Multi Document Summarization Based On Cross-Document Relation Using Voting Technique

Yogan Jaya Kumar
Faculty of Information and Communication Technology
University Teknikal Malaysia Melaka
Melaka, Malaysia

Naomie Salim, Albaraa Abuobieda, Ameer Tawfik
Faculty of Computing
University Teknologi Malaysia
Skudai, Johor

***Abstract*—News articles which are available through online search often provide readers with large collection of texts. Especially in the case of news story, different news sources reporting on the same event usually returns multiple articles in response to a reader's search. In this work, we first identify cross-document relations from un-annotated texts using Genetic-CBR approach. Following that, we develop a new sentence scoring model based on voting technique over the identified cross-document relations. Our experiments show that incorporating the proposed methods in the summarization process yields substantial improvement over the mainstream methods. The performances of all methods were evaluated using ROUGE—a standard evaluation metric used in text summarization.**

***Index Terms*—Multi document summarization, Cross-document relation, Machine learning, Case-based reasoning, Genetic algorithm, Voting Technique.**

I. Introduction

As far as text summarization is concerned, many related research studies have been reported in academia [1-3]. Mainly, the studies revolve around extractive summarization: the important sentences are identified and directly extracted from the document, i.e. the final summary consists of original sentences. Generally, statistical and linguistic features of sentences are used to determine the importance of sentences.

Another concern which arises along is the size of the texts collection which needs to be summarized. For example, online news surfing provides readers with many articles related to a particular event; as it involves multiple news sources. Thus, the need for multi document summarization is deemed necessary for condensing the multi source texts into a shorter version.

Of course, documents which are related to the same topic usually contain semantically related textual units. Inspired by this fact, in this paper, we investigate the utility of cross-document relations for identifying highly relevant sentences to be included in the summary. The study on cross-document relations can be associated with Radev, who claimed that inter-document relationships can be based on CST (Cross-document Structure Theory) model [4]. The CST model describes semantically related textual units such as words, phrases or sentences from topically related documents.

In this work, we first describe an efficient, supervised learning method for identifying the relations between sentences directly from un-annotated documents. Our technique incorporates the integration of the genetic learning algorithm and the case base reasoning (CBR) model that is tailored to the task of classification. Then, based on the identified cross-document relations, we implement the proposed voting technique to the sentence scoring model to select the highly ranked summary sentences.

II. Related Works

If we look back at previous approaches concerning text summarization, we can observe that there are two methods which are relatively common in multi document summarization studies, namely the cluster based method and graph based method. The cluster based method which was pioneered by Radev et al. uses clustering technique, to generate sentence clusters [5]. High ranking sentences from each cluster are then selected to be included in the summary.

For the graph based method, its fundamental theory is supported by the links that exist between sentences. These links exist based on some measured similarity between the sentences. Sentences with high similarity weights (with respect to other sentences in the documents) will be ranked top for summary sentence selection. A popular graph based ranking algorithm is Google's PageRank which has been traditionally used in Web-link analysis and social networks [6].

As stated earlier in Section 1, the CST model defines the cross-document relations that exist between topically related documents. Following this, a number of researchers have addressed the benefits of CST for summarization task. In the work presented by Zhang et al., they replace low-salience sentences with sentences that maximize total number of CST relations in the final summary [7]. Similarly, Jorge and Pardo worked on CST relations for content selection methods to produce preference-based summaries [8]. However the major limitation of the above works is that the CST relations need to be manually annotated by human experts; which is a drawback for an automatic summarization system.

Our work, in contrast, treats this limitation by identifying the relations between sentences directly from un-annotated

978-1-4673-6232-0/13/$31.00 ©2013 IEEE

documents. Moreover, our voting model is designed to rank sentence based on the identified cross-document relation.

Although there have been some attempts to learn the CST relations in texts, to our knowledge, only two interrelated works with evaluations were based on English texts, where the authors applied boosting classification algorithm to identify the presence of CST relations between sentences [9, 10]. However their classifier showed poor performance in classifying most of the CST relations; obtaining average values of 45% precision, 31% recall, and 35% f-measure.

Besides English texts, CST parsing had also been studied for Brazilian Portuguese texts and Japanese texts [11, 12]. For Brazilian text, the authors experimented with three types of classifiers namely, the multi-class, hierarchical, and binary classifiers; and obtained a general accuracy of 41.58%, 61.50% and 70.51% respectively on unbalanced data. For Japanese text, the authors however attempted only two relations, i.e. *Equivalence* and *Transition* and obtained an f-measure of 75.50% and 45.64% respectively using a SVM classifier.

In our study, we believe that the performance of the classifier should be promising enough in order to see its impact in the summarization system. This is essential because the performance of the classifier would certainly have direct implication on the final results of the system.

III. Overview Of Approach

In this section, we present the overall architecture of our proposed approach. As highlighted in Fig. 1, there are two main phases; which includes cross-document relation (CST relation) identification and sentence scoring using voting technique. First we describe the cross-document relation identification. The voting technique implementation will be described later in Section IV. As we mentioned earlier in this paper, we will investigate the utility of cross-document relations or CST relations for identifying highly relevant sentences to be included in the summary. We have considered four types CST relations, namely *Identity*, *Subsumption*, *Description* and *Overlap*; as they cover most of the other relations in the CST model. Refer to Table 1.

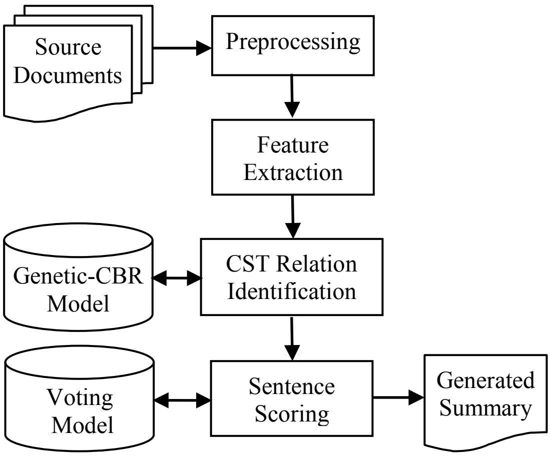

Fig. 1. General architecture of the proposed approach.

TABLE I. Description of CST relations used in this work.

Relations	Description
Identity	The same text appears in more than one location.
Subsumption	S1 contains all information in S2, plus additional information not in S2.
Description	S1 describes an entity mentioned in S2.
Overlap (partial equivalence)	S1 provides facts X and Y while S2 provides facts X and Z.

Relying on manually annotated text for cross-document relation identification can consume time and resources. This has motivated us to automatically identify the four aforementioned relations to facilitate our multi document summarization task. Here, we propose a Genetic-CBR approach for the identification task.

"Case-Based Reasoning (CBR) is the usual name given to problem solving methods which make use of specific past experiences. It is a form of problem solving by analogy in which a new problem is solved by recognizing its similarity to a specific known problem, then transferring the solution of the known problem to the new one" [13]. We could also regard CBR as a type of supervised learning method as it finds solutions for new problems based on existing solutions.

The general process of CBR consists of four major phases, namely *Retrieve*, *Reuse*, *Revise*, and *Retain* that links to a central repository called the casebase [14]. When a new case (problem) is received, the CBR model will first retrieve the most similar cases from the casebase (where previous solved cases are stored) and the solution from the retrieved cases will be reused for the new case. If no similar cases are found in the casebase, the solution for the new case will be revised and retained into the casebase as a new solved case.

Each case in our casebase represents an example of sentence pair with its known cross-document relationship type. Next we describe the features that represent each sentence pair:

Cosine similarity – cosine similarity is used to measure how similar two sentences (*S*) are. Here the sentences are represented as word vectors with *tf-idf* as its element (*i*) value:

$$cos(S_1,S_2)=\frac{\sum S_{1,i}\cdot S_{2,i}}{\sqrt{\sum(S_{1,i})^2}\cdot\sqrt{\sum(S_{2,i})^2}}. \tag{1}$$

Word overlap – this feature represents the measure based on the number of overlapping words in the two sentences. This measure is not sensitive to the word order in the sentences:

$$overlap(S_1,S_2)=\frac{\#commonwords(S_1,S_2)}{\#words(S_1)+\#words(S_2)}. \tag{2}$$

Length type – this feature gives the length type of the first sentence when the lengths of two sentences are compared.

$$\begin{aligned} lengtype(S_1) = 1 \quad & \text{if} \quad length(S_1) > length(S_2), \\ \text{-1} \quad & \text{if} \quad length(S_1) < length(S_2), \\ 0 \quad & \text{if} \quad length(S_1) = length(S_2). \end{aligned} \tag{3}$$

NP similarity – this feature represents the noun phrase (NP) similarity between two sentences. The similarity between the NPs is calculated according to Jaccard coefficient as defined as in the following equation:

$$NP(S_1, S_2) = \frac{NP(S_1) \cap NP(S_2)}{NP(S_1) \cup NP(S_2)}. \tag{4}$$

VP similarity – this feature represents the verb phrase (VP) similarity between two sentences. The similarity between the VPs is calculated according to Jaccard coefficient as defined as in the following equation:

$$VP(S_1, S_2) = \frac{VP(S_1) \cap VP(S_2)}{VP(S_1) \cup VP(S_2)}. \tag{5}$$

To determine the relationship type for a new case, the model will compare the feature vector of the new case with existing cases in the casebase. If the similarity value is less than the threshold value, the model will revise the new case solution as “No relation” and retain the revised case into the casebase. In our implementation, we propose a weighted cosine similarity measure to compute the similarity between two cases; Eq. 6. An example is given in Table 2.

$$wcos(X, Y) = \frac{\sum_{k=1}^{5} w_k x_k \times w_k y_k}{\sqrt{\sum_{k=1}^{5} (w_k x_k)^2} \times \sqrt{\sum_{k=1}^{5} (w_k y_k)^2}}. \tag{6}$$

In order to obtain the feature weights, we have integrated feature weighting using genetic algorithm. The details on the genetic learning process can be found in our previous work [15]. Fig. 2 illustrates the overall process flow of Genetic-CBR.

TABLE II. AN EXAMPLE OF SIMILARITY MEASURE BETWEEN CASES.

Input features	**New case**	**Case 1**	**Case 2**
Cosine Similarity	0.63	0.23	0.44
Word Overlap	0.51	0.36	0.34
Length Type	1	0	1
NP Similarity	0.42	0.27	0.55
VP Similarity	0.47	0.16	0.36
Similarity with new case		0.68	0.97

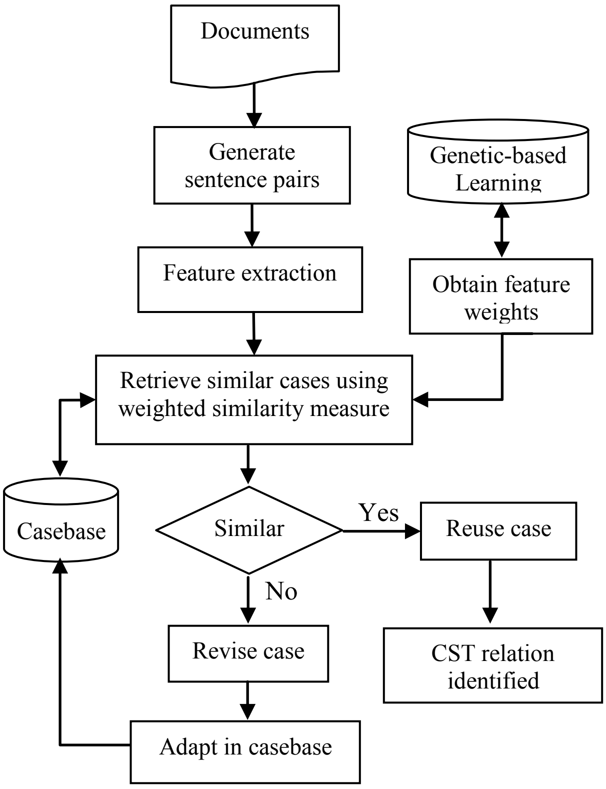

Fig. 2. Genetic-CBR approach for cross-document relation identification.

IV. VOTING TECHNIQUE APPROACH

In this section, we will introduce a new sentence scoring mechanism based on voting technique. Voting techniques were first proposed for expert search task [16]. Expert search has been part of the retrieval task in the Text Retrieval Conferences (TREC) Enterprise tracks since 2005, as a platform to evaluate expert search approaches [17]. Recently, voting techniques have also been employed for thread retrieval in online forum [18]. In this work, we incorporate a novel adaptation of the voting technique to score the sentences based on the cross-document relations identified by our Genetic-CBR classifier.

As described earlier, our work considers four types of cross-document relations, i.e. identity, subsumption, description and overlap. From the document set, we can build an affinity (adjacency) matrix ***M*** representing the relations between sentences s_i and s_j. ***M*** is defined as follows:

$$M_{i,j} = \begin{cases} \text{Rel}_{sen}(s_i, s_j), & \text{if i} \neq j, \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $\text{Rel}_{sen}(s_i, s_j)$ specifies the relationship type between sentence s_i and s_j; i.e. either identity (*I*), subsumption (*S*), description (*D*) and overlap (*O*). The initial score for each sentence is given as follows:

$$InitialScore\left(s_i\right) = \frac{\left|M_i\right|}{\left|S\right|-1} \tag{8}$$

where S is the set of sentences in the document set. Next we show a simple example for computing the initial score of sentences; refer to Fig. 3. Given the affinity matrix $\boldsymbol{M}$, representing the relations between sentences in a document set (with $|S|= 5$), the cumulative sum of relations for each sentence is computed first. Then the initial score for each sentence is obtained using Eq. 8. The diagonal values are all set to zero as is represents a reflexive relation; i.e. the sentences are related to themselves. For instance, in Fig. 3, the initial score for sentence 1, s_1 is 0.75 and the initial score for sentence 3, s_3 is 0.25.

$$M = \begin{array}{c|ccccccc} & s_1 & s_2 & s_3 & s_4 & s_5 & \underline{cum.\ sum} & \underline{ini.\ score} \\ \hline s_1 & 0 & \boldsymbol{S} & 0 & \boldsymbol{D} & \boldsymbol{O} & 3 & 3/4 \\ s_2 & \boldsymbol{D} & 0 & \boldsymbol{S} & \boldsymbol{I} & 0 & 3 & 3/4 \\ s_3 & 0 & 0 & 0 & 0 & \boldsymbol{O} & 1 & 1/4 \\ s_4 & \boldsymbol{S} & \boldsymbol{I} & 0 & 0 & 0 & 2 & 2/4 \\ s_5 & \boldsymbol{O} & 0 & \boldsymbol{O} & 0 & 0 & 2 & 2/4 \end{array}$$

Fig. 3. An example of initial score computation.

In our work, we look at the directionality of the relations as the basis to vote a sentence. To vote a sentence, we consider two relations; subsumption and description, since their directionality is 1-way (while the others having 2-way direction). Based on these two cross-document relations, we set two conditions to vote the sentences:

Condition 1: If s_i subsumes s_j, then we vote s_i; as it contains all the information in s_j plus additional information not in s_j.

Condition 2: If s_i describes s_j, then we vote s_j; as an entity in s_j is being described by another sentence, thus reflecting its relevance.

The voting technique proposed for this work is based on Eq. 9 below. Using voteCombMAX – our score aggregation technique – the initial score of a sentence is updated based on the two abovementioned conditions.

$$score\_sentence_{voteCombMAX} = InitialScore + max\{vote\} \tag{9}$$

We consider two forms of evidence to aggregate the votes of sentence s_i; first, the number of sentences related to sentence s_i; and second, the maximum of scores of sentences voting for sentence s_i. If a sentence receives votes from both the abovementioned conditions, then the average of its maximum scores will be taken. At the same time we also filter the sentences which are being subsumed by other sentences. Note that the filtered sentences are ignored at this stage.

Once the scores of all sentences have been computed, we re-rank the sentences using their updated scores. We keep sentences that pass the cutoff length greater than 9 and eliminate redundant sentences based on a cutoff similarity value of 0.7. Finally, high ranking sentences are selected until the desired summary length is met.

V. Experiments And Results

We first perform feature weighting to find the optimal weights for the features. The optimal weights obtained were 0.18374, 0.94211, 0.81638, 0.61879 and 0.00631, representing the weights for cosine similarity, length type, word overlap, noun phrase similarity and verb phrase similarity, respectively [15]. We then use these results for the weighted cosine similarity function in our Genetic-CBR classifier. The performance of our classifier has been reported earlier in [15]. We used the evaluation measures commonly used in classification tasks – Precision, Recall and F-measure. The Genetic-CBR classifier obtained good classification results (with average 85.76% precision, 84.02% recall and 84.47% f-measure) It also performed better than neural network (NN) and support vector machine (SVM) which are two popular machine learning techniques commonly used for classification tasks [19].

Then, using the optimized Genetic-CBR classifier together with the sentence voting technique, we evaluated our proposed summarization model. We use the Document Understanding Conference (DUC) 2002 document sets (D061j, D062j, D073b, D077b, D079a, D083a, D085d, D089d, D091c, D092c, D097e, D103g, D109h and D115i) corresponding to natural disaster news stories. The evaluation results were obtained using ROUGE: Recall-Oriented Understudy for Gisting Evaluation [20]. ROUGE measures the quality of a system generated summary by comparing it to a human model summary. As stated in the literature, there are two mainstream approaches towards multi document summarization tasks, i.e. using cluster based method and graph based method. We built the comparison models based on these two methods. For the graph based method, we rank the sentences using the popular PageRank algorithm. The cluster based method employs the widely used k-means clustering algorithm to generate clusters; sentences in each cluster are then ranked similar to graph based scoring.

Table 3-6 shows the comparison between the proposed model (using voting technique) and the other methods based on ROUGE measures. Fig. 4-7 visualizes the results. The findings demonstrate that the proposed model achieved highest score among all comparison models (H1 is excluded for this comparison as it is a human benchmark and was expected to give best results). We believe that considering only the similarity between sentences (as in graph based method) as evidence will not provide good ranking for the sentences. We also need to consider the type of relations that exist among them and rank them based on those relations. For example descriptive sentences are considered less important to be included in a summary, but sentences that are described by other sentences in the documents are considered important.

TABLE III. Summarization results comparison based on average recall, precision and F-measure using ROUGE-1.

Method	AVG-R	AVG-P	AVG-F
H1	0.39419	0.39402	0.39283
Voting	0.30977	0.31422	0.31077
Cluster Based	0.28493	0.29736	0.28995
Graph Based	0.28885	0.30672	0.29627

TABLE IV. Summarization results comparison based on average recall, precision and F-measure using ROUGE-2.

Method	AVG-R	AVG-P	AVG-F
H1	0.18393	0.1838	0.18332
Voting	0.10893	0.1093	0.10874
Cluster Based	0.08528	0.08876	0.08667
Graph Based	0.07537	0.07917	0.0769

TABLE V. Summarization results comparison based on average recall, precision and F-measure using ROUGE-S.

Method	AVG-R	AVG-P	AVG-F
H1	0.1433	0.14312	0.14149
Voting	0.08399	0.08553	0.08353
Cluster Based	0.06924	0.07384	0.07049
Graph Based	0.06815	0.07581	0.07065

TABLE VI. Summarization results comparison based on average recall, precision and F-measure using ROUGE-SU.

Method	AVG-R	AVG-P	AVG-F
H1	0.14744	0.14725	0.1456
Voting	0.08776	0.08942	0.08732
Cluster Based	0.07284	0.07774	0.07419
Graph Based	0.07183	0.0799	0.07447

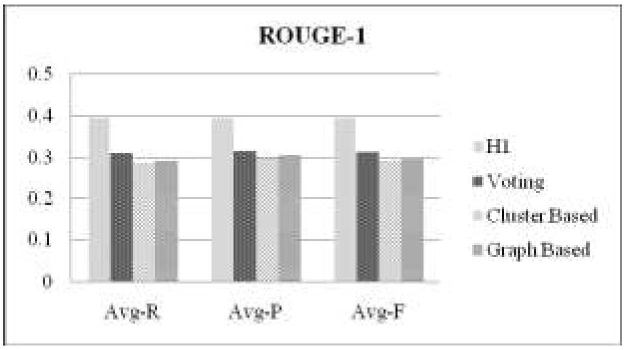

Fig. 4. Summarization results comparison based on average recall, precision and f-measure using ROUGE-1.

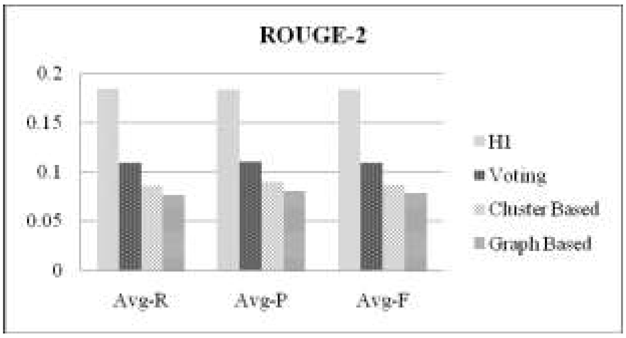

Fig. 5. Summarization results comparison based on average recall, precision and f-measure using ROUGE-2.

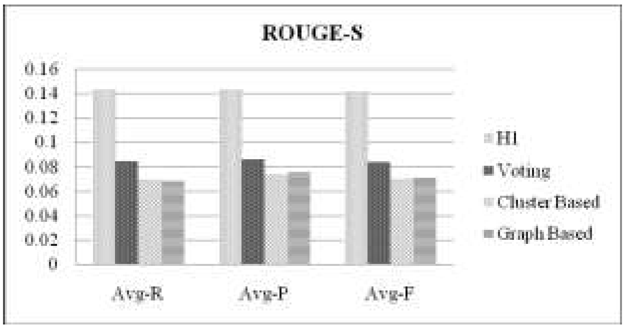

Fig. 6. Summarization results comparison based on average recall, precision and f-measure using ROUGE-S.

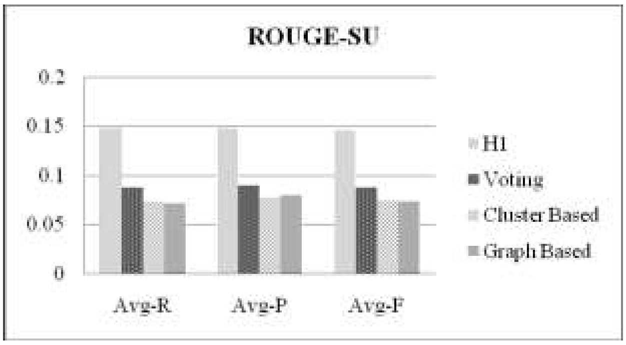

Fig. 7. Summarization results comparison based on average recall, precision and f-measure using ROUGE-SU.

VI. Conclusion

In this paper, we have proposed a multi document summarization model by investigating the utility of cross-document relations (CST relations) to identify highly relevant sentences to be included in the summary. In literature, CST related summarization studies were based on manually annotated relations by human experts. In this work we have filled this gap by automatically identifying the CST relations between sentences from un-annotated text documents. We achieve this by using a classifier named Genetic-CBR which integrates genetic learning algorithm to the case base reasoning model. The proposed classifier obtained good classification results (with average 85.76% precision, 84.02% recall and 84.47% f-measure); making it promising to be integrated into our summarization model. Following that, we develop a new sentence scoring model based on voting technique over the CST relations identified by our classifier. Here, we vote the sentences based on the type of relations they hold with other sentences.

The overall performance of our proposed model was evaluated using the dataset obtained from DUC 2002 whereby its performance was assessed using four ROUGE measures. We also made comparisons with the mainstream methods: cluster based method and graph based method. The experimental findings showed that the proposed model produced better results.

ACKNOWLEDGMENT

This research is supported by Universiti Teknikal Malaysia Melaka (UTeM), Ministry of Higher Education (MOHE) Malaysia and Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under Research University Grant Category (VOT Q.J130000.7826 .4F011)

REFERENCES

[1] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques," Journal of Emerging Technologies in Web Intelligence, vol. 2, pp. 258-268, 2010.

[2] Y. J. Kumar and N. Salim, "Automatic multi document summarization approaches," Journal of Computer Science, vol. 8, pp. 133-140, 2011.

[3] A. Nenkova and K. McKeown, "Automatic summarization," Foundations and Trends in Information Retrieval, vol. 5, pp. 103-233, 2011.

[4] D. R. Radev, "A common theory of information fusion from multiple text sources step one: cross-document structure," presented at the Proceedings of the 1st SIGdial workshop on Discourse and dialogue - Volume 10, Hong Kong, 2000.

[5] D. R. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents," Information Processing & Management, vol. 40, pp. 919-938, 2004.

[6] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," Computer networks and ISDN systems, vol. 30, pp. 107-117, 1998.

[7] Z. Zhang, S. Blair-Goldensohn, and D. R. Radev, "Towards CST-enhanced summarization," presented at the Eighteenth national conference on Artificial intelligence, Edmonton, Alberta, Canada, 2002.

[8] M. L. d. R. C. Jorge and T. A. S. Pardo, "Experiments with CST-based multidocument summarization," presented at the Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, Uppsala, Sweden, 2010.

[9] Z. Zhang, J. Otterbacher, and D. Radev, "Learning cross-document structural relationships using boosting," presented at the Proceedings of the twelfth international conference on Information and knowledge management, New Orleans, LA, USA, 2003.

[10] Z. Zhang and D. Radev, "Combining labeled and unlabeled data for learning cross-document structural relationships," presented at the Proceedings of the First international joint conference on Natural Language Processing, Hainan Island, China, 2005.

[11] M. L. d. R. C. Jorge and T. A. S. Pardo, "Automatic identification of multi-document relations," presented at the PROPOR 2012 PhD and MSc/MA Dissertation Contest, 2012.

[12] Y. Miyabe, H. Takamura, and M. Okumura, "Identifying cross-document relations between sentences," presented at the 3rd International Joint Conference on Natural Language Processing, 2008.

[13] R. Bareiss, Exemplar based knowledge acquisition: a unified approach to concept representati on, classification, and learning: Academic Press Professional, Inc., 1989.

[14] A. Aamodt and E. Plaza, "Case-based reasoning: foundational issues, methodological variations, and system approaches," AI Commun., vol. 7, pp. 39-59, 1994.

[15] Y. J. Kumar, N. Salim, and A. Abuobieda, "A Genetic-CBR approach for cross-document relationship identification," in Advanced Machine Learning Technologies and Applications, ed: Springer, 2012, pp. 182-192.

[16] C. Macdonald and I. Ounis, "Voting techniques for expert search," Knowl. Inf. Syst., vol. 16, pp. 259-280, 2008.

[17] N. Craswell, A. P. de Vries, and I. Soboroff, "Overview of the TREC 2005 Enterprise Track," presented at the Trec, 2005.

[18] A. T. Albaham and N. Salim, "Adapting voting techniques for online forum thread retrieval," in Advanced Machine Learning Technologies and Applications, ed: Springer, 2012, pp. 439-448.

[19] S. B. Kotsiantis, "Supervised machine learning: a review of classification techniques," presented at the Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies, 2007.

[20] C.-Y. Lin, "ROUGE: a package for automatic evaluation of summaries," 2004.