Refining Term Weights of Documents Using Term Dependencies

Hee-soo Kim
Dept. of Computer Engineering
Ajou University
Suwon, Gyeonggi-do 443-749
Republic of Korea
+82 31 219 2442
heemanz@ajou.ac.kr

Ikkyu Choi
Dept. of Computer Engineering
Ajou University
Suwon, Gyeonggi-do 443-749
Republic of Korea
+82 31 219 1839
ikchoi@ajou.ac.kr

Minkoo Kim
Dept. of Computer Engineering
Ajou University
Suwon, Gyeonggi-do 443-749
Republic of Korea
+82 31 219 2437
minkoo@ajou.ac.kr

ABSTRACT
When processing raw documents in Information Retrieval (IR) System, a *term-weighting scheme* is used to calculate the importance of each term which occurs in a document. However, most *term-weighting schemes* assume that a term is independent of the other terms. Term dependency is an indispensable consequence of language use [1]. Therefore, this assumption can make the information of a document being lost. In this paper, we propose new approach to refine term weights of documents using term dependencies discovered from a set of documents. Then, we evaluate our method with two experiments based on the vector space model [2] and the language model [3].

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval models*

General Terms
Algorithms, Experimentation

Keywords
Indexing, Text Mining, Term-weighting scheme

1. INTRODUCTION
Term-weighting schemes are used to calculate the importance of each term which occurs in a document. These *term-weighting schemes* generally assume that a term is independent of other terms. However, terms are interrelated with other terms which co-occur in a document. The term dependency is an indispensable consequence of language use [1]. Therefore, the use of *term-weighting scheme*, which is based on the assumption of term-to-term independence, may make the information of a document being lost. Loss of this information may lower the performance of IR system. We assumed that the use of term dependencies is a factor which affects the correctness of term weights. Then, we use term dependencies for improving performance of IR system.

In order to compute term dependencies, we adopt the method of *mining association rules* proposed by Rakesh Agrawal and Ramakrishman Srikant [4]. *Mining association rules*, which is one of the Data Mining techniques, is used to discover relationships among sets of items in a set of customers' transactions. The association rule is represented as the form of "$X \rightarrow Y$," where X and Y are sets of items and $X \cap Y$ is an empty set. The association rule also has *support* and *confidence* values:

$$Support(X \rightarrow Y) = \Pr(X \cup Y)$$

$$Confidence(X \rightarrow Y) = \Pr(Y \mid X) = \frac{\Pr(X \cup Y)}{\Pr(X)}$$

where, $\Pr(X)$ is the probability that a transaction includes all items in a set X. In a previous study [5], term dependencies by *mining association rules* were used to select representative terms of document class. We have an interest in *confidence* value of an association rule because it can be considered to be a measurement of term dependency in IR system.

2. REFINING TERM WEIGHTS
The proposed method consists of two steps. The first step is to discover term dependencies from a set of total documents using mining association rules. In order to use *mining association rules*, we can consider documents and terms to be transactions and items, respectively. When finding term associations, we consider that term dependencies consist of only two terms. There are a large number of term pairs. For example, following pairs can exist in a document which contains terms of A, B and C: $\{A\}$ and $\{B\}$, $\{A, B\}$ and $\{C\}$, $\{B, C\}$ and $\{A\}$, and so on. It is impossible to consider all term dependencies among sets of terms because the time complexity for finding term dependencies exponentially increases. Therefore, *confidence* value of "$X \rightarrow Y$" is the following equation:

$$Confidence(X \rightarrow Y) = \frac{\text{the number of documents containing } X \text{ and } Y}{\text{the number of documents containing } X}$$

where, X and Y are exclusive terms.

In the second step, term weights of all documents are updated by term dependencies. Our basic idea is that terms are mutually affected by other terms in a document. First, we make a term association graph of each document according to the term dependencies. At this time, terms in the document and term dependencies between terms are represented as nodes and links

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

such as Figure 1. t_i is the *i*th term in a set of documents and $c_{i,j}$ is the *confidence* value of "$t_i \rightarrow t_j$".

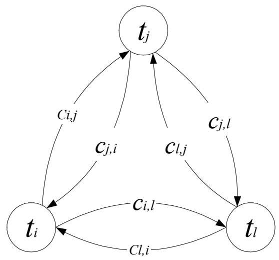

Figure 1. Part of term association graph

Then term weights of documents are updated by following equations:

D : a set of documents

d_k : the kth document

t_i : the ith term of D

$c_{i,j}$: the confidence value of "$t_i \rightarrow t_j$"

$w_{i,k}$: the ith term weight of d_k

dl_k : the length of d_k

$$w'_{i,k} = \frac{\sum_{i \neq j, t_j \in d_k} w_{j,k} \cdot c_{j,i}}{dl_k}$$

$$newWeight_{i,k} = \alpha \cdot w_{i,k} + \beta \cdot w'_{i,k} \qquad (where, \alpha + \beta = 1)$$

Each term weight is transferred to the other term weight as much as *confidence* value between two terms. Consequently, $w'_{i,k}$ becomes the term weight influenced by other terms. Thus, the refined term weight is computed by linear combination with the original term weight and the dependency-based term weight.

3. EXPERMENTS

In our experiments, we use 224680 documents of the 'department of energy (DOE)' which is offered by Text Retrieval Conference (TREC). Also, we use ten topics (65, 66, 68, 82, 83, 96, 102, 111, 134 and 135) of TREC-1 as queries. In order to evaluate the refined term weights, we compare retrieval performances based on the refined term weights and the original term weights in vector space model and language model, respectively. In vector space model and language model, term weights are respectively computed by *tf-idf scheme* and by the probabilities that those terms occur in a document. Thus our method modified term weights based on term-to-term independence. Table 1 and 2 shows experimental results with vector space model and language model, respectively.

Table 1. Averages of the precisions in Vector Space Model

Used term weights	10 docs	20 docs	30 docs
baseline(α=1.00, β=.00)	0.310	0.260	0.197
refined weights(α=.75, β=.25)	0.430	0.380	0.333
refined weights(α=.50, β=.50)	0.450	0.375	0.353
refined weights(α=.25, β=.75)	0.460	0.360	0.283
refined weights(α=.00, β=1.00)	0.370	0.290	0.223

Table 2. Averages of the precisions in Language Model

Used term weights	10 docs	20 docs	30 docs
baseline(α=1.00, β=.00)	0.410	0.375	0.313
refined weights(α=.75, β=.25)	0.440	0.375	0.303
refined weights(α=.50, β=.50)	0.430	0.385	0.303
refined weights(α=.25, β=.75)	0.420	0.380	0.307
refined weights(α=.00, β=.1.00)	0.430	0.360	0.297

In vector space model, we indicate that retrieval performances based on the refined weights are better than retrieval performance based on the original term weights. In language model, the results are better than baseline at 10 docs. However, the use of term dependencies lowers retrieval performances than that of baseline at 30 docs.

4. CONCLUSIONS AND FUTURE WORK

We proposed a novel method to refine term weights of documents using term dependencies discovered from a set of total documents. This method used mining association rules to find term dependencies and then accordingly refined term weights of documents. In our experiments, we evaluated the proposed method by comparing retrieval performances in vector space model and language model. These results indicate that the use of term dependencies makes term weights more accurate. However, we will need to experiment with the large documents set since the used set of documents was rather small. In the future, we will also propose more specific method to use term dependencies.

5. ACKNOWLEDGMENTS

This work is performed as a part of National Research Laboratory supported by Ministry of Science and Technology in Korea (M10302000087-03J0000-044000).

6. REFERENCES

[1] Nikolaos Nanas, Victoria Uren, and Anne De Roeck. Building and Applying a Concept Hierarchy Representation of a User Profile. In 26th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003

[2] Ricardo Baeza-Yates and Berthier Ribeiro-Neto: Modern Information Retrieval. 19-71 and 117-134, Addison Wesley, 1999.

[3] Jay M. Ponte and W. Bruce Croft. A Language Modeling Approach to In Information Retrieval. In 21th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998.

[4] Rakesh Agrawal, Tomasz Imielinski and Arun Swami: Mining Association Rules between Sets of Items in Large Databases. In Proc. of the ACM SIGMOD Conference, 207-216, Washington, D.C., 1993

[5] Shian-Hua Lin, Chi-Sheng Shih, Meng Chang Chen, Jan-Ming Ho, Ming-Tat Ko and Yueh-Ming Huang: Extracting Classification Knowledge of Internet Documents with Minig Term Associations: A Semantic Approach. In 21th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998.