Update Summarization using Semi-Supervised Learning Based on Hellinger Distance

Dingding Wang
Department of Computer & Electrical Engineering and Computer Science
Florida Atlantic University
777 Glades Road
Boca Raton, FL 33431
wangd@fau.edu

Sahar Sohangir
Department of Computer & Electrical Engineering and Computer Science
Florida Atlantic University
777 Glades Road
Boca Raton, FL 33431
ssohangir2014@fau.edu

Tao Li
School of Computing and Information Sciences
Florida International University
11200 SW 8 Street
Miami, FL 33199
taoli@cs.fiu.edu

ABSTRACT

Update summarization aims to generate brief summaries of recent documents to capture new information different from earlier documents. In this paper, we propose a new method to generate the sentence similarity graph using a novel similarity measure based on Helliger distance and apply semi-supervised learning on the sentence graph to select the sentences with maximum consistency and minimum redundancy to form the summaries. We use TAC 2011 data to evaluate our proposed method and compare it with existing baselines. The experimental results show the effectiveness of our proposed method.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Update Summarization; Semi-Supervised Learning; Hellinger Distance

1. INTRODUCTION

Popular online publishers and social media users produce huge amount of text data every day, so it is critical to extract the most important and up-to-date information to help users quickly understand these documents. Thus update document summarization has been receiving more and more attention, which aims to generate a query-relevant summary of multiple articles, under the assumption that the user has already read the earlier articles. Update summarization provides a useful way to make users stay in the know of developing and evolving events. For example, in the event of the spread of the Ebola virus in 2014, the earlier articles reported the Ebola outbreak in West Africa, then there was news about the infected cases found in Europe and the US, and later reports showed that no new cases was diagnosed in the US after December 2014. In this event, since people highly concerned about the development of disease control, a timely updated event summary will help people understand the situation quickly and conveniently.

The problem of update summarization was introduced in Document Understanding Conference (DUC) by National Institute of Standards and Technology (NIST) in 2007 and was a main task of the summarization track in Text Analysis Conference (TAC) through 2008 $\sim$ 2011. Given a topic q, it is required to summarize a set of document B under the assumption that the reader has already read and summarized an earlier set of documents A. Both the summaries of document sets A and B should focus on the given topic and the summary of B must be the least redundant with the summary of A.

The existing research on update summarization mainly focuses on query-relevant sentence ranking, graph optimization, and model-based analysis using cosine similarity [12]. For example, Boudin et al. [1] used Maximal Marginal Relevance (MMR) to rank the sentences and selected the top-ranked sentences to form the summaries. Delort et al. [2] proposed a topic model to identify the novelty in the document collection. Shen et al. [11] applied a minimum dominating set approximation to find the most important sentences on the sentence similarity graph. Wan [15] proposed a co-ranking algorithm to solve the problem. Li et al. [6] proposed a complex three-level hierarchical dirichlet process model to select sentences. Wang et al. [16] proposed an incremental hierarchical clustering based summarization approach to update summaries in real time.

In this paper, we propose a new similarity measure which is based on Hellinger distance to better capture the sentence relationships than the Euclidean distance based cosine similarity. We apply the MMR strategy to generate the summary for the earlier document set, and then propose a label propagation approach using the Green's function to determine the sentence importance in the later document set. The redundancy in the update summary is also eliminated in the final selection procedures. In the experiments, we use TAC 2011 dataset to evaluate the proposed method and compare the results with existing update summarization systems.

CIKM'15, October 19–23, 2015, Melbourne, Australia.

DOI: http://dx.doi.org/10.1145/2806416.2806628.

2. THE PROPOSED METHODOLOGIES

2.1 Hellinger Distance Based Similarity Calculation

In IR, each document is represented by a nonnegative vector: $x = (x_1, x_2, \ldots, x_m)^T$ where x_t relates to the frequency of term t in the document and m is the size of the vocabulary. Because x is nonnegative, we may review them as probability and normalize each document to $\sum_{i=1}^m x_i = 1$.

One of the most widely used similarities between two documents x and y is the cosine similarity, which can be directly derived from Euclidean distance as follows. Assuming each document is normalized to 1 in L_2 norm: $\sum_{i=1}^m x_i^2 = 1$. The Euclidean distance between x and y is

$$d_{\text{Euclid}}(x,y) = \Big[\sum_i (x_i - y_i)^2\Big]^{\frac{1}{2}} = \Big[2 - 2\sum_i x_i y_i\Big]^{\frac{1}{2}} \quad (1)$$

The constant 2 here is unimportant. From Eq.(1), the Euclidean distance corresponds directly to the cosine similarity in Eq.(2).

$$\cos(x,y) = \frac{\sum_{i=1}^m x_i y_i}{\sqrt{\sum_{i=1}^m x_i^2}\sqrt{\sum_{i=1}^m y_i^2}}. \quad (2)$$

Since the word-document associations are nonnegative, it is expected that they are better treated with probabilistic approaches. However Euclidean distance is generally not a good metric for dealing with probabilities. Thus in this paper, we propose a new similarity, the square-root cosine (sqrt-cos) similarity, based on Hellinger distance which is more appropriate for solving IR problems such as measuring query relevance.

The hellinger distance between probabilities x and y is defined as follows.

$$H(x,y) = \Big[\sum_i^m (\sqrt{x}_i - \sqrt{y}_i)^2\Big]^{\frac{1}{2}} = \Big[2 - 2\sum_i^m \sqrt{x_i y_i}\Big]^{\frac{1}{2}}, \quad (3)$$

since $\sum_i x_i = 1$ and $\sum_i y_i = 1$.

Hellinger distance has two important properties which makes it a better distance measure in IR tasks. (a.) It is a metric because it is symmetric and satisfies triangle inequality:

$$H(x,y) = H(y,x),\ H(x,z) \le H(x,y) + H(y,z).$$

(b.) Helinger distance relates closely to the widely used KL divergence (also called information gain, or relative entropy)

$$\text{KL}(x,y) = \sum_i^m x_i \log \frac{x_i}{y_i} \quad (4)$$

They are special cases of the α-distance between two probability distributions [4, 9]

$$D_\alpha(x,y) = \frac{1}{\alpha(1-\alpha)} \sum_{i=1}^m \Big[x_i^\alpha y_i^{1-\alpha} - \alpha x_i + (\alpha - 1) y_i\Big] \quad (5)$$

Helinger distance is a special case at $\alpha = 1/2$:

$$H(x,y) = \sqrt{\frac{1}{2} D_{\frac{1}{2}}(x,y)}.$$

while the KL divergence is the case at

$$D_1(x,y) = \lim_{\alpha \to 1} D_\alpha(x,y) = \text{KL}(x,y),$$

$$D_0(x,y) = \lim_{\alpha \to 0} D_\alpha(x,y) = \text{KL}(y,x). \quad (6)$$

Therefore, Hellinger distance can be viewed as the symmetric middle point of KL divergence.

Assuming each document is normalized to 1 in L_1 norm: $\sum_{i=1}^m x_i = 1$, the Hellinger distance leads naturally to the SqrtCos similarty in Eq.(7).

$$\text{SqrtCos}(x,y) = \frac{\sum_{i=1}^m \sqrt{x_i y_i}}{(\sum_{i=1}^m x_i)(\sum_{i=1}^m y_i)}. \quad (7)$$

In recent research, there is a trend in IR is to use binary weighting instead of the traditional term frequency (**tf_idf**). We point out that Hellinger distance and SqrtCos similarity bridge between these two (extreme) situations. For example, if the frequency of word A is 4 and the frequency of word B is 1, in **tf_idf** weighting their relative importance is 4:1. In binary weighting, their relative importance is 1:1. In Hellinger distance, their relative importance is $\sqrt{4} : 1$. Therefore, Hellinger distance can be alternatively viewed as a compromise between tf_idf and binary weighting.

In order to utilize the advantages of Hellinger distance as discussed above, in this paper we use SqrtCos to calculate the pairwise sentence similarity to generate summaries for both the earlier document set A and the later coming set B.

2.2 Generating the Query-Relevant Summary for Document Set A using revised MMR

In order to summarize document set A, we revise Maximal Marginal Relevance (MMR) which has been successfully applied in query-relevant multi-document summarization systems in the following two ways: (1) The proposed SqrtCos similarity based on Hellinger distance will be used as the similarity measure. (2) The average similarity is used to determine the redundancy between a candidate sentence and the selected sentences. We use the average similarity instead of the maximum similarity in the original MMR for document summarization because we prefer to exclude sentences which are similar to more than one sentences in the selected sentence set.

Thus the respective incremental algorithm optimizes the following condition:

$$\max_{x_j \in A - S_{m-1}} [SqrtCos(x_j; q) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} SqrtCos(x_j; x_i)], \quad (8)$$

where S_{m-1} is current selected sentence set containing $m-1$ sentences, x_i is a sentence in S, and x_j is the candidate for the $m-th$ sentence to be selected.

2.3 Generating Update Summaries for Document Set B

Once we have the selected sentences from document set A, we can construct a sentence graph based on the pair wise SqrtCos similarities among sentences in A and B as shown in Figure 1. The sentences from A are labeled as 1 to represent sentences in the summary of A and 0 to represent sentences not selected. The question marks represents the sentence labels to be assigned in document set B. Once the sentence graph is constructed, the sentence selection problem can be treated as a label propagation from labeled data (i.e., sentences in A) to unlabeled data (i.e., sentences in B). In its simplest form, label propagation is like a random walk on

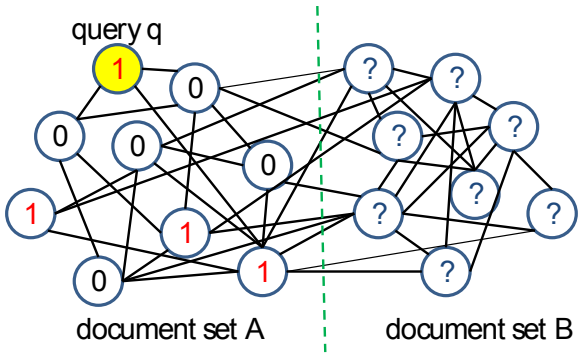

Figure 1: Label Propagation on Sentence Similarity Graph.

a graph [14]. There are different approaches to solve label propagation problems including using the diffusion kernel [5, 13], the harmonic nature of the diffusive function, etc. In this paper, in order to keep the coherency and consistency of the generated summaries, we emphasize the global and coherent nature of label propagation and apply the Green's function of the Laplace operator to solve the problem [3, 10].

2.3.1 *Label Propagation using Green's Function*

Given a graph with edge weights W, the *combinatorial Laplacian* is defined to be $L = D - W$, where D is the diagonal matrix consisting of the row sums of W; i.e., $D = \text{diag}(W\mathbf{e})$, $\mathbf{e} = (1 \cdots 1)^T$. Green's function for a generic graph is defined as the inverse of $L = D - W$. We construct Green's function using eigenvectors of L:

$$L\mathbf{v}_k = \lambda_k \mathbf{v}_k, \quad \mathbf{v}_p^T \mathbf{v}_q = \delta_{pq}, \tag{9}$$

where $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are the eigenvectors of L, $\lambda_1, \ldots, \lambda_n$ the eigenvalues of L, such that $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and such that the inner product of $\mathbf{v}_i$ and $\mathbf{v}_j$ is 1 if $i = j$ and 0 otherwise. We assume the graph is connected (otherwise we deal with each connected component one at a time). The first eigenvector is a constant vector $\mathbf{v}_1 = \mathbf{e}/\sqrt{n}$ whose associated eigenvalue is 1. After discarding this zero-mode, Green's function is defined as the positive definite part of L

$$G^{(1)} = L_+^{-1} = \frac{1}{(D-W)_+} = \sum_{i=2}^{n} \frac{\mathbf{v}_i \mathbf{v}_i^T}{\lambda_i}. \tag{10}$$

Note that Green's function can also be defined on the generalized eigenvectors of the Laplacian matrix:

$$L\mathbf{u}_k = \zeta_k D\mathbf{u}_k, \quad \mathbf{u}_p^T D\mathbf{u}_q = \mathbf{z}_p^T \mathbf{z}_q = \delta_{pq}. \tag{11}$$

where $0 = \zeta_1 \leq \zeta_2 \leq \cdots \leq \zeta_n$ are the eigenvalues and the zero-mode is again the first eigenvector $\mathbf{u}_1 = \mathbf{e}/\sqrt{n}$. Then we have

$$G^{(2)} = \frac{1}{(D-W)_+} = \sum_{k=2}^{n} \frac{\mathbf{u}_k \mathbf{u}_k^T}{\zeta_k}. \tag{12}$$

2.3.2 *Sentence Selection Procedures*

In our sentence graph, edge weights W represent the pairwise SqrtCos similarities among topic q, sentences in document sets A and B. The sentence selection problem is illustrated in Figure 1. Let y_0 represent the partial labels obtained from document set A, we compute the complete labels as the linear influence propagation:

$$y = Gy_0, \tag{13}$$

where G is the Green's function built from the constructed sentence graph.

Once we obtain the sentence labels, we only keep the sentences with label "1" in document set B which indicates the sentences are relevant to the given topic and contents in the earlier document set. If the total length of the sentences labeled to be "1" is longer than the required summary length, we will eliminate sentences which are most similar to the sentences in the summary of A and least similar to the sentences in B and the given query.

3. EXPERIMENTS

3.1 Data Set

In the experiments, we use TAC 2011 update summarization dataset for evaluating our method and comparing it with existing methods. In this dataset, there are 44 topics and for each topic there are 10 newswire articles in both document set A and B (representing the earlier collection of documents and the later documents respectively). A list of aspects for each topic is given and the task requires to generate a 100-word summary for both document set A and B. When generating the summary for document set B, the assumption is that the user has already read the earlier articles. Summaries generated by human labelers are provided in this task for evaluation.

3.2 Evaluation Measures

In the evaluation, we will compare the results by different methods with the human created summaries using Rouge toolkit (version 1.5.5) [7]. it is widely applied by Document Understanding Conference(DUC) and TAC for document summarization performance evaluation. It measures the quality of a summary by counting the unit overlaps between the candidate summary and a set of reference summaries. Several automatic evaluation methods are implemented in ROUGE, such as ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-SU. ROUGE-N is an n-gram recall ROUGE-L uses the longest common subsequence (LCS) statistics, while ROUGE-W is based on weighted LCS and ROUGE-SU is based on skip-bigram plus unigram. Intuitively, the higher the ROUGE scores, the more similar the two summaries.

3.3 Baselines

In the experiments, we use the following widely used update summarization methods as the baselines.

- Lead: The method selects the leading sentences in the documents to form the summary until its length reaches the required length.

- Centroid: The method applies the MEAD algorithm [8] to extract sentences according to the following three parameters: centroid value, positional value, and first sentence overlap. Both Lead and Centroid are standard methods provided by NIST.

- MMR: The method uses MMR like criterion to select sentences and at the same time reduce the redundancy among the selected sentences [1].
- CoRank: The method proposes a co-ranking process to rank sentences based on predefined update and consistency scores [15].
- DomSet: The method uses minimum dominating set to select sentences [11].

3.4 Experimental Results

Table 1 shows the Rouge scores of our method and the baseline methods using TAC 2011 data.

Methods	Rouge-1	Rouge-2	Rouge-SU
Lead	0.294	0.057	0.094
Centroid	0.283	0.059	0.091
MMR	0.347	0.075	0.115
CoRank	0.368	0.088	0.127
DomSet	0.359	0.083	0.120
Our Method	0.373	0.091	0.136

Table 1: Update summarization performance comparison on TAC 2011 data using ROUGE evaluation methods.

From the results, we have the following observations. (1) The original MMR outperforms the baselines with straightforward strategies such as Lead and Centroid. It is because MMR maximizes the relevance of the selected sentences with the topic and also reduces the redundancy among the sentences. (2) More advanced ranking methods like CoRank and graph-based methods like DomSet outperform MMR because they either take into consideration the consistency or utilize the overall relationships among sentences. (3) Our proposed method has the best results and significantly outperforms the original MMR because we use the Hellinger distance based similarity measure which deals with probabilities better. We also use label propagation with Green's function into the update summary generation so that the advantages of semi-supervised learning methods can be applied directly.

4. CONCLUSION AND FUTURE WORK

In this paper, we propose a new update summarization method which first uses a similarity measure based on Helliger distance to capture the semantics among documents and then applies a semi-supervised method using label propagation with Green's function to generate the update summaries. Experiments on TAC 2011 data show the effectiveness of the proposed method. In this paper, we just use the given topic description as a single query to find related contents in the documents, and the learning problem is a binary classification problem. In the future, we can further detect the aspects in the topics and transfer the problem into multi-class semi-supervised learning problems to obtain more accurate results for better coverage and consistency.

Acknowledgement

The work is partially supported by National Science Foundation under grant CNS-1126619, IIS-1213026, and CNS-1461926.

5. REFERENCES

[1] F. Boudin, M. El-Beze, and J. Torres-Moreno. The lia update summarization systems at tac 2008. In *Proceedings of TAC 2008*, 2008.

[2] J.-Y. Delort and E. Alfonseca. Dualsum: A topic-model based approach for update summarization. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 214–223, 2012.

[3] C. Ding, H. D. Simon, R. Jin, and T. Li. A learning framework using green's function and kernel regularization with application to recommender system. In *Proceedings of ACM SIGKDD 2007*, pages 260–269, ACM.

[4] A. O. Hero, B. Ma, O. Michel, and J. Gorman. Alpha divergence for classification, indexing and retrieval. Technical report, University of Michigan, 2001.

[5] R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *In Proceedings of the ICML*, pages 315–322, 2002.

[6] J. Li, S. Li, X. Wang, Y. Tian, and B. Chang. Update summarization using a multi-level hierarchical dirichlet process model. In *Proceedings of COLING 2012*, pages 1603–1618, 2012.

[7] C.-Y. Lin and E. Hovy. From single to multi-document summarization: A prototype system and its evaluation. In *Proceedings of ACL*, 2002.

[8] D. Radev, H. Jing, M. Stys, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, pages 919–938, 2004.

[9] A. Renyi. On measures of entropy and information. *Proc. Fourth Berkeley Symp. Math. Stat. and Probability*, 1960.

[10] B. Shao, D. Wang, T. Li, and M. Ogihara. Music recommendation based on acoustic features and user access patterns. *Trans. Audio, Speech and Lang. Proc.*, 17(8):1602–1611, 2009.

[11] C. Shen and T. Li. Multi-document summarization via the minimum dominating set. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 984–992, 2010.

[12] A. Singhal. Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, page 24(4): 35ÍC43, 2001.

[13] A. J. Smola and R. Kondor. Kernels and regularization on graphs. *Learning Theory and Kernel Machines Lecture Notes in Computer Science*, pages 144–158, 2003.

[14] M. Szummer and T. Jaakkola. Partially labeled classification with markov random walks. In *Proceedings of NIPS*, pages 945–952, 2001.

[15] X. Wan. Update summarization based on co-ranking with constraints. In *Proceedings of COLING 2012: Posters*, pages 1291–1300, 2012.

[16] D. Wang and T. Li. Document update summarization using incremental hierarchical clustering. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 279–288, 2010.