A Novel Approach to Semantic Indexing Based on Concept

Bo-Yeong Kang
Department of Computer Engineering
Kyungpook National University
1370, Sangyukdong, Pukgu, Daegu, Korea(ROK)
comeng99@hotmail.com

Abstract

This paper suggests the efficient indexing method based on a concept vector space that is capable of representing the semantic content of a document. The two information measure, namely the information quantity and the information ratio, are defined to represent the degree of the semantic importance within a document. The proposed method is expected to compensate the limitations of term frequency based methods by exploiting related lexical items. Furthermore, with information ratio, this approach is independent of document length.

1 Introduction

To improve the unstable performance of a traditional keyword-based search, a Web document should include both an index and index weight that represent the semantic content of the document. However, most of the previous works on indexing and the weighting function, which depend on statistical methods, have limitations in extracting exact indexes(Moens, 2000). The objective of this paper is to propose a method that extracts indexes efficiently and weights them according to their semantic importance degree in a document using concept vector space model.

A document is regarded as a conglomerate concept that comprises by many concepts. Hence, an n-dimensional concept vector space model is defined in such a way that a document is recognized as a vector in n-dimensional concept space. We used lexical chains for the extraction of concepts. With concept vectors and text vectors, semantic indexes and their semantic importance degree are computed. Furthermore, proposed indexing method had an advantage in being independent of document length because we regarded overall text information as a value 1 and represented each index weight by the semantic information ratio of overall text information.

2 Related Works

Since index terms are not equally important regarding the content of the text, they have term weights as an indicator of importance. Many weighting functions have been proposed and tested. However, most weight functions depend on the statistical methods or on the document's term distribution tendency. Representative weighting functions include such factors as term frequency, inverse document frequency, the product of the term and inverse document frequency, and length normalization(Moens, 2000).

Term frequency is useful in a long document, but not in a short document. In addition, term frequency cannot represent the exact term frequency because it does not include anaphoras, synonyms, and so on. Inverse document frequency is inappropriate for a reference collection that changes frequently because the weight of an index term needs be recomputed. A length normalization method is proposed because term frequency factors are numerous for long documents, and negligible for short ones, obscuring the real importance of terms. As this approach also uses term frequency function, it has the same disadvantage as term frequency does.

Hence, we made an effort to use methods based on the linguistic phenomena to enhance the indexing performance. Our approach focuses on proposing concept vector space for extracting and weighting indexes, and we intend to compensate limitations of the term frequency based methods by employing lexical chains. Lexical chains are to link related lexical items

in a document, and to represent the lexical cohesion structure of a document(Morris, 1991).

3 Semantic Indexing Based on Concept

Current approaches to index weighting for information retrieval are based on the statistic method. We propose an approach that changes the basic index term weighting method by considering semantics and concepts of a document. In this approach, the concepts of a document are understood, and the semantic indexes and their weights are derived from those concepts.

3.1 System Overview

We have developed a system that performs the index term weighting semantically based on concept vector space. A schematic overview of the proposed system is as follows: A document is regarded as a complex concept that consists of various concepts; it is recognized as a vector in concept vector space. Then, each concept was extracted by lexical chains(Morris, 1988 and 1991). Extracted concepts and lexical items were scored at the time of constructing lexical chains. Each scored chain was represented as a concept vector in concept vector space, and the overall text vector was made up of those concept vectors. The semantic importance of concepts and words was normalized according to the overall text vector. Indexes that include their semantic weight are then extracted.

The proposed system has four main components:

- Lexical chains construction
- Chains and nouns weighting
- Term reweighting based on concept
- Semantic index term extraction

The former two components are based on concept extraction using lexical chains, and the latter two components are related with the index term extraction based on the concept vector space, which will be explained in the next section.

3.2 Lexical Chains and Concept Vector Space Model

Lexical chains are employed to link related lexical items in a document, and to represent the lexical cohesion structure in a document(Morris, 1991). In accordance with the accepted view in linguistic works that lexical chains provide representation of discourse structures(Morris, 1988 and 1991), we assume that

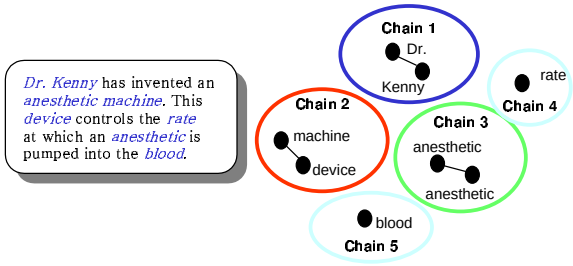

Figure 1: Lexical chains of a sample text

each lexical chain is regarded as a concept that expresses the meaning of a document. Therefore, each concept was extracted by lexical chains.

For example, Figure 1 shows a sample text composed of five chains. Since we can not deal all the concept of a document, we discriminate representative chains from lexical chains. Representative chains are chains delegated to represent a representative concept of a document. A concept of the sample text is mainly composed of representative chains, such as chain 1, chain 2, and chain 3. Each chain represents each different representative concept: for example man, $machine$ and $anesthetic$.

As seen in Figure 1, a document consists of various concepts. These concepts represent the semantic content of a document, and their composition generates a complex composition. Therefore we suggest the concept space model where a document is represented by a complex of concepts. In the concept space model, lexical items are discriminated by the interpretation of concepts and words that constitute a document.

Definition 1 (Concept Vector Space Model) *Concept space is an n-dimensional space composed of n-concept axes. Each concept axis represents one concept, and has a magnitude of C_i. In concept space, a document T is represented by the sum of n-dimensional concept vectors, $\vec{C}_i$.*

$$\vec{T} = \sum_{i=1}^{n} \vec{C}_i \tag{1}$$

Although each concept that constitutes the overall text is different, concept similarity may vary. In this paper, however, we assume that concepts are mutually independent without consideration of their similarity. Figure 2 shows the concept space version of the sample text.

3.3 Concept Extraction Using Lexical Chains

Lexical chains are employed for concept extraction. Lexical chains are formed using WordNet and asso-

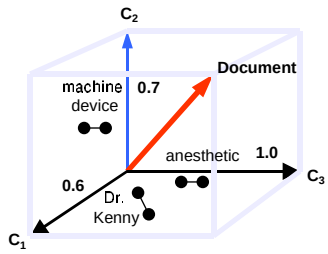

Figure 2: The concept space version of the sample text

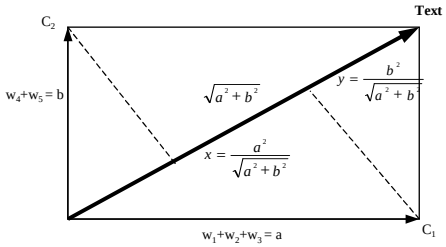

Figure 3: Vector space property

ciated relations among words. Chains have four relations: synonym, hypernyms, hyponym, meronym. The definitions on the score of each noun and chain are written as definition 2 and definition 3.

Definition 2 (Score of Noun) *Let $NR^k_{N_i}$ denotes the number of relations that noun N_i has with relation k. $SR^k_{N_i}$ represents the weight of relation k. Then the score $S_{NOUN}(N_i)$ of a noun N_i in a lexical chain is defined as:*

$$S_{NOUN}(N_i) = \sum_k (NR^k_{N_i} \times SR^k_{N_i}) \quad (2)$$

where $k \in$ set of relations.

Definition 3 (Score of Chain) *The score $S_{CHAIN}(Ch_x)$ of a chain Ch_x is defined as:*

$$S_{CHAIN}(Ch_x) = \sum_{i=1}^{n} S_{NOUN}(N_i) + penalty \quad (3)$$

where $S_{NOUN}(N_i)$ is the score of noun N_i, and $N_1, ..., N_n \in Ch_x$.

Representative chains are chains delegated to represent concepts. If the number of the chains was m, chain Ch_x, should satisfy the criterion of the definition 4.

Definition 4 (Criterion of Representative Chain) *The criterion of representative chain, is defined as:*

$$S_{CHAIN}(Ch_x) \geq \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} S_{CHAIN}(Ch_i) \quad (4)$$

3.4 Information Quantity and Information Ratio

We describe a method to normalize the semantic importance of each concept and lexical item on the concept vector space. Figure 3 depicts the magnitude of the text vector derived from concept vectors C_1 and C_2. When the magnitude of vector C_1 is a and that of vector C_2 is b, the overall text magnitude is $\sqrt{a^2 + b^2}$.

Each concept is composed of words and its weight w_i. In composing the text concept vector, the part that vector C_1 contributes to a text vector is x, and the part that vector C_2 contributes is y. By expanding the vector space property, the weight of lexical items and concepts was normalized as in definitions 5 and definition 6.

Definition 5 (Information Quantity, Ω) *Information quantity is the semantic quantity of a text, concept or a word in the overall document information. Ω_T, Ω_C, Ω_W are defined as follows. The magnitude of concept vector C_i is $S_{CHAIN}(Ch_i)$:*

$$\Omega_T = \sqrt{\sum_k C_k^2} \quad (5)$$

$$\Omega_{C_i} = \frac{C_i^2}{\sqrt{\sum_k C_k^2}} \quad (6)$$

$$\Omega_{W_j} = \Omega_T \times \Psi_{W_j|T} = \frac{W_j \cdot C_i}{\sqrt{\sum_k C_k^2}} \quad (7)$$

The text information quantity, denoted by Ω_T, is the magnitude generated by the composition of all concepts. Ω_{C_i} denotes the concept information quantity. The concept information quantity was derived by the same method in which x and y were derived in Figure 3. Ω_{W_j} represents the information quantity of a word. $\Psi_{W_j|T}$ is illustrated below.

Definition 6 (Information Ratio, Ψ) *Information ratio is the ratio of the information quantity of a comparative target to the information quantity of a text, concept or word. $\Psi_{C|T}$, $\Psi_{W|C}$ and $\Psi_{W|T}$ are defined as follows:*

$$\Psi_{W_j|C_i} = \frac{S_{NOUN}(W_j)}{S_{CHAIN}(C_i)} = \frac{|W_j|}{|C_i|} \quad (8)$$

$$\Psi_{C_i|T} = \frac{\Omega_{C_i}}{\Omega_T} = \frac{C_i^2}{\sum_k C_k^2} \quad (9)$$

$$\Psi_{W_j|T} = \Psi_{W_j|C_i} \times \Psi_{C_i|T} = \frac{W_j \cdot C_i}{\sum_k C_k^2} \quad (10)$$

The weight of a word and a chain was given when forming lexical chains by definitions 2 and 3. $\Psi_{W_j|C_i}$ denotes the information ratio of a word to the concept in which it is included. $\Psi_{C_i|T}$ is the information ratio of a concept to the text. The information ratio of a word to the overall text is denoted by $\Psi_{W_i|T}$.

The semantic index and weight are extracted according to the numerical value of information quantity and information ratio. We extracted nouns satisfying definition 7 as semantic indexes.

Definition 7 (Semantic Index) *The semantic index that represents the content of a document is defined as follows:*

$$\Omega_{W_j} \geq \beta \cdot \frac{1}{m} \sum_{i=1}^{m} (\Omega_{W_i}) \quad (11)$$

Although in both cases information quantity is the same, the relative importance of each word in a document differs according to the document information quantity. Therefore, we regard information ratio rather than information quantity as the semantic weight of indexes. This approach has an advantage in that we need not consider document length when indexing because the overall text information has a value 1 and the weight of the index is provided by the semantic information ratio to overall text information value, 1, whether a text is long or not.

4 Experimental Results

In this section we discuss a series of experiments conducted on the proposed system. The results achieved below allow us to claim that the lexical chains and concept vector space effectively provide us with the semantically important index terms. The goal of the experiment is to validate the performance of the proposed system and to show the potential in search performance improvement.

4.1 Standard TF vs. Semantic Indexing

Five texts of Reader's Digest from Web were selected and six subjects participated in this study. The texts were composed of average 11 lines in length(about five to seventeen lines long), each focused on a specific topic relevant to $exercise$, $diet$, $holiday$ $blues$,$yoga$, and $weight\ control$. Most texts are related to a general topic, $exercise$. Each subject was presented with five short texts and asked to find index terms and weight each with value from 0 to 1. Other than that, relevancy to a general topic, $exercise$, was rated for each text. The score that was rated by six subjects is normalized as an average.

Table 1: Manually extracted index terms and relevancy to $exercise$

Text	Index		Rel.
Text1	exercise(0.39)	back(0.3)	0.64
	pain(0.175)		
Text2	diet(0.56)	exercise(0.31)	0.55
Text3	yoga(0.5)	exercise(0.25)	0.45
	mind(0.11)	health(0.1)	
Text4	weight(0.46)	control(0.18)	0.26
	calorie(0.11)	exercise(0.11)	
Text5	holiday(0.432)	humor(0.23)	0.099
	blues(0.15)		

Table 2: Percent Agreement(PA) to manually extracted index terms

	T1	T2	T3	T4	T5	Avg.
PA	0.79	1.0	0.88	0.79	0.83	0.858

The results of manually extracted index terms and their weights are given in Table 1. The index term weight and the relevance score are obtained by averaging the individual scores rated by six subjects. Although a specific topic of each text is different, most texts are related to the $exercise$ topic. The percent agreement to the selected index terms is shown in Table 2(Gale, 1992). The average percent agreement is about 0.86. This indicates the agreement among subjects to an index term is average 86 percent.

We compared these ideal result with standard term frequency(standard TF, S-TF) and the proposed semantic weight. Table 3 and Figures 4-6 show the comparison results. We omitted a few words in representing figures and tables, because standard TF method extracts all words as index terms. From Table 3, subjects regarded $exercise$, $back$, and $pain$ as index terms in Text 1, and the other words are recognized as relatively unimportant ones. Even though $exercise$ was mentioned only three times in Text 1, it had considerable semantic importance in the document; yet its standard TF weight did not represent this point at all, because the importance of $exercise$ was the same as that of $muscle$, which is also mentioned three times in a text. The proposed approach, however, was able to

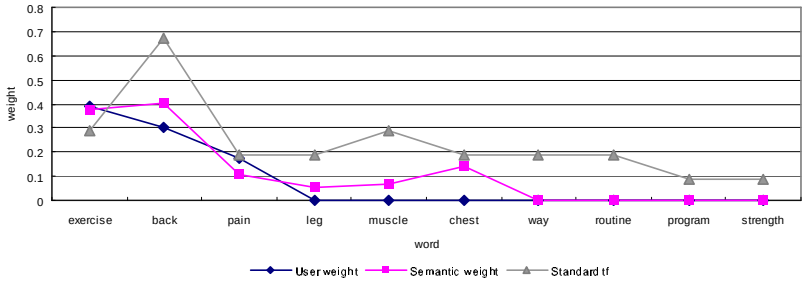

Figure 4: Weight comparison of Text1

Table 3: Weight comparison of Text 1

Text 1			
Word	Subject Weight	Standard TF	Semantic Weight
exercise	0.39	0.29	0.3748
back	0.3	0.67	0.4060
pain	0.175	0.19	0.1065
chest	0.0	0.19	0.1398
leg	0.0	0.19	0.0506
muscle	0.0	0.29	0.0676
way	0.0	0.19	0.0
routine	0.0	0.19	0.0
program	0.0	0.09	0.0
strength	0.0	0.09	0.0

differentiate the semantic importance of words. Figure 4 shows the comparison chart version of Table 3, which contains three weight lines. As the weight line is closer to the subject weight line, it is expected to show better performance. We find from the figure that the semantic weight line is analogous to the manually weighted value line than the the standard TF weight line is.

Figures 5 and 6 show two of four texts(Text2, Text3, Text4, Text5). Figures on the other texts are omitted due to space consideration. In Figure 5, *pound* is mentioned most frequently in a text, consequently, standard TF rates the weight of *pound* very high. Nevertheless, subjects regarded it as unimportant word. Our approach discriminated its importance and computed its weight lower than *diet* and *exerciese*. From the results, we see the proposed system is more analogous to the user weight line than the standard TF weight line.

Table 4: Weight comparison to the index term *exercise* of five texts.

Text	Subject	TF	LN	S-TF	Proposed	Rel.
1	0.39	3	0.428	0.29	0.3748	0.64
2	0.31	3	0.75	0.375	0.2401	0.55
3	0.25	1	0.33	0.18	0.1320	0.45
4	0.11	1	0.125	0.11	0	0.26
5	0	1	0.2	0.12	0	0.09

4.2 Applicability of Search Performance Improvements

When semantically indexed texts are probed with a single query, *exercise*, the ranking result is expected to be the same as the order of the relevance score to the general topic *exercise*, which was rated by subjects.

Table 4 lists the weight comparison to the index term *exercise* of five texts, and the subjects' relevance rate to the general topic *exercise*. Subjects' relevance rate is closely related with the subjects' weight to the index term, *exericise*. The expected ranking result is as following Table 5. TF weight method hardly discerns the subtle semantic importance of each texts, for example, Text1 and Text2 have the same rank. Length normalization(LN) and standard TF discern each texts but fail to rank correctly. However, the proposed indexing method provides better ranking results than the other TF based indexing methods.

4.3 Conclusion

In this paper, we intended to change the basic indexing methods by presenting a novel approach using a concept vector space model for extracting and weighting indexes. Our experiment for semantic indexing supports the validity of the presented approach, which is capable of capturing the semantic importance of

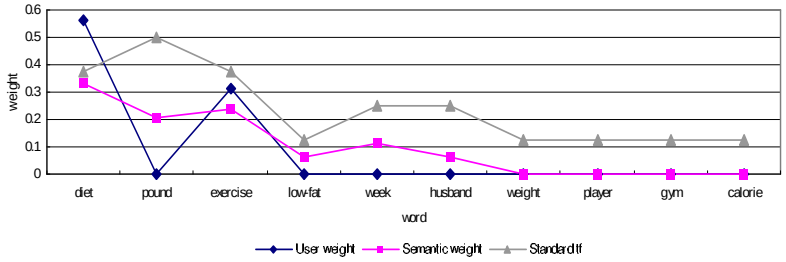

Figure 5: Weight comparison of Text2

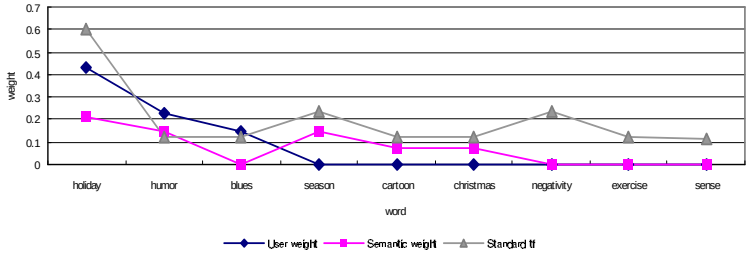

Figure 6: Weight comparison of Text5

Table 5: Expected ranking results to the query *exercise*

Rank	Rel.	Subject	TF	LN	S-TF	Proposed
1	Text1	Text1	Text1 Text2	Text2	Text2	Text1
2	Text2	Text2	Text3 Text4 Text5	Text1	Text1	Text2
3	Text3	Text3		Text3	Text3	Text3
4	Text4	Text4		Text5	Text5	Text4 Text5
5	Text5	Text5		Text4	Text4	

a word within the overall document. Seen from the experimental results, the proposed method achieves a level of performance comparable to major weighting methods. In an experiment, we didn't compared our method with inverse document frequency(IDF) yet, because we will develop more sophisticated weighting method concerning IDF in future work.

References

R. Barzilay and M. Elhadad, Using lexical chains for text summarization, Proc. ACL'97 Workshop on Intelligent Scalable Text Summarization(1997).

M.-F. Moens, Automatic Indexing and Abstracting of Document Texts, Kluwer Academic Publishers(2000).

J. Morris, Lexical cohesion, the thesaurus, and the structure of text, Master's thesis, Department of Computer Science, University of Toronto(1988).

J. Morris and G. Hirst, Lexical cohesion computed by thesaural relations as an indicator of the structure of text, Computational Linguistics 17(1)(1991) 21-43.

W. Gale, K. Church, and D. Yarowsky, Extimating upper and lower bounds on the performance of word-sense disambiguation programs. In Proceedings of the 30th annual Meeting of the Association for Computational Linguistics(ACL-92)(1992) 249-256.

Reader's Digest Web site, http://www.rd.com