Automatic Text Summarization in Engineering Information Management

Jiaming Zhan[1], Han Tong Loh[1], Ying Liu[2], and Aixin Sun[3]

[1] Department of Mechanical Engineering, National University of Singapore, Singapore 119077
{jiaming,mpelht}@nus.edu.sg
[2] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong SAR, China
mfyliu@inet.polyu.edu.hk
[3] School of Computer Engineering, Nanyang Technological University, Singapore 639798
axsun@ntu.edu.sg

Abstract. In today's knowledge-intensive engineering environment, information management is an important and essential activity. However, existing researches of Engineering Information Management (EIM) mainly focused on numerical data such as computer models and process data. Textual data, especially the case of free texts, which constitute a significant part of engineering information, have been somewhat ignored, mainly due to their lack of structure and the noisy information contained in them. Since summarization is a process to distill important information from source documents and at the same time remove irrelevant and redundant information, it could address the obstacles for handling textual data in EIM. Moreover, text summarization could address the increasing demand to integrate information from multiple documents and reduce the time in acquiring useful information from massive textual data in the engineering domain. This paper discusses in detail the need to apply text summarization in EIM and introduces a case study in summarizing multiple online customer reviews.

Keywords: Automatic Text Summarization, Engineering Information Management, Online Customer Reviews.

1 Textual Information Within Engineering Domain

Information management is an important and essential activity in today's knowledge-intensive engineering environment. Information lies at the core of a modern engineering environment, comprising not only numerical data but also textual data. Effective and efficient information management is one of the key factors by which industrial and engineering performance can be greatly improved [1].

However, most existing Engineering Information Management (EIM) applications focus on the handling of numerical data. Textual data, such as technical papers, patent documents, e-mails and customer reviews, which constitute a significant part of engineering information, have been somewhat ignored. There are probably three major reasons for this lack of attention:

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 347–350, 2007.

- Numerical data are well structured and organized in databases, making them relatively easy to handle. In comparison, textual data are usually stored as unstructured free texts or semi-structured data so that there is a greater level of difficulty in handling textual databases.
- Compared to the clean and purified numerical data, textual data contain a lot of noisy and redundant information.
- Most existing EIM applications have focused on design and manufacturing phases in which numerical information dominates.

As with numerical data, textual data offer a wealth of information in engineering activities, especially with the explosive growth of enterprise Intranet and the Internet [2, 3]. There has been an increasing demand of advanced techniques to reduce the time in acquiring useful information and knowledge from massive textual data, commonly appearing in technical papers, patent documents, reports, white papers, e-mails, Web pages, and notes from call centers [4].

2 The Need of Text Summarization Within EIM

Because of the rich information involved in textual data, how to utilize and how to discover knowledge from them effectively and efficiently has become a concern. However, only a few studies have been reported on textual information management within engineering domain, due to the aforementioned limitations. These existing studies can be divided into two major areas: information indexing & searching [5] and automatic text classification [6]. On the other hand, another important issue, i.e. integrating information from multiple textual sources and extracting useful information to fulfill users' requirements, has not yet been covered by previous studies.

Due to the current overload of engineering information, even with the powerful classification and searching tools, users often encounter a huge amount of retrieved documents for any given topic. Users have to screen these documents manually, which often takes a lot of time, until they satisfactorily identify documents relevant to their specific purposes. In such context, a summarization system, which can integrate the information from retrieved documents and facilitate the searching process, is much needed. The retrieved documents, regarding the same query, must share much common information which is interesting to readers. Besides, in some documents there must exist some unique information. Therefore, the summarization system should be able to integrate the common and unique information from all documents. At the same time, this summarization system should be able to exclude the redundant and noisy information across the documents.

Summarization is a process to distill the most important information from source documents and at the same time remove irrelevant and redundant information. Moreover, the output of a summarization system would be a well structured text compared to the unstructured source documents. Therefore, automatic text summarization could probably address the aforementioned limitations for handling textual information in EIM.

The first implementation of automatic text summarization can be traced back to 1950s [7]. During the last decade, there has been increasing interest with Multi-Document Summarization (MDS), as an outcome of the capability to collect large sets of documents online [8, 9]. The most popular MDS approach is clustering-summarization which separates a document set into non-overlapping clusters of documents and summarizes each cluster. However, when applied to the real-world engineering document sets, the number of clusters is difficult to determine, and moreover, topics often overlap with each other and are not perfectly distributed in non-overlapping clusters of documents [3].

3 Case Study: Summarizing Online Customer Reviews

This case study aims to investigate the domain of customer reviews, which constitute a typical kind of documents used in EIM. Some work has been reported dealing with the vast amount of customer reviews [10, 11]. All these work focused on opinion mining which was to discover the reviewers' orientations, whether positive or negative, regarding various features of a product, e.g. weight of a laptop and picture quality of a digital camera. However, opinion mining is not enough to cover all the important information from customer reviews and there is a desire to apply summarization techniques to identify the significant topics from multiple customer reviews [3].

In this case study, we propose a summarization approach based on the topical structure, which consists of a list of significant topics that are extracted from a document set [3]. The summarization performance was compared with the approaches of opinion mining and clustering-summarization. The data sets used in the experiment included five sets from Hu's corpus [10] and three sets from Amazon.com. These document sets were moderate-sized with 40 to 100 documents per set. The compression ratio of summarization, i.e. the length ratio of summary to original text, was set to 10%. Summarization performance was evaluated according to users' responsiveness. Human assessors were required to give a score for each summary based on its content and coverage of important topics in the review set. The score was an integer between 1 and 5, with 1 being the least responsive and 5 being the most responsive. In order to reduce bias in the evaluation, three human assessors from different backgrounds joined the scoring process. For one set, all the peer summaries were evaluated by the same human assessor so that the hypothesis testing (paired t-test) could be performed to compare the peer summaries.

Table 1 shows the average responsiveness scores of opinion mining, clustering-summarization and our approach based on all the review sets. Table 2 presents the results of paired t-test between our approach and other methods. It could be found that our approach based on topical structure performed significantly better than other peer methods.

Table 1. Average responsiveness scores

	Responsiveness score
Opinion mining	2.9
Clustering-summarization	2.3
Our approach	4.3

Table 2. Hypothesis testing (paired t-test)

Null hypothesis (H0): There is no difference between the two methods. Alternative hypothesis (H1): The first method outperforms the second one.	
	P-value
Our approach vs. opinion mining	1.91×10-3
Our approach vs. clustering-summarization	2.43×10-4

4 Conclusion

This paper reviews the existing EIM applications, with the focus on textual information management, and addresses the need to apply automatic text summarization in EIM. Moreover, a case study to summarize multiple online customer reviews is introduced. This work might enrich the research of EIM since it examined the textual information which has been somewhat ignored in the previous studies. In the future, we will apply the summarization techniques to other types of documents in engineering domain, such as technical papers and patent documents.

References

1. Hicks, B.J., Culley, S.J., McMahon, C.A.: A study of issues relating to information management across engineering SMEs. International Journal of Information Management 26(4), 267–289 (2006)
2. Liu, Y.: A concept-based text classification system for manufacturing information retrieval. Ph.D. Thesis, National University of Singapore (2005)
3. Zhan, J., Loh, H.T., Liu, Y.: Automatic summarization of online customer reviews. In: Proceedings of the 3rd International Conference on WEBIST, Barcelona, Spain (2007)
4. Blumberg, R., Atre, S.: The problem with unstructured data. DM Review (2003)
5. Fong, A.C.M., Hui, S.C.: An intelligent online machine fault diagnosis system. Computing and Control Engineering Journal 12(5), 217–223 (2001)
6. Menon, R., Loh, H.T., Keerthi, S.S., Brombacher, A.C., Leong, C.: The needs and benefits of applying textual data mining within the product development process. Quality and Reliability Engineering International 20(1), 1–15 (2004)
7. Luhn, H.P.: The automatic creation of literature abstracts. IBM Journal of Research and Development 2(2), 159–165 (1958)
8. Mani, I., Bloedorn, E.: Summarizing similarities and differences among related documents. Information Retrieval 1(1-2), 35–67 (1999)
9. Radev, D.R., Jing, H., Styś, M., Tam, D.: Centroid-based summarization of multiple documents. Information Processing and Management 40(6), 919–938 (2004)
10. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD, Seattle, WA, pp. 168–177 (2004)
11. Popescu, A.-M., Etzioni, O.: Extracting product features and opinions from reviews. In: Proceedings of HLT/EMNLP 2005, Vancouver, Canada, pp. 339–346 (2005)