Multi-Document Summarization by Information Distance

Chong Long† Minlie Huang† Xiaoyan Zhu† § Ming Li‡
† State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, China
‡School of Computer Science, University of Waterloo, Canada
§ Corresponding Author: zxy-dcs@tsinghua.edu.cn

***Abstract*—We are now living in a world where information is growing and updating quickly. Knowledge can be acquired more efficiently with the help of automatic document summarization and updating techniques. This paper describes a novel approach for multi-document update summarization. The best summary is defined as one of which has the minimal information distance to the entire document set. And the best update summary has the minimal conditional information distance to a document cluster given that a prior document cluster has already been read. We propose two methods to approximate information distance between two documents, one by compression and the other by the coding theory. Experiments on the DUC 2007 dataset[1] and the TAC 2008 dataset[2] have proved that our method closely correlates with the human-written summaries and outperforms LexRank in many categories under the ROUGE evaluation criterion.**

***Keywords*-Data Mining; Text Mining; Kolmogorov Complexity; Information Distance**

I. Introduction

Automated summarization dates back to the 1950's [1]. In recent years, since web contents grow in an increasing speed, people need to have a concise overview of a large set of articles in a short time. So document summarization, aiming at generating brief and understandable summaries, has quickly become a hot research topic. Document updating technique is also very helpful for people to acquire new information or knowledge by eliminating out-of-date or redundant information. Multi-document update summarization is introduced by Document Understanding Conference (DUC) in 2007. It aims to produce a summary describing the majority of information content from a set of documents under the assumption that the user has already read a given set of earlier documents. This type of summarization has been proved extremely useful in tracing news stories: only new and update contents should be summarized if we have already known something about the story.

In a news service website called "NewsBlaster[3]", news articles are grouped into several topics, and a great number of articles have two summaries: one is the whole story of the topic, and the other one tells readers what have happened recently. For example, there are ten news articles about the development of Australia's uranium mine project in its Kakadu National Park and the protests and obstacles encountered. A good summary should contain four aspects: (1) What is the project going on? (2) What is the attitude of the government? (3) Where are the protests and obstacles coming from? (4) How does the government deal with these problems? As exemplified, a good summary is expected to preserve the information contained in the documents as much as possible, and at the same time keep the information as novel as possible [2].

Information distance is based on the theory of Kolmogorov complexity. It is now widely accepted as an information theory for individual objects parallel to Shannon's information theory which is defined on an ensemble of objects. In this paper, we propose a novel document summarization approach based on the theory of information distance among many objects. In order to deal with update summarization, we will extend the information distance theory to conditional information distance among many objects. Finally summaries are generated according to our newly developed theory.

II. Related Work

Generally speaking, there are mainly two different kinds of document summarization methods: extraction-based and abstraction-based. Here we focus on extractive summarization. Most extractive summarization studies have focused on NLP and statistical machine learning techniques. Carbonell and Goldstein proposed to use Maximal Marginal Relevance (MMR) , which aims to select summary sentences relevant to the user query and least similar to previously chosen ones [3]. Radev *et al.* described an extractive multi-document summarizer which extracts a summary from multiple documents based on the document cluster centroids [4]. Researchers have also proposed a number of machine learning methods to extract the sentences [5]. Most recently, graph-based methods have been proposed for document summarization, such as LexRank [6] and TextRank [7].

Different from all the previous summarization methods, we will propose a novel summarization approach based on the information distance theory. In the next section, this

[1] http://duc.nist.gov/
[2] http://www.nist.gov/tac/
[3] http://newsblaster.cs.columbia.edu/

theory will be reviewed and our extended theory will be introduced.

III. Theory

Fix a universal Turing machine U. The Kolmogorov complexity [8] of a binary string x conditioned to another binary string y, $K_U(x|y)$, is the length of the shortest (prefix-free) program for U that outputs x with input y. It can be shown that for a different universal Turing machine U', for all x, y

$$K_U(x|y) = K_{U'}(x|y) + C,$$

where the constant C depends only on U'. Thus $K_U(x|y)$ can be simply written as $K(x|y)$. We write $K(x|\epsilon)$, where ϵ is the empty string, as $K(x)$. It has also been defined in [9] that the energy to convert between x and y to be the smallest number of bits needed to convert x to y and vice versa. That is, with respect to a universal Turing machine U, the cost of conversion between x and y is:

$$E(x, y) = \min\{|p| : U(x, p) = y,\ U(y, p) = x\} \quad (1)$$

It is clear that $E(x, y) \leq K(x|y) + K(y|x)$. From this observation, the following theorem has been proved in [9]:

Theorem 1: $E(x, y) = \max\{K(x|y), K(y|x)\}$.

Thus, the max distance was defined in [9]:

$$D_{\max}(x, y) = \max\{K(x|y), K(y|x)\}. \quad (2)$$

This distance is shown to satisfy the basic distance requirements such as positivity, symmetricity, and triangle inequality is admissible [9].

Here for an object x, we can measure its information by Kolmogorov complexity $K(x)$; for two objects x and y, their shared information can be measured by information distance $D(x, y)$. In [10], the authors generalize the theory of information distance to more than two objects. Similar to Equation 1, given strings $x_1, \ldots, x_n$, they define the minimal amount of thermodynamic energy needed to convert x_i to x_j for all i and j:

$$E_m(x_1, \ldots, x_n) = \min\{|p| : U(x_i, p, j) = x_j\} \quad (3)$$

Then it is proved in [10] that:

Theorem 2: Modulo to an $O(\log n)$ additive factor,

$$\begin{aligned}&\min_i K(x_1 \ldots x_n|x_i)\\ &\quad \leq E_m(x_1, \ldots, x_n)\\ &\quad\quad \leq \min_i \sum_{k \neq i} D_{\max}(x_i, x_k)\end{aligned} \quad (4)$$

In update summarization, the summary should contain new information which former documents have not mentioned, so Equation 3 is extended to (for all i and j):

$$E_m(x_1, \ldots, x_n|c) = \min\{|p| : U(x_i, p, j|c) = x_j\} \quad (5)$$

where c is the conditional sequence that is given for free to compute from sequence x to y and from y to x. Similar to Equation 4:

Theorem 3: Modulo to an $O(\log n)$ additive factor,

$$\begin{aligned}&\min_i K(x_1 \ldots x_n|x_i, c)\\ &\quad \leq E_m(x_1, \ldots, x_n|c)\\ &\quad\quad \leq \min_i \sum_{k \neq i} D_{\max}(x_i, x_k|c)\end{aligned} \quad (6)$$

Given n objects and a conditional sequence c, the left-hand side of Equation 6 may be interpreted as the most comprehensive object that contains the most information about all of the others. The right-hand side of the equation may be interpreted as the most typical object that is similar to all of the others.

IV. Summarization Approach

We have developed the theory of conditional information distance among many objects. In this section, a new summarization model will firstly be built based on our new theory, and then we are going to develop a method to approximate Kolmogorov complexity and information distance through two different ways.

A. Modeling

1) Modeling Traditional Summarization: The task of traditional multi-document summarization can be described as follows: given n documents $B = \{B_1, B_2, \ldots, B_n\}$, the task requires the system to generate a summary S of B. According to our theory, the conditional information distance among $B_1, B_2, \ldots, B_n$ is $E_m(B)$.

However, it is very difficult to compute E_m. Moreover, E_m itself does not tell us how to generate a summary. Equation 4 has provided us a feasible way to approximate E_m: the most comprehensive object and the most typical one are the left and right of Equation 6, respectively. The most comprehensive object is long enough to cover as much information in B as possible, while the most typical object is a concise one that expresses the most common idea shared by those objects. Since we aim to produce a short summary to represent the general information, the right-hand side of Equation 4 should be used. The most typical document is the B_j such that

$$\min_j \sum_{i \neq j} D_{\max}(B_i, B_j)$$

However, B_j is far from enough to be a good summary. A good method should be able to select the information from B_1 to B_n to form a best S. We view this S as a document in this set. Since S is a short summary, it does not contain extra information outside B. The best traditional summary S_{trad} should satisfy the constraint as:

$$S_{trad} = \arg\min_S \sum_i D_{\max}(B_i, S) \quad (7)$$

In most applications, the length of S is confined by $|S| \leq \theta$ (θ is a constant integer) or $|S| \leq \alpha \sum_i |B_i|$ (α is a constant real number between 0 and 1).

2) Modeling Update Summarization: Given a set of earlier m articles $A = \{A_1,A_2,\ldots,A_m\}$, the update summarization task is to summarize new contents presented by a document set $B = \{B_1,B_2,\ldots,B_m\}$. This earlier article set A can be viewed as a precondition. Thus this task can be well modeled by the conditional version of information distance. The best summary S_{best} should satisfy the constraint as follows:

$$S_{best} = \arg\min_S \sum_i D_{\max}(B_i, S|A) \tag{8}$$

If $m = 0$ ($A = \phi$), it will be a traditional multi-document summarization problem. If $m > 0$ ($A \neq \phi$), it will be a multi-document update summarization problem. Therefore, the traditional summarization can be viewed as a special case of formula 8.

According to [11], from Equation 8 we can get:

$$D_{\max}(B_i, S|A) = D_{\max}(B_i^A, S|A) = D_{\max}(B_i^A, S)$$

where B_i is mapped to B_i^A under the condition of A. Then for a document B_i and a document set A, B_i^A is a set of B_i's sentences ($B_{i,k}$s) which are different from all the sentences in A_1 to A_m:

$$B_i^A = \{B_{i,k} | \forall\ sen \in \bigcup_i A_i', D_{\max}(B_{i,k}, sen) > \varphi\}$$

where A_i' is the sentence set of a document A_i and φ is a threshold. Note that φ is the only parameter to be specified in our approach and it is only related to update summarization clusters. We tune it on the B cluster of the DUC 2007 dataset under the ROUGE-1 criterion.

We have already developed a framework for summarization. However, the problem is that neither $K(.)$ nor $D_{max}(.,.)$ is computable. Two methods can be used in the approximation and the computation of information distance: one by compression and the other by the coding theory. In the next several sub-sections, we will discuss how to use these two methods to do the approximations, respectively.

B. Approximation by Compression

In [12], the authors proposed to approximate Kolmogorov complexity through a compressor C. The boundary case is $C = K$ if C is powerful enough. Now we have to use a real-world reference compressor C which approximates the information distance $E(x, y)$. The compression distance $E_C(x, y)$ is defined as

$$E_C(x, y) = C(xy) - \min\{C(x), C(y)\}.$$

Here $C(xy)$ denotes the compressed size of the concatenation of x and y. $C(x)$ denotes the compressed size of x, and $C(y)$ denotes the compressed size of y. Then $E_C(x, y)$ is just an approximation of D_{max}:

$$D_{max}(x, y) = E(x, y) \approx E_C(x, y) \tag{9}$$

C. Approximation by the coding theory

The compression method is a language-independent summarization method. It is easy to be implemented and it can summarize documents written in any other language without any modifications. However, this method only uses morphological features of a sentence. The semantic meanings of terms or phrases have been heavily neglected by simple compression. Alternatively, we can use frequency count, and use Shannon-Fano code [13] to encode a phrase which occurs in probability p in approximately $-\log p$ bits to obtain a short description.

Coding-theory-based approximation method can deal with a sentence in word and phrase granularities. Therefore, firstly we divide a sentence into semantic elements; then information distance between two sentences is estimated through their semantic element sets.

1) Semantic Element Extraction: In a document, each word or entity contains a certain amount of information, and the information varies according to the word or entity's importance. Such words or entities are called “semantic elements”, and “elements” for short in this paper. There are two types of elements: (1) named entities such as person, organization, time, and location, containing a large portion of meaningful information; and (2) common words except stop-words. For example, the meaningful elements of the sentence “George W. Bush was born on July 6th, 1946” are “George W. Bush” (person), “born” and “July 6th, 1946” (time).

First, we recognize entities about a person, location, organization and other names with Stanford Named Entity Recognition (NER)[4]. We also extract entities about a date or time. Totaly five types of name entities are recognized. Second, words or phrases with the same meanings are normalized into one entity through coreference resolution. For example, “George W. Bush” and “George Bush” are normalized to the same entity; ‘May 15th, 2008”, “May 15, 2008” and “5/15/2008” are recognized as the same date.

2) Information Distance Approximation: Next we will take several steps to do the approximations. Although some steps contain rough approximations, we will investigate the influence of our estimations with extensive experiments in Section V-E.

Let $M = \{M_1, M_2, \ldots\}$ and $N = \{N_1, N_2, \ldots\}$ to be two sets of sentences. After those steps mentioned in Section IV-C1, each sentence M_i (or N_j) has an element set M_i^* (or N_j^*). According to Equation 2,

$$D_{\max}(M, N) = \max\{K(M|N), K(N|M)\},$$

[4] http://nlp.stanford.edu/ner/index.shtml

Table I
SUMMARY OF DATASETS

Year	2007	2008
# Topics	10	48
# Clusters for each Topic	3	2
# Documents	250	960
# Manual Summaries for each Cluster	4	4

then

$$\begin{aligned} K(M|N) &\approx K(\textstyle\bigcup_i M_i^* \setminus \bigcup_j N_j^*), \\ K(N|M) &\approx K(\textstyle\bigcup_j N_j^* \setminus \bigcup_i M_i^*). \end{aligned} \tag{10}$$

The Kolmogorov complexity of an element set W can be computed by the sum of the complexities of all its elements:

$$K(W) = \sum_{w \in W} K(w)$$

According to the the coding theory, the complexity of an element w can be computed by its probability [8], which can usually be approximated by its document frequency in the corpus:

$$K(w) = -\log P(w) \approx -\log df(w) \tag{11}$$

Although the approximation method based on the coding theory contains semantic information, there are mainly three steps may lead to an approximation bias during the process of generating a summary:

1. when the complexity between two sentences is computed through their elements' complexities in Equation 10;

2. when an element's complexity is estimated by its document frequency in Equation 11;

3. and when E_m is approximated by the right-hand side of Equation 6.

In Section V-E we will analyze how these approximations affect our system's performance.

V. EXPERIMENTAL RESULTS

In this section, our summarization method will be evaluated on the DUC 2007 and the TAC 2008 datasets.

A. Datasets

In DUC 2007 and TAC 2008, one of those tasks is "update summarization". The task requires participants to write a short summary of a set of newswire articles, under the assumption that the user has already read a given set of earlier articles. The length of the summary should be no more than 100 words. Table I shows the statistics of the datasets. In DUC 2007, there are 10 topics, each of which has three clusters: A, B and C. Each cluster has about 10 news articles. Summarizing on cluster A is a traditional summarization task. Summarizing on cluster B is required to generate an update summary with already read A. The summary from cluster C should be updated with cluster A and B. In TAC 2008, there are 48 topics and two clusters (A and B) for each topic. The update task is similar with that in DUC 2007. For each cluster, there are four standard summaries written by four different people, respectively. The score on a cluster is the mean of the scores on four manual summaries. The overall score on a dataset is just the mean of the scores on its clusters.

B. Preprocessing

During the preprocessing process, we need to filter out those sentences which are impossible to be a part of a summary. The top 10% entities with the highest document frequency on a document set are viewed as "the topic set". The sentences which don't contain any entity of the topic set are eliminated and the remaining ones are called "candidate sentences". After preprocessing we search for the best combination of the candidate sentences to make Equation 8 minimal.

In our datasets described in Section V-A, each set has on average 9.6 documents with less than 300 sentences. Averagely less than 70 candidate sentences are selected. As the length of the summary is less than 100 words, averagely there are three to four sentences per summary in average. Therefore, there are totally less than $C_4^{70} < 10^6$ different combinations. It is a small number and our approach can generate the summaries in real time. The exhaustive enumeration is simple but can get the optimal result. We will develop a heuristic search algorithm to generate longer summaries for larger document sets in our future work.

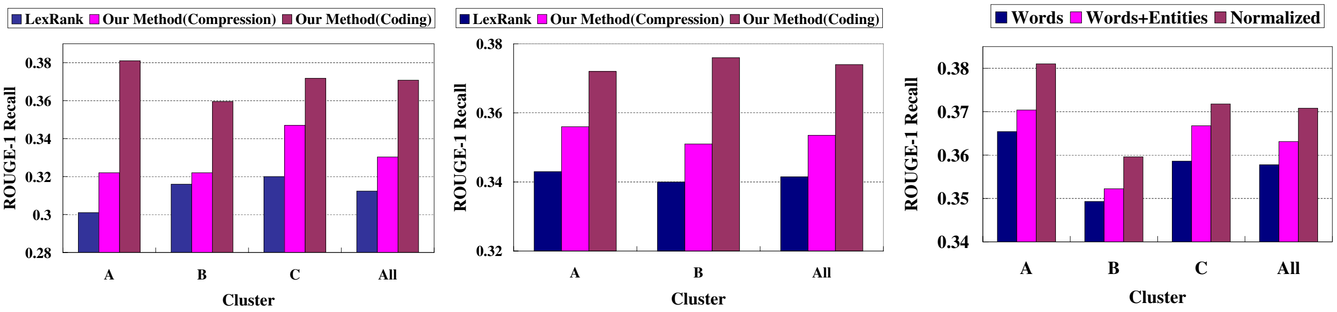
C. Evaluation Metrics

ROUGE toolkit [14] is used for evaluation. It measures the quality of a summary by counting the overlapping units such as the n-gram, word sequences and word pairs between the candidate summary and the reference summary. The ROUGE toolkit reports separate F-measure scores for 1, 2, 3 and 4-gram, and also for longest common subsequence co-occurrences. We use ROUGE-1 recall to evaluate our results.

D. Results on the DUC 2007 and the TAC 2008 datasets

Firstly we will show the results of our system on all datasets. We take the popular summarization method, LexRank [6] implemented in MEAD, as our baseline. We run the LexRank package on each set's "candidate sentences" (as described in Section V-B), with the same preprocessing step as our method does. Then our multi-document update summarization approach is implemented, approximated by compression and the coding theory, respectively. The results on the DUC 2007 dataset and the TAC 2008 dataset are shown on0 the left side and the middle of Figure 1, respectively.

From these two figures we can get three conclusions: (1) our methods, even simply approximated by compression, always outperform the LexRank method; this means that the proposed framework is safe and sound, and has already exhibited potentials for generating good summaries even with



Results on the DUC 2007 dataset

Results on the TAC 2008 dataset

Comparison of Different Elements

Figure 1. Comparisons

a very simple compression technique problem effectively; (2) the results of approximation by the coding theory are much better than those of compression. This phenomenon implies that semantic information is very important while approximating information distance between sentences; and (3) our system performs almost equally well on traditional clusters (Cluster A) and update clusters (Cluster B and C) by virtue of that the framework is universal to two cases.

E. Estimation Analysis

We noticed in Section IV-C2 that there are mainly three approximating biases. Next we will analyze them.

1) Different Elements: For the first bias, we check our method on estimating the distance among sentences through their elements. Two important steps taken from a sentence M to its element set M^* are recognizing words or phrases as entities and grouping them through coreference resolution. Here three different methods are compared on the DUC 2007 dataset, as shown on the right side of Figure 1: (a) "Words" means to treat every word as an element. For example, "George Bush" is viewed as two different elements; (b) "Words+Entities": after entities are recognized, phrases such as "George Bush" are viewed as one element. Other words such as "born" are still in the element set; (c) "Normalized": after reference resolution, "George Bush" and "George W. Bush", "May 15,2008" and "5/15/2008" are normalized into one element, respectively. Through this figure we can see that after entity recognition and reference resolution, the performance has been improved remarkably, where we may conjecture it has got a more accurate approximation of information distance.

2) Complexity Estimation: Pyramids [15] are used to study the second bias. Manually built pyramids have provided us a good way to investigate how human being write summaries with important semantic units. A pyramid represents the opinions of multiple human summary writers, each of whom has written a reference summary for the input set of documents. Each semantic unit (usually a short sentence or a phrase) is called a Summary Content Unit (SCU).

For example, in the pyramids provided by DUC 2007 and TAC 2008, words or phrases contributed to the reference summaries are named "contributors". They are collected and grouped according to manual annotations. The following is a SCU example in an XML format:

```
<scu uid="22" label="Euro was scheduled
 to be launched on January 1, 1999">
  <contributor label="Jan. 1, 1999 date
   for introduction of the euro
   approached">
  </contributor>
  <contributor label="(the Euro) will
   go into effect on January 1, 1999">
  </contributor>
  <contributor label="Euro was scheduled
   to be launched on January 1, 1999">
  </contributor>
  <contributor label="scheduled January
   1999 introduction of the euro">
  </contributor>
</scu>
```

This SCU has four contributors from four different reference summaries. The most important elements, "January 1999" and "Euro", exist in all four contributors. The more frequently an element exists in contributors, the more important it is. In a pyramid, we take every contributor defined in the XML file as a document. Each element T in this pyramid has a weight $w(T)$, computed by T's document frequency on the contributors. Compared with Equation 11, $w(T)$ might be a more accurate approximation of $K(T)$ in that $w(T)$ was weighted by the human-annotated SCUs. Let $K(T) = w(T)$, we get a group of better results, which is closer to human summaries.

In Figure 2 we have three groups of results on each topic of the DUC 2007 dataset: the brown ones are the average ROUGE-1 scores of human written summaries provided by the organizers. Each set has four reference summaries written by four different people. The blue ones are the results of our approach with $K(T) = w(T)$. The pink ones

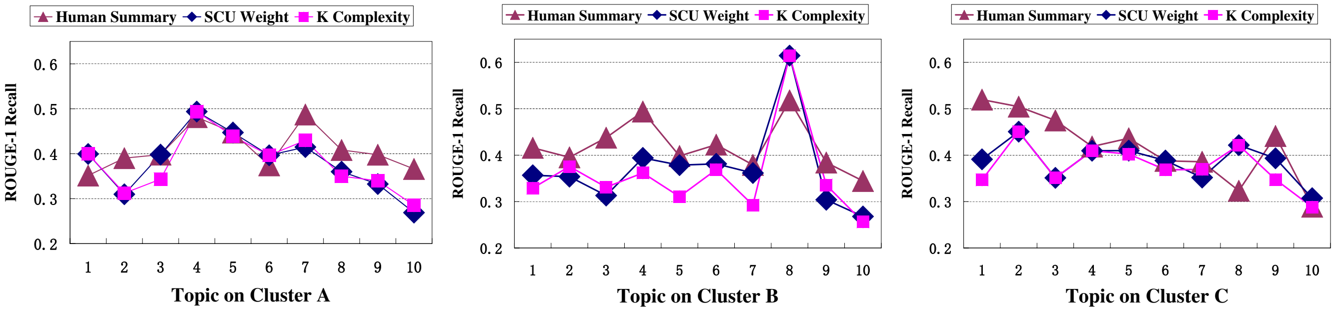


Figure 2. Comparisons on the DUC 2007 Dataset

are the results of our proposed approach with Equation 11. From this figure we have two observations: the first one is that the approximation of $K(T)$ by Equation 11 has very close results with that by $K(T) = w(T)$. The latter takes into account the important semantic units which must be concerned when assessors write summaries. Thus we believe this might be a right way to approach the unit weights of "perfect" summaries. The second one is that our ROUGE-1 scores are close to those of the human written summaries.

As to the third bias,we will study it in theory to find out the upper bound of the difference between E_m and $\min_S \sum_i D_{\max}(B_i, S|A)$.

VI. Conclusion and Future Work

In this paper, we have proposed a novel document summarization framework based on the theory of information distance. We proposed two methods to approximate information distance between two documents, one by compression and the other by the coding theory. Experiments show that our approach performs well on the DUC 2007 and the TAC 2008 datasets. In future work, we will further improve our approach mainly in two ways: firstly, better approximation of information distance will be studied; then a heuristic method will be developed in order to find the best summary more effectively.

Acknowledgment

The work was supported by NSFC under grant No.60803075, the National Basic Research Program (973 project in China) under grant No.2007CB311003, and Microsoft joint project "Opinion Summarization toward Opinion Search". The work was also supported by IRCI project from the International Development Research Center, Canada.

References

[1] H. P. Luhn, "The automatic creation of literature abstracts." *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159–165, 1958.

[2] X. Wan, J. Yang, and J. Xiao, "Manifold-ranking based topic-focused multi-document summarization," in *IJCAI*, 2008, pp. 2903–2908.

[3] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *SIGIR*, August 1998, pp. 335–336.

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

[5] D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen, "Document summarization using conditional random fields," in *IJCAI*, 2007.

[6] G. Erkan and D. R. Radev, "Lexpagerank: Prestige in multi-document text summarization," in *EMNLP*, 2004.

[7] R. Mihalcea and P. Tarau, "A language independent algorithm for single and multiple document summarization," in *IJCNLP*, 2005.

[8] M. Li and P. M. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*. Springer-Verlag, 1997.

[9] C. H. Bennett, P. Gács, M. Li, P. M. Vitányi, and W. H. Zurek, "Information distance," *IEEE Transactions on Information Theory*, vol. 44, no. 4, pp. 1407–1423, July 1998.

[10] C. Long, X. Zhu, M. Li, and B. Ma, "Information shared by many objects," in *CIKM*, 2008, pp. 1213–1220.

[11] X. Zhang, Y. Hao, X. Zhu, and M. Li, "Information distance from a question to an answer," in *SIGKDD*, August 2007.

[12] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, "An information-based sequence distance and its application to whole mitochondrial genome phylogeny," *Bioinformatics*, vol. 17, no. 2, pp. 149–154, 2001.

[13] R. L. Cilibrasi and P. M. Vitányi, "The google similarity distance," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 370–383, March 2007.

[14] C.-Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *HLT-NAACL*, 2003, pp. 71–78.

[15] A. Nenkova, R. Passonneau, and K. Mckeown, "The pyramid method: Incorporating human content selection variation in summarization evaluation," *ACM Transactions on Speech and Language Processing*, vol. 4, no. 2, 2007.