in the country.

Table no. 1

Nepal Earthquake rehabilitation information.

The extent of destruction in Nepal
The department reports about 380 000 people have lost their houses on 28th April 2015.
The department reports that about 4 million people required shelter.
The rescue program and providing tents for affected people began from 29th April 2015.

A list of important conceptual sentences regarding the damage in Nepal is shown in the table above.This provides a summary on the destruction in Nepal. Such a summary provides a primary overview of how the property and habitat of people was influenced by the earthquake, and accordingly, domain analysts will take suitable measures by contacting the corresponding department and coming up with a rehabilitation program to cope with the disaster.

In disaster management, over thousands of news reports are released by various mass media sources during the disaster, which give detailed information about various events regarding the disaster and the time span can vary from days to months depending upon the severity of the disaster.In such case, it is difficult for the domain experts to find important information overall or the most relevant information as per the query. Hence multidocument summarization techniques can be used to extract meaningful information from multiple reports.

An ontology on disaster management is provided by domain experts.. Such an ontology has a lot of conceptual information about the document set. We can use this ontology for summarization purposes. An important question here is how to utilize the ontology to obtain summaries, i.e., summary with non redundant sentences.

In this paper, we explore the possibility of employing the ontology into summarization problems in the domain of disaster management. We will represent a sentence as a sentence vector using ontology. We then proceed further from two directions i.e generic summarization and query-focused summarization. In generic summarization, we will study centroid-based sentence selection approach by making use of different vector space models, and utilize the ontology to reduce information redundancy. In query-focused summarization, we employ ontology based query expansion to optimize the final summary results.

II. Related Work

A. Generic Summarization

In generic summarization, a score is assigned to every sentence, the sentences are then ranked according to this score, and the sentences that are ranked at the top are selected as the summary based. Both unsupervised and supervised methods are proposed for the purpose of analysis of the information present in the document set, and find top ranked sentences into the summary based on statistical and conceptual features. For example, MEAD uses centroid based method.In this method the scores for the sentences are computed based on features of the sentences.

But, many existing methods tend to ignore concept-based information in the sentences. Such conceptual information is very important for good quality summaries.However such techniques cannot be applied for domain specific summarization.In such cases detailed semantic relationship is ignored

B. Query-Focused Summarization

Information about a given query is condensed and converted into summaries,in query focused summarization.Extraction of the sentences suiting user’s need is done. Many methods that are used for generic summarization can also be used to incorporate the query information .A robust summarization system was developed within the GATE architecture .This system uses robust components for semantic tagging and co-reference resolution.

III.PROPOSED SYSTEM

A.UPLOADING DATA:

The first step in document summarization is collection of documents that are to be summarized. Now,in our example various news reports regarding Nepal earthquake is to be summarized. The mobile devices will upload various documents to the server .The further computation for summarization purpose will be carried out on this server.

B.SENTENCE MAPPING:

Sentence mapping is one of the most important step in document summarization.The importance of each sentence is calculated using the concept of term frequency and inverse document frequency.

Term frequency and inverse document frequency can be obtained using cosine similarity algorithm.

C.COSINE SIMILARITY ALGORITHM:

In cosine similarity algorithm ,each sentence is converted into a vector.Then the similarity between these vectors is measured.This is done by measuring the cosine value of the angle between them.One of the benefits of using cosine similarity algorithm is that it is very efficient to evaluate.

D.CLUSTER CREATION:

Once the importance of various terms in the document is obtained,now we have to group similar documents.For clustering of documents we will be using inverse document frequency concept.Here each document will be assigned with a particular inverse document frequency.The documents are then sorted in descending order according to the value IDF.Then using some threshold values,these documents will be clustered using IDF.

E.SUMMARIZATION:

After clustering of documents is done, summarization is done using term frequency in the clustered.

After this step a rough draft of document summarization is ready.

F.REDUCE REDUNDANCY:

We will be using weight-sum approach to remove conceptually same but grammatically different sentences i.e to reduce redundancy in the final summary.

The basic block diagram of our proposed system is as shown below:

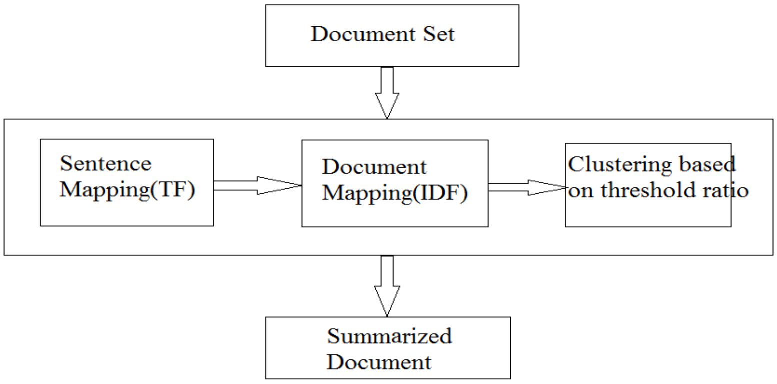

Fig 1.:Block diagram of proposed system.

IV.Mathematical Model

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

$$tf_{ij} = n_{ij \div \sum k n_{k,j}}$$

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

here $n_{i,j}$ is the number of occurrences of term ti in sentence sj .

The denominator$\sum_k n_{k,j}$ is the sum of number of occurrences of all the terms in sentence sj .

V.Conclusion

Thus In this paper,we have explored a feasible approach for summarization of multiple documents.In this approach we upload multiple documents from mobile devices.We use cosine similarity algorithm to find term frequency and inverse document frequency.This inverse document frequency is used to cluster similar documents.Based on IDF and TF summarization is done. After that to reduce redundant data ,we are using weight-sum approach.Hence the final summary is obtained.

VI.References

1. H. Hsu, C. Tsai, M. Chiang, and C. Yang, “Topic generation for web document summarization,” inProc. IEEE SMC, 2008, pp. 3702–3707.
2. 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.
3. An Empirical Study of Ontology-Based Multi-Document Summarization in Disaster Management Lei Li and Tao Li-IEEE transactions on systems, man, and cybernetics: systems, vol. 44, no. 2, february 2014.