Automatic Summarization for Financial News Delivery on Mobile Devices

Christopher C. C. Yang and Fu Lee Wang
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China
{yang, flwang}@se.cuhk.edu.hk

ABSTRACT
Wireless access with mobile devices is a promising addition to the WWW and traditional electronic business. Mobile devices provide convenience and portable access to the huge information space on the Internet. It is desire to access the most updated financial information through mobile devices in order to make critical and urgent decision for most of the investors. In this paper, we present a financial news delivery system on mobile devices based on the fractal summarization model. *Fractal summarization* is developed based on the fractal theory. It generates a brief skeleton of summary at the first stage, and the details of the summary on different levels of the document are generated on demands of users. Such interactive summarization reduces the computation load in comparing with the generation of the entire summary in one batch by the traditional summarization, which is ideal for wireless access.

Index Terms: Document summarization, financial news delivery, fisheye view, fractal view, handheld devices, mobile commerce.

1. INTRODUCTION

Access to the Internet through mobile phones and other handheld devices is growing significantly in recent years. A lot of user-centered m-services applications, such as web surfing, e-mail checking, and stock price quoting, have been developed. However, m-services should not be limited to user-centered applications but extended to knowledge management. There is a large amount of financial news generated in the Internet everyday. With a fast paced economy, an m-commerce organization must gain advantage by accessing the most updated and accurate financial information available and make decision as fast as possible. As a result, it is desire to have financial news delivery through mobile devices so that investors can retrieve relevant information anywhere any time. Because of the huge volume of the news generated everyday, most of news delivery services provide summarization tools to support users in searching relevant information through Web browser on PC platforms, such as Lycos Financial Feed System with summarization system from Diyatech and YellowBrix with Inxight's Summarizer. Unfortunately, there are many shortcomings associated with mobile devices, such as limited screen size, narrow network bandwidth, small memory capacity and low computing power. Summarizers for PC platform are not adaptable to mobile devices directly. In order to reduce the information displayed and downloading time, a WAP gateway is setup to summarize the news for users to preview its major content. The wireless handheld devices can conduct interactive navigation with the gateway through wireless network to retrieve the summary piece by piece. In this paper, we present a financial news delivery system on mobile devices based on the fractal summarization model. In addition, information visualization techniques are presented to reduce the visual loads.

2. FRACTAL SUMMARIZATION MODEL

Traditional automatic text summarization is the selection of sentences from the source document based on their significance to the document [2][9]. The selection is based on the salient features of document, such as thematic, location, title, and cue features.

- The thematic feature is first identified by Luhn [9], the *tfidf* (term frequency inverse document frequency) [12] method is currently most widely used approach. The system calculates the *tfidf* score for each term in the document first, and the thematic weight of sentence is calculated as the sum of *tfidf* score of its constituent words.
- The significance of sentence is indicated by its location based on the hypotheses that topic sentences tend to occur at the beginning or in the end of documents or paragraphs [2]. Therefore, the location weight of sentence can be calculated by a simple function of its ordinal location in the document.
- The title feature is proposed based on the hypothesis that the author conceives the title as circumscribing the subject matter of the document [2]. A dictionary of heading keywords with *tfidf* weights is automatically constructed from the heading sentences of document first. The heading weight of sentence is calculated as the sum of heading weight of its constituent words.
- The cue feature is proposed by Edmundson [2] based on the hypothesis that the probable relevance of a sentence is affected by the presence of pragmatic words. A pre-stored dictionary of cue phrase with cue weights is used for calculation of cue weight. The cue weight of sentence is calculated as the sum of cue weight of its constituent words.

Typical summarization systems obtain the sentence weights by computing the weighted sum of the weights of all the features [2][8]. The sentences with sentence weight higher than a threshold value are selected as part of the summary. It has been proved that the weighting of different features does not have any substantial effect on the average precision [8]. The maximal weights of each feature are normalized to one in our system.

The traditional summarization models consider the source document as a sequence of sentences. However, many studies [3][5] of human abstraction process have shown that the human abstractors extract the topic sentences according to the document structure from the top level to the low level until they have extracted sufficient information. On the other hand, it is believed that the document summarization on handheld devices must make use of "tree view" [1] and "hierarchical display", which is not suitable for a sequence of sentences. *Fractal summarization model* is developed based on the fractal theory [10]. In fractal summarization, the important information is captured from the source text by exploring the hierarchical structure and salient features of the document. A condensed version of the document that is informatively close to the original is produced iteratively using the contractive transformation in the fractal theory. Similar to the fractal geometry, large document has a hierarchical structure with several levels, chapters, sections, subsections, paragraphs, sentences, terms, words and characters. At the lower abstraction level of a document, more specific information can be obtained. Although a document is not a true mathematical fractal object since a document cannot be viewed in an infinite abstraction level, we may consider a document as a *prefractal* [4]. The lowest abstraction level in our consideration is a term. The fractal summarization model applies a similar technique as fractal image compression [7]; it generates the summary by a simple recursive deterministic algorithm based on the iterated representation of a document. In fractal summarization, the user specifies the *compression ratio* of summarization. The default value of compression ratio is 4%, because high-compression ratio summary can achieve a reasonable high precision [13] and it can save network bandwidth. The system will use the compression ratio to calculate the total number of sentences to be extracted as the summary and the source document is segmented into range blocks of text according to the document structure (Figure 1). The system will calculate the sentence weight as the traditional summarization and allocate the quota of sentence to each range block proportionally to the sum of sentence weight in the range block. Each range block is then iteratively partitioned to child blocks and the quota is propagated down the summarization tree according to the sum of sentence weights in the child blocks until a contractive mapping is found to transform the text block to less than five sentences by traditional summarization methods, because it is proven that the optimal length of summary by extraction of fixed number of sentences is three to five sentences [6]. The summaries generated by fractal summarized are remained as tree structure, which are suitable for hierarchical display on handheld devices. Experiments have shown that the fractal summarization outperforms the traditional summarization [14].

3. FRACTAL SUMMARIZATION OF FINANCIAL NEWS AT YAHOO! NEWS

Fractal summarization model summarize the documents based on hierarchical document structure. In addition to large text document, a lot of other documents also exhibit hierarchical document structure, such as web-site and newspaper. The model is applied to summarize the financial news downloaded from Yahoo! News. Because a large volume of news articles is generated everyday, categorization is required for easy searching and browsing of news. For example, there are twenty-one categories in the Yahoo! News, each of them will be subdivided into subcategories. Each subcategory contains around ten news articles, each news article may contain more than one section, and each section contains few paragraphs, each paragraph contains few sentences. As we are interested with financial news, we will focus on Yahoo! News-'Business' category only.

Figure 1 illustrates the fractal summarization of Yahoo! News-'Business' category. Fractal summarization generates a brief skeleton of summary at the first stage, and the details of the summary at different levels of the news tree are generated on demands of users. The system will first show a card contains with 6 subcategories of 'Business' category (Figure 2a), it gives user a general idea how the news articles are organized, and the user can select subcategory to obtain more details. Such interactive summarization reduces the computation load in comparing with the generation of the entire summary in one batch by the traditional automatic summarization, which is ideal for m-services.

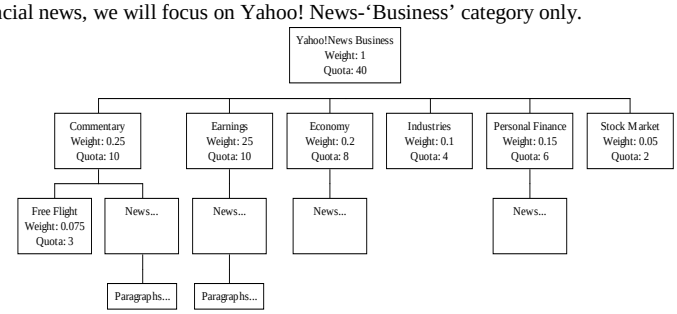
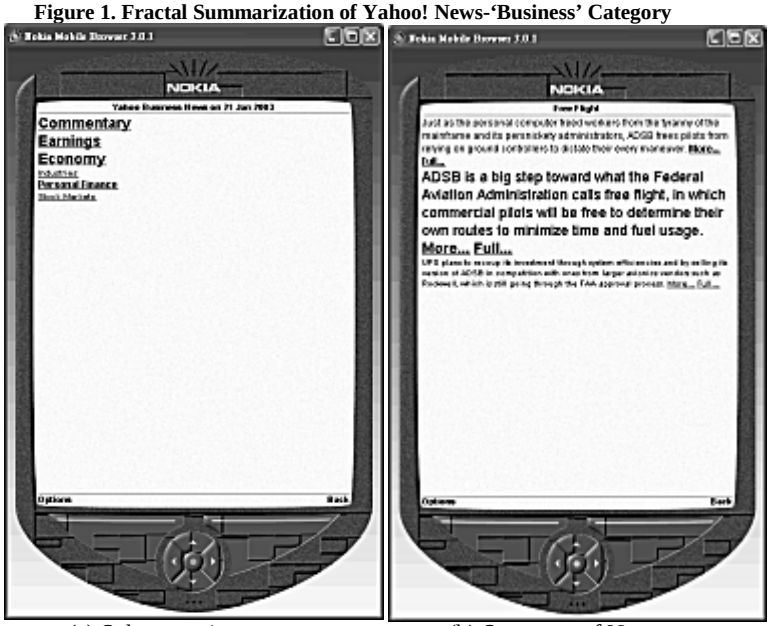


Figure 1. Fractal Summarization of Yahoo! News-'Business' Category

Given a card of a summary node, there may be too many sentences or child-nodes to be visualized or displayed in the small screen of the hand held devices. In our system, the size of objects depends on the significance of the objects. The 3-scale font mode available for WML is utilized. The prototype system using Nokia Handset Simulator is presented on Figure 2. As shown in Figure 2a, 3 subcategories of Business category are displayed in large font, which means that they are more important; and the rest are in normal font or small font according to their importance. When the user click the anchor link of subcategory, the WAP gateway will delivery a card depends on the quota allocated. If a large quota is allocated to the subcategory, the system will show another card containing of index of news article. However, if the quota is less than 5 sentences, the system will show a card with the summary of all news articles in the subcategory (Figure 2b). In the summary page, when the user clicks the anchor link 'More' at end of each sentence, the system will generate the summary for the corresponding news articles with compression ratio 20%, because it has been proved that extraction of 20% sentences can be as informative as the full text of the source document [11]. On the other hand, the user can clicks the anchor link 'Full' to view the full text of the news articles.



(a) Subcategories (b) Summary of News

Figure 2. Screen of WAP Summarization System

4. REFERENCES

[1] Buyukkokten O. et al., 2001. "Accordion Summarization for End-Game Browsing on PDAs and Cellular Phones". *Human-Computer Interaction Conf. 2001 (CHI 2001)*. Washington.

[2] Edmundson H. P., 1968. "New Method in Automatic Extraction". *Journal of the ACM*, 16(2) 264-285.

[3] Endres-Niggemeyer B. et al., 1995. "How to Implement a Naturalistic Model of Abstracting". *Info. Pro. & Man. 31(5) 631-674.*

[4] Feder J., 1988. *Fractals*. Plenum, New York.

[5] Glaser B. G. et al., 1967. "The discovery of grounded theory; strategies for qualitative research". Aldine de Gruyter, New York.

[6] Goldstein J. et al., 1999. "Summarizing text documents: Sentence selection and evaluation metrics". In *Proc. of SIGIR*, 121-128.

[7] Jacquin A., 1993. "Fractal image coding: A review". In *Proc. of the IEEE, 81(10) 1451-1465.*

[8] Lam-Adesina M. et al., 2001. "Applying summarization Techniques for Term Selection in Relevance Feedback", In *Proc. of SIGIR 2001, 1-9*.

[9] Luhn H. P., 1958. "The Automatic Creation of Literature Abstracts". *IBM Journal of R & D*, 159-165.

[10] Mandelbrot B., 1983. The fractal geometry of nature, New York: W.H. Freeman.

[11] Morris G. et al., 1992. "The effect and limitation of automated text condensing on reading comprehension performance". *Info. Sys. Research, 17-35*.

[12] Salton G. et al., 1988. "Term-Weighting Approaches in Automatic Text Retrieval", *Info. Pro. & Man., 24, 513-523*.

[13] Teufel S. et al., 1998. "Sentence Extraction and rhetorical classification for flexible abstracts", *AAAI Spring Sym. on Intel. Text Summarization*, Stanford.

[14] Yang C. C., and Wang, F. L., 2002. "Document Summarization on Handheld Device" In *Proc. of Workshop on e-Business WEB2002,* Barcelona.