Fractal Summarization: Summarization Based on Fractal Theory

Christopher C. Yang
Department of Systems Eng. & Eng. Management
Chinese University of Hong Kong
Shatin, Hong Kong SAR, China
yang@se.cuhk.edu.hk

Fu Lee Wang
Department of Systems Eng. & Eng. Management
Chinese University of Hong Kong
Shatin, Hong Kong SAR, China
flwang@se.cuhk.edu.hk

ABSTRACT
In this paper, we introduce the fractal summarization model based on the fractal theory. 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. User evaluation has shown that fractal summarization outperforms traditional summarization.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing– *Abstracting Methods.*

General Terms: Algorithms, Experimentation, Theory.

Keywords: Document summarization, fractal summarization.

1. INTRODUCTION
As the Internet is growing exponentially, huge amount of information are available online. It is difficult to identify the relevant information. The information-overloading problem can be reduced by automatic summarization. Many summarization models have been proposed previously. None of the models are entirely based on document structure, and they do not take into account of the fact that the human abstractors extract sentences according to the hierarchical document structure. Document structure can be described as fractals that are some mathematical objects with high degree of redundancy. In the past, fractal theory has been widely applied in the area of digital image compression, which is similar to the text summarization in the sense that they both extract the most important information from the source and reduce the complexity of the source. The fractal summarization model is the first effort to apply fractal theory to document summarization. It generates the summary by a recursive deterministic algorithm based on the iterated representation of a document. The fractal summarization highly improves the divergence of information coverage of summary and it is robust and transparent, the user can easily control the compression ratio, and the system generates a summary that maximize the information coverage and minimize the distance of summary from the source document.

2. TRADITIONAL SUMMARIZATION
Related research has shown that human abstractors use readymade text passages [3], 80% sentences in abstract were closely matched with source document [10]. As a result, traditional automatic text summarization is selection of sentences from source document based on the salient features of document, such as thematic, location, title, and cue features [1][12].

- The thematic feature is first identified by Luhn [12], the *tfidf* (term frequency inverse document frequency) [14] is most widely used. The *tfidf* scores for the terms are calculated first, and the thematic weight of sentence is calculated as the sum of *tfidf* score of its constituent terms.
- 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 [1]. The location weight of sentence is calculated by a simple function of its ordinal location in the document.

SIGIR'03, July 28-August 1, 2003, Toronto, Canada.
ACM 1-58113-646-3/03/0007

- The title feature is proposed based on the hypothesis that the author conceives the title as circumscribing the subject matter of the document [1]. A dictionary of heading terms with weights is automatically constructed from the heading sentences. The heading weight of sentence is calculated as the sum of heading weight of its constituent terms.
- The cue feature is proposed by Edmundson [1] based on the hypothesis that the probable relevance of a sentence is affected by the presence of pragmatic words. The cue weight of sentence is calculated as the sum of cue weight of its constituent terms from a pre-stored cue dictionary.

Typical summarization systems obtain the sentence weights by computing the weighted sum of the weights of all the features [1][11]. 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 [11]. In our system, the maximum weight of each feature is normalized to one.

3. FRACTAL SUMMARIZATION
Many studies [2][5] in human abstraction have shown that the human abstractors extract the topic sentences according to the document structure from top level to low level until they have extracted sufficient information. *Fractal Summarization Model* is proposed based on document structure and fractal theory. Fractals are mathematical objects that have high degrees of redundancy [13]. Similar to the geometrical fractal, large document has a hierarchical structure with multiple levels, chapters, sections, subsections, paragraphs, sentences, and terms. At the lower abstraction level, more specific information can be obtained. Although a document is not a true fractal object since it cannot be viewed in an infinite abstraction level but it can be considered as *prefractal* that is a fractal structure in its early stage with finite recursions [4].

Fractal view is fractal-based method for controlling information displayed [9]. The fractal tree is extended to any logical tree. The fractal value of root of a tree is set to 1, and the fractal value is propagated to other nodes by dividing the fractal value of parent node by the number of child nodes and assigning the value to the child node as their fractal value. A threshold value is chosen to control the amount of information displayed, the nodes with a fractal value less than the threshold value will be hidden.

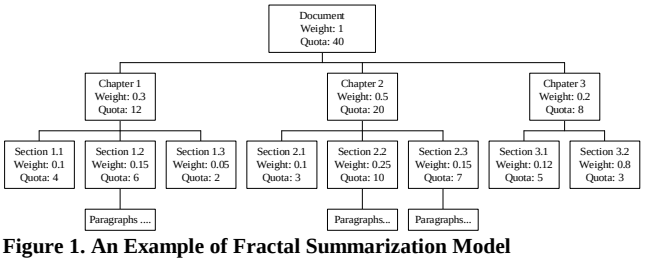

Figure 1. An Example of Fractal Summarization Model

The Fractal Summarization Model is developed based on the models of fractal view and fractal image compression [8]. The source document is partitioned into range-blocks according to document structure and represented as a fractal tree (Figure 1). The fractal value of each node is calculated as the sum of sentence weights of the sentences under the range-block. User may choose a *compression ratio* to specify the ratio of sentences to be extracted as the summary. The sentence quota of the summary can be calculated accordingly and it will be propagated to the child-nodes directly proportional to their fractal values. Figure 1 illustrates an example; the total sentence quota of root node is 40. There are three child-nodes under the root node with fractal values 0.3, 0.5 and 0.2, therefore the child nodes are allocated with sentence quota of 12, 20 and 8 sentences respectively. For each child node, the quota will be

shared by grandchild nodes. As it was proven that the optimal length of summary for summarization by extraction of fixed number of sentences is 3 to 5 sentences [6], if quota of a node exceeds 5 sentences (the default threshold value in our system), the system will process its child nodes iteratively until the quota is less than threshold. For example, the Section 1.1 and Section 1.3 are allocated with 4 and 2 sentences; the system extracts sentences from these sections as part of summary. However, the Section 1.2 is allocated with 6 sentences, then, the system continues to process its paragraphs. The details of algorithm are shown as below:

Fractal Summarization Algorithm

1. Choose a Compression Ratio and Threshold Value.
2. Calculate the total Sentence Quota of the summary.
3. Partition the document into range blocks.
4. Transform the document into fractal tree.
5. Set the current node to the root of the fractal tree.
6. **Repeat**

6.1 **For** each child node under current node,
Calculate the fractal value of child node.

6.2 Allocate Quota to child nodes in proportion to fractal values.

6.3 **For** each child nodes,
If the quota is less than threshold value
Select the sentences in the range block by extraction
Else
Set the current node to the child node
Repeat Step 6.1, 6.2, 6.3

7. **Until** all the child nodes under current node are processed.

The traditional automatic summarization techniques adopt the traditional salient features, but they consider the document as a sequence of sentences. In the fractal summarization, the traditional salient features are adopted and the hierarchical fractal structure is also considered. The fractal values of the nodes in the hierarchical fractal structure are computed based on the traditional salient features.

- Among the keyword features proposed previously, the *tfidf* score of keyword is the most widely used approach; however, it does not take into account of the document structure in the traditional summarization. Most researchers assume that the weight of a term remains the same over the entire document. However, Hearst claimed that a term should carry different weight in different location of a full-length document [7]. For example, a term is considered as more important in a range-block than other range-blocks if the term appears in the range-block more frequently than other range-blocks. In fractal summarization model, the *tfidf* of a term in a range block is defined as proportional to the term frequency within a range-block and inversely proportional to the frequency of range-block containing the term.
- Traditional summarization systems assume that the location weight of a sentence is static, however the fractal summarization calculate the location weight based on which document-level we are looking at. For example, if we consider the first and second sentences on the same paragraph at the paragraph-level, the first sentence is much important to the paragraph, however, the difference is insignificant if we are looking at the whole document. In the fractal summarization model, we calculate the location weight for a range-block by traditional methods, all sentences inside a range-block will receive same position weight.
- At different abstraction level, some headings should be hidden and some headings are emphasized. For example, only the document heading is considered if we look at the document-level. However, if we look at the chapter-level, we consider the document heading and chapter heading, and we consider chapter heading as more important since the main concept of this chapter is represented by the chapter heading. Therefore, the significance of the heading is inversely proportional to its distance from the sentence. Propagation of fractal value [9] is a promising approach to calculate the heading weight for a sentence.
- When human abstractors extract the sentences, they pay more attention to the range block with heading contains some bonus word such as "conclusion", since they consider it as a more important part and more sentences are extracted. The cue feature of heading sentence is classified as rhetorical feature [15]. We proposed to consider the cue feature not only in sentence-level. Give a document tree, we examine the heading of each range-block and adjust their quota accordingly. This procedure can be repeated to sub range-blocks until sentence-level.

4. EXPERIMENTAL RESULT

A user evaluation has been conducted using the Hong Kong Annual Report 2000 as the corpus. Five subjects were asked to evaluate the quality of summaries produced by the traditional summarization technique and the fractal summarization. Summaries generated by both techniques on 23 documents are assigned to subjects in random order. The precision, measured as the ratio of sentences accepted by the user as part of summary, is shown in Table 1. The precision of the fractal summarization is higher that that of the traditional summarization. The fractal summarization can achieve up to 88.75% precision and 84% on average, while the traditional summarization can achieve up to 77.5% precision and 67% on average. Besides, it is believed that a full-length text document contains a set of subtopics [7] and a good quality summary should cover as many subtopics as possible. In the experiment, the traditional summarization model extracts most sentences mainly from few chapters and no sentence is extracted from some chapters, but fractal summarization model extracts the sentences distributed in all chapters. Therefore, the fractal summarization model produces a summary with a wider coverage of information subtopics.

Table 1. Precision of Two Summarization Models

User ID	Fractal Summarization	Traditional Summarization
1	81.25%	71.25%
2	85.00%	67.50%
3	80.00%	56.25%
4	85.00%	63.75%
5	88.75%	77.50%

5. CONCLUSION

The fractal summarization model based on the statistical data and the structure of documents has been proposed. Thematic feature, location feature, heading feature, and cue features are adopted. Experiments have been conducted and the results show that the fractal summarization outperforms the traditional summarization.

6. REFERENCES

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

[2] Endres-Niggemeyer B. et al. How to Implement a Naturalistic Model of Abstracting. Info. Processing & Management. 31 631-674, 1995.

[3] Endres-Niggemeyer B. SimSum: an empirically founded simulation of summarizing. Info. Processing & Management, 36, 659-682. 2000.

[4] Feder J. Fractals. Plenum, N.Y., 1988.

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

[6] Goldstein J. et al. Summarizing text documents: Sentence selection and evaluation metrics. In Proc. SIGIR'99, 121-128, 1999.

[7] Hearst M. A. Subtopic Structuring for Full-Length Document Access. In Proc. SIGIR'93, 56-68, 1993.

[8] Jacquin. A. E. Fractal image coding: A review. In Proc. of IEEE, 81(10) 1451-1465, 1993.

[9] Koike, H. Fractal Views: A Fractal-Based Method for Controlling Information Display, ACM Tran. on IS, 13(3) 305-323, 1995.

[10] Kupiec J. et al. A Trainable Document Summarizer. Proc. SIGIR'95, 68-73, Seattle, USA. 1995.

[11] Lam-Adesina M. et al. Applying summarization Techniques for Term Selection in Relevance Feedback, In Proc. SIGIR'01, 1-9, 2001.

[12] Luhn H. P. The Automatic Creation of Literature Abstracts. IBM J. of R&D, 159-165, 1958.

[13] Mandelbrot B. The fractal geometry of nature. Freeman, N.Y. 1983.

[14] Salton G. et al. Term-Weighting Approaches in Automatic Text Retrieval. Info. Processing & Management, 24, 513-523, 1988.

[15] Teufel S. et al. Sentence Extraction and rhetorical classification for flexible abstracts, AAAI'98 Spring Sym., Stanford, 1998.