# A Model for Information Retrieval Agent System Based on Keywords Distribution

Jae-Woo LEE
*Dept. of Computer Science, Kyungbok College,*
*131, Sinpyeong-ri, Pocheon-si, 487-717, Gyeonggi-do, Korea*
*It21c@korea.ac.kr*

## Abstract

*Information retrieval is one of the most important technologies at present. We can always get many information in the Internet or distributed computing systems using various information retrieval models. For searching proper information that we need, it is necessary to construct efficient information retrieval agent systems helping many web clients' requests. In this paper, we propose a simple new model for information retrieval agents based on many terms or keywords distribution in a document or distributed database. For the key paragraph extraction we use meaningful term's frequency and the key word distribution characteristics in a document, and those terms are selected by using stemming, filtering stop-lists, synonym for search meaningful terms in a document. The agent receives a web client's information retrieval request and extracts key paragraph with frequency and distribution using the keywords of the client, and then the agent constructs profile of the documents with the keywords, key paragraph, address of the document browsing. And then we can search many documents or knowledge easily using the profile for information retrieval and browse the document.*

## 1. Introduction

The Internet or distributed computing systems enable us to search various knowledge or information at anywhere. They are one of the most important properties in our current information society. In the Internet or distributed computing systems, information retrieval is one of the most important technologies for searching proper information that we need. We can always get many information in the Internet or distributed computing systems using various information retrieval models. For searching proper information that we need, it is necessary to construct efficient information retrieval agent systems helping many web clients' requests. Those information retrieval agent systems can help many clients and web servers for efficient information management. But there are various information in our result of searching the information in the Internet or distributed database systems and it is not easy to search proper information that we need. Because thus information searching only can be enable by using information retrieval systems that are constructed well by applying information technology such as indexing.

Information retrieval systems are based, either directly or indirectly, on models of the retrieval process. These retrieval models specify how representations of text documents and information needs should be compared in order to estimate the likelihood that a document will be judged relevant. The estimates of the relevance of documents to a given query are the basis for the document rankings that are now a familiar part of information retrieval systems. Examples of simple models include the probabilistic or Bayes classifier model and the vector space model. Many others have been proposed and are being used [1].

For efficient information retrieval and knowledge discovery in the Internet or distributed computing systems, various algorithms are required for analyzing documents. Generally those are classified three kinds of algorithms, retrieval algorithms, filtering algorithm and indexing algorithms. Especially indexing algorithms are constructing data structure for searching information exactly. Automatic indexing is one of the most important data for efficient information retrieval. And we should consider documents or data structure besides such algorithms, we should select proper algorithm according to such data structure [2, 3].

As we above mentioned before, information retrieval is one of the most important technologies at present. We can always get many information in the Internet or distributed computing systems using

2007 International Conference on Multimedia and Ubiquitous Engineering(MUE'07)
0-7695-2777-9/07 $20.00 © 2007 IEEE

various information retrieval models. For searching proper information that we need, it is necessary to construct efficient information retrieval agent systems helping many web clients' requests. In this paper, we propose a simple new model for information retrieval agents based on many terms or keywords distribution in a document or distributed database. For the key paragraph extraction we use meaningful term's frequency and the key word distribution characteristics in a document, and those terms are selected by using stemming, filtering stop-lists, synonym for search meaningful terms in a document. The agent receives a web client's information retrieval request and extracts key paragraph with frequency and distribution using the keywords of the client, and then the agent constructs profile of the documents with the keywords, key paragraph, address of the document browsing. And then we can search many documents or knowledge easily using the profile for information retrieval and browse the document.

In our proposed model for information retrieval agents, we used an idea that an important keyword in documents is distributed partly in each paragraph. So we use frequency and distribution characteristics about keyword in documents. First, we extract important words in a electronic document or web page by using stemming, filtering stop-lists, synonym, etc. Those extracted important terms will be candidate index about a document. Next, we examine distributed characteristics of those important words, candidate index. And then we can extract important paragraph by using a keyword location and frequency in a document. An important keyword can be presented partly in documents. We can represent a document briefly as extracted paragraph and keyword and browse the document using location of keyword in document.

Contents of our paper is as following. In section 2, we introduce briefly about information retrieval and define our proposed model for information retrieval agents based on keywords distribution. In section 3, we explain proposed agent's procedure about key paragraph extraction and automated indexing including profile of a document. We also explain summary of document and document browsing. In section 4, we use an example for more ease understanding our agent systems. Finally, in conclusion we summarize our agent system and plan future works.

## 2. Information Retrieval and Proposed Model for Agent System

### 2.1. Information Retrieval

For efficient information retrieval, it is important that keywords are defined very well as appropriate terms about all documents in the Internet or distributed computing systems. It maybe proper to constructing these indexing manually, but that requires enormous time and labor. Perhaps, only the creators of the documents can those indexing the documents. Because of these reasons, there are many automatic indexing algorithms about documents in many papers and researches. We should consider various agents on constructing information retrieval systems, and also define conceptual model, file structure, query operation, term operation and hardware. Information retrieval systems are based on those various environment factors, such as hardware, algorithms, etc [1,2,3].

Relating to information retrieval systems, conceptual model is about matching query language to any documents like Boolean operation, probabilistic, string search and vector space, etc. File structure is research area about document database structure, flat file, inverted file, signature files, pat tree, graphs and hashing, etc. Query operation is about specifying user's need for information retrieval, feedback, parse, Boolean and cluster, etc. Terms operation is about logical processing of terms, stemming, weight, thesaurus, stop list and truncation, etc. Document operation is about handling document, parse, display cluster, rank, sort, field mask and assign IDs, etc. Hardware is about processing environment, von Neumann, parallel, magnetic disk, etc[4,5,6,7].

And there are three kinds of algorithms, retrieval algorithms, filtering algorithm and indexing algorithms. Retrieval algorithms are extracting important or meaningful information from database, knowledgebase or documents. Those are classified to sequential scanning and searching indexed text files. Filtering algorithms are simplifying result of text or information for efficient information retrieval. Using stop word or special condition, eliminate unsuitable result of retrieval. Indexing algorithms are constructing data structure for searching information exactly. Indexing is one of the most important data for efficient information retrieval. Those algorithms use inverted files, signature files and trees, etc [2, 3, 8, 9].

For automated indexing there are various algorithms, stemming, stop list, thesaurus, etc. Stemming is the process of matching morphological term variants for using general terms. Stemming is for reducing the size if index files and improving information retrieval system's performance. Stop lists is defined that eliminate worthless terms in documents, many of the most frequently occurring words in documents, such as 'the', 'of', 'is', etc. in English. Thesaurus is machine readable document database typically contains a list of terms including single word or a phrase. And those are designed for subject areas and domain dependent [1, 4, 5, 6, 7, 8, 9].

### 2.2. Our Proposed Agent Model for Information Retrieval

Our proposed model for information retrieval agent systems is based on many terms or keywords distribution in a document or distributed database. The model is composed of two agents systems, the one is about analyzing documents of server systems and the other is constructing profiles for documents browsing.

As the analyzing documents agent, for key paragraph extraction we use meaningful term's frequency and the key word distribution characteristics in a document, and those terms are selected by using stemming, filtering stop-lists, synonym for search meaningful terms in a document. The agent receives a web client's information retrieval request and extracts key paragraph with frequency and distribution using the keywords of the client. And constructing profiles agent constructs profile of the documents with the keywords, key paragraph, address of the document browsing. And then many clients can search many documents or knowledge easily using the profile for information retrieval and browse the document. Thus our agent model is shown in Figure 1.

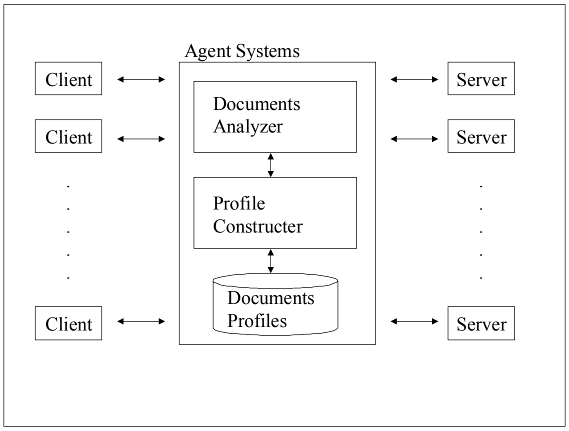

**Figure 1.** A model for information retrieval agent systems

## 3. Agent Systems Model Based on Keywords Distribution

### 3.1. Agent-1 : Documents Analyzing for Extraction Keyword

A document is composed of many terms and important words are spread out documents. In those documents, we want to get a meaningful or worth terms and the word indexed. It is important for efficient information retrieval or knowledge discovery that indexing is defined very well by appropriate terms about all documents. But it is not easy to extract proper terms about a document, it maybe is proper to constructing these indexing manually even though enormous time and labor is required for such indexing and only a creator of the documents can do those tasks. In this paper, we propose a new method for key paragraph extraction, which compute term frequency and key word distribution of each term selected by using stemming, filtering stop-lists, synonym for search meaningful terms in a document. Using the extracted paragraph with frequency and distribution, we construct profile with the index, key paragraph of the document. And then we can search many documents or knowledge easily using the profile for information retrieval and browsing document. After we extract worth terms using stemming, filtering stop-lists, synonym, etc. we define location of the terms in document. The criterion of term's location can be defined various type, document line or sentence, etc. Our proposed indexing algorithm is as following.

$t_i$ : *ith meaningful term in a document*
$f_i$ : *frequency of ith term in a document*
$l_{ij}$ : *if ith term appears in jth location of the document, 1, otherwise 0*
$d_i$ : *if ith term's frequency is greater than a criterion, 1, otherwise 0*

1. Extract worth terms in documents using stemming, filtering stop-lists, synonym, etc. This processing is scanning a document for searching meaningful terms($t_i$). And examine the term's location of document, where location can be defined line or sentence.

2. When extract meaningful term, compute terms frequency($f_i$), location of document($l_{ij}$) respectively.

3. Create table about each terms and frequency. And compute summation of term frequency in each location of document.

After extract meaningful terms, some terms can be eliminated by criterion of frequency($D$). For example, when a term is appear less than three times, the term is treated as worthless term. When a term $t_i$ is defined worth term or not, we use denotation $d_i = \{0,1\}$ , meaningful *1*, worthless *0*. When we need to extract more keyword, we can change criterion of frequency($D$). As shown in table 1, there are extracted terms and location, summation of frequency.

**Table 1.** Term's frequency and distribution in document

| Extracted keywords | Frequency | Relative location in document($l_{ij}$) | Applying ($f_i > D$ ) |
|---|---|---|---|
| $t_1$ | $f_1$ | $l_1$ $l_4$ $l_5$ $l_6$ $l_9$ | $d_1 = 1$ *or* $0$ |
| $t_2$ | $f_2$ | $l_2$ $l_8$ | $d_2 = 1$ *or* $0$ |
| $t_3$ | $f_3$ | $l_3$ | $d_3 = 1$ *or* $0$ |
| $t_4$ | $f_4$ | $l_3$ | $d_4 = 1$ *or* $0$ |
| $t_5$ | $f_5$ | $l_5$ $l_7$ $l_8$ | $d_5 = 1$ *or* $0$ |
| $t_6$ | $f_6$ | $l_5$ | $d_5 = 1$ *or* $0$ |
| … | … | … | … |
| $t_i$ | $f_i$ | $l_{ij}$ … | $d_i = 1$ *or* $0$ |
| | | … | |
| $t_n$ | $f_n$ | $l_m$ | $d_n = 1$ *or* $0$ |
| | $\sum$ | $s_1$ $s_2$ $s_3$ $s_4$ …… $s_j$ $s_m$ | |

4. The region where summation of frequency is greater than a criterion is extracted key paragraph and terms in the key paragraph can be defined as keywords.

Summation of frequency($s_j$) is computed as

$$S_j = \sum_i (d_i \times l_{ij})$$

where $d_i = 0$ or $1$ and $l_{ij} = 0$ *or* $1$.

We can select region as key paragraph where summation of frequency is greater than a criterion on paragraph. And terms in the selected paragraph are defined keywords.

### 3.2. Agent-2 : Document Profiles for Summary of the Document and Browsing

In section 3.1, we extract key paragraph and keywords, index. A document can be represented by key paragraph and keywords in briefly. And if we make profile by using the keyword and paragraph, we always can get information about the document easily. Especially we can define address($a_j$) in a document with keyword position for document browsing.

Using the extracted paragraph with frequency and distribution, we construct profile with the index, key paragraph of the document. And then we can search many documents or knowledge easily using the profile for information retrieval and browsing document.

**Table 2.** A document profile with address in document

| Extracted keywords | Frequency | Relative location in document($l_j$) | Address |
|---|---|---|---|
| $\boldsymbol{t_1}$ | $f_1$ | $l_1$ $l_4$ $l_5$ $l_6$ $l_9$ | $a_1$ |
| | | | |
| | | | |
| | | | |
| $\boldsymbol{t_5}$ | $f_5$ | $l_5$ $l_7$ $l_8$ | $a_5$ |
| | | | |
| … | … | … | … |
| $t_i$ | $f_i$ | $l_{ij}$ … | $a_j$ |
| | | … | |
| $t_n$ | $f_n$ | $l_m$ | $a_m$ |
| ***selected paragraph*** | | $\boldsymbol{s_4}$ $\boldsymbol{s_5}$ $\boldsymbol{s_6}$ $\boldsymbol{s_7}$ $\boldsymbol{s_8}$ … $s_j$ $s_m$ | |

As shown in table 2 the terms in selected paragraph is linked at address of documents. Using this table we can search many information or knowledge easily for information retrieval and browsing document.

## 4. An Application and Document Browsing

### 4.1. Processing of Extracting Keyword and Key Paragraph

We explain example of a document for extracting keywords and paragraphs. We show a document about "information retrieval" as shown in table 3. We assume a location of terms as 1 sentence.

1. Extract worth terms in documents using stemming, filtering stop-lists, synonym, etc. This processing is scanning a document for searching meaningful terms($t_i$), 'text', 'knowledge', 'information', etc. And examine the term's location of document, where location can be defined line or sentence.

2. When extract meaningful term, compute terms frequency($f_i$), location of document($l_i$) respectively, term 'text' is 7 and term 'information' is 3, etc.

3. Create table about each terms and frequency. And compute summation of term frequency in each location of document as shown in table 4.2.

After extract meaningful terms, some terms can be eliminated by criterion of frequency($D$) where we assume *D=2*.

4. The region where summation of frequency is most is extracted key paragraph, sentence 7, 12 and

2007 International Conference on Multimedia and Ubiquitous Engineering(MUE'07)
0-7695-2777-9/07 $20.00 © 2007 IEEE

terms in the key paragraph are defined keyword, worth terms ‘text’, ‘information’, and ‘document’.

**Table 3.** A document sample about information retrieval

| Location | Title of Document : Information Retrieval |
|---|---|
| 1 | Written as well as spoken text is a very important means of communicating human thoughts and knowledge. |
| 2 | In our current information society, we are overwhelmed with electronic textual documents. |
| 3 | Document collections are constantly growing and their content is constantly evolving. |
| 4 | Information retrieval and selection systems are becoming of increasing importance. |
| 5 | They must help us to find documents or information relevant to our needs. |
| 6 | Written text is considered as an intricate cognitive phenomenon. |
| 7 | The cognitive process of creating and understanding natural language text is complex and not yet completely understood. |
| 8 | However, it is clear that besides coding and decoding linguistic signs, it involves additional cognitive processes. |
| 9 | Communication through natural language text is basically ostensive and inferential. |
| 10 | The creator ostensively signals his or her communicative goals. |
| 11 | The inferential character of understanding natural language is one of the factors that makes an automated understanding of text a difficult operation. |
| 12 | The inferences refer to knowledge that is shared by the text’ s creator and user and that is not made explicit in the text. |
| 13 | The inferences also refer to the individual cognitive state of the user and allow determining the meaning of a text to the individual user. |

## 4.2. Document Browsing

After extract key word and key paragraph we constructed the profile of the document using created table as shown in table 4. In table 4 the term ‘text’, ‘language’ and ‘inference’ is located in sentence 1,7,12, so we can browse the document using profile like table 4. For more refined browsing the document, we can construct systems with feedback.

As shown in figure 2 we can represent the logical structure of the document, so we can browse the document in briefly. And if we need to get more information in detail, we use address on keyword, ‘text’.

**Table 4.** A document profile of the sample document

| Terms | $f_i$ | Relative location in document($l_{ij}$) | | | | | | | | | | | | | $a_i$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | *7* | ***8*** | ***9*** | 10 | 11 | 12 | 13 | |
| text | **7** | **1** | | | | | **1** | **1** | | **1** | | **1** | **1** | **1** | $\mathbf{a_1}$ |
| means | 1 | 1 | | | | | | | | | | | | | |
| human | 1 | 1 | | | | | | | | | | | | | |
| thought | 1 | 1 | | | | | | | | | | | | | |
| knowledge | **2** | **1** | | | | | | | | | | | **1** | | |
| information | **3** | | **1** | | **1** | **1** | | | | | | | | | |
| society | 1 | | 1 | | | | | | | | | | | | |
| document | **3** | | **1** | **1** | | **1** | | | | | | | | | |
| collection | 1 | | | 1 | | | | | | | | | | | |
| content | 1 | | | 1 | | | | | | | | | | | |
| retrieval | 1 | | | | 1 | | | | | | | | | | |
| selection | 1 | | | | 1 | | | | | | | | | | |
| system | 1 | | | | 1 | | | | | | | | | | |
| importance | 1 | | | | 1 | | | | | | | | | | |
| need | 1 | | | | | 1 | | | | | | | | | |
| phenomenon | 1 | | | | | | 1 | | | | | | | | |
| process | **2** | | | | | | | **1** | **1** | | | | | | |
| language | **3** | | | | | | | **1** | | **1** | | **1** | | | $\mathbf{a_7}$ |
| coding | 1 | | | | | | | | 1 | | | | | | |
| decoding | 1 | | | | | | | | 1 | | | | | | |
| sign | 1 | | | | | | | | 1 | | | | | | |
| communication | 1 | | | | | | | | | 1 | | | | | |
| creator | 1 | | | | | | | | | | 1 | | | | |
| goal | 1 | | | | | | | | | | 1 | | | | |
| character | 1 | | | | | | | | | | | 1 | | | |
| factor | 1 | | | | | | | | | | | 1 | | | |
| operation | 1 | | | | | | | | | | | 1 | | | |
| inference | **2** | | | | | | | | | | | | **1** | **1** | $\mathbf{a_{12}}$ |
| state | 1 | | | | | | | | | | | | | 1 | |
| user | 1 | | | | | | | | | | | | | 1 | |
| meaning | 1 | | | | | | | | | | | | | 1 | |
| ***selected*** $\sum(d > 3)$ | | 2 | 2 | 1 | 1 | 2 | 1 | **3** | 1 | 2 | | 2 | **3** | 2 | |

Title of document : Information Retrieval

Key Paragraph : sentence 7, 12

> The cognitive process of creating and understanding natural language text is complex and not yet completely understood. The inferences refer to knowledge that is shared by the text’ s creator and user and that is not made explicit in the text.

Keyword : Text, Information, Document, Language

Content of Document : Address

Written as well as spoken text is …

The cognitive process of creating …

The inferences refer to knowledge …

**Figure 2.** The Structure of Document for Profile

## 5. Conclusion

Information retrieval is one of the most important technologies at present. In the Internet or distributed computing systems, it is necessary to construct efficient information retrieval agent systems helping many web clients' requests for searching proper information that we need. In this paper, we propose a simple new model for information retrieval agents based on many terms or keywords distribution in a document or distributed database. For the key paragraph extraction we use meaningful term's frequency and the key word distribution characteristics in a document, and those terms are selected by using stemming, filtering stop-lists, synonym for search meaningful terms in a document. The agent receives a web client's information retrieval request and extracts key paragraph with frequency and distribution using the keywords of the client, and then the agent constructs profile of the documents with the keywords, key paragraph, address of the document browsing. And then we can search many documents or knowledge easily using the profile for information retrieval and browse the document.

In the future, we will further research to represent various documents and classify various types of documents for agent systems. It will be very important that we define structure of document in detail and indexing algorithm for information retrieval systems.

## References

[1] W. Bruce Croft, Advances in Information Retrieval : Recent Research From the Center for Intelligent Information Retrieval, Kluwer Academic Publishers, 2002

[2] William B. Frakes and Ricardo Baeza-Yates, Information Retrieval : Data Structures & Algorithms, Prentice Hall, 1992

[3] Marie-Francine Moens, Automatic Indexing and Abstracting of Document Texts, Kluwer Academic Publishers, 2000

[4] Aho, A., and M. Corasick, "Efficient String Matching : An Aid to Bibliographic Search," Communication of the ACM, Vol. 18, No. 6, 1975, pp. 333-340

[5] Faloutsos, C., and S. Christodoulakis, "Signature Files : An Access Method for Documents and its Analytical Performance Evaluation," ACM Transactions on Office Information Systems, Vol. 2, No. 4, 1984, pp. 267-288

[6] Fox, C., "A Stop List for General Text," SIGIR Forum, Vol 24, No. 1, 1990, pp. 19-35

[7] Harman, D., "How Effective is Suffixing?," Journal of the American Society for Information Science, Vol. 42, No. 1, 1991, pp. 7-15

[8] Bookstein, A. and D. R. Swanson, "Probabilistic Models for Automatic Indexing," Journal of the American Society for Information Science, Vol. 25, No. 5, 1974, pp. 312-318

[9] Salton, G. and C. S. Yang, "On the Specification of Term Values in Automatic Indexing," Journal of Documentation, Vol. 29, No. 4, 1973, pp. 351-372