IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A Survey of Multi-Document Summarization

Shafna T T[*1], Chitharanjan K[2]

[*1, 2] Dept. Computer Science & Engineering, SCT College of Engineering, Trivandrum, India

shafnateetee@gmail.com

Abstract

This paper describes a survey of Multi-DocumentSummarizers (MDS) built into detection of new information. MDS aims at extraction of information from a set of documents written about same topic and helps to familiarize themselves with information content in large cluster of documents. There are several strategies for selecting interesting and informative sentences from set documents. Generic and Topic-based Multi-Document Summarization are the two main strategies. This paper goes through different approaches in each strategy.

Keywords: Multi Document Summarizer, Generic, Topic Focused, Relevance Propagation

Introduction

The explosion of WWW with the development of information systems and computer hardware capabilities, has led to a vast board of information. It has become virtually impossible for a person to read this large cluster of documents and understand them. We summarize to put similar ideas from the document set in a more concise and comprehensive form. When the summarization is done by a machine, it is called Automatic Summarization. Key benefits of Automatic Summarization are that it is unbiased because it represents information extracted from sources algorithmically, without any editorial touch or subjective human intervention. Automatic Multi-Document Summarization can be of twotypes based on their relation to the source- Extractive Summarization and Abstractive Summarization [1]. Extractive Summarization selects the important sentences from the document to form summary. Abstractive Summarization paraphrases using noval sentences. In Extractive Summarization sentences are usually ranked on the basis of scores calculated by predefined features such as Frequency-Inverse sentence frequency, sentence or token position and number of keywords. Extractive summary are better than automatic summary because in abstractive problems like semantic representation, inference and language generation is harder than sentence extraction in extractive summarization.

There is also a need for the automatic creation of generic and topic-oriented summaries. A generic or topic-general summary is a summary that truly reflects the main content of the original text. A viewpoint-oriented summary summarizes text according to a certain viewpoint, which might express the information need of the user in a retrieval system. Various approaches are there for each type of summary. In Generic, we deals with how generic summary is created: different steps involved in the process. In Topic based MDS, different approaches are discussed first. Then we provide detailed study of steps in its process of summary creation.

Multi-Document Summarization have different levels of approaches namely Morphological Level, Syntactic Level and Semantic Level. A comparative study of these levels is done in the last section.

Generic MDS

The generic summarization technologies [2] aim at processing texts from a variety of sources without the need of a priori knowledge acquisition about the collection, and use only knowledge resources that are or can be made generally available. In summarization, it is important to exploit the discourse structure of text and sentences in order to detect main content. Important content in a single discourse based on linguistic theories of topic is detected and comment or focus and on patterns of thematic progression in texts. Hierarchical topic trees of individual texts are built, and perform topic segmentation and summarization at various levels of topical detail. In order to reduce the content of an individual sentence, the parsed sentence is analyzed. The parser allows detecting the main grammatical constructions from the dependent ones and gives an indication of the semantic relationships between content items. Sentence reduction is especially needed when the summaries are in the form of headlines. Finally, when summarizing multiple documents, it is important to detect redundant content. This can be done with statistical techniques that cluster the lexical and syntactic features of sentences.

A. Pre-processing of the texts

The pre-processing of the texts focuses on an initial cleaning, a refined tokenization in single and compound words and on grammatical analysis. The removal of tags, text between certain tags, very short sentences, parenthetical text and of direct speech comes under initial cleaning. Next step is text tokenization and lexical analysis. The most important functionalities regard sentence break detection and detection of compound terms and then they are grammatically analyzed.

B. Detection of important sentences at different levels of topical detail

The text processing consists of the following steps: (1) detection of the main topic of a sentence; (2) computation of the term distributions of a text; (3) construction of the hierarchical topic tree. The main topic of a sentence is the word or word group that reflects the aboutness or the topical participant of that sentence. The distribution of each term in a text, which gives us information on term frequency, co occurrence and proximity, is computed. Knowing the topics of sentences and term distributions allows us computing topic shifts, nested topics (i.e., subtopics of another topic) and semantic returns (i.e., a topic is suspended at one point and resumed later in the discourse) and in finding the topic tree.

A widely used model for term weighting is the tf-idf method, where the term weight is increased based on its frequency in a document and decreased based on its frequency across the document set. To measure these effects, the term frequency[3] in a document is defined as

$$\text{Tft,d}= \frac{\text{number of occurrence of term t}}{\text{Number of occurrence of term d}}$$

And the inverse document frequency as

$$\text{Idft}= \log \frac{N}{\text{dft}}$$

where N is the total number of documents, and idft is the document frequency, i.e. the number of documents in which the term t occurs.

The hierarchical table of content or topic tree is gradually built and corrected as more evidence becomes available. It indicates the more general and more detailed subtopics of a text. For each topic, the text segment that covers the topic is represented by its boundaries, i.e., it's begin and end positions in the text in terms of character positions. Each topic is described with one or more terms extracted. A topic tree allows zooming in and out into the content of a text. It can be exploited in two different ways for automatic summarization. A selected number of the topical levels can be chosen for the basis of the summary. Or the topics with the highest coverage (i.e., that cover the largest segments as indicated by the character pointers in the text) can be selected. Both approaches are flexible because different levels of topical detail can be chosen. Moreover, the tree allows selecting certain topical viewpoints and computing their salience. As a next step, sentences that introduce the topics, i.e., the first sentence of the corresponding topical segment in the text or phrases that relate the topics can be extracted from the text to form the summary.

C. Sentence compression

For summarization tasks in which brevity is crucial, it is often necessary to reduce the length of these sentences to their relevant content. Two main parameters are considered in the sentence condensation process: (1) syntactic and morph syntactic information; (2) and the presence of important topic terms in the sentence. The aim is to find the semantic relationships between the topic terms. A condensed summary sentence (also called a headline) is constructed by outputting the phrase with the smallest length that spans the largest amount of topic terms in its respective clause. From a linguistic point-of-view, the sentence condensation method results in truncated simple clauses or phrases that indicate a semantic relation between two or more topic terms. The reduction scheme also causes the phrasal context pre-modifying the first and post-modifying the last topic term to be removed, except when they are part of a collocation compound term. When deleting relative clauses without co-reference resolution, we might miss important content clauses that refer to topical entities. However, deletion of embedded clauses seems justified in most cases. When appositive clauses contain sufficient information relevant to the summarization task, they have a fair chance of ending up in the summary as an independent clause. Non-restrictive relative clauses usual function as a gloss of a noun phrase and can be deleted safely. Restrictive relative clauses contain information that affects the interpretation of the main clause.

D. Detection of redundant content

When summarizing multiple documents, the texts should be condensed in a very sizeable way and redundant content should be eliminated. Use the clustering to detect sentences similar in content and select the most representative sentence (medoid) of each cluster of related sentences. It regard nonhierarchical (partitioning) methods that are based on the selection of representative objects (i.e., medoids). A candidate medoid attracts the most similar sentences from the set of remaining sentences based on a criterion or constraint of cluster goodness.

The problem that both algorithms try to solve can be seen as an optimization problem. The mathematical models for the algorithms use the following notations:

- The set of n objects (i.e., sentences) to be clustered is denoted by X ={x1; x2; . . . ; xn}.
- The similarity between objects xi and xj (also called objects i and j) is denoted by s(i,j). It was computed as the cosine between the term vectors of the sentences.
- A solution to the model is determined by two types of decisions:
- The selection of objects as representative objects in clusters: yi is defined as a A–1 variable, equal to 1 if and only if object i is selected as the representative object (i = 1; . . . ; n).
- The assignments of each object j to one of the selected representative objects: zij is a A–1 variable, equal to 1 if and only if object j is assigned to the cluster of which i is the representative object.

By following these steps, Generic summary of the document set is generated.

Topic Focused MDS

In this system, multi document summary is produced based on user's query. The response time of the system is expected to be minimal for practical purpose. Query-based summaries [4] are constructed as an answer to an information need expressed by a user's query, where: Indicative summaries point to information of the document, which helps the user to decide whether the document should be read or not; Informative summaries provide all the relevant information to represent the original document.

A. Approaches based on Document Graphs

Jagadeeshet. al. presents an extractive multi-document summarization method, that represents the documents as graphs [5]. The document graph is produced from a plain text document, by first tokenizing, then parsing it into NPs. The relations are generated following heuristic rules. A centric graph is produced from all source documents and guides the summarizer in its search for candidate sentences to be added to the output summary. The query-based summarization is done in three ways:

a. The centric graph of the documents is compared with the concepts in the query;
b. The graph of the document and a graph of the query are generated and the similarity between each sentence and the query are measured, the best sentences ordered chronologically according to their appearance in the input documents produce the summary;
c. A query modification technique is used by including the graph of a selected sentence to the query graph.

The best results come from summarizer (b).

The method in [6] shows how answers to questions can be improved by extracting more information about the topic with summarization techniques, based on text analysis for query-based single document extracts. The RST (Rhetorical Structure Theory) is used to create a graph representation of the document - a weighted graph in which each node represents a sentence and the weight of an edge represents the distance between two sentences. If a sentence is relevant to an answer, a second sentence is evaluated as relevant too, based on the weight of the path between the two sentences. The approach is of two steps. First the relations between sentences are defined in a discourse graph. Then, a graph search algorithm is used to extract the most salient sentences from the graph for the summary. The sentences with the cheapest path from the entry point are selected.

B. Approaches using linguistics

The approach in [7] is based on HMM (Hidden Markov Model) for sentence selection within a document and a question answering algorithm for generation of a multi-document summary. The developed system CLASSY makes use of linguistics, patterns with lexical cues for sentence and phrase elimination. Typographic cues like title paragraph and other specific paragraphs are used to detect the topic description and obtain question-answering capability. In a separate pre-processing step a named entity identifier ran on all document sets, generates lists of entities for the categories of location, person, date, organization, and evaluates each topic description looking for keywords. After all linguistic processing, and query terms generated, HMM model is used to score the individual sentences classifying them as summary and non-summary sentences.

The approach in [8] is a multi-document summarizer that uses query-interpretation to analyze the given user profile and topic narrative for document clusters before creating the summary. It is based on basic elements, a head modifier relation triple representation of document content which is created by using a parser to produce a syntactic parse tree and a set of 'cutting rules' to extract just the valid basic elements from the tree. Scores are assigned to the sentences based on their basic elements, and then standard filtering and redundancy removal techniques are applied before generating the summaries which consists in outputting the topmost sentences until the required sentence limit is reached.

C. Machine-learning approaches

In the approach of [9] information retrieval techniques are combined with summarization techniques in producing the summary extracts. This approach incorporates a new notion of sentence importance independent of query into the final scoring. The sentences are scored using a set of features from all sentences, normalized in a maximum score and the final score of a sentence is calculated using a weighted linear combination of the individual feature values. The top scoring sentences are selected for the summary until the summary length reaches the desired limit. A new feature - Information Measure - captures the sentence importance based on the distribution of its constituent words in the domain corpus. The formula consists of two parts:

a. Query dependent ranking of a document/sentence;
b. The explicit notion of importance or prior of a document/sentence.

This allows query independent forms of evidence to be incorporated into the ranking process.

Fast Sum [10] is based on word-frequency features of clusters, documents and topics. Summary sentences are ranked by a regression Support Vector Machine. The method involves sentence splitting, filtering candidate sentences and computing the word frequencies in the documents of a cluster, topic description and the topic title. All sentences in the topic cluster are ranked for summarizability. The topic contains a topic title and a topic description. The former is a list of key words or phrases describing the topic, and the later contains the query or queries. The features used are word-based and sentence-based. Word-based features are computed based on the probability of words for the different containers. Sentence-based features include the length and position of the sentence in the document. Because of adopting Least Angle Regression, a new approach for selecting features, Fast Sum can rely on a minimal set of features leading to fast processing times, e.g. 1250 news documents per 60 seconds.

Topic Focused MDS Steps

The overall summarization framework is developed as follows. Theme Cluster Identification groups the sentences in the documents into a number of theme clusters. Sentence Ranking evaluates the significance of the sentences in relation to the given query by propagating the query relevance via the mutual reinforcement between sentences and theme clusters. Sentence Extraction involves redundancy control that removes the sentences containing much duplicate information and chooses the novel sentences to form the summaries.

A. Theme Cluster Identification

K-means Clustering (KC) [11], Spectral Clustering (SC) [12]and Affinity Propagation (AP) [13] are the clustering techniques used. K-means Clustering (KC) is a partition-based clustering algorithm. It randomly selects K sentences as the initial centroids of the K clusters and then iteratively assigns all sentences to the closest cluster and recomputed the centroid of each cluster until the centroids do not change. The similarity between the sentence and the cluster centroid is computed by the standard cosine measure. Spectral Clustering (SC) is a graph based clustering algorithm. It clusters the sentences using the top eigenvectors of graph Laplacian, which is defined on the sentence affinity matrix and then finds the best cut of the graph so that the predefined criterion function can be optimized. The optimized graph cut corresponds to the clusters of the sentences. Many criterion functions, such as the ratio cut [14], the normalized cut [15] and the min-max cut [16] have been proposed along with the corresponding eigen-problem for finding optimal solutions. The above two algorithms require to predefine the cluster number. Affinity Propagation (AP) is different from the above two clustering algorithms in that it does not have to predefine the cluster number. It is also graph based. The algorithm takes each sentence as a vertex in a graph and considers all the vertices as potential exemplars. Then it recursively transmits the real valued messages along edges of the graph until a good set of exemplars and corresponding clusters emerges.

B. Sentence Ranking

Manifold ranking is a semi-supervised learning that explores the relationship among all the data points in the feature space[17].

It has two versions regarding the different tasks:

a. To rank the data points, or
b. To predict the labels of the unlabeled data points.

The manifold-ranking based summarization approach consists of two steps:

1. The manifold-ranking score is computed for each sentence in the manifold-ranking process where the score denotes the biased information richness of a sentence;
2. Based on the manifold-ranking scores, the diversity penalty is imposed on each sentence and the overall ranking score of each sentence is obtained to reflect both the

biased information richness and the information novelty of the sentence.

The sentences with high overall ranking scores are chosen for the summary. The definitions of biased information richness and information novelty are given as below:

Biased Information Richness:

Given a sentence collection χ ={xi | 1<i<n} and a topic T, the biased information richness of sentence xi is used to denote the information degree of the sentence xi with respect to both the sentence collection and T, i.e. the richness of information contained in the sentence xi biased towards T.

Information Novelty:

Given a set of sentences in the summary R={xi | 1<i<m}, the information novelty of sentence xi is used to measure the novelty degree of information contained in the sentence xi, with respect to all other sentences in the set R.. The underlying idea of the proposed approach is that a good summary is expected to include the sentences with both high biased information richness and high information novelty.

C. Sentence Extraction and Redundancy Control

In multi-document summarization, the number of the documents to be summarized can be very large. This makes information redundancy problem appear to be more serious in multi-document summarization than in single-document summarization. Redundancy removal becomes an inevitable process. Since our focus in this study is the design of effective (sentence) ranking algorithms, we propose a sentence selection method which is a derivative of MMR. At the beginning, we choose the first sentence from the ranking list into the summary. Then we exam the next one and compare it with the sentence(s) already included in the summary. Only the sentence that is not too similar to any sentence in the summary (i.e., the cosine similarity between them is lower than a threshold) is selected into the summary. This process is repeated until the length of the sentences in the summary reaches the length limitation. The threshold is set to A.9 in this paper.

Relevance Propagation Model

The manifold-ranking based summarization approach constructs a weighted graph that explicitly represents both query and sentences as vertices. The pre-specified positive ranking score of query is then propagated to nearby vertices via the graph iteratively until a global stable state is achieved. At the end, all the sentences are ranked according to their final scores, with a larger score indicating a higher chance to be extracted. However, this approach performed relevance propagation among homogeneous objects (i.e., sentences). The information beyond the sentence level is totally ignored.

Actually, in a given document set, there usually exist a number of themes (or topics) with each theme represented by a cluster of highly related sentences [18],[19]. The theme clusters are of different size and especially different importance to assist the users in understanding the content in the whole document set. For example, the theme cluster which is relevant to the query is usually more important than the theme cluster which is irrelevant to the given query. So the cluster level information is supposed to have great influence on sentence ranking. Based on the above analysis, we argue that the ranking score of a sentence depends not only on its relevance to the given query, but also on the relevance of its belonging cluster to the query. We apply mutual reinforcement principle to query-focused sentence and theme cluster ranking, i.e., “A sentence should be ranked higher if it is contained inthe theme cluster which is more relevant to the given query while a theme cluster should be ranked higher if it contains many sentences which are more relevant to the given query.”The above principle is similar to the principle by Zhato detect key terms and generic summary sentences [20].

A. RARP ALGORITHM

The Reinforcement After Relevance Propagation(RARP)

Algorithm performs the internal relevance propagation in the sentence set and the cluster set separately until the stable states of both are achieved. The obtained sentence and cluster ranking scores are then updated via external mutual reinforcement until all the scores are converged.

PSEUDO-CODE OF RARP ALGORITHM

Algorithm 1: RARP(S,C,q,Ys,Yc)

Input: Sentence set S, query q, cluster theme set C, vectors Ys andYc,Tre=0.0001

Output: The ranking vectors of Fs an Fc

```
Compute the ranking score vector Fs(0) and Fc(0) of S and C
Construct the matrix Lcs
k=0;δ=0
Repeat
Fs(k+1)=θFs(0) +(1- θ) Lcs Fc(k)
Fc(k+1)=θ Fc(0) +(1- θ) LcsFs(k)
δ=max(Fs(k+1)- Fs(k) ,Fc(k+1)- Fc(k))
```

```
k=k+1
Until $\delta$<Tre
Fs=Fs(k); Fc=Fc(k)
Return
```

B. RDRP ALGORITHM

It alternatively performs one round of internal relevance propagation in the sentence set (or the cluster set), and one round of external mutual reinforcement to update the current ranking scores of the cluster set (or the sentence set). Thewhole process is iterated until an overall global stable state is reached.

PSEUDO-CODE OF RDRP ALGORITHM

Algorithm 2: RDRP(S,C,q,Ys,Yc)

Input: Sentence set S, query q, cluster theme set C, vectors Ys and Yc, Tre=0.0001

Output: The ranking vectors of Fs an Fc

```
Construct the matrix Lcs ,Lss and Lcc
Compute the ranking score vector Fs(0) and Fc(0) of S and C
k=0;$\delta$=0
Repeat
    Fs(k+1)=$\eta$Lss L csFc(k) +(1- $\eta$) Y s(0)
    Fc(k+1)= $\eta$L cc L csFs(k) +(1- $\eta$) Y c(0)
    $\delta$=max(Fs(k+1)- Fs(k) ,Fc(k+1)- Fc(k))
    k=k+1
Until $\delta$<Tre
Fs=Fs(k); Fc=Fc(k)
Return
```

The overall performance of RDRP is superior to that of RARP when employing the same clustering algorithm. The success of RDRP is attributed to its capability of incorporating query relevance into mutual reinforcement.

Comparative Study of MDS Approaches

Multi Document Summarization (MDS) approaches can be classified on the basis of how deep the processing of documents is done to retrieve data.

A. Morphological level

A summarizer is said to work at Morphological Level if it treats documents as a set of words or a bag of words without understanding their meaning. The units of comparison are phrases, sentences or paragraphs. Tf *idf are used for weighting individual term and sentences. All centroid based, MMR based summarization falls into this category. The Dice Coefficient, Jaccard and Cosine similarity is used to find the weight of term in the documents and their similarity Salton matched paragraphs within and across document in terms of similarity metric. The similarity metric used was the Cosine Similarity Coefficient. Paragraphs we reconnected to many other paragraphs with a similarity above threshold were considered salient, since they would likely contain topics discussed in other documents. This approach is prune to using up compression very quickly because of the emphasis on paragraph. It does not address redundancy problem characteristics of MDS.

Ando et al has used a vector space model, which takes advantage of the method similar to Latent Semantic Analysis, to reduce the dimensionality of the vector space. Analysis is to derive semantic similarity between terms based on their occurrences in common contexts. Carbonell et al [21] used Maximal Marginal Relevance(MMR) for ranking the passages. In query focused

MDS, the passages are drawn from multiple documents and ranked. The redundancy is controlled by a single parameter. For maximum diversity. This parameter is set to zero.

Mani and Bloedorn [22] take into account the cohesion relations among terms in topic based MDS. The relations include proximity, co-reference, synonymy and hyponymy.

It is very robust, but falls short of dealing with redundancy because only some case of information equivalence will be caught by similarity metric. Certainly if two passages were being compared, but whenever one passage use content word, the other passage use synonym for it, the passage would come out as

C. Syntactic level

A summarizer may compare text units across documents using their syntactic phase. Here each sentence is compared with syntactic paraphrase of the other.

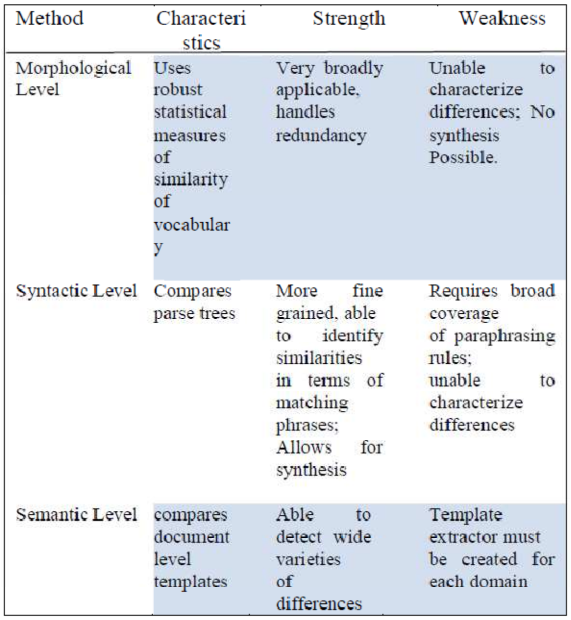
Barzilay et al[23] discuss number of paraphrasing rules including active versus passive forms, omission of head NP in pseudo partiative, ordering of syntactic components in sentence, classifier vs appositive expressions. 85% of paraphrasing was achieved by syntactic and lexical transformations alone, indicating that surface transformations accounted for much of the dissimilarity between informationally equivalent sentences. Typically, language generation systems have access to a full semantic representation of the domain. A content planner selects and orders propositions from an underlying knowledge base to form text content. A sentence planner determines how to combine propositions into a single sentence, and a sentence generator realizes each set of combined propositions as a sentence, mapping from concepts to words and building syntactic structure.

In this level, content planning operates over full sentences, producing sentence fragments. Thus, content planning straddles the border between interpretation and generation. Sentence generation begins with phrases. Our task is to produce fluent sentences that combine these phrases, arranging them in novel contexts. In this process, new grammatical constraints may be imposed and paraphrasing may be required.

D. Semantic Level

A summarizer is said to work in semantic level when they consider named entities, relations among them and events. They identify semantic level elements in each document; these elements are then matched to provide semantic level similarities and differences.

Kathleen McKeon and Dragomir R. Radev[24],[25]introduces a methodology for developing summarization systems, identifies planning operators for combining information in a concise summary, and uses empirically collected phrases to mark summarized material. The content planner (i.e., the module which determines what information to include in the summary) and the linguistic component (i. e.,the module which determines the words and surface syntactic form of the summary) of our system is developed. Operators which are used to combine information is identified; which includes techniques for linking information together in a related way (e. g., identifying changes, similarities, trends) as well as making generalizations. Phrases that are used to mark summaries are identified and used these to build the system lexicon.



Method	Characteristics	Strength	Weakness
Morphological Level	Uses robust statistical measures of similarity of vocabulary	Very broadly applicable, handles redundancy	Unable to characterize differences; No synthesis Possible.
Syntactic Level	Compares parse trees	More fine grained, able to identify similarities in terms of matching phrases; Allows for synthesis	Requires broad coverage of paraphrasing rules; unable to characterize differences
Semantic Level	compares document level templates	Able to detect wide varieties of differences	Template extractor must be created for each domain

Conclusion

This paper presented an overview of a variety of multi document summarization. Both Generic and Query-based summarization approaches implemented in different levels were covered. Generic summarization provides general summary of the multi-documents while Query based

summarization provides summary based on specific query. Then we provide comparison of different approaches of MDS.It is found that semantic level approach is more stronger than other levels

References

[1] D. Wang, S. H. Zhu, T. Li, Y. Chi, and Y. H. Gong, "Integrating document clustering and multi document summarization," ACM Trans. Knowl. Disc. From Data, vol. 5, no. 3, 2011.

[2] Marie-Francine Moens , Roxana Angheluta, Jos Dumortier," Generictechnologies for single- and multi-document summarization," InformationProcessing and Management ,vol 41, pp 569–586,2005

[3] Katariina Nyberg ,Tapani Raiko, Teemu Tiinanen, Eero Hyvönen,"Document classification utilising ontologies and relations betweendocuments" ACM Proceedings of the Eighth Workshop on Mining andLearning with Graphs Pages 86-93,2010

[4] Damova Mariana; Koychev Ivan ," Query-Based Summarization: Asurvey," In Proceedings of S3T'2010, Varna, Bulgaria, September, 2010, pp.142-147

[5] Ahmed A. Mohamed, Sanguthevar Rajasekaran (2006). Query-BasedSummarization Based on Document Graphs. In Proceedings of IEEEInternational Symposium on Signal Processing and Information Technology,pp.408-410, Vancouver,Canada, 2006

[6] Wauter Bosma (2005). Query-Based Summarization using RhetoricalStructure Theory. In: Ton van der Wouden, Michaela Poss, H. Reckman andC. Cremers, ed.,15th Meeting of CLIN, 2005

[7] John M. Conroy, Judith D. Schlesinger, Jade Goldstein Stewart (2005).CLASSY Query-Based Multi-Document Summarization. In DUC 05Conference Proceedings, Boston, USA

[8] Koychev I., Nikolov, R. and Dicheva D.: SmartBook: The NewGeneration e-Book, Proc. of BooksOnline'09 Workshop, in conjunction withECDL 2009, Corfu,October 2, 2009

[9] Jagadeesh J, Prasad Pingali, Vasudeva Varma . Capturing Sentence Priorfor Query-Based Multi-Document Summarization. In RIAO, http://dblp.unitrier.de,2007

[10] Frank Schilder, Ravikumar Kondadadi," FastSum: Fast and accuratequerybased multi-document summarization." In Proceedings of the 46thmeeting of the Association for Computational Linguistics, Columbus, Ohio

[11] K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review,"ACMComput. Surv., vol. 31, no. 3, pp. 264–323, Sep. 1999.

[12] U. V. Luxburg, "A tutorial on spectral clustering," Statist. Comput., vol.17, no. 4, pp. 395–416, Aug. 2007.

[13] J. F. Bredan and D. Delbert, "Clustering by passing messages betweendata points," Science, vol. 315, no. 5814, pp. 972–976, Jan. 2007.

[14] L. Hagen and A. B. Kahng, "New spectral methods for ratio cutpartitioning and clustering," IEEE Trans. Comput. Aided-Design Integr.Circuits Syst., vol. 11, no. 9, pp. 1074–1085, Sep. 1992.

[15] J. Shi and J. Malik, "Normalized cuts and image segmentation," in Proc.10th CVPR Conf., 1997, pp. 888–905. [16] H. Q. Ding, X. F. He, H. Y. Zha, M. Gu, and H. Simon, "A min-max cutfor graph partitioning and data clustering," in Proc. 1st ICDM Conf.,2001, pp.107–114.

[16] X. J.Wan, J. W. Yang, and J. G. Xiao, "Manifold-ranking based topic focused multi-document summarization," in Proc. 18th IJCAI Conf.,2007, pp. 2903–2908.

[17] S. Harabagiu and F. Lacatusu, "Topic themes for multi-document summarization," in Proc. 28th SIGIR Conf., 2005, pp. 202–209.

[18] H. Hardy, N. Shimizu, T. L. Ting, G. B. Wise, and X. Zhang, "Cross document summarization by concept classification," in Proc.25th SIGIRConf., 2002, pp. 121–128.

[19] H. Y. Zha, "Generic summarization and key phrase extraction using mutual reinforcement principle and sentence clustering," in Proc. 25th SIGIRConf., 2002, pp. 113–120.

[20] J. Carbonell and J. Goldstein. The use of mmr: Diversity-based rerankingand reordering documents and producing summaries. In The Proceedings of21st Annual International ACMSIGIR Conference on Research and Development in Information Retrieval, pages 335–336,1998.

[21] Inderjeet Mani and Eric Bloedorn. Multi-document summarization bygraph search and matching.1997.

[22] Regina Barzilay and Michael Elhadad. Using Lexical Chains for Text Summarization. In Inderjeet Mani and Mark T. Maybury, editors, Advances inAutomatic Text Summarization, pages 111–121. The MIT Press, 1999.

[23] Kathleen R. McKeown and Dragomir R. Radev. Generating summaries of multiple news articles.In Proceedings, ACM Conference on Research and Development in Information Retrieval SIGIR'95, pages 74–82, Seattle, Washington, July 1995

[24] Kathleen R. McKeown, Regina Barzilay, David Evans, VasileiosHatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, BarrySchiffman, and Sergey Sigelman. Tracking and summarizing news on a dailybasis with columbias newsblaster. In Human Language Technology.