Automatic Classification
Using Concept Knowledge of Web Documents

Sang-Ho Choi[a], Sa-Joon Park[a], Su-Cheol Hwang[b] and Ki-Tae Kim[a]

[a]Dept.of Computer Science and Engineering
University of Chung-Ang, Seoul 156-756, S. Korea
Tel : 82-2-8205304, Fax 82-2-8205304, E-mail : shchoi;phdjoon;ktkim@ailab.cse.cau.ac.kr

[b]School of Computing Information and Engineering
College of Inha Tech, Incheon 402-752, S. Korea
Tel : 83-32-8702331, Fax 82-32-8702519, E-mail :schwang@inhatc.ac.kr

ABSTRACT

In order to classify web documents, we suggest a method using concept knowledge of category. In our study, the concept relations between keywords are extracted using hyperlink information and after the extracted keywords are classified into each category, these are used as an index. Then TFIDF for each category is extended to determine index weight value. The system is constructed for experimenting and estimating, which is consist of web robot, indexer, concept knowledge database for each category and the document classifier. Our system to be applied the extended TFIDF method shows an accuracy of 88% in automatic classifying of web documents.

Keywords
Concept Knowledge, Document Classification, TFIDF

1.0 INTRODUCTION

With the use of the Internet spreading widely around the world, anyone can use the Internet search engines to look for the information they need. The two main methods of information search are keyword search and directory search. In order to enable directory searches, the web documents must first be classified and stored in separate directories. The more precisely the documents are classified and stored the more accurate are the search results. Today, the quality of the search results, which relies on accuracy, is considered more important rather than the quantity. Up to the present, a classifier manually classified the web documents and organized them into different categories. This process takes up much manpower and time. Due to this, studies are underway for developing an automatic classification method. This study proposes to use category concept knowledge for classification. For this, concept knowledge must first be constructed for each category. Keywords, which represent each category, are used for constructing the concept knowledge. The keywords applied here hold concepts based on the document titles and hyperlinks. Words are extracted from the anchor text materials of the received web document's title and hyperlink and the index words are extracted through morpheme analysis. Among the index words the unnecessary words are deleted and the carefully selected words are used in constructing concept-based keywords for each category. These concept-based keywords are applied in executing the automatic classification algorithm, which determines which category a web document belongs to. The document processed by the algorithm is organized into categories and the document index words are extracted into new concept-based keywords to construct concept-based category information.

2.0 BACKGROUND

2.1 Document Classification Method

Document classification refers to the process of allocating new documents to the appropriate category. The general classification method is the statistical method using words within the document. The most widely used two statistical methods apply the Vector similarity and the Bayesian probability respectively.

The classification method applying the Vector similarity express the document and each of the classification categories in the vector form and calculate the similarity between the document and each category by the angle between the two vectors. The document is classified into the category with the highest similarity(Frakes,1992).

The classification method using Bayesian probability combines Bayesian rule with probability. When words come in the classifying document, the appearing event of each word is assumed to be independent. The probability of this document to belong to a certain category is dependent on the probability which is obtained from combining words included in document with category (Lewis, 1998).

2.2 Automatic Indexing

The index is a list of important items, terms, people's names and place names that appear within the text and indicates the page numbers where one could find them. In other words, it is an extraction of the representative words in a text.

In the past, indexes were made by librarians. During the 1950s the number of documents increased rapidly but the number of librarians to make indexes for the documents did not increase as much as, so they could not cover the whole task. As so, studies were carried out in order to develop automatic indexing technology to solve the problem(Salton ,1987).

Automatic indexing is a process of extracting index words from a document. The process consists of producing index word nominees by analyzing information materials through morpheme analysis, processing disuse words among the nominees and selecting index words by extracting special index words.

3.0 EXTRACTING CONCEPT-BASED KEYWORDS

When some web documents are written for homepages, anchor tag <A> is generally used in order to link to the other documents. As soon as the user reading the document clicks this link, the screen switches to the designated part. The words, phrases, or sentences used here are called anchor texts. The hyperlink, consisting of a link and anchor text, has characteristics of abstract, relevant, hierarchical and universal.

By using the connection of hyperlink between web documents, the correlation between the documents can be analogized(Park, 2000). If the correlations between hyperlinks are restricted to those of documents within a certain field, the concept-based keywords of the domain can be extracted.

3.1 Keyword Extraction Using Hyperlinks

Two web documents are connected by a hyperlink. The hyperlink indicates the relation between the two documents through anchor texts, as shown in figure 1.

The anchor text contents are extracted by the keywords of the two web documents, which are connected according to the features of the hyperlink.

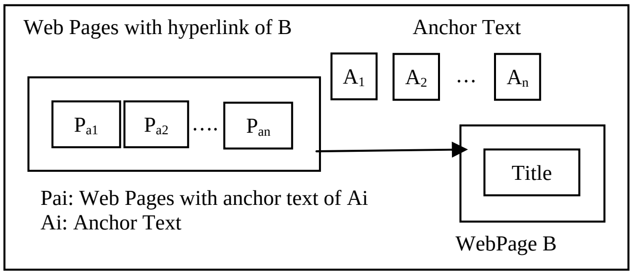
Web Pages with hyperlink of B
Anchor Text
A1
A2
…
An
Pa1
Pa2
….
Pan
Title
Pai: Web Pages with anchor text of Ai
Ai: Anchor Text
WebPage B

Figure 1: Keywords from the hyperlink connecting web document

When web page P is connected to web document B by anchor text A, anchor text A is parceled into words and stemmed to determine the keywords of document B.

After extracting the concept-based keywords using hyperlinks the web document has more than one keyword. The concept relation is produced by using the keyword of each document and links of the hyperlink. The link shows a hierarchical structure or the references for the contents. Using the link the hierarchical relation between documents and the references are converted into an abstract form of the hierarchical relation between keywords and references. The concept relation is then shown through these abstract forms.

3.2 Extracting the Concept-based Keyword

More than one web document have a certain keyword in common and these documents are connected by the hyperlink information. The web documents that are connected to a web document containing a certain keyword are classified into groups that contain the same keywords like figure 2. And the document classified for each keyword is abstracted to relation between keywords using link relations like figure 3.

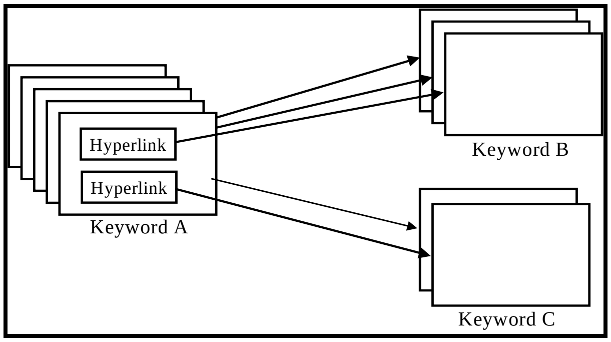

Figure 2: A connection between web documents using hyperlink

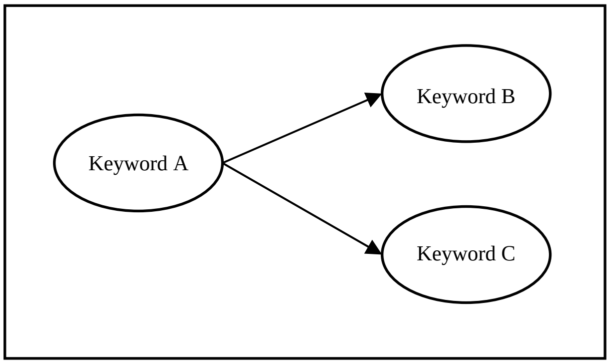

Figure 3: Concept relation consisted of keywords extracted from web

The web documents including keyword A and the documents containing keywords B and C respectively are all connected through the hyperlink. Since the connected documents are relevant to each other they are connected and expressed through the associative relation. As so, if the keywords contained in each document are connected to the hyperlink connection, the abstract concept relation between keywords can be obtained.

In this study, the relation between the keywords obtained through the above process is referred to as the concept relation between keywords.

3.3 Category Indexing

The keywords extracted based on concept need to be indexed by category. Also when classifying documents a method for evaluating weight value is necessary as a standard for measurement. In the present study, we have extended TFIDF(Term Frequency Inverse Document Frequency) which is one of methods estimating index weighting value. TFIDF is a method for calculating the weight value of word w in document d. According to TFIDF, the importance of a word is in proportion to its term frequency in a document and is in inverse proportion to the document frequency of all the documents containing the word(Salton,1987).

Because indexing is performed by categories rather than by documents, the indexing of the documents must be extended to categories. Indexing is processed by using the extended TFIDF shown below.

$$TFIDFc(k,C) = \sqrt{\mathrm{TF(k,C)} * \frac{N}{DF(k)}} \qquad (1)$$

TFIDFc(k,C) : TFIDF for keyword k from category C
TF(k, C) : The frequency of keyword k in category C
DF(k) : The document frequency of documents containing keyword k within category C
N : The total number of documents in category C

The weight value of TFIDF in category C is obtained through the indexing of keywords by figuring out keyword k's frequency and document frequency. More information value is added to words that appear more frequently within the category and play a crucial role while the number of documents the word appears is smaller.

3.4 Document Classification

Document classification is the process of determining the category of each document. In this paper, the documents are classified into a certain category by using the concept knowledge constructed through the web document collector and indexer.

The process of document classification of figure 4 is as follows; before classifying the documents a concept knowledge database is constructed for each formerly constructed category. Each category database regulates the size of the index words since they vary in size. After the normalization is complete the documents are input and classified into categories. The documents are classified by inputting the document into the index word extractor and extracting the document's keywords. Then the extracted index words are applied to the concept knowledge of each category. The total weight value of the keyword in each category is calculated by using TFIDF. The category with the largest weight value is where the document should be classified in.

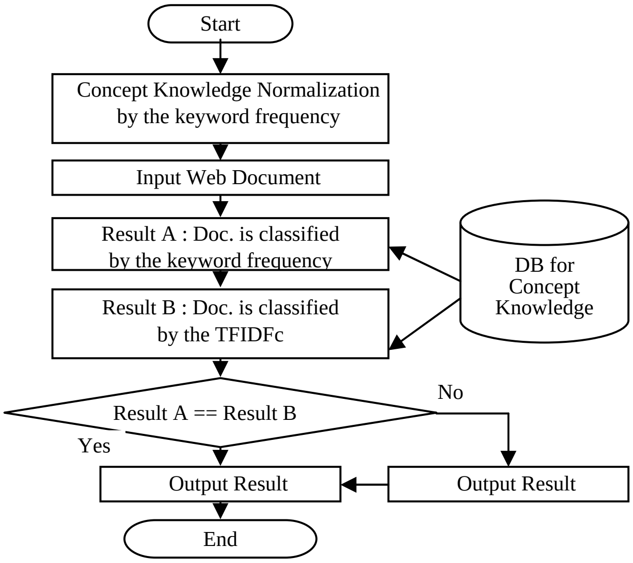
Start
Concept Knowledge Normalization by the keyword frequency
Input Web Document
Result A : Doc. is classified by the keyword frequency
Result B : Doc. is classified by the TFIDFc
DB for Concept Knowledge
Result A == Result B
No
Yes
Output Result
Output Result
End

Figure 4:Concept knowledge DB of Each Category

4.0 THE SYSTEM STRUCTURE

Figure 5 shows the structure of system to use for classifying web documents.

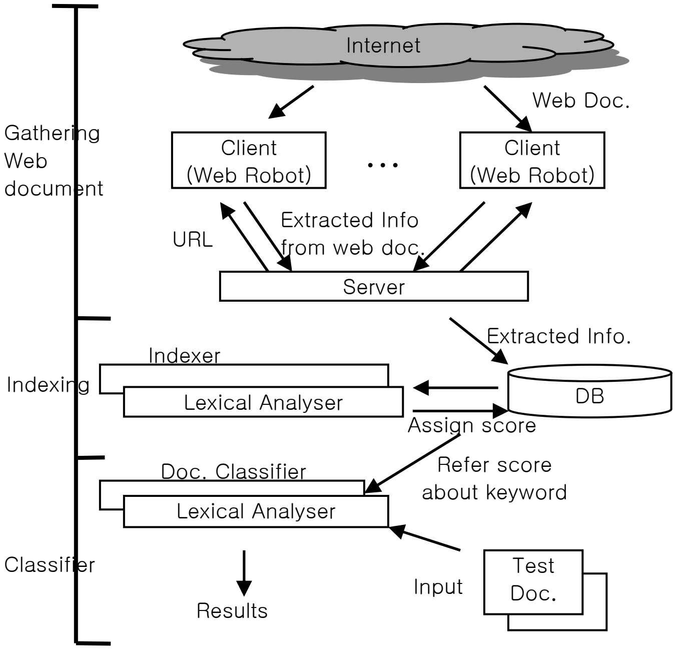

Figure 5: The structure of` system for classifying web documents

The system is largely composed of web robot, indexer, concept knowledge database for each category and document classifier.

The web robot has multiple clients connected to a single server that collects web documents according to the instructions from the server. The indexer extracts index words from the collected web documents and executes indexing to construct concept knowledge in each category's database. The database of each category stores the concept knowledge that is extracted from the collected documents. The document classifier uses the concept knowledge constructed by the web document collector and indexer to classify each of the input documents.

5.0 EXPERIMENT AND RESULT

5.1 Testing Environment

In order to construct databases for each category, web documents were collected from each of the categories in the internet portal site Yahoo Korea[Yahoo Korea]'s web directory. The categories are economy, government and entertainment and the documents are obtained from these categories. Web documents from the starting web document to the documents reached after clicking five hyperlinks were collected. The number of web documents gathered from each of the three categories is economy 3000, government 900 and entertainment 2000 like table 1. The categories differ in size. So the process of regulation is necessary.

Table 1: Collecting Conditions of Web Document for Constructing of Concept knowledge

Web Directory No. 1	
Name	Business and Economy
Starting URL Address	http://kr.dir.yahoo.com/business_and_economy/
Number of Web Documents Collected	3000(83 among 3083 discarded)
Depth of Search	5 levels
Web Directory No. 2	
Name	Government
Starting URL Address	http://kr.dir.yahoo.com/government/
Number of Web Documents Collected	900
Depth of Search	5 levels
Web Directory No. 3	
Name	Entertainment
Starting URL Address	http://kr.dir.yahoo.com/entertainment/
Number of Web Documents Collected	2000(44 among 2044 discarded)
Depth of Search	5 levels

5.2 Evaluation and Results

For testing document classification, articles from politics, economy, and entertainment sections were randomly extracted from newspapers. These articles matched the three main categories economy, government and entertainment. The total number of articles extracted is 500. Among them 250 articles dealt with economy, 120 with government and 130 with entertainment. Table 2 indicates the number of articles for testing.

Table 2: Classification of Newspaper Articles Used in the Test

Category	Number of Newspaper Articles Used in Test
Economy (Business and Economy)	250
Government	120
Entertainment	130
Total	500

Newspaper articles amounting up to 500 were input into the system implemented in this study in order to check the document classification result. In this case where the documents were classified according to simply word frequency, 120 out of 500 documents were classified incorrectly to show an accuracy of 76%. When the documents were classified by TFIDF, 80 out of 500 documents were classified incorrectly to show an accuracy of 84%. However when the method proposed in this study was applied, only 60 documents were misplaced. As a result, our suggested method shows the highest accuracy.

Table 3: Comparison of Accuracy

Classification Method	Number of Incorrect Classification	Accuracy
Simple Word Frequency	120	76 %
Simple TFIDF	80	84 %
Proposed Method	60	88 %

5.3 Evaluation

According to the result of the test in section 5.2, document classification using simple word frequency showed the poorest performance scoring 76% in accuracy. Classification applying simple TFIDF showed improved performance scoring 84%. However, the proposed method showed still better performance scoring 88% in accuracy. The categories applied in this test were economy, government and entertainment. Among these three, economy and government are very closely related to each other. Due to this there were many cases where the document belonging to the economy category was misplaced in the government category. Meanwhile, the documents belonging to the entertainment category were classified accurately since entertainment is relatively less related to the other two categories.

6.0 CONCLUSION AND FUTURE STUDIES

In previous studies, there were many tasks to be done manually in order to construct concept knowledge. For this reason, the construction of concept knowledge was the most difficult processes in constructing a system. This study has attempted to solve this problem by developing a method of automatic concept knowledge construction making manual work unnecessary. In order to classify documents, the semantic of the subject the document is purported to deliver must be understood. Therefore, the study proposes to use the hyperlink information as well as the statistical method for applying the semantic of the connections in classifying documents into categories. Index words were extracted by the index word extractor, which uses the internet web directory and morpheme analyzer. The extracted index words were then constructed into concept knowledge for each category. The document classifier used the concept knowledge and the extended TFIDF for each category in order to classify the documents. The test results where the proposed method and algorithm scored 88% in accuracy verified the possibility of automatic classification.

While the present study used only the document title and hyperlink to construct the concept knowledge, further studies should search for a way to use the keywords within the full-text also. Additionally, further our studies are also required in finding a way to extract the characteristics of the relations between keywords by using the extracted concept relation.

7.0 REFERENCES

Frakes, W. and Baeza-Yates,R. (1992)., Information Retrieval. Prentice Hall

Lewis, David D. (1998). Naïve (Bayes) at forty: The independence assumption in information retrieval. Proceedings of ECML-98. 10th European Conference on Machine Learning.

Park, S., Kim,S.K, Hwang,S.C.,Kim,K.T .(2000). Obtaining Web Information from Expert Search Engines by Using Concept Graphs. Spring 2000, Academic Presentations (B) 27 (1) pp. 295-297, Korea Information Science Society.

Salton, G., and Buckley, C. (1987). Term weighting approaches in automatic text retrieval. Tech Report 87-881 Dept. of Computer Science, Cornell University.