E-commerce Webpage Summarization by Using Statistical Similarity Measures

Jantima Polpinij

Department of Computer Science
Faculty of Informatics, Mahasarakham University
Mahasarakham 44150 THAILAND

Jantima.p@msu.ac.th

Abstract

More and more information has been made available and accessible due to increasing of e-commerce websites. In order to assist users find useful information, automatic text summarization systems have been made available. They help users access to the information contents quicker. Text summarization is the problem of condensing a source text into a shorter version preserving its information content. In this paper, we apply automatic text summarization to an e-commerce webpage summarization based on statistic similarity measures. Our system consists of two main processes: topic identification and webpage summarization. The first process is to compute topic identification with term frequencies and weight of word, while the second process summarizes the text by extracting the most significant sentences and paragraphs. Finally, the experimental results of sentence and paragraph extraction show en effective accuracy after testing by the precision (*P*).

Keywords**:** *e-commerce, text summarization, similarity measure*

1. Introduction

With the growth of electronic commerce (e-commerce), it is becoming increasingly common, more and more information has been made available and accessible. The information explosion has resulted in a well recognized information overload problem. Text summarization can help users access to the information content in a shorter period of time because it consists of reducing the size of a text while preserving its information content.

Therefore, this paper presents an automatic text summarization system based on a statistical approach to help users access to the information content in a shorter period of time. The system of summarization consists of two processes: topic identification and text extraction summary. The first process (topic identification) is done by a word segmentation to find out term frequencies and then using Mixed Minimum and Maximum (MMM) model to identify the rank of topics. Then, in the second process, we use the identified topic to extract the significant paragraph or sentence in the text as its summary. However, our approach does not require the external knowledge other than the document itself, and be able to summarize a general text document.

The rest of this paper is organized as follows. In Section 2, it is literature review. We describe the research methodology based on text summarization technique in Section 3. Subsequently, some experimental results will be proved. The conclusion is provided in the final section.

2. Literature Reviews

In general, text summarization can be broadly classified into two approaches: abstraction and extraction. The abstraction requires hard machinery from Natural Language Processing (NLP), including linguistic grammars, lexical parsing, and summary generation [1]. The extraction can be easily viewed as the process of selecting the most significant (i.e. sentence, paragraph) from the original text and concatenating them into a shorter form. Although on many debate that extraction approach makes text hard to read due to the insufficient coherence, it also depends on the objective of summarization. However, most of recent researches in this area are based on extraction [2]. Text summarization approaches can be broadly classified into four approaches: *Statistic Approach* [3], *Linguistic Approach* [4], *Passage Extraction* [5], and *Discourse Structure* [5].

In the previous studies of text summarization, their approaches can be found in [6]. The first computational paper on automated extraction, Luhn [3] describes a simple technique that uses term frequencies to weigh sentences, which are then extracted to form a summary. Luhn's technique is

in *Statistic Approach* area that is used to identify keywords of the document. The basic idea of statistic approach by Luhn has influenced the research field of text summarization. Subsequent works have demonstrated the success of Luhn's approach [7, 8]. A particular type of term aggregation and normalization of Luhn is applied by [9] that has been supplanted by the use of stemming. In addition, absolute term frequency is less useful than the term frequencies, which are normalized to take into account the document length and frequency in a collection [10]. Furthermore, several methods of scoring the relevance of sentences or passages and combining the scores are described in [5, 11, 12, 13, and 14]. General advances in text summarization will therefore require method of Natural language processing (NLP) on the document. The linguistic features depend on what type or style of the document. It can also be use for document case. Some linguistic approaches are applied in [4]. In their work, fusion techniques are used to match a particular subtopic in each document.

On the one hand, due to increasing of the number of e-commerce websites, many researches proposed text summarization to apply for accessing the information content in a short time. According to Wu et al. [15], a summarization technique is applied for mobile-commerce (m-commerce). This is because it is difficult to effectively present a large table of information on small devices as they tend to have limited display and processing capabilities. Hence, large tables need to be reduced and summarized in order to be properly displayed on small devices. In 2005, Chen and Sairamesh [16] proposed a summarization technique to extract crucial information in high volume business data that efficiently are critical for enterprises to make timely business decisions and adapt accordingly. Then Liu [17] also applied text summarization to be a very important function in e-business-intelligence service. This is because he believed that human beings have proven to be extremely capable summarizers, while computer based automated abstracting and summarizing has proven to be extremely challenging tasks.

3. Research Methodology

In this section, it is to describe the details of each module. The overall architecture is shown in Figure 1. This system consists of two main processes. First, Thai text processing is run to tokenize a given text into meaningful words. This process is so called the topics identification specific. The second process is the web-page summarization using a similarity measure technique.

3.1 The Topic Identification

This process is used for finding a feature of document with filtering input to determine the most important topics. In general, a text can have many topics and the topic extraction process can be parameterized to include more or fewer of them to produce longer or shorter summaries. In the system, the topic identification process can be broadly divided into 3 sub-processes: word segmentation, term frequency and weighting, and the topic identification specific.

Word segmentation is the first and obligatory task in natural language processing because word is a basic unit in linguistics. After word segmentation processing, the term frequency and weighing are used for refining a feature for a document.

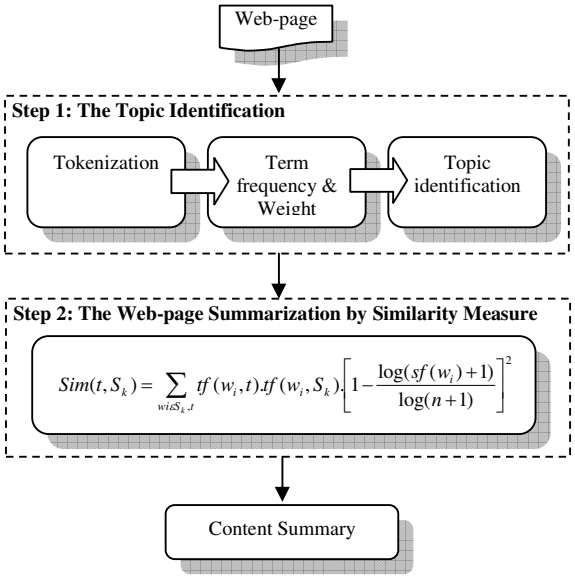

Figure 1. The Research Methodology

The statistical analysis is based on the theory that a term's frequencies can determined its utility as an index term. In general, high frequency terms and very low frequency terms deem to be poor candidates because the basic operation is only that of counting words. Therefore, the word can be assigned *weights* that represent the value of the word as an indexing term. Weights can be determined in a variety ways. We applied *TF-ISF* (*Term Frequency- Inverse Sentence Frequency*) to

measure each word. Of course, this technique is an adaptation of conventional *TF - IDF* (*Term Frequency - Inverse Document Frequency*) [18]. Suppose *S* is the number of the sentences in document, *SF* is the number of sentence that consists of weighting term words, and *TF* be the frequency of term word in a document. Hence, *ISF* is used to refine a feature of document and calculated by equation follows:

$$ISF = 1 + log\ (\ |S| / SF\ ) \qquad (1)$$

When a document is passed through the segmentation process, each word in the document is calculated for its frequency value.

Finally, the topic identification is the technique based on word frequency to produce the overall ranking. Then, the top-rank term words are returned. Generally, a document contains many words. We consider only the range of the highest weight of words calculated by *TF-ISF* for topic identification. Therefore, the range of topic identification can be shown as follows:

Range of Topic Identification

$$= [Min_{\ Approximated} * 100, Max\ (TF\text{-}ISF)] \qquad (3)$$

In this work, it was applied a *Mixed Min and Max* (*MMM*) to fine the range of topic identification. This model is based on the concept of fuzzy sets proposed by Zadeh [19]. The *MMM* model has been developed by Fox and Sharat [20]. Each index term has a fuzzy set associated with it. The word document weight of a document with respect to an index term *A* is considered to be the degree of membership of the word document in the fuzzy set associated with *A*. The degree of membership for union and intersection are defined fuzzy set theory as follows:

$$d_{A \cup B} = \max\ (d_A, d_B) \qquad (4)$$

$$d_{A \cap B} = \min\ (d_A, d_B) \qquad (5)$$

The MMM model attempts to soften the Boolean operation by considering the queried-document similarity as a linear combination of the *min* and *max* document weighting. So, a document *D* can be given with index-term weight d_{A1}, d_{A2}, .. d_{An} for terms A_1, A_2,...,A_n, and the following queries:

$$Q_{or} = (A_1\ or\ A_2\ or\ \ldots\ or\ A_n) \qquad (6)$$

$$Q_{and} = (A_1\ and\ A_2\ and\ \ldots\ and\ A_n) \qquad (7)$$

The queried-document similarity in the MMM model is computed as follows:

$$SIM\ (Q_{or}, D) = C_{or1} * \max\ (d_{A1}, d_{A2}, .. d_{An}) + C_{or2} * \min\ (d_{A1}, d_{A2}, .. d_{An}) \qquad (8)$$

$$SIM\ (Q_{and}, D) = C_{and1} * \min\ (d_{A1}, d_{A2}, .. d_{An}) + C_{and2} * \max\ (d_{A1}, d_{A2}, .. d_{An}) \qquad (9)$$

Where, C_{or1} and C_{or2} are "soften" coefficients of "*or*"operator, and C_{and1}, C_{and2} are softness coefficients of "*and*" operator. To give the maximum of the document weight more importance while considering "*or*" query and the minimum of the document more importance while considering "*and*" query. In general, they have $C_{or1} > C_{or2}$ and $C_{and1} > C_{and2}$. For simplicity it is generally assumed that $C_{or1} = 1 - C_{or2}$ and $C_{and1} = 1 - C_{and2}$. The best performance usually occurs with C_{and1} in the range [0.5, 0.8] and $C_{or1} > 0.2$ [19]. In this our experiment, we select 0.5 of coefficients.

This work used only the minimum of MMM. The similarity of the node can be calculated as follows:

$$SIM\ (Q,D) = C_{op} * max + (1\text{-}C_{op}) * min \qquad (10)$$

In addition, this work also used term words weighting to extract the relevant paragraph. Let *w* denote term word weighting and transform (w_{A1}, w_{A1}, .., w_{An}) into (d_{A1}, d_{A2}, .. d_{An}). So, it can get:

$$SIM\ (Q,D) = C_{op} * maximum\ (TF\text{-}ISF) + (1\text{-}C_{op}) * minimum\ (TF\text{-}ISF) \qquad (11)$$

3.2 The Web-page Summarization by Similarity Measure

Many systems for text summarization have been proposed by using similarity measures [21] between text spans (sentences or paragraphs) and topic identification. Representative sentences or paragraphs can be than selected by comparing the score of a given document to a present threshold. This work uses the extraction of relevant to a given query by Kanus et al. [22] and it also uses a *TF-ISF* representation and compute the similarity between sentences S_k and query *t* (*topic identification*) as follows:

$$Sim(t, S_k) = \sum_{w_i \in S_k, t} tf\ (w_i, t).tf\ (w_i, S_k).\left[1 - \frac{\log(sf\ (w_i) + 1)}{\log(n+1)}\right]^2 \qquad (13)$$

Where, let *TF* (*w*,*d*) be the frequency of term *t* in the document, *SF*(*w*) be the sentence frequency of the term *w*, and *n* be the total number of documents in the collection. When they are calculated by *TF-ISF*, they sometime had shown the highest score. They cannot be the topics because they are less significance on a document. Afterwards, sentence S_k and topic identification *t* are processed by equation (13). In each document, a threshold is than estimated from data selecting the most relevant sentences. The approach of sentences extraction method is a variation of the above method where the topic identification is enriched before computing the similarity. Since topic identification and sentences may be very short, this allows computing more meaningful similarities. Topic identification expansion appeared to be very important in our experiments.

4. The Experimental Results

This work does not use any standard dataset but we collect data from some of e-commerce websites that still are available such as Amazon and eBay. Our collection approximates 200 documents. Each document size ranges from 1 to 2 pages per document, and each page contains about 3 to 5 paragraphs.

In addition, we evaluate the results of summarization by using the information retrieval standard. A common performance measure for our system evaluation is *precision* (*P*) [23]. It is the proportion of retrieved documents that are relevant.

In our experiment, it is to use similarity for extracting the most relevant sentences and paragraphs. We consider compression rates of 20%, 25%, 30%, and 40%. Then, Table 1 reports the results obtained by the automatic summarizer that are compared with summaries of manually produced by human. The performance is expressed in terms of precision is expressed in percentage (%).

Based on Table 1, it can be explained from the experiment that, the values of precision for all data sets is significantly higher with the rate of 40% than with the compression rate of 20%, 25%, and 30%. This is because, with the larger the compression rate, the larger number of sentences or paragraphs can be easily selected for the summary. That means larger probability of a sentences or paragraph has selected by a summarizer matches with a sentences or paragraph belonging to the extractive summary.

Table1. The Results of Web-page Summarization

Extraction	Compression rate (%)	Consistent Relevance between Human & the Automatic Summarizer (%)
Sentence	20	65
	25	70
	30	70
	40	75
Paragraph	20	85
	25	80
	30	90
	40	100

On the one hand, the results are indicated that the degree of dissimilarity between human summaries and system summaries in our experiments. Then, summary based on paragraph has easily selected by human and an automatic summarizer to coincide more than summary based on sentence.

5. Conclusion

Due to increasing of e-commerce websites, the number of information has been made available and accessible. It leads to a problem of information overload. Text summarization is believed that it can help users access to the information content in a shorter period of time. This is because it consists of reducing the size of a text while preserving its information content. This paper present applying automatic text summarization to e-commerce webpage summarization based on statistic similarity measures. The method consists of two main processes: topic identification and webpage summarization. The topic identification is to compute and extract keywords with term frequencies and weight of word. Meanwhile, the webpage summarization is to extract the most significant sentences and paragraphs. Eventually, the experimental results of sentence and paragraph extraction show en effective accuracy after testing by the precision (P). However, although overall of text summary can help users understand the content on the webpage in short time, it still lacks of polished gist. Therefore, the system should be improved with syntactic and semantic analysis approaches.

6. References

[1] Hahn, U. & Mani, I. (2000) The Challenges of Automatic Summarization. IEEE Computer. 33 (11): 29-35.

[2] Goldstein, J.; Kantrowitz, M.; Mittal, V. & Carbonell, J. (1999) Summarizing Text Documents: Sentence selection and evaluation metrics. In Proceeding of the 22nd ACM SIGIR. 121-128.

[3] Luhn, H. P. (1958) The Automatic Creation of Literature Abstarcts. IBM Journal of Research and Development. 159-165.

[4] Kan, M.Y. & Klavans, J.L. (2002) Using Librarian Techniques in Automatic Text Summarization for Information Retrieval. Proceedings of the Joint Conference on Digital Libraries (JCDL 2002), Portland, Oregon, USA: July 2002. pp. 36-45.

[5] Hovy, E.H. & Lin, C.Y. (1998) Automating Text Summarization in SUMMARIST. In I. Mani and M. Maybury (eds), Advances in Automated Text Summarization. Cambridge: MIT Press.

[6] Mani, I., Maybury M.T. (eds) (1999) Advances in Automated Text Summarization. Cambridge: MIT Press.

[7] Buyukkokten, O.; Garcia-Molina, H. & Paepcke, A. (2001) Seeing the whole parts: Text Summarization for web browsing on handheld devices. WWW10.

[8] Lam-Adesina, A.M. & Jones, G. J. F. (2001) Applying summarization techniques for term selection in relevance feedback. In Proceeding of the 24th ACM SIGIR.

[9] Frakes, W.B. (1992) Stemming Algorithms. In W.B. Frakes and R. Baeza-Yates. Information Retrieval-Data Structions and Algorithms. Prentice Hall. 131-160.

[10] Harman, D. (1992) Ranking Algorithms. In W.B. Frakes and R. Baeza-Yates. Information Retrieval-Data Structions and Algorithms. Prentice Hall. 363-392.

[11] Mike, S.; Itoh, E.; Ono, K. & Sumita, K (1994) A Full-Text Retrieval System with Dynamic Abstract Generation Function. In Proceedings of the 17th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR-94). 152—161.

[12] Kupiec, J.; Pedersen, J. & Chen, F. (1995) A Trainable Document Summarizer. In Proceedings of the 18th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR). 68-73. Seattle. WA.

[13] Aone, C.; Okurowski, M.E.; Gorlinsky, J. & Larsen, B. (1997) A Scalable Summarization System using Robust NLP. Proceedings of the Workshop on Intelligent Scalable Text Summarization, 66—73. In Proceeding of ACL/EACL Conference. Madrid. Spain.

[14] Strzalkowski, T. (1998) In I. Mani and M. Maybury (eds). Advances in Automated Text Summarization. Cambridge: MIT Press.

[15] Wu, K.L.; Chen, S.K. & Yu, P.S. (2002) Dynamic Refinement of Table Summarization for M-Commerce. Proceedings of the 4th IEEE Int'l Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems (WECWIS).

[16] Chen, M. & Sairamesh, J. (2005) Ranking-based Business Information Processing: Applications to Business Solutions and eCommerce Systems. Proceedings of the Seventh IEEE International Conference on E-Commerce Technology (CEC).

[17] Liu, S. (2005) Enhancing E-Business-Intelligence-Service: A Topic-Guided Text Summarization Framework. Proceedings of the Seventh IEEE International Conference on E-Commerce Technology (CEC).

[18] Joachims, T. (1999) Transductive Inference for Text Classification using Support Vector Machines. In Proceedings of the International Conference on Machine Learning (ICML).

[19] Zedeh, L. A. (1965) Fuzzy Sets. Information and Control. Vol 8. 338-353.

[20] Fox, E. A., Sharat, S. (1986) A comparison of two methods for soft Boolean interpretation in information retrieval. TR-86-1. Virginia Tech. Department of Computer Science.

[21] Amini, M. R. 1995. Interactive Learning for Text Summarization. University of Paris. France.

[22] Kanus, D.; Mittendorf, E.; Schauble, P. & Sheridan, P. (1994) Highlighting Relevant Passages for Users of the Interactive SPIDER Retrieval System. In Proceedings of the TREC-4.

[23] Baeza-Yates, R. & Ribeiro-Neto, B. (1999) Modern information retrieval. The ACM press.