Hierarchical Summarization of Large Documents

Christopher C. Yang
College of Information Science and Technology, Drexel University, Philadelphia, PA 19104.
E-mail: chris.yang@ischool.drexel.edu

Fu Lee Wang
Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong SAR, China

Many automatic text summarization models have been developed in the last decades. Related research in information science has shown that human abstractors extract sentences for summaries based on the hierarchical structure of documents; however, the existing automatic summarization models do not take into account the human abstractor's behavior of sentence extraction and only consider the document as a sequence of sentences during the process of extraction of sentences as a summary. In general, a document exhibits a well-defined hierarchical structure that can be described as fractals—mathematical objects with a high degree of redundancy. In this article, we introduce the fractal summarization model based on the fractal theory. The important information is captured from the source document by exploring the hierarchical structure and salient features of the document. A condensed version of the document that is informatively close to the source document is produced iteratively using the contractive transformation in the fractal theory. The fractal summarization model is the first attempt to apply fractal theory to document summarization. It significantly improves the divergence of information coverage of summary and the precision of summary. User evaluations have been conducted. Results have indicated that fractal summarization is promising and outperforms current summarization techniques that do not consider the hierarchical structure of documents.

Introduction

As the Internet is growing exponentially, huge amounts of information are available online. It is difficult to identify the relevant information to satisfy the information needs of users. The problem of information overloading can be reduced by automatic summarization together with conventional information search engines to efficiently access the relevance of retrieved documents. Many summarization models have been proposed previously (Edmundson, 1969; Luhn, 1958). Existing summarization models consider the document as a sequence of sentences. Although some summarization systems calculate the sentence scores partially based on the document structure, they do not summarize the document based on the importance of the components in the document structure and their propagated properties. Document structures can be described as *fractals* that are mathematical objects with a high degree of redundancy (Mandelbrot, 1983). *Fractal theory* has been widely applied in the area of digital image compression (Jacquin, 1993) and *information visualization* (Koike, 1995; Yang, Chen, & Hong, 2003), which is similar to text summarization in the sense that they both extract the most important information content from the source. The *fractal summarization model* is the first attempt to apply fractal theory to document summarization.

Related research has shown that human abstractors extract the sentences based on the outline of a document (Endres-Niggemeyer, Maier, & Sigel, 1995; Glaser & Strauss, 1967). Individuals search for topic sentences starting from the top level to the lower levels of documents until sufficient information has been found. The current summarization models do not take into account of the fact that the human abstractors extract sentences according to the hierarchical document structure. The proposed fractal summarization model extracts the summary by a recursive deterministic algorithm based on the iterated representation of a document. The original document is represented as a *fractal tree* according to its document structure. The system extracts the sentences from the top level to lower levels. It computes the fractal value of each node in the fractal tree and determines the number of sentences to be extracted from each node. This process is iteratively applied to the child-nodes, and the key sentences are extracted utilizing statistical methods and salient features of documents.

For the purpose of this study, users' evaluations have been conducted. Results obtained from these evaluations have indicated that fractal summarization outperforms the current summarization techniques that do not utilize the hierarchical structure of documents. The fractal summarization model

Received July 12, 2006; revised October 3, 2007; accepted October 3, 2007

 DOI: 10.1002/asi.20781

significantly improves the divergence of information coverage of a summary. In addition, since the fractal summarization system is robust and transparent, users can easily control the compression ratio and interact with the summarization system. Consequently, the system will extract a summary that will maximize the information coverage and minimize the distance of summary from the source document.

The article is organized as follows. We introduce the sentence extraction methods in current automatic summarization which are based on some widely used salient features of documents. Next, we introduce the fractal theory and its application on fractal view for controlling the amount of information displayed. The following section proposes the fractal summarization model based on the statistical data and the hierarchical structure of documents. Thematic, location, heading and cue features are adopted. Then we will present the results of fractal summarization. Finally, we discuss the significance and findings of this work.

Automatic Text Summarization Using Salient Features

Related research has shown that human abstractors use readymade text passages from source document for summarization (Endres-Niggemeyer, 2002). Eighty percent of the sentences in the manmade abstracts were closely matched with sentences in source documents (Kupiec, Pedersen, & Chen, 1995). As a result, selection of representative sentences is considered as a good approximation of summarization (Aminin & Gallinari, 2002). The existing automatic text summarization is mainly the selection of sentences from the source document based on their significance in the document using statistical techniques and techniques based on surface domain-independent linguistic analyses (Aminin & Gallinari, 2002; Luhn, 1958; Radev & McKeown, 1998). The statistical approach of selection of sentences is conducted based on the salient features of the document. The thematic, location, heading, and cue features are the most widely used extraction features.

- The thematic feature was first identified by Luhn (1958). Edmundson (1969) proposed that one assign each term in the document a thematic weight based on its term frequency. The thematic weight of a sentence is calculated as the sum of thematic weight of its constituent keywords (Edmundson, 1969). In information retrieval, absolute term frequency by itself is considered as less useful than term frequency normalized to the document length and term frequency in the collection (Salton & Buckley, 1988). As a result, the *tfidf* (Term Frequency × Inverse Document Frequency) method is proposed to calculate the thematic weight of keyword (Salton & Buckley, 1988). Most recent summarization systems use the *tfidf* score to compute the sentence score based on thematic feature for sentence $k(s_k)$, $SS_T(k)$.

$$SS_T(k) = \sum_{t_i \in s_k} w_{ij}$$

where w_{ij} is the *tfidf* score of term $i(t_i)$ in document j

$$w_{ij} = tf_{ij} \times \log_2\left(\frac{N \times |t_i|}{n}\right)$$

where tf_{ij} is the term frequency of term i in document j, N is the total number of documents in the corpus, n is the number of documents in the corpus in which term i occurs and $|t_i|$ is the length of term i.

- The significance of a sentence is indicated by its location (Baxendale, 1958) based on the hypotheses that representative sentences tend to occur at the beginning or at the end of documents or paragraphs (Edmundson, 1969). Edmundson (1969) proposed to assign positive weights to sentences according to their ordinal position in the document; for example, the sentences in the first and last paragraphs and the first and last sentences of the paragraphs are more important. There are several functions proposed to calculate the location weight of sentences; they commonly assign the sentences at the beginning and at the end of a document relatively higher location weights. Alternatively, the preference of sentence location can be stored in a list called Optimum Position Policy, and the sentence will be selected based on their order in the list (Lin & Hovy, 1997). The sentence score based on location feature for sentence $k(s_k)$ is computed as follows:

$$SS_L(k) = \frac{1}{\min(k, K - k + 1)}$$

where K is the total number of sentences in the document.

- The heading feature is proposed based on the hypothesis that the author conceives the heading as circumscribing the subject matter of the document. When the author partitions the document into major sections, he summarizes it by choosing the appropriate headings (Baxendale, 1958). The weight of a heading is very similar to the keyword approach. A heading glossary is a list consisting of all words in the titles, headings, and subheadings. Positive weights are assigned to the heading glossary, where the heading words will be assigned a weight relatively prime to the heading words. The heading weight of a sentence is calculated by the sum of heading weights of its constituent words. The sentence score based on the heading feature for sentence $k(s_k)$, $SS_L(k)$, is computed as the sum of the *tfidf* scores of heading keywords that appears in s_k.

$$SS_H(k) = \sum_{t_i \in s_k \cap Heading(S_k)} w_{ij}$$

where $Heading(s_k)$ is the heading of s_k.

- The cue phrase approach was proposed by Edmundson (1969) based on the hypothesis that the probable relevance of a sentence is affected by the presence of pragmatic words such as "significant," "impossible," and "hardly." A cue dictionary is preconstructed to identify the cue phrases, which comprises of three subdictionaries: (a) bonus words, which are positively relevant; (b) stigma words, which are negatively relevant; and (c) null words, which are irrelevant. The sentence score based on the cue feature for sentence $k(s_k)$, $SS_C(k)$, is computed as the sum of the cue weights of

DOI: 10.1002/asi

constituent terms t_i of s_k.

$$SS_C(k) = \sum_{t_i \in s_k} w_{cue}(t_i)$$

where $w_{cue}(t_i)$ is the cue weight of term t_i in the cue dictionary.

Typical summarization systems obtain the sentence significance score for sentence k, $SSS(k)$, by computing the weighted sum of the sentence scores based on all the features (Edmundson, 1969; Lam-Adesina & Jones, 2001).

$$SSS(k) = a_1 \times SS_T(k) + a_2 \times SS_H(k) + a_3 \times SS_L(k) + a_4 \times SS_C(k)$$

where a_1, a_2, a_3, and a_4 are positive integers to control the weighting of four components.

The sentences with sentence significance scores 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 of the summarization system (Lam-Adesina & Jones, 2001).

Fractal Theory and Fractal View

Fractals are mathematical objects that have a high degree of redundancy (Mandelbrot, 1983). These objects are made of transformed copies of themselves or part of themselves. Mandelbrot (1983) was the first researcher who investigated the fractal geometry and developed the fractal theory. In his well-known example, the length of the British coastline depends on the measurement scale. The larger the scale is, the smaller the value of the length of the coastline and the higher the abstraction level. The British coastline includes bays and peninsulas. Bays include subbays, and peninsulas include subpeninsulas. Using fractals to represent these structures, abstraction of the British coastline can be generated with different abstraction levels. Fractal theory is grounded in geometry and dimension theory. Fractals are independent of scale and appear equally detailed at any level of magnification. Such property is known as self-similarity. Any portion of a self-similar fractal curve appears identical to the whole curve. If we shrink or enlarge a fractal pattern, its appearance remains unchanged.

Fractal view is a fractal-based method for controlling information displayed (Koike, 1995). Fractal view provides an approximation mechanism for the observers to adjust the abstraction level and therefore control the amount of information displayed. At a lower abstraction level, more details of the fractal objects can be viewed.

A physical tree is a classical example of fractal objects. A tree is made of subtrees; each of them also is a tree. By changing the scale, different levels of abstraction views are obtained. The idea of a fractal tree can be extended to any logical tree. The degree of importance of each node is represented by its fractal value. The fractal value of focus is set to 1. Regarding the focus as a new root, we propagate the fractal value to the other nodes with the following expression:

$$Fv_{focus} = 1$$
$$Fv_{child\text{-}node\ of\ x} = Fv_x \, \mathrm{CN_x}^{-1/D}$$

where Fv_x is the fractal value of node x; C is a constant between 0 and 1 to control rate of decay; N_x is the number of child-nodes of node x; and D is the fractal dimension. The main idea of propagation of fractal value is that the fractal value of the parent-node is shared by its child-nodes with a constant to control the rate of decay.

In fractal view, 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.

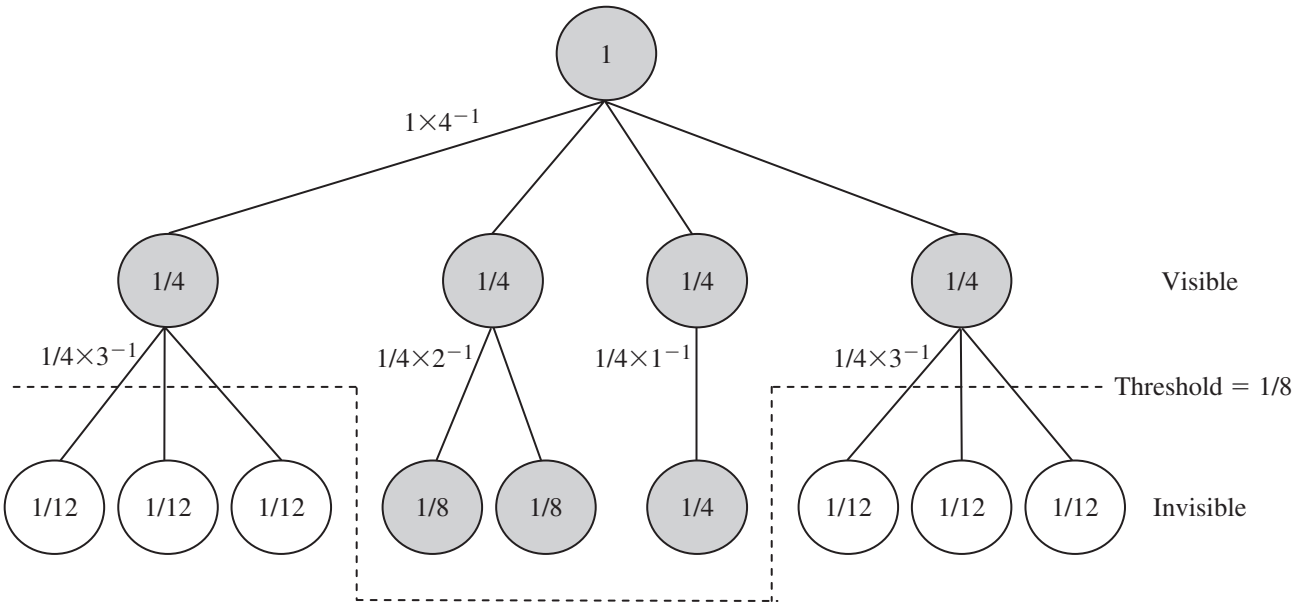
Figure 1 illustrates an example of propagation of the fractal values in a fractal tree, and the fractal view generated at different threshold values. In Figure 1a, the threshold value is chosen as 1/8; eight nodes with fractal value larger than or equal to the threshold value is visible to users. In Figure 1b, the threshold value is increased to 1/4; the number of nodes with fractal value larger than or equal to the threshold value decreased to six. Therefore, only six nodes are visible to users, and other nodes become invisible. By changing the threshold value, the user can adjust the amount of information displayed.

Fractal Summarization

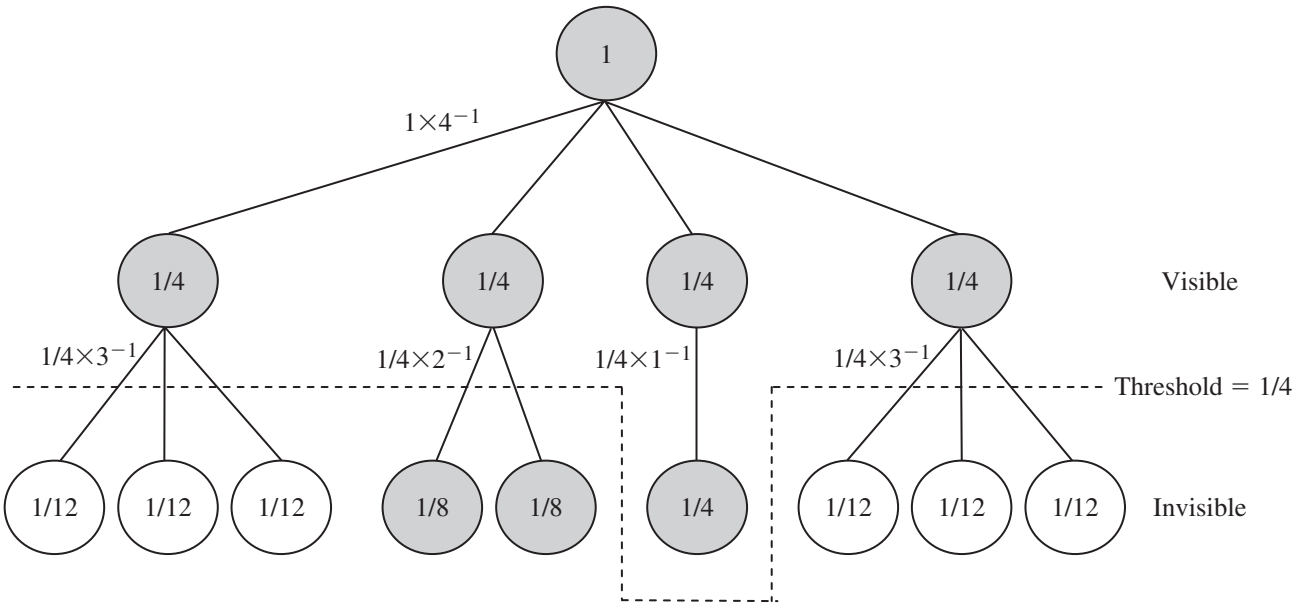
Many studies (Endres-Niggemeyer et al., 1995; Glaser & Strauss, 1967) of human abstraction have shown that the human abstractors extract the topic sentences according to the document structures from the top level to the low level until they have extracted sufficient information. A *fractal summarization model* is proposed to extract a summary based on document structure and fractal theory. Similar to the geometrical fractal, a large document has a hierarchical structure with several levels, chapters, sections, subsections, paragraphs, sentences, and terms (Figure 2). At lower abstraction levels, more specific information can be obtained. A document is not a true fractal object since it cannot be viewed at an infinite abstraction level; however, it can be considered as *prefractal*; that is, fractal structure at its early stage with finite recursions (Feder, 1988). The smallest units are terms within a document in the fractal summarization model.

The fractal summarization model was developed based on fractal view and fractal image compression (Jacquin, 1993). The source document is first partitioned into range-blocks according to document structure, and the document is then transformed into a fractal tree as its natural document structure (Figure 3). Each range-block in the document is represented by a node in the fractal tree. The fractal value of each node is calculated as the sum of sentence weights of the sentences under the range-block. A user may choose a *compression ratio* to specify the ratio of sentences to be extracted as the summary. The total number of sentence quotas of the

DOI: 10.1002/asi



(a) Threshold = 1/8



(b) Threshold = 1/4

FIG. 1. An example of the propagation of fractal values and different fractal views generated at different threshold values.

summary can be calculated accordingly, and it will be propagated to the child-nodes directly proportional to their fractal values.

Figure 3 illustrates the fractal summarization. The total sentence quota of the 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 a sentence quota of 12, 20, and 8 sentences, respectively. The system then processes its child-nodes one by one, and then the sentence quota of the child-nodes will be propagated to grandchild-nodes according to the fractal value of grandchild-nodes. For example, the quota of chapter 1 will be shared by its three child-nodes. As it was proven that the optimal length of a summary for summarization by extraction of a fixed number of sentences is three to five sentences (Goldstein, Kantrowitz, Mittal, & Carbonell, 1999), if the quota of a child-node exceeds five sentences (i.e., the default threshold value in our system), the system will process its child-nodes

DOI: 10.1002/asi

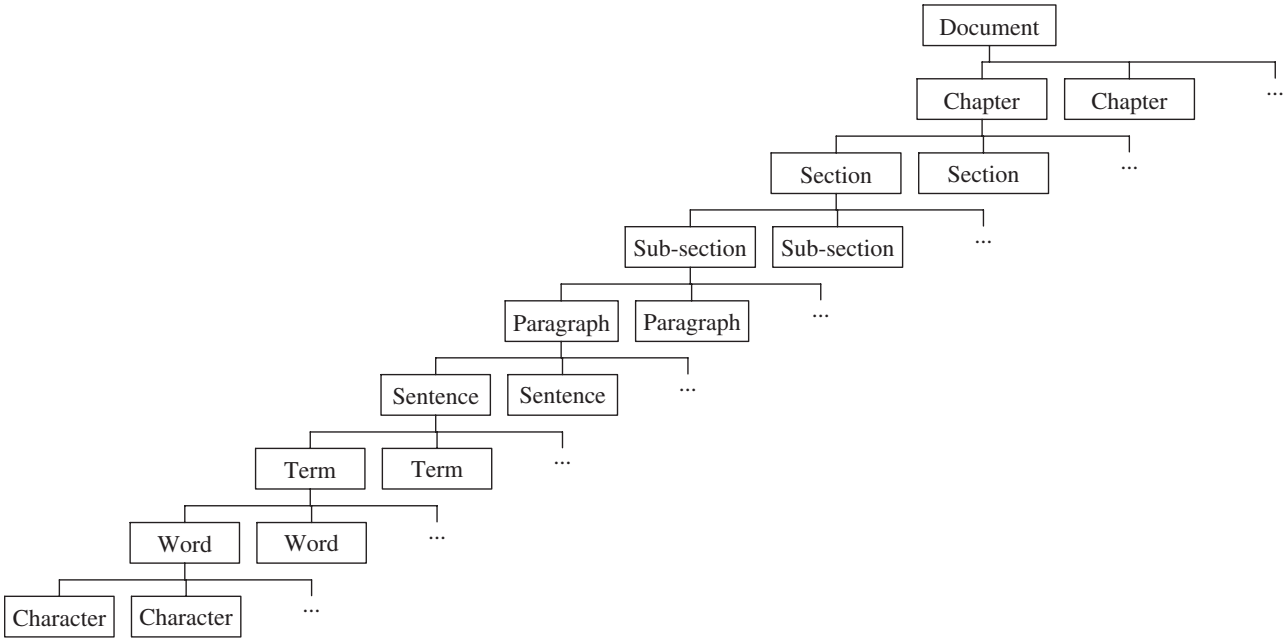


FIG. 2. Hierarchical structure of a large document.

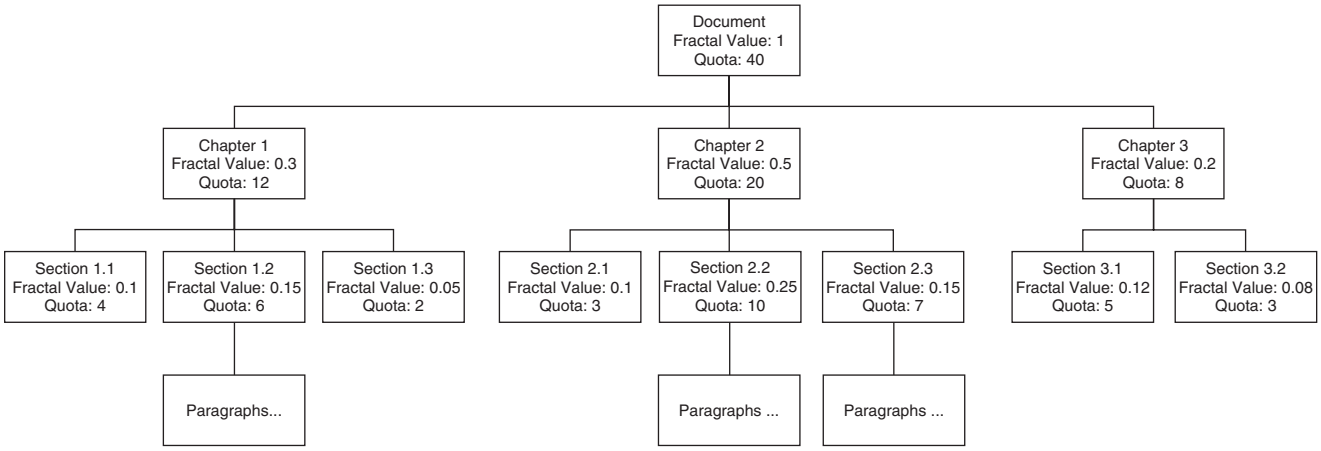


FIG. 3. An example of the fractal summarization model.

iteratively until the quota is less than the threshold. For example, Sections 1.1 and 1.3 are allocated with four and two sentences, respectively; the system will transform the nodes into the corresponding number of topic sentences by extraction of sentences using a statistical method such that the distance of topic sentences and the sentences under the range-block is minimized. The distance between the summary and the document can be calculated by Manhattan distance (Black, 2006) between the thematic, title, location, and cue weight of the topic sentences against the source document. In the example, since Section 1.2 is allocated with six sentences, which is larger than the threshold value, the system continues to process its paragraphs. The details of the fractal summarization algorithm are presented next:

Algorithm 1. Fractal Summarization Algorithm

1. Choose a compression ratio to specify the ratio of sentences extracted.
2. Choose the threshold value of the maximum number of sentences extracted from each node.
3. Calculate the total sentence quota of the summary.
4. Partition the document into range-blocks according to the document structure.
5. Transform the document into a fractal tree.
7. Set the current node to the root node of the fractal tree and initialize its fractal value to 1.
8. *Repeat*.

9.1. For each child-node under the current node,
Calculate the fractal value of child-node.

DOI: 10.1002/asi

$$Fv(child\_of\_x) = Fv(x)C\left(\frac{RBSS(child\_of\_x)}{\sum_{y \in children\_of\_x} RBSS(y)}\right)^{\frac{1}{D}}$$

where $Fv(x)$ is the fractal value of range-block x
$RBSS(y)$ is the range-block significance score of range-block y
y is any child of x
C is a constant, $0 < C \leq 1$
D is the fractal dimension, $0 < D \leq 1$

9.2. Allocate Quota to child-nodes in proportion to fractal values.
9.3. *For* each child-node
If the quota is less than the threshold value
Select the sentences in the range-block by extraction of sentences
Else
Set the current node to the child-node
Repeat Steps 9.1, 9.2, and 9.3
10. *Until* all the child-nodes under the current node are processed.

The fractal value propagation is an important computation in the algorithm. Unlike Koike (1995), the propagation does not depend on the number of child nodes but on the RBSS of child-nodes.

When $D = 1$,

$$Fv(child\_of\_x) = Fv(x)C\left(\frac{1}{\frac{RBSS\ (child\_of\_x)}{\sum_{y \in children\_of\_x} RBSS(y)}}\right)^{-\frac{1}{D}}$$

$$= Fv(x)C\left(\frac{RBSS(child\_of\_x)}{\sum_{y \in children\_of\_x} RBSS(y)}\right)^{\frac{1}{D}}$$

The fractal dimension D plays an important role in a physical tree. The logical tree is a conceptual extension of the physical tree (Koike, 1995). The fractal dimension is the measurement of complexity of a fractal object (Mandelbrot, 1983). The fractal dimension of a logical tree was developed by Koike (1995), but the range of fractal dimension was not specified. The fractal dimension is set as 1 in the applications of fractal tree, such as fractal view (Koike, 1995). If the fractal dimension is greater than 1, the sum of the fractal values of child-nodes is greater than the fractal value of their parent-node. However, it is not desired in the fractal value propagation. To avoid such a condition, the range of D is set as [0, 1]. In our fractal summarization model, $D = 1$. The sum of the fractal values of child-nodes is equal to the fractal value of the parent-node.

The *RBSS* of range-block r, $RBSS(r)$, is computed as the sum of the fractal sentence scores based on the thematic, location, heading, and cue features for all the sentences within range-block r. We conducted an experiment to investigate the weighting of extraction features (Wang, 2003). Results showed that the summarization with equal weighting of all extraction features performed the best. Therefore, the $RBSS(r)$ is calculated as the unweighted sum of individual extraction features in our research. The fractal sentence scores are different from the sentence scores introduced in Section 2 since fractal sentence scores are computed based on the hierarchical structure of the document in addition to the salient features. We shall introduce the fractal sentence scores that are utilized in the $RBSS(r)$ in the following subsection.

The fractal summarization algorithm is recursive in structure. A recursive algorithm solves a given problem by calling themselves recursively one or more times to deal with closely related subproblems (Cormen, Leiserson, & Rivest, 1989). The fractal summarization algorithm selects topic sentences by a divide-and-conquer technique. The system calculates the fractal value of each child-node and allocates the quota to the child-node based on the fractal value. For each child-node, if the quota allocated is larger than a threshold value, the system will repeat the same process to the child-node. Basically, Steps 9.1, 9.2, and 9.3 are recursive. To extract topic sentences from a range-block, the system breaks the range-block into several subrange-blocks that are similar to the original one, but smaller in size. The system extracts sentences from each subrange-block recursively, then combines these extracted sentences to create a solution to the original problem. Therefore, the algorithm is recursive.

The fractal summarization algorithm also is deterministic. A deterministic system is a system in which no randomness is involved, and it produces the same output for a given starting condition (Ito, 1987). The system does not use any random variables, and it is not timing-sensitive. Given a fixed set of significance score functions, training corpus, and a document, the system will always return the same set of extracted sentences if the user inputs a fixed value of a compression ratio. Because the range-block significance scores *RBSS(r)* and the fractal value *Fv(r)* for each range-block in the document are purely functional, the quota allocation and sentence-extraction process are deterministic. Finally, the sentences are extracted based on their significance scores. In summary, the algorithm is deterministic.

The time complexity of the proposed algorithm is linear. An algorithm is called linear if the time it requires is directly proportional to the size of the input (Weiss, 1997). For a document with n sentences with k levels in the document tree (i.e., tree depth), the system needs to iterate for at most k levels. For each level, the system needs to calculate sentence scores for n sentences, to calculate the fractal value of each range-block, and to allocate the quota accordingly; however, the number of range-blocks is far less than the number of sentences. In conclusion, the time complexity is $O(kn)$.

Thematic Feature in Fractal Summarization

Among the thematic features proposed previously, the *tfidf* score of keyword is the most widely used approach. However, it does not take into account the document

DOI: 10.1002/asi

structure in the current summarization techniques; therefore, modification of the *tfidf* formulation is derived to capture the document structure and reflect the significance of a term within a range-block.

Many researchers assume that the weight of a term remains the same over the entire document. According to the formulation of the *tfidf* score, it will remain constant within a document. However, Hearst (1993) claimed that a term should carry different weights in different locations of a full-length document. For example, if a term appears only once in chapter A while it appears frequently in chapter B, the term is obviously more important in chapter B than it is in chapter A. This idea can be extended to other document levels. At the document level, a specific term inside a document should carry the same weight; however, at the chapter-level, a specific term inside a chapter should carry the same weight, but the specific term in two different chapters may carry different weights, and so on. As a result, the *tfidf* score should be modified to different document levels instead of the whole document. In the fractal summarization model, the *tfidf* score should be proportional to the term frequency within a range-block, but inversely proportional to the frequency of range-blocks containing the term. That is,

$$w_{ir} = tf_{ir} \times log_2\left(\frac{N' \times |t_i|}{n'}\right)$$

where tf_{ir} is the frequency of term t_i in range-block r, N' is the number of range-blocks in the document, n' is the number of range-blocks in the corpus in which term t_i occurs, and $|t_i|$ is the length of the term t_i.

Adopting the new formulation of the *tfidf* score, the score may vary a lot in different locations of a document. Furthermore, the *tfidf* score of a term at a specific location may vary when the document is viewed at different abstraction levels. Taking the term, "Hong Kong," in the first chapter, first section, first subsection, first paragraph, and first sentence of the Hong Kong Annual Report 2000 as an example (see Table 1), the *tfidf* scores at different document levels have significant differences: The maximum value is 3,528 at the document level, and the minimum value is 6 at the sentence level.

The fractal sentence score based on thematic features for sentence $k(s_k)$ in range-block r is computed as follows:

$$FSS_T(k, r) = \sum_{t_i \in s_k} w_{ir}$$

$$RBSS_T(r) = \sum_{k \in r} FSS_T(k, r)$$

Location Feature in Fractal Summarization

Current summarization systems assume that the location weight of a sentence is static, where the location weight of a sentence is fixed. However, the fractal summarization model adopts a dynamic approach: The sentence score based on location feature depends on which document level we are considering.

TABLE 1. The *tfidf* score of the term "Hong Kong" at different document levels.

	Term frequency	Text block frequency	No. of text block	*tfidf* Score
Document level	1,113	1	1	3,528
Chapter level	70	23	23	222
Section level	69	247	358	256
Subsection level	16	405	794	66
Paragraph level	2	787	2,580	10
Sentence level	1	1,113	8,053	6

The location of the sentence within a document reflects the significance of the sentence. For example, the sentences at the beginning and the end of a document are usually more important than the others. If we consider the first and second sentences on the same paragraph at the paragraph level, the first sentence has more impact on the paragraph than does the second one; however, the difference of the importance of two consecutive sentences is insignificant at the document level. Therefore, the importance of the sentence due to its location should depend on the document level at which the sentence is located.

In the fractal summarization model, we calculate the location weight for a range-block instead of an individual sentence; all the sentences within a range-block will receive the same location weight. The location weight of a range-block is $1/p$, where p is the shortest distance of the range-block to the first or last range-block under same parent range-block. The fractal sentence score based on location of any sentences in range-block r is computed as follows:

$$FSS_L(k, r) = \frac{1}{\prod_{\substack{y \in path\ from\ k\ to \\ z\ (exluding\ r)}} \begin{bmatrix} \min\ (d(y, first\ child \\ of\ y's\ parent), \\ d(y, last\ child \\ of\ y's\ parant)) \end{bmatrix}}$$

$$RBSS_L(r) = \frac{1}{\min(d(r, first), d(r, last))}$$

where *first* is the first range-block in the same level of r; *last* is the last range-block in the same level or *r;* $d(r, first)$ and $d(r, last)$ are the distance between range-block r and the first range-block and the distance between range-block r and the last range-block, respectively.

Considering the previous example of a generic fractal summarization model (Figure 3), the quotas assigned to each range-block are changed as presented in Figure 4, if only the location feature is considered.

Heading Feature in Fractal Summarization

In fractal summarization, the sentence score based on the heading feature of a sentence is dynamic and depends on which document level we are considering in the document.

DOI: 10.1002/asi

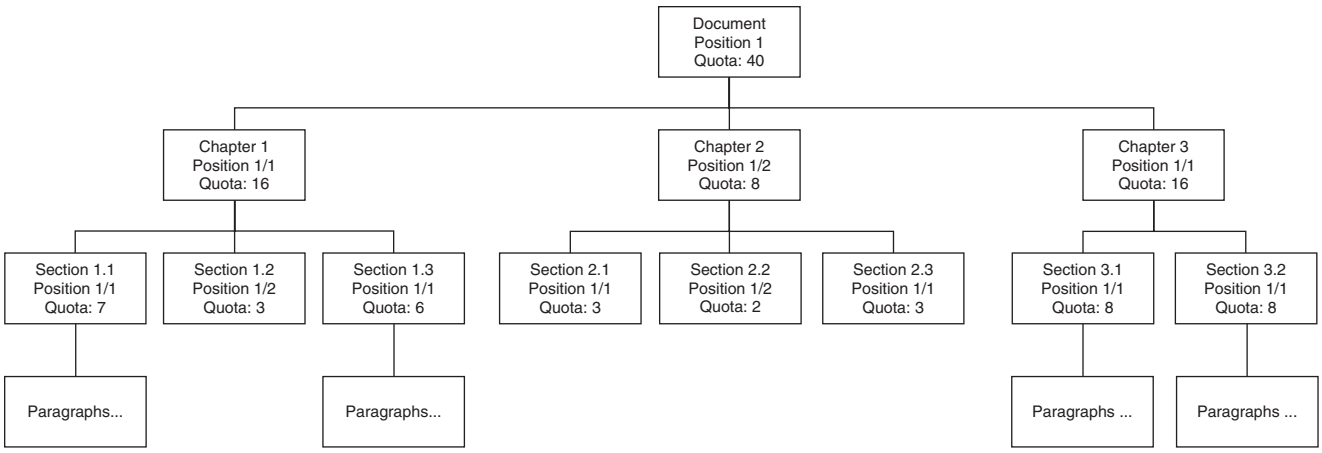


FIG. 4. Fractal summarization with location feature only. (a) Heading feature at the document level. (b) Heading feature at the section level. (c) Heading feature at the subsection level.

At different abstraction levels, some headings should be hidden while some should be emphasized.

The fractal sentence score based on the heading feature for sentence $k(s_k)$ in range-block r, $FSS_h(k, r)$ is:

$$FSS_H(k, r) = \sum_{y \in path\ from\ root\ to\ r} \frac{\sum_{t_i \in s_k \cap heading(y)} w_{iy}}{\prod_{q \in path\ from\ y\ to\ r} m_q}$$

where w_{iy} is the *tfidf* score of term i in range-block y
m_q is the number of children of range-block q

$$RBSS_H(r) = \sum_{k \in r} FSS_H(k, r)$$

Taking the first sentence from the first chapter, first section, first subsection, and first paragraph in a large document as an example, if we consider at the document level, only the document heading should be considered. However, if we consider at the chapter level, then we should consider the document heading as well as the chapter heading. Since the main topic of this chapter is represented by the chapter heading, the terms appearing in the chapter heading should have a greater impact on the sentence. Most internal nodes above the paragraph level in the document tree usually associate with a heading, and there are two types of headings: structural and informative. Structural headings indicate the structure of the document only, but not any information about the content of the document. For example, "Introduction," "Overview," and "Conclusion" are structural headings. The informative headings give us an abstract of the content of the branch; they help us to understand the content of the document and are used for calculation of heading weights. On the other hand, structural headings can be easily isolated by string matching with a dictionary of those structural headings, and they will be used for cue features later in this article. The terms in the informative headings are very important in extracting the sentences for summarization. Given a sentence in a paragraph, the headings of its corresponding subsection, section, chapter, and document should be considered. The significance of a term in the heading also is affected by the distance between the sentence and the heading in terms of depth in the hierarchical structure of the document. Propagation of fractal value (Koike, 1995) is a promising approach to calculate the heading weight for a sentence.

The first sentence of this section is illustrates the propagation of the heading weight. As shown in Figure 5, the sentence "In fractal summarization, the sentence score based on the heading feature of a sentence is dynamic and depends on which document level we are considering in the document" is located in the subsection entitled "Heading Feature in Fractal Summarization"; the heading of this section is "Fractal Summarization", and the heading of the document is "Hierarchical Summarization of Large Documents". To compute the heading weight of the sentence, we shall propagate the weight of the terms that appear in both the sentence and the heading based on the distance between the heading and the sentence and the degrees of the heading node. On the other hand, the visibility of a heading sentence is determined by the document level at which we are looking.

For example, if we are considering the significance of the sentence at the document level, only the heading of the document is visible. Therefore, the heading weight of the sentence is calculated as:

$$w_{heading} = w_{heading\ in\ document}$$

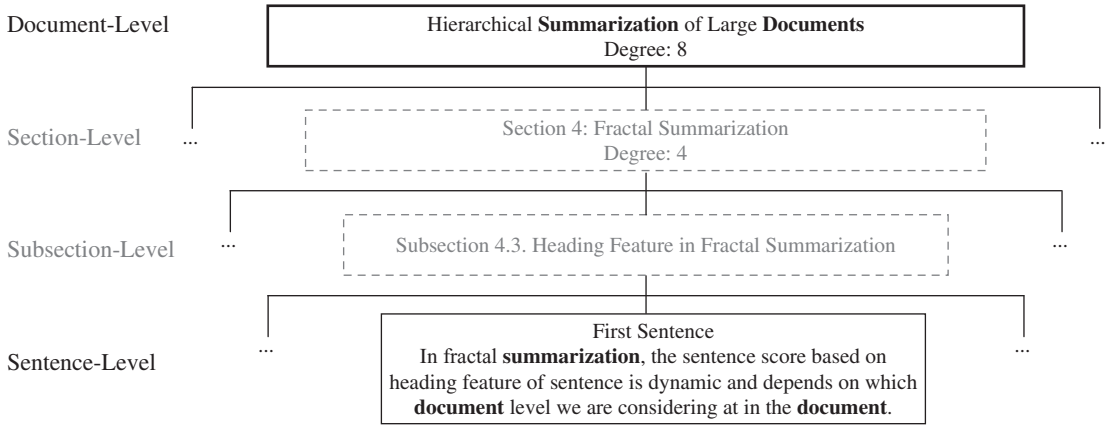
where $w_{heading\ in\ document} = (w_{\text{"}summarization\text{"}} + w_{\text{"}document\text{"}})in\ heading_{document}$

However, if we are considering the sentence at the section level, both the heading of the document and the heading of the section are visible. Therefore, the heading weight of the sentence is calculated as:

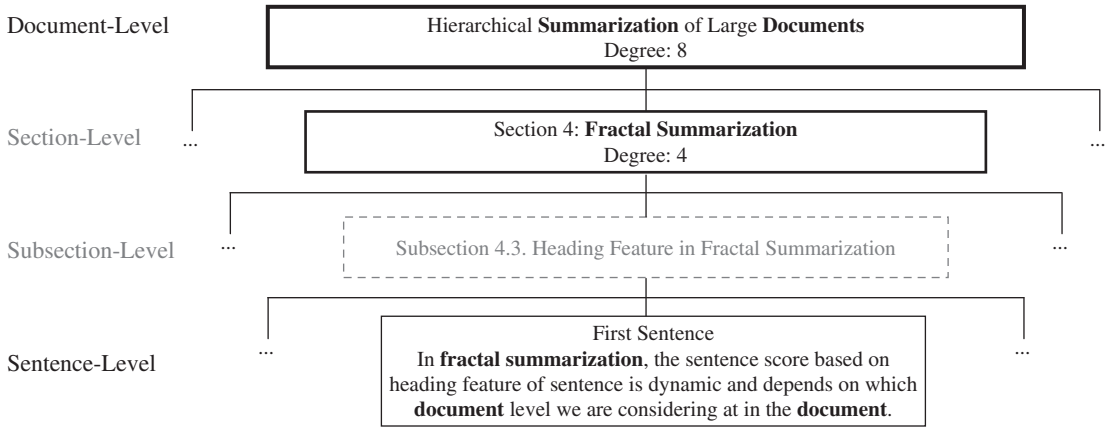
$$w_{heading} = w_{heading\ in\ document} + w_{heading\ in\ section}$$

where $w_{heading\ in\ document} = (w_{\text{"}summarization\text{"}} + w_{\text{"}document\text{"}})$ in $heading_{document}/8$

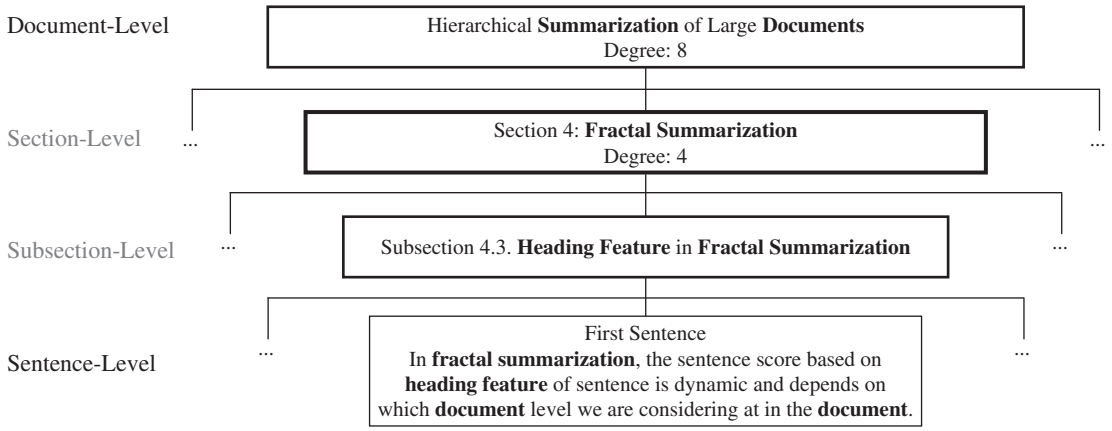
DOI: 10.1002/asi



(a) Heading Feature at Document-Level



(b) Heading Feature at Section-Level



(c) Heading Feature at Subsection-Level

FIG. 5. Example of heading feature in fractal summarization.

$$w_{heading\ in\ section} = (w_{\text{“}fractal\ summarization\text{”}}) \text{in } heading_{section}$$

Finally, if we are considering the sentence at the subsection level, all the headings—including the headings of the document, the section, and the subsection, are visible. Therefore, the heading weight of the sentence is calculated as:

$$w_{heading} = w_{heading\ in\ document} + w_{heading\ in\ section} + w_{heading\ in\ subsection}$$

where

$$w_{heading\ in\ document} = (w_{\text{“}summarization\text{”}} + w_{\text{“}document\text{”}})\ in\ heading_{document}/(8 \times 4)$$

DOI: 10.1002/asi

$w_{heading\ in\ section} = (w_{\text{“}fractal\ summarization\text{”}})\ in\ heading_{section}/4$

$w_{heading\ in\ subsection} = (w_{\text{“}heading\ feature\text{”}} + w_{\text{“}fractal\ summarization\text{”}})\ in\ heading_{subsection}.$

Cue Feature in Fractal Summarization

When human abstractors extract the sentences from a document, they will follow the document structure to search the topic sentences. During the extraction of information, the abstractors will pay more attention to the range-block with a heading that contains some bonus words such as “Conclusion” since they consider it as a more important part in the document and extract more information for those important parts. The cue feature of a heading sentence is usually classified as a rhetorical feature (Teufel & Moens, 1998).

As a result, we propose to consider the cue feature not only at the sentence level but also at other document levels. Given a document tree, we will examine the headings of each range-block by the method of cue feature and adjust their quotas of the entire range-block accordingly. This procedure can be repeated to the subrange-blocks until it reaches the sentence level.

$$RBSS_c(r) = \sum_{k \in r} FSS_C(k, r)$$

The $RBSS(r)$ is then computed as the sum of the normalized values of $RBSS_T(r)$, $RBSS_L(r)$, $RBSS_H(r)$, and $RBSS_C(r)$. The individual feature score of a range-block is divided by the maximal feature score of all the sibling nodes of the range-block; hence, the feature scores are normalized such that the maximum score of each feature is 1.

Unfortunately, there is no cue dictionary which has been commonly accepted as a standard. Various cue dictionaries with different cue scores have been developed in the literature (Edmundson, 1969; Kupiec et al., 1995). To eliminate the impact of a cue dictionary, the cue feature is currently disabled in our experiment.

Visualization of Fractal Summarization

In current automatic summarization systems, the sentences extracted are concatenated together linearly as a summary. As most users read the document according to the document structure, a linear structure summary violates the reading pattern of human users. It also is difficult for users to explore information on a linear summary. The summary extracted by fractal summarization is represented in a hierarchical tree structure, and a summary in a hierarchical view allows users to navigate information according to the tree structure. In addition, the fractal summarization can be further enhanced by fractal view to adjust the tree structure of summary based on the user-selected topics. Such interaction supports users to construct a summary that is tailored to the interests of the users.

Figure 6 shows a screen capture of the fractal summarization system. The sentences extracted are organized according to the document structure. It provides a clear picture on

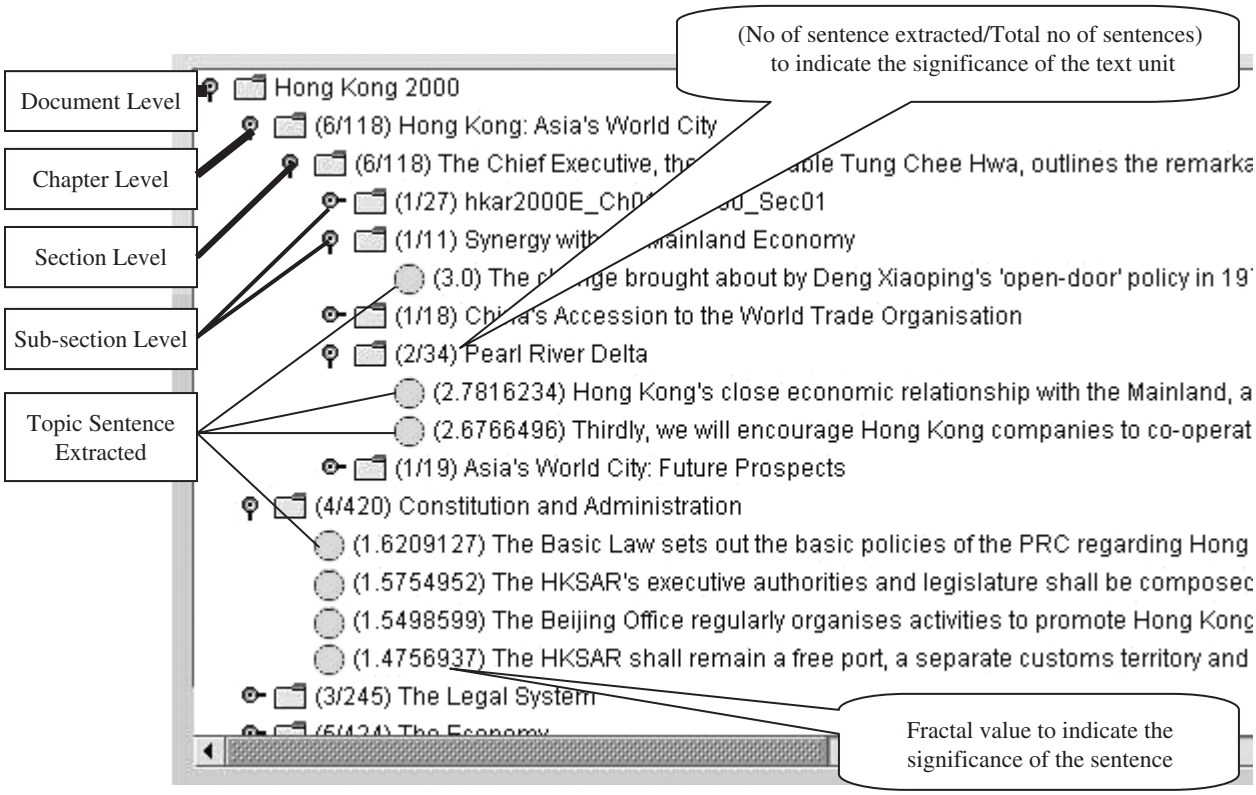


FIG. 6. Fractal summarization system.

DOI: 10.1002/asi

the global structures of documents and the allocation of information in each chapter. Users can easily explore the summary according to the document structure and retrieve the source document when necessary. On the other hand, if the sentences are concatenated together linearly, users may have to read the sentences from the beginning to the end of the summary to understand it. Given that a summary is organized according to the document structure as illustrated in Figure 6, users can read the heading of the range-block and decide whether to extend the range-block to a lower level to obtain further details. If the user considers the summary of a certain range-block less useful, he or she may collapse the range-block into a single node. It helps users to filter the irrelevant information in the summary. A tree structure summary can easily provide a customized view of the summary for users.

On the other hand, the fractal summarization system helps users to identify the important text units by providing a high degree of system transparency. As shown in Figure 6, the system shows the number of sentences extracted from each range-block and the total number of sentences contained by the range-block. By reading these number, the user can approximately know the significance of each range-block and determine which range-block to further expand to get the detailed information. Moreover, the system also provides fractal value alongside individual sentences extracted. The user can know the relative significance of the sentences extracted and can quickly identify the most important text units.

Experiments

The fractal summarization model is a novel summarization model which performs summarization based on the hierarchical structure of a document in the same manner as does a human abstractor. Although the current summarization models employ the same similar salient features as fractal summarization, they consider a document as a sequence of sentences without considering the hierarchical structure of documents. In this study, two experiments were conducted to compare the performance of fractal summarization models against the nonhierarchical summarization model. The first experiment utilized the Hong Kong Annual Report, a large document with 23 chapters and over 8,000 sentences. In the second experiment, the TIPSTER Text Summarization Evaluation was employed.

Previous studies of automatic summarization (Kupiec et al., 1995; Teufel & Moens, 1997) have identified an upper limit for the precision of the summarization systems; the performance improves with additions of extraction features until it reaches the upper boundary after additions of three or four salient features. After reaching the upper boundary, further additions of salient features will not improve precision. On the contrary, this factor may sometimes even decrease the aggregated precision of the system (Kupiec et al., 1995). Therefore, both summarization models (i.e., fractal summarization and nonhierarchical summarization) in our experiment adopt the three most widely used summarization features: thematic, location, and heading.

The nonhierarchical summarization model and the fractal summarization model each adopt the techniques described earlier. It has been proven that the weighting of different summarization features does not have any substantial effect on the average precision in some information retrieval applications (Lam-Adesina & Jones, 2001). In the present study, some experiments have been conducted to investigate the impact of weightings with summarization features on overall precision of automatic summarization. The results indicated that the summarization system with equal weighting of summarization features performs the best. In the current experiment, the maximum value of each summarization feature has been normalized to 1, and the total weight of sentences calculated as the sum of scores of all summarization features without weightings. In nonhierarchical summarization, the sentences are extracted in a liner space while the fractal summarization model extracts the sentences according to the hierarchical structure of the document.

Hong Kong Annual Report

As the fractal summarization model is designed to summarize a large document with an explicit hierarchical document structure, we selected the Hong Kong Annual Report for the first experiment. The Hong Kong Annual Report is an official document which is available to the public. It has a rich hierarchical structure to organize different topics into chapters, sections, and subsections in the report. We applied the fractal summarization and the nonhierarchical summarization on this report.

Figure 7 shows the number of sentences extracted from each chapter by two techniques. As shown in the figure, the nonhierarchical summarization model extracted sentences mainly from chapters 4, 5, and 6; however, it did not extract any sentence from some other chapters. In contrast, the fractal summarization model extracted the sentences distributed more evenly from each chapter. A full-length text document contains a set of subtopics (Hearst, 1993). A good informative summary should cover as many subtopics as possible because an informative summary aims at summarizing what the source text says as much as possible. This concept was supported by Nomoto and Matsumoto (2001), who stated that a good summary should find diverse topic areas in the text and reduce redundant information content. In a nonhierarchical summary model, where a document is treated as a sequence of sentences, most sentences will tend to be extracted from the most important text unit, leading to a redundancy of information and a lack of the topic diversity that exists in the source. The fractal summarization model takes into account the hierarchical structure of the document and extracts the sentences distributed evenly from the source. Hence, fractal summarization produces a summary with wider coverage of information subtopics than does the nonhierarchical summarization model, and it can be considered as a better summary in terms of information coverage.

DOI: 10.1002/asi

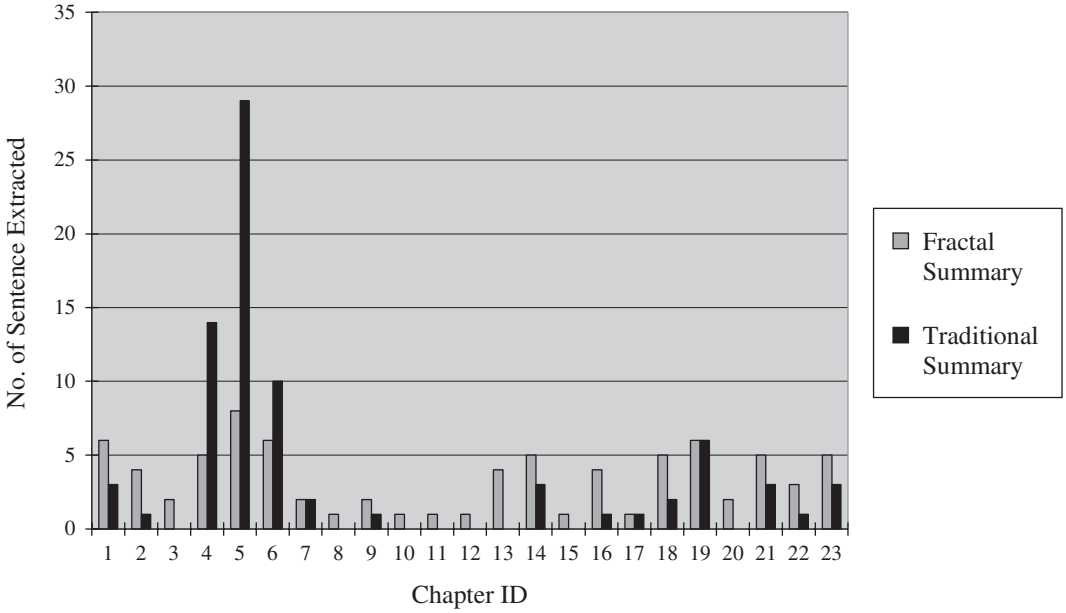


FIG. 7. Number of sentences extracted from each chapter of the Hong Kong Annual Report 2000.

A user evaluation of 10 participants was conducted to evaluate the performance of the summarization techniques. All participants were college graduates residing in Hong Kong. They had a general knowledge about the contents of the Hong Kong Annual Report and were qualified to judge the quality of the summaries. To eliminate any bias of individual effect of hierarchical navigation, participants were divided into two groups. The first group evaluated the quality of summaries in an online environment (Figure 6) where all summaries (extracted by fractal summarization and nonhierarchical summarization) were organized in a tree structure. The second group evaluated the quality of summaries in an offline environment where all summaries (extracted by fractal summarization and nonhierarchical summarization) were concatenated as a linear summary. The summaries were delivered without informing the users how they were extracted. During the user evaluation, both the fractal summary and the nonhierarchical summary of the Hong Kong Annual Report were presented to each participant in random order, and he or she was asked to accept or reject each sentence in the summaries.

We measured the performance by precision, an intrinsic measurement of summaries (Jing et al., 1998). The performance of a summarization system is usually measured by precision only (Kupiec et al., 1995; Teufel & Moens, 1998), as the measurement of recall is limited by the compression ratio of a summarization system. The precision of a summary is computed as the ratio of sentences accepted by a user to the total number of sentences in the summary. That is,

$$precision = \frac{no.\ of\ sentences\ accepted\ by\ the\ use\ as\ part\ of\ the\ summary}{no\ of\ sentences\ in\ the\ summary}$$

The average precision for each participant is shown in Table 2. As shown in the table, there is no significant difference between the precision of the two groups. The hierarchical navigation can help a user to quickly explore the summary and search for information; however, it does not add the value of the content of the summaries. Therefore, the precision of summaries is not affected. The results show that all participants considered the summaries extracted by the fractal summarization as better than those extracted by the nonhierarchical summarization. Fractal summarization can achieve up to a 91.3% precision, and achieves 85.1% on average, while nonhierarchical summarization can achieve up to a maximum of 77.5% precision, and achieves 67.0% on average. A one-tailed t test of paired data analysis shows

TABLE 2. Precision of summaries of the Hong Kong Annual Report (compression ratio = 1%).

User ID	Summary format	Fractal summarization model	Nonhierarchical summarization model
1	Tree-structured summary	81.25%	71.25%
2	Tree-structured summary	85.00%	67.50%
3	Tree-structured summary	80.00%	56.25%
4	Tree-structured summary	85.00%	63.75%
5	Tree-structured summary	88.75%	77.50%
6	Linear-structured summary	81.25%	61.25%
7	Linear-structured summary	91.25%	76.25%
8	Linear-structured summary	86.25%	58.75%
9	Linear-structured summary	85.00%	65.00%
10	Linear-structured summary	87.50%	72.50%
Mean (All)		*85.13%*	*67.00%*
Mean (Tree-Structured Summary)		84.00%	67.25%
Mean (Linear-Structured Summary)		86.25%	66.75%

DOI: 10.1002/asi

that the precision of the fractal summarization model significantly outperforms nonhierarchical summarization at a 99% confidence level.

The preliminary experiment shows that the concept of document hierarchy is useful to automatic summarization. The first experiment suggests that the fractal summarization has outperformed the nonhierarchical summarization, which has not considered the hierarchical structure of documents for summarization of large documents such as the Hong Kong Annual Report. The result of this experiment is the first gauge of the effectiveness of the model. Subsequent larger scale experiments are presented to demonstrate the effectiveness of the proposed model. The standard categorization task of TIPSTER Text Summarization Evaluation (SUMMAC) was conducted in the second experiment to evaluate the proposed model.

SUMMAC

SUMMAC is the first large-scale, developer-independent evaluation of automatic text summarization systems (Mani et al., 1999). The documents for the TIPSTER evaluation are documents drawn from Text Retrieval (TREC; Voorhees & Harman, 1997) CDs 4 and 5. Extrinsic evaluation (Morris, Kasper, & Adams, 1992) tasks based on activities typically carried out by information analysts in the U.S. government are defined for experiments. In our experiment, the *categorization* task was selected. This categorization task focuses on generic summaries. The evaluation aimed to discover whether a generic summary could effectively present enough information to allow an analyst to quickly and correctly categorize a document.

In SUMMAC, summaries are extracted at a relatively low compression ratio. The compression ratio for the categorization task was 10%. In the categorization task, the summaries of the documents were presented to the participants with a topic pair, and they were asked to determine if the document was relevant to the topic. The contingency table for the categorization task is shown in Table 3. The system was evaluated in three measurements.

$$\text{Precision} = TP/(TP + FP)$$
$$\text{Recall} = TP/(TP + FN)$$

$$\text{F}-\text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Performances for summarization systems developed by the participants are shown in Table 3 (Mani et al., 1999).

TABLE 3. Classification task contingency table.

	Participant's judgment	
Ground truth	Relevant	Irrelevant
Relevant	TP (True positive)	FN (False negative)
Irrelevant	FP (False positive)	TN (True positive)

TABLE 4. Categorization accuracy for summaries with fixed length.

Systems	Precision	Recall	F-Score
CIR	0.68	0.35	0.43
IBM	0.63	0.37	0.44
NMSU	0.69	0.34	0.43
Surrey	0.69	0.31	0.39
Pen	0.66	0.29	0.38
ISI	0.71	0.35	0.44
IA	0.67	0.33	0.41
BT	0.70	0.33	0.41
NTU	0.68	0.33	0.43
SRA	0.73	0.37	0.45
LN	0.68	0.37	0.45
Cornell/SabIR	0.52	0.36	0.42
GE	0.69	0.33	0.42
CGI/CMU	0.69	0.33	0.42
Mean	*0.67*	*0.34*	*0.42*

We followed the standard TIPSTER setting to conduct the classification task using fractal summarization. For each topic, 100 documents are selected from the TIPSTER corpus. Ten percent of the sentences are extracted from each document by fractal summarization. Given a set of sentences extracted, the participants were asked to classify the documents as relevant or irrelevant to the topic based on the sentences extracted. Fourteen sites have participated in the categorization task; their performances are summarized in Table 4.

There was a slight difference in the setting of the TIPSTER experiment and that of our experiment. The TIPSTER categorization task involved 24 participants while only 15 participants were involved in our experiment. It has been shown that the unanimous agreement between participants is relatively weak (Carletta et al., 1997; Mani et al., 1999). On the other hand, it has been shown that the system performance rankings remain relatively stable even lack of agreement in the relevance judgment (Voorhees, 1998). As a result, the system accuracy of the TIPSTER evaluation can be used to benchmark our system even though the participants are different. In our study, 50 documents randomly selected of the 100-documents corpus were assigned to each participant for the classification task. The results are shown in Table 5.

As shown in Tables 4 and 5, the fractal summarization system has a similar precision to the other summarization systems; however, it has a higher recall. In other words, there are less FN (False Negative) documents (i.e., participants classified relevant documents as irrelevant documents). The fractal summarization model extracts sentences distributively over the document; therefore, it has a better representation of the overall information in the document. The fractal summarization extracts the important information from the document; therefore, the probability for the participant to classify a relevant document as irrelevant is lower than that for the other summarization systems. It is believed that the F-score can better measure the performance of an information

DOI: 10.1002/asi

TABLE 5. Categorization accuracy for fractal summaries with fixed length (by extraction of sentences).

Participant	Precision	Recall	F-Score
1	0.88	0.70	0.78
2	0.67	0.44	0.53
3	0.83	0.63	0.71
4	0.50	0.33	0.40
5	0.67	0.57	0.62
6	0.63	0.71	0.67
7	0.57	0.50	0.53
8	0.71	0.63	0.67
9	0.71	0.56	0.63
10	0.50	0.67	0.57
11	0.67	0.75	0.71
12	0.67	0.57	0.62
13	0.71	0.63	0.67
14	0.67	0.50	0.57
15	0.83	0.63	0.71
Mean	*0.68*	*0.59*	*0.63*

system (Rijsbergen, 1979). The fractal summarization achieved an F-score of 0.63 while other summarization systems achieved a mean F-score of 0.42. Therefore, the fractal summarization system outperforms the other summarization systems.

Traditionally, automatic summarization is extraction of sentences from source documents. Research also has been conducted to generate summary by extraction of text units at other document levels (Mitra, Singhal, & Buckley, 1997; Salton, Allan, Buckley, & Singhal, 1994; Salton, Singhal, Buckley, & Mitra, 1996; Salton, Singhal, Mitra, & Buckley, 1999; Strzalkowski, Wang, & Wise, 1998). The fractal summarization model was originally designed for document summarization by extraction of sentences. Theoretically, fractal summarization can be used to extract text units at any document level. We therefore conducted an experiment to investigate the performance of the summarization system by extraction of clauses as summaries. A sentence may be composed of two or more subsentences that use punctuation as delimiters. These subsentences are clauses which are the children of sentences in the hierarchical structure in our fractal summarization model.

The fractal summarization by extraction of clauses is similar to fractal summarization by extraction of sentences. Only a simple extension is required to the proposed system. The system calculates the individual feature score for each clause by using the approaches described earlier. The overall scores of clauses are then calculated. The system calculates the RBSS as the sum of clause scores instead of sentence scores. The system then allocates a quota of clauses to be extracted into range-blocks iteratively by propagation. When the quota allocated is less than a threshold value, clauses with the highest scores will be extracted and concatenated as summary.

If the summary is generated by the extraction of paragraphs, then the compression ratio is defined as the number of paragraphs in the summary to the number of paragraphs in the source document (Mitra et al., 1997). There are other definitions of the compression ratio because other summarization systems may use other document units during extraction. Similarly, the compression ratio is defined as the ratio of clauses extracted by the summarization system.

The classification task and fractal summarization were employed. The 15 participants involved in the previous experiment were asked to participate. The summarization system extracts 10% of the clauses from the source documents by fractal summarization. The remaining 50 documents are assigned to each participant for classification task. The results are shown in Table 6.

As shown in Table 6, the precision, recall, and F-score are all slightly improved when the summarization system extracts clauses instead of sentences. It is commonly believed that the F-score is a better measurement of the performance of an information retrieval system. One-tailed *t* test of paired data analysis shows that the F-score of the fractal summarization by extraction of clauses significantly outperforms fractal summarization by extraction of sentences at a 98% confidence level. The results show that summarization by extraction of clauses performs better than summarization by extraction of sentences when fractal summarization is employed.

The fractal summarization model is the first summarization model based on hierarchical structure of a document. Advanced summarization techniques take the document structure into consideration to compute the probability of a sentence to be included in the summary. Rhetorical structure of texts has been applied to automatic summarization (Marcu, 1997; Ono, Sumita, & Miike, 1993; Miike, Itoh, Ono, & Sumita, 1994). Because there are some fundamental differences between the two techniques, it is impossible to compare the results of the two techniques. The differences are as follows:

- The rhetorical structure-based summarization techniques assume that the relationship between text units form a binary tree structure (Marcu, 1997); however, a large document may have a more complicated tree structure. For example, there are many chapters in the Hong Kong Annual Report,

TABLE 6. Categorization accuracy for fractal summaries (by extraction of clauses).

Participant	Precision	Recall	F-Score
1	0.80	0.67	0.73
2	1.00	0.57	0.73
3	0.75	0.86	0.80
4	0.57	0.44	0.50
5	0.71	0.63	0.67
6	0.80	0.50	0.62
7	0.71	0.63	0.67
8	0.86	0.75	0.80
9	0.63	0.63	0.63
10	0.67	0.57	0.62
11	0.71	0.56	0.63
12	0.75	0.75	0.75
13	0.83	0.56	0.67
14	0.67	0.50	0.57
15	0.86	0.67	0.75
Mean	*0.75*	*0.62*	*0.67*

DOI: 10.1002/asi

and each of them addresses issues in different areas. The chapters are equally important, and they may not fit properly into a rhetorical structure tree. The fractal summarization model is capable of summarizing a document with any number of child-nodes.

- The linguistic rules to determine the logical relation between text units vary across languages. Therefore, the rhetorical structure-based summarization for an English document cannot compare with summarization for other languages (Marcu, 1997). The document hierarchy is language independent. The fractal summarization model has been implemented for English and Chinese (Wang & Yang, 2003).
- The rhetorical structure-based summarization requires a comprehensive rhetorical structure analysis and intensive human interactions (Marcu, 1997; Ono et al., 1993). The fractal summarization model is fully automated; therefore, it is more desired.
- The time complexity for the fractal summarization model is linear. It is more suitable for large documents. Although the time complexity is not stated in rhetorical structure-based summarization (Marcu, 1997; Marcu, 1999; Ono et al., 1993), there may be a large number of rhetorical structure trees, which then must be selected by a constraint-satisfaction procedure (Marcu, 1999). The rhetorical structure-based summarization has been implemented on only small-size documents. Ono et al. (1993) tested documents with about 90 to 175 sentences. Marcu (1999) tested documents with 161 to 725 words. The document tested in our study contains 8,000 sentences (about 200,000 words). These two techniques are not comparable in their document sizes.

Conclusion

In this article, a novel summarization model based on fractal theory has been presented. The fractal summarization model was developed based on the statistical data and the structure of documents. Thematic features, location features, heading features, and cue features are adopted. To illustrate the feasibility of the model, experiments of fractal summarization were conducted. The fractal theory was widely used in digital image processing and coding. Applying the fractal theory in document summarization is the first effort in automatic text summarization. Results of this research have shown that such an approach is promising and provides a new direction in automatic summarization.

The fractal summarization extracts the summary of a document by a recursive algorithm based on the hierarchical document structure. In the summarization process, the salient features are adopted. User evaluation was conducted. The fractal summarization model achieved 85.05% precision on average and up to 91.25% precision while the nonhierarchical summarization achieved 67.00% precision on average and up to a maximum of 77.50% precision. Additionally, the fractal summarization has wider information coverage. This can be shown by the result that the nonhierarchical summarization mainly extracts topic sentences from a few chapters while the fractal summarization model extracts sentences evenly from all chapters. An experiment based on the SUMMAC corpus also showed that fractal summarization achieves higher precision and recall than do other summarization systems as reported in the literature. In addition, the fractal summarization system is robust and interactive. Users can easily control the compression ratio and select the range-block on which to focus. Such interaction allows users to explore specific topics in the documents.

Recent research in automatic summarization has focused on information organization. Documents were grouped into document sets before summarization (Nobata, Sekine, Uchimoto, & Isahara, 2003; Radev, Jing, Stys, & Tam, 2004). It has been shown that the hierarchical summarization of multiple documents organized in a hierarchical structure significantly outperforms other multidocument summarization systems without using the hierarchical structure (Wang, Yang, & Shi, 2006). A set of documents can be organized into a hierarchical structure by different classifications. Future research will focus on developing an algorithm to automatically organize a large set of documents into a hierarchical structure which can be utilized for multidocument hierarchical summarization.

References

Aminin, M., & Gallinari, P. (2002). The use of unlabeled data to improve supervised learning for text summarization, In Proceedings of the 25th annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'02) (pp. 105–112). New York: ACM Press.

Baxendale, P.B. (1958). Machine-made index for technical literature—An experiment. IBM Journal of Research and Development, 2(4), 354–361.

Black, P.E., (2006). Manhattan distance, in Dictionary of Algorithms and Data Structures, Paul E. Black, ed., Gaithersburg, MS: U.S. National Institute of Standards and Technology. 31 May 2006. (Last accessed 17 January 2008) Available from: http://www.nist.gov/dads/HTML/manhattanDistance.html

Carletta, J., Isard, A., Isard, S., Jowtko, J.C., Doherty-Sneddon, G., & Anderson, A.H. (1997). The reliability of a dialogue structure coding scheme. Computational Linguistics, 23(1), 13–32.

Cormen, T.H., Leiserson, C.E., & Rivest, R.L. (1989). Introduction to algorithms. Cambridge, MA: MIT Press.

Edmundson, H.P. (1969). New methods in automatic extraction. Journal of the ACM, 16(2), 264–285.

Endres-Niggemeyer, B. (2002). SimSum: An empirically founded simulation of summarizing. Information Processing and Management, 36(4), 659–682.

Endres-Niggemeyer, B., Maier, E., & Sigel, A. (1995). How to implement a naturalistic model of abstracting: Four core working steps of an expert abstractor. Information Processing and Management, 31(5), 631–674.

Feder, J. (1988). Fractals. New York: Plenum Press.

Glaser, B.G., & Strauss, A.L. (1967). The discovery of grounded theory: Strategies for qualitative research. New York: deGruyter.

Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (1999). Summarizing text documents: Sentence selection and evaluation metrics. In Proceedings of the 22nd annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'99) (pp. 121–128). New York: ACM Press.

Hearst, M.A. (1993). Subtopic structuring for full-length document access. In Proceedings of the 16th annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'93) (pp. 56–68). New York: ACM Press.

Ito, K. (1987). Turing machines. Encyclopedic dictionary of mathematics (2nd ed., Vol. 1, pp. 136–137). Cambridge, MA: MIT Press.

Jacquin., A.E. (1993). Fractal image coding: A review. Proceedings of the IEEE, 81(10), 1451–1465.

Jing, H., Barzilay, R., McKeown, K., & Elhadad, M. (1998). Summarization evaluation methods: Experiments and analysis. In Proceedings of the

DOI: 10.1002/asi

AAAI Spring Symposium on Intelligent Text Summarization (pp. 60–68). Menlo Park, CA: AAAI.
Koike, H. (1995). Fractal views: A fractal-based method for controlling information display. ACM Transaction on Information Systems, 13(3), 305–323.
Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. In Proceedings of the 18th annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95) (pp. 68–73). New York: ACM Press.
Lam-Adesina, M., & Jones, G.J.F. (2001). Applying summarization techniques for term selection in relevance feedback. In Proceeding of the 24th annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01) (pp. 1–9). New York: ACM Press.
Lin, C.Y., & Hovy, E.H. (1997). Identifying topics by position. In Proceedings of the Applied Natural Language Processing Conference (ANLP-97) (pp. 283–290). San Francisco: Kaufmann.
Luhn, H.P. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2), 159–165.
Mandelbrot, B. (1983). The fractal geometry of nature. New York: Freeman.
Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., & Sundheim, B. (1999). The TIPSTER SUMMAC Text Summarization Evaluation. In Proceedings of the 9th Conference of the European Chapter of the Association for Computation Linguistics (EACL-1999) (pp. 77–85). Morristown, NJ: ACL.
Marcu, D., (1997). From discourse structures to text summaries. In Proceedings of the Workshop on Intelligent Scalable Text Summarization (82–88). Morristown, NJ: ACL.
Marcu, D. (1999). Discourse trees are good indicators of importance in text. In I. Mani & M. Maybury (Eds.), Advances in automatic text summarization (pp. 123–136). Cambridge, MA: MIT Press.
Miike, S., Itoh, E., Ono, K., & Sumita, K., (1994). A full-text retrieval system with a dynamic abstract generation function. In Proceedings of the 17th annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94) (pp. 152–161), New York: Springer-Verlag.
Mitra, M., Singhal, A., & Buckley, C. (1997) Automatic text summarization by paragraph extraction. In Proceedings of ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization (pp. 31–36). Morristown, NJ: ACL.
Morris, G., Kasper, G.M., & Adams, D.A. (1992). The effect and limitation of automated text condensing on reading comprehension performance. Information System Research, 3(1), 17–35.
Nobata C., Sekine S., Uchimoto K., Isahara H. (2003). A summarization system with categorization of document sets. In Proceedings of the Third Proceedings of the Third NTCIR Workshop on research in information Retrieval, Automatic Text Summarization and Question Answering (pp. 33–38). Tokyo: National Institute of Informatics.
Nomoto, T., & Matsumoto, Y. (2001). A new approach to unsupervised text summarization. In Proceedings of the 24th annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'2001) (pp. 26–34). New York: ACM Press.
Ono, K., Sumita, K., & Miike S. (1994). Abstract generation based on rhetorical structure extraction. In Proceedings of the 15th Conference on Computational linguistics (COLING' 94) (pp 344–348). Morristown, NJ: ACL.
Radev, D.R., Jing, H., Stys, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. Information Processing and Management, 40, 919–938.
Radev, D.R., & McKeown, K.R. (1998). Generating natural language summaries from multiple online sources. Computational Linguistics, 24(3), 469–500.
Rijsbergen, C.J. van. (1979). Information retrieval. London: Butterworths.
Salton, G., Allan, J., Buckley, C., & Singhal, A. (1994). Automatic analysis, theme generation, and summarization of machine-readable texts. Science, 264, 1422–1426.
Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5), 513–523.
Salton, G., Singhal, A., Buckley, C., & Mitra, M. (1996). Automatic text decomposition using text segments and text themes. In Proceedings of the 7th ACM Conference on Hypertext (Hypertext'96) (pp. 53–65). New York: ACM Press.
Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1999). Automatic text structuring and summarization. In I. Mani & M. Maybury (Eds.), Advances in automatic text summarization (pp. 241–255). Cambridge, MA: MIT Press.
Strzalkowski, T., Wang, J., & Wise, B. (1998). A robust practical text summarization. In Proceedings of the AAAI Spring Symposium on Intelligent Text Summarization (pp. 26–33). Menlo Park, CA: AAAI.
Teufel, S., & Moens, M. (1997). Sentence extraction as a classification task. In Proceedings of the Workshop of Intelligent and Scalable Text summarization (pp. 56–68). Morristown, NJ: ACL.
Teufel, S., & Moens, M. (1998). Sentence extraction and rhetorical classification for flexible abstracts. In Proceedings of the AAAI Spring Symposium on Intelligent Text Summarization (pp. 89–97). Menlo Park, CA: AAAI.
Voorhees, E.M., (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In Proceedings of the 21st annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98) (pp. 315–323), New York: ACM Press.
Voorhees, E.M., & Harman, D. (1997). Overview of the 5th Text REtrieval conference (TREC-5). In D. Harman (Ed.), TREC-5, Proceedings of the Fourth Text Retrieval Conference (pp. 1–28). Washington, DC: Government Printing Office.
Wang, F.L., (2003). Fractal Summarization, August 2003, Ph.D. Thesis, Hong Kong: Chinese University of Hong Kong.
Wang, F.L., & Yang, C.C., (2003). Automatic summarization of Chinese and English parallel documents, International Conference of Asian Digital Libraries (ICADL 2003), Kuala Lumpur, Malaysia, Lecture Notes in Computer Science, 2911, 46–61.
Wang, F.L., Yang, C.C., & Shi, X. (2006). Multi-document summarization for terrorism information extraction. IEEE Intelligence and Security Informatics Conference (ISI 2006), San Diego, CA. Lecture Notes in Computer Science, 3975, 602–608.
Weiss, M.A.. (1997). Data structures and algorithm analysis in C. Reading, MA: Addison-Wesley.
Yang, C.C., Chen, H., & Hong, K. (2003). Visualization of large category map for Internet browsing. Decision Support Systems, 35(1), 89–102.

DOI: 10.1002/asi