경영정보학연구
제16권 제4호
2006년 12월

개념 네트워크를 이용한 정보 검색 방법*

허 원 창**, 이 상 진***

Document Retrieval using Concept Network

Wonchang Hur, Sangjin Lee

The advent of KM(knowledge management) concept have led many organizations to seek an effective way to make use of their knowledge. But the absence of right tools for systematic handling of unstructured information makes it difficult to automatically retrieve and share relevant information that exactly meet user's needs. we propose a systematic method to enable content-based information retrieval from corpus of unstructured documents. In our method, a document is represented by using several key terms which are automatically selected based on their quantitative relevancy to the document. Basically, the relevancy is calculated by using a traditional TFIDF measure that are widely accepted in the related research, but to improve effectiveness of the measure, we exploited 'concept network' that represents term-term relationships. In particular, in constructing the concept network, we have also considered relative position of terms occurring in a document. A prototype system for experiment has been implemented. The experiment result shows that our approach can have higher performance over the conventional TFIDF method.

Keywords : Document Management, Information Retrieval, Concept Network, Knowledge Management

---

* 이 논문은 2005~2006년도 인하대학교의 지원에 의하여 연구되었음. (INHA-34086)
** 교신저자, 인하대학교 경영학부
*** 삼성네트웍스 기업솔루션사업팀

Ⅰ. 서 론

데이터베이스의 등장은 기업에 존재하는 방대한 량의 정보를 실시간으로 저장하고 검색할 수 있도록 하였다. 이러한 정보공유를 통해 기업은 운영의 효율을 기하며, 보다 정확한 정보에 근거한 의사결정을 통해 경영활동을 할 수 있었다. 지식경영(Knowledge Management)의 개념이 대두되면서, 많은 기업들이 기업 내의 유형/무형의 고부가가치 지식들을 효과적으로 공유하거나 활용하고자 노력하고 있으나, 아직은 데이터베이스와 같은 편리한 메커니즘으로 그러한 지식을 활용할 수 있는 기술적 방안이 부족하다고 할 수 있다[Alvai, 1999; Awad, 2004].

일반적인 정보의 존재 형태는 정보가 나타내는 개념을 표현하는 방식에 따라 <표 1>과 같이 크게 두 가지로 구분해 볼 수 있다.

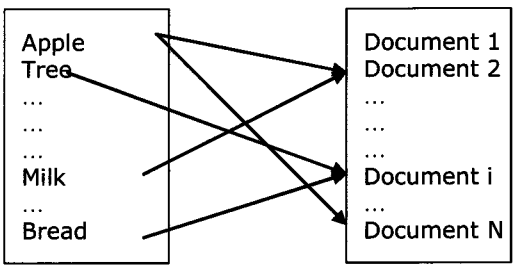
<표 1> 정보의 존재 형태

구조화된 정보	비구조화된 정보
체계화된 표현 형식 (Schema)	자율적인 표현 형식 (자연어)
의미(Semantics)과 표현 형식(Syntax)의 밀접한 연관성	의미와 형식의 비관련성
검색 도구(질의어)의 존재	검색 도구(질의어) 부재
자동화된 정보 추출, 처리	효과적 정보 추출, 처리 방법 부재
데이터베이스, STEP, SGML 등	HTML, Text, Multimedia 데이터 등

데이터베이스, STEP(STandard for the Exchange of Product data), 그리고 SGML(Standard Generalized Markup Language) 등은 표현하고자 하는 정보를 미리 정의되어 있는 체계화된 표현 형식(데이터스키마)에 맞추어 기술하게 되어 있으며 이는 정보시스템을 통한 자동화된 관리를 가능하게 한다. 하지만, 높은 부가가치를 창출하는 지식이나, 조직원의 학습능력을 고양시킬 수 있는 지식 등은 특정한 스키마로 구조화할 수 없는 경우가 많다[Awad, 2004; Tiwana, 2002]. 이러한 정보가 유형의 형태로 존재하기 위해서는 주로 자연어로 된 문서로 존재할 수밖에 없으며, 이러한 경우 데이터베이스에서 사용하는 SQL(Structured Query Language)과 같이 구조적인 질의어를 통해 접근하는 것이 불가능하다.

따라서 이와 같이 비구조적인 형태로 저장된 정보를 효과적으로 공유하는 시스템이 가능하기 위해서는 정보의 형식이나 메타정보(meta information)을 이용한 검색방법이 아닌, 정보의 내용에 기반을 둔 검색방법이 가능해야 하며, 이는 문서와 질의의 표현방식, 문서와 질의와의 의미적 연관성 계산 등을 기술적으로 지원할 수 있는 체계적 방법을 요구한다[Gopal, 1999].

현재 널리 사용되고 있는 내용기반 정보검색은 대부분 주제어(keyword)를 이용하여 문서와 질의를 표현하고, 해당 주제어들과 문서에 포함된 어휘들 간의 문자열 매칭(string matching)을 통해, 해당 주제어가 포함된 문서를 찾아내는 방식을 사용한다. <그림 1>은 문자열 매칭 기반의 주제어 검색방식에서 사용하고 있는 문서의 색인 구조를 나타낸다[Frakes, 1992; Grossman, 1999].



<그림 1> Inverted File

이러한 정보검색 방식은 주제어가 담고 있는 개념과 관계없이, 문자열의 동일성이라는 형식적인 요소만을 고려하기 때문에, 다중적인 의미를 갖는 주제어를 이용한 검색이나, 방대한 검색결과의 우선순위 부여 등을 해결하기 위한 적절한

해결책이 되지 못한다. 특히 주어진 주제어와 유사한 의미를 갖는 다른 주제어들은 검색에 있어서 중요한 정보임에도 불구하고 적극적으로 고려되지 못한다.

본 연구에서는, 주제어들 간에 존재하는 의미적 연관성(semantic relations)을 정량화하는 방법을 제시하고, 이를 사용하여 원하는 정보를 문서의 내용에 기반을 두어 검색할 수 있는 기술적 방식을 제안한다.

Ⅱ. 문서의 표현 방법

자연어로 된 문서를 표현하는 방법은, <표 2>와 같이 주제어를 이용한 표현 방식과 문자열을 이용한 표현 방식이 있다[Frakes, 1992]. 다개국어로 구성된 문서나 복잡한 기호와 수식이 포함된 문서의 검색을 위해 사용되는 N-gram(N개의 문자열)과 같은 특수한 예를 제외하고, 대부분의 지식관리를 위한 정보시스템에서는 주제어 기반 표현 방식을 사용한다. 이는 자연어로 기술된 문서로부터 필요한 개념을 추출하기 위한 유일한 단서가 바로 주제어이기 때문이다. Luhn H.[1960]은 그의 오랜 저서에서 "문서를 표현하는 가장 중요한 요소는 이를 구성하고 있는 어휘들이며 이러한 어휘들의 분포나 빈도수와 같은 통계 자료는 문서의 개념을 표현하는 가장 효과적이자 유일한 자료"라고 이야기하고 있다[Frakes, 1992].

<표 2> 문서의 내용기반 표현방법

	주제어 기반 표현	문자열 기반 표현
방식	주요한 어휘의 집합	N개의 연속된 문자 집합
특징	개념적 표현 어휘의 개념에 기반을 둔 표현	형식적 표현 어휘를 문자의 집합으로 표현
장단점	언어에 종속적 문서의 개념 표현 가능 어휘의 변형처리 작업	언어에 독립적 문서의 개념 표현 불가 추가 작업 불필요

물론, 인위적으로 부가한 문서의 속성(주제어, 제목, 저자, 분야 등)을 이용하여 문서를 표현하는 방식도 있을 것이나, 이를 위해서는 이러한 메타정보를 기존 정보에 추가해야 하는 방대한 량의 수작업을 필요로 하게 된다. 따라서 효과적인 주제어 기반의 문서 표현 방식을 위해서는 컴퓨터에 의한 자동적인 표현이 가능해야 한다.

자동화된 문서의 표현방법으로 가장 널리 사용되는 방법은 문서집합에 빈번하게 등장하는 n개의 주제어를 추출하여 문서를 n차원의 벡터로 표현하는 방법이다[Lewis, 1992]. 이를 위해서는 n개의 주제어들 각각이 해당 문서에 대하여 가지는 연관성을 적절한 방식으로 수치화해야 한다. 다음에서는 이러한 연관성의 수치화방식으로 가장 많이 사용되는 TFIDF 표현기법에 대하여 설명한다.

2.1 TFIDF 표현

TFIDF 표현기법은 TF(Term Frequency)와 IDF (Inverse Document Frequency)를 이용하여 문서와 특정 주제어 간의 연관성을 정량화하는 방법이다. 이는 문서에 존재하는 어휘들의 점유도(Exhaustivity)와 집중도(Specificity)의 두 가지 개념의 정량화에 기반을 둔 방식이다. 즉, 일반적으로 문서에 등장하는 어휘들의 분포를 조사해 보면 해당 문서에서 중요한 어휘일수록 그 문서에의 점유도와 집중도가 높은 값을 갖는다는 통계적 가정에 기반을 둔 접근방법이다[Faloutsos, 1994; Frakes, 1992; Grossman, 1999; Combarro, 2005].

여기서 점유도란, 한 어휘가 표현하는 개념이 해당 문서가 담고 있는 전체 개념에서 차지하는 비중을 의미하는 것으로써, 한 문서에서 자주 사용되는 어휘는 점유도가 높다고 할 수 있다. 일반적으로 점유도의 측정은 주로 어휘의 한 문서 내에서의 빈도수나 밀집도(Density)와 같은 통계 자료를 사용한다. 한편 집중도란 한 어휘가 표현

하는 해당 문서에서만 갖게 되는 개념의 구체성을 의미한다. 즉, 여러 문서에 골고루 사용되는 언어는 특정 문서에의 집중도가 낮다고 할 수 있다. 집중도의 측정은 여러 문서에서의 어휘의 출현 분포(Distribution)와 같은 통계자료를 사용한다.

점유도와 집중도의 개념에 근간한 TFIDF 실질적인 계산방법은 가중치의 적용방식, 정규화 방법, 로그(log) 스케일의 적용 방식 등에 따라 매우 다양하다[Singhal, 1996 : Sebastiani, 2002]. 다음의 식 (1)은 가장 대표적으로 사용되는 TFIDF 계산 방식이다.

$$tfidf(t_i, d_j) = tf(t_i, d_j) \times \log(\frac{|D|}{df(t_i)}) \quad (1)$$

위의 식에서 어휘의 점유도는 $tf(t_i, d_j)$ (TF: Term Frequency)로 표현되었으며 이는 어휘 t_i가 문서 d_j에 나타난 빈도수를 의미한다. 어휘의 집중도는 문서집합의 크기(문서의 개수)인 $|D|$를 어휘 t_i가 등장하는 문서의 개수인 $df(t_i)$ (DF: Document Frequency)로 나눈 값으로 표현되었으며 이는 어휘가 특정한 문서에 집중된 정도를 나타낸다. 어휘의 집중도는 일반적으로 식에서 보는 바와 같이 log 스케일을 사용한다. 이는 문서집합의 크기가 큰 경우 IDF가 TF에 비해 상대적으로 큰 영향을 미치는 것을 조절하는 효과가 있다.

2.2 문서 - 어휘 행렬

TFIDF 측정치를 사용하면 문서집합 D에 대하여, 주제어 집합 T의 각 주제어를 사용하여 식 (2)와 같은 문서-어휘 행렬 M_D (|T|X|D|) 으로 표현할 수 있다.

$$M_D = \begin{pmatrix} r(t_1, d_1) & r(t_1, d_2) & \cdots & r(t_1, d_{|D|}) \\ r(t_2, d_1) & r(t_2, d_2) & \cdots & r(t_2, d_{|D|}) \\ \cdots & \cdots & \cdots & \cdots \\ r(t_{|T|}, d_1) & r(t_{|T|}, d_2) & \cdots & r(t_{|T|}, d_{|D|}) \end{pmatrix},$$

$$r(t_i, d_j) \in [0,\ 1], 1 \le i \le |T|, 1 \le j \le |D| \quad (2)$$

식 (2)에서 r_{ij}는 어휘 t_i가 문서 d_j에 가지는 의미적 연관성(relevance)의 크기를 의미하며 식 (1)의 TFIDF를 사용하여 식 (3)과 같은 코사인(cosine) 정규화를 통해 계산된다.

$$r(t_i, d_j) = \frac{tfidf(t_i, d_j)}{\sqrt{\sum_{s=1}^{|T|}(tfidf(t_s, d_j))^2}} \quad (3)$$

이러한 문서집합의 표현방식은, 문서를 주제어휘들이 구성하는 벡터공간에 수치화하여 표현할 수 있기 때문에, 의사결정나무나 인공신경회로망 같은 통계적 기법을 사용하여, 문서의 분류, 검색 등의 기능을 가능하게 한다.

Ⅲ. 개념 네트워크(Concept Network)를 이용한 문서의 표현

TFIDF를 이용한 문서-어휘 행렬의 표현방식은 문서와 어휘간의 관련성을 표현하기 위한 방법으로, 일반적으로 어휘들 사이에 존재하는 관련성은 고려하지 않는다. 어휘간의 의미적 연관성을 표현하는 대표적인 방식은 개념 네트워크(concept network)를 사용하는 것이다[Chen S., 1999]. 개념 네트워크의 꼭지점(node)은 주제 어휘들을 나타내며 꼭지점 사이를 연결하는 선분(edge)은 어휘간의 의미적 관련성을 나타낸다. 어휘간의 의미적 관련성은 보통 0~1사이의 값을 갖는 퍼지 값(fuzzy value)로 계산될 수 있다. 이러한 개념 네트워크는 최종적으로 하나의 정방행렬로 표현될 수 있다.

본 연구에서는 이러한 개념 네트워크를 계산하는 새로운 방법을 제시한다. 그리고 이를 이용하여 문서 - 어휘 행렬을 새롭게 정의하였다. 제시된 방법의 가장 큰 특징은 두 어휘가 발생하는 회수의 상관관계와 함께, 문서 내에서의 위치관

계를 함께 고려한다는 점이다. 이어지는 내용에서는 이러한 어휘간 위치관계를 이용하여 개념 행렬을 계산하는 방법을 자세히 설명한다.

3.1 어휘의 발생위치를 고려한 개념 행렬 (Concept Matrix)

개념 네트워크에서의 개념(concept)이란 사용자 혹은 시스템이 정의하는 문서 집합의 주제 어휘들을 일컫는다. 이러한 주제 어휘들 간의 의미적 연관성을 정량화하기 위해서는 두 어휘가 문서집합의 각 문서에서 등장하는 패턴을 통계적으로 수치화하는 과정이 필요하다. 이를 위해 가장 널리 사용되는 통계수치는 어휘의 동시발생 횟수(co-occurrence)이다. 한편, 일반적으로 한 문서는 여러 개의 단락으로 구성되며, 각 단락은 여러 개의 문장으로 구성된다. 문서의 구조는 문서가 기술하고 있는 내용에 기반을 둔 것이며, 어휘들 간의 의미적 연관관계 또한 이러한 문서의 구조에 영향을 받게 된다. 즉, 어휘간의 의미적 연관관계를 추정하기 위해서는 어휘의 등장 횟수와 더불어 어휘의 발생위치의 관계 또한 주요한 통계 수치로 활용될 수 있다. 본 연구에서는 어휘간의 동시발생 회수와 어휘의 위치관계를 고려하여 어휘간의 의미적 관련성을 추정하기 위한 수식을 고안하였다. 이는 다음의 수식 (4)와 같이 표현된다.

$$f(t_i \rightarrow t_j) = \frac{\sum_{k=1}^{|D|}\left\{co-tfidf((t_i,t_j),d_k)\times\log\left(\frac{|d_k|}{dist((t_i,t_j),d_k)}\right)/\log(|d_k|)\right\}}{\sum_{k=1}^{|D|} tfidf(t_i, d_k)}$$

where,

$$co-tfidf((t_i, t_j),d_k) = co-tf((t_i, t_j), d_k) \times \log\left(\frac{|D|}{co-df(t_i,t_j)}\right)$$

$$dist((t_i,t_j),d_k) = Average_{pos_i}\left\{Min_{pos_j}\{|pos_i(t_i)-pos_j(t_j)|\}\right\} \quad (4)$$

식 (4)에서 *co-tfidf*는 *co-tf*와 *co-df*로 표현된다. $cotf((t_i, t_j), d_k)$는 두 어휘가 공통으로 해당 문서에 가지는 연관성은 두 어휘가 동시에 등장한 회수를 나타내며, $codf(t_i, t_j)$는 두 어휘의 동반등장이 발생한 문서의 개수를 나타낸다. 따라서 *co-tfidf*는 두 어휘를 하나의 쌍으로 하여 *tfidf*를 계산한 것이다. *dist*는 두 어휘의 거리를 나타낸다. 계산을 위해서 군집분석(cluster analysis)에서 활용하는 군집간의 거리 계산방식을 활용하였다. *dist*의 계산식에서 $pos_i(t_i)$는 어휘 t_i가 발생한 위치의 집합이다. 따라서 식은 두 어휘의 모든 발생위치 간의 평균 최단거리를 의미한다.

*f*의 계산은 *co-tfidf*와 *dist*를 이용한다. 식에서 분자는 두 어휘를 결합한 개념이 각 문서에 대해 갖는 연관성의 합이고, 분모는 어휘 t_i가 단독으로 각 문서에 가지는 연관성의 합이다. 따라서 $f(t_i \rightarrow t_j)$는 어휘 t_i가 단독으로 사용될 때와 어휘 t_i와 t_k가 공통으로 고려되었을 때, 각 문서와의 연관성을 비율로 나타낸 것이다. 이를 사용하면 서로 다른 두 쌍의 어휘의 연관성을 비교할 수 있다. 예를 들어, $f(t_i \rightarrow t_j) > f(t_i \rightarrow t_k)$가 성립한다면, 어휘 t_i는 t_k에 비해 t_j와 더 연관성이 높다고 할 수 있을 것이다.

다음의 <표 3>은 두 어휘의 발생거리를 계산하는 간단한 예를 보여준다. 표에서 w의 값은 *dist*값을 이용하여 식 (4)에서 분자의 log항 이후의 값을 계산한 것이다.

표에서 두 어휘 t_1과 t_2는 문서 d_1과 d_2에서 발생빈도는 모두 동일하지만, 발생위치는 서로 다르다. 두 어휘는 d_1에서는 다소 떨어진 위치에서 발생하였으며, d_2에서는 가까운 위치에서 발생하였고 식 (4)에 의해 계산된 위치에 따른 가중치는 d_2에서 모두 더 높게 나타났다. 이는 발생 빈도가 동일하게 높지만 두 어휘가 문서의 구조상에서 서로 먼 거리에서 발생한다면 이는 의미적 관련

<표 3> 어휘 T1과 T2의 발생위치와 가중치 계산

문서(크기)	어휘	빈도	발생위치(pos)	dist ($t_1 \rightarrow t_2$)	dist ($t_2 \rightarrow t_1$)
d_1(150)	t_1	3	(8, 22, 58)	{|8-4|+|22-16|+|58-16|}/3=19.3, w = 0.409	{|2-8|+|4-8|+|16-22|}/3 = 5.3 w = 0.667
	t_2	3	(2, 4, 16)		
d_2(90)	t_1	3	(2, 8, 16)	{|2-4|+|8-6|+|16-14||/3 = 2 w = 0.868	{|4-2|+|6-8|+|14-16|||/3=2 w = 0.868
	t_2	3	(4, 6, 14)		

성이 상대적으로 낮고, 반대로 발생 빈도는 낮지만 두 어휘 사이의 발생 거리가 가깝다면 의미적 관련성이 상대적으로 높다는 가정을 수치로 반영하는 것이다.

종합하면, 식 (4)는 두 개의 어휘로 복합된 하나의 개념을 TFIDF 표현방법에 적용한 값(co-tfidf)과, 어휘간의 발생위치를 사용하여 측정한 거리(dist)를 사용하여 계산된 것이다. 따라서 식에 의하면, 두 어휘가 동시에 발생하는 경우가 빈번하며, 그런 현상이 특정한 문서에 집중될 경우(연관성은 없으나, 한 어휘가 일반적으로 자주 사용되는 어휘일 경우의 문제를 해결하기 위한 방법), 그리고 그 발생위치가 가까울 경우 연관성의 값이 높다지게 된다. 최종적으로 문서를 표현하는 데 사용된 각 어휘들 간의 연관성을 사용하여 다음 식 (5)에 보는 바와 같은 개념 행렬(concept matrix, M)을 계산할 수 있다.

$$M_T = \begin{pmatrix} f(t_1 \rightarrow t_1) & f(t_1 \rightarrow t_2) & \cdots & f(t_1 \rightarrow t_{|T|}) \\ f(t_2 \rightarrow t_1) & f(t_2 \rightarrow t_2) & \cdots & f(t_2 \rightarrow t_{|T|}) \\ \cdots & \cdots & \cdots & \cdots \\ f(t_{|T|} \rightarrow t_1) & f(t_{|T|} \rightarrow t) & \cdots & f(t_{|T|} \rightarrow t_{|T|}) \end{pmatrix},$$

$$f(t_i \rightarrow t_j) \in [0, 1], 1 \le i \le |T|, 1 \le j \le |T| \qquad (5)$$

3.2 개념행렬을 이용한 문서-어휘 행렬 표현

개념 행렬이 유효한(valid) 퍼지 관계를 표현하기 위해서는 식 (6)의 이행 속성(transitivity)을 만족하여야 한다. 이행 속성이란, 어휘 i와 j, j와 k의 연관성이 각각 주어졌을 때, i와 k의 연관성은 독립적으로 할당될 수 없고, 기존에 주어진 두 개의 연관성 값에 의해 간접적인 제약을 받는다는 것을 의미한다. 따라서 어휘 j와 어휘 k의 관련성은 모든 다른 어휘 i를 매개로한 각각의 관련성의 최소값들 중에서, 가장 큰 값보다 커야 한다. 즉, 식 (6)의 부등식을 만족해야 한다[Zimmermann, 1991].

$$f(t_i \rightarrow t_k) \ge Max_i[Min(f(t_i \rightarrow t_j), f(t_j \rightarrow t_k))],$$
$$0 \le i,\ j,\ k \le m \qquad (6)$$

이러한 퍼지 이행 속성(transitivity)을 만족하는 행렬을 이행폐쇄 행렬(transitive closure matrix)이라고 하며, 이는 식 (7)과 같이 구할 수 있다.

$$M^T = M^{T+1} = M^{T+2} = M^2 = M \otimes M =$$
$$\begin{pmatrix} \bigcup_{i=1\text{-}m} f_{1i} \cap f_{i1} & \bigcup_{i=1,.,m} f_{1i} \cap f_{i2} & \cdots & \bigcup_{i=1\text{-}m} f_{1i} \cap f_{im} \\ \bigcup_{i=1\text{-}m} f_{2i} \cap f_{i1} & \bigcup_{i=1,.,m} f_{2i} \cap f_{i2} & \cdots & \bigcup_{i=1\text{-}m} f_{2i} \cap f_{im} \\ \cdots & \cdots & \cdots & \cdots \\ \bigcup_{i=1\text{-}m} f_{mi} \cap f_{m1} & \bigcup_{i=1\text{-}m} f_{mi} \cap f_{i2} & \cdots & \bigcup_{i=1\text{-}m} f_{mi} \cap f_{im} \end{pmatrix} \qquad (7)$$

$$where\ ,\ f_{ij} = f(t_i \rightarrow t_j)$$

이행폐쇄는 모든 퍼지관계가 이행 속성을 만족하는 상태로, 서로 다른 두 개의 어휘 사이의 관련성이 제 3의 어휘에 의해 변화되지 않는 상태를 말한다. 이러한 이행폐쇄 행렬은 퍼지연산에 의해 계산된다. 퍼지연산(⊗)은 앞서 언급한 식 (6)을 일반화 한 것으로 어휘 j와 k의 관련성을 모든 어휘 i에 대해서 어휘 i와 j, 어휘 i와 k의

관련성의 최소값(∩)들 중, 가장 큰 값(∪)으로 설정하도록 한다.

이렇게 구해진 이행폐쇠행렬은 다음의 식 (8)에서 문서-어휘 행렬에 적용될 수 있다. 즉, 위의 식에 의해 사용자에 의해 설정된 문서와 주제(어휘)와의 직접적인 관계 뿐 아니라, 어휘 사이의 관련성에 의해 간접적인 관계까지 고려한 새로운 문서-어휘 행렬을 구할 수 있다.

$$D^* = D \otimes M^T \qquad (8)$$

식 (8)은 문서 i에 대하여, 어휘 j와 어휘 k가 가지는 관련성은, 두 어휘 간에 존재하는 개념적 연관성에 독립적이지 않다는 사실을 반영한 것이다. 4장에서는 이러한 문서 - 어휘 행렬의 표현 방식이 기존 방식에 비하여 보다 정확한 검색결과를 제시하는 것을 실험을 통하여 제시하였다.

Ⅳ. 실험결과 및 분석

본 연구에서 제안된 문서 표현 방식의 장점을 검증하기 위하여 구성된 문서 집합을 대상으로 검색을 수행하였다. 검색 성능의 비교를 위하여 대표적인 인터넷 검색 업체인 Google의 검색 방식(Google), 기존의 TFIDF 표현방식(Tfidf), 그리고 본 연구에서 제안된 방식(C-Tfidf)에 따라 각각 검색 결과를 측정하였다. Google의 검색 성능을 비교한 것은, 본 실험을 통해 제시된 방법론이 최적의 검색 성능을 보이는 것을 증명하기 위한 것은 아니며, 다만 상용 검색시스템과의 비교를 통한 참고자료로써의 의미를 제시하기 위한 것이다.

<표 4> 문서집합과 전문가의 검색결과

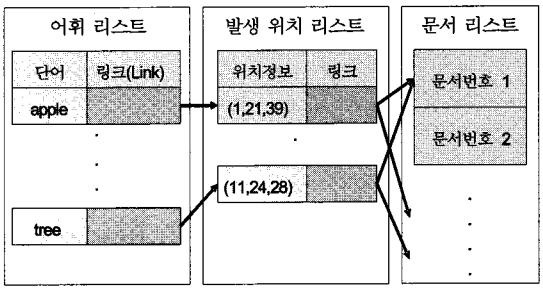
검색어	관련 문서번호	개수
concurrency control	**003**, 032, **068**, **120**, 135, 144, 173, 262, 305, 332, 384, 421, 552, 639, **697**, 768, 834, **839**	18
information retrieval	**056**, **209**, 515, 548, 595, 640, 863, 905, 975	9
knowledge base	**068**, 076, 082, 177, 190, 275, 353, 378, 391, 492, 496, **513**, 663, 679, 690, 751, 758, 763, 773	19
knowledge discovery	046, 118, 193, 195, **209**, 222, 227, 373, 575, 691, **697**, 813, 866, 884, 898, 911, 923, 939, 952, 957, 962, 970	22
machine learning	166, 329, 336, 699, 824, 877	6
multimedia	064, 081, 284, 303, 361, 386, **406**, 425, 433, 504, 508, 533, 612, 855, 872, 889, 926	17
neural network	**120**, 183, 211, 340, 448, 683, 978	7
query optimization	210, 321, 338, 429, 493, 641, 650, 793, 900, 915, 941	11
security	**003**, 016, **056**, 067, 085, 228, 274, 292, **406**, **513**, 524, 555, 661, 668, **839**	15
temporal database	001, 021, 036, 277, 297, 389, 464, **513**, 563, 566, 583, 591, 746	13
총계 (127 + 중복검색 10)		137

4.1 문서집합 및 실험설계

실험에 사용된 문서집합은 저널 'IEEE Knowledge and Data Engineering'에 1995년 ~ 1999년까지 출간된 총 350여 편의 논문들 중 10여 가지 주제로 선택된 127개의 논문을 사용하여 구성하였다. 문서집합의 구성에 정확성을 기하기 위하여 2인의 전문가가 개별 문서의 내용을 직접 확인하여 주제어와 관련성이 없다고 판단되는 문서는 제외하는 방식을 취하였다. 최종적으로 구성된 문서집합은 다음과 같다. 진하게 표시된 문서는 여러 주제에 중복적으로 관련되는 문서이다. 각 문서는 대략 200~300 단어의 어휘로 구성되어 있다.

검색 실험을 위해 프로토타입 검색시스템을 개발하였다. 검색시스템은 문서들의 전처리(preprocessing) 및 인덱스 생성기능을 수행한다. 전처리 과정에서는 문서를 읽어서 불필요한 어휘들을 제거하고 중심이 되는 주요어휘만을 추출하는 과정을 거친다. 이는 크게, 각 어휘의 품사를 파악하고 이를 통해 불필요한 문법 어휘들을 제거하는 과정과, 어휘 변화를 처리하기 위하여 모든 어휘를 기본형으로 변형하는 어미 처리 과정을 거친다. 불필요한 어휘의 처리는 문서에 포함된 어휘 중에서 너무 자주 발생하는 무의미한 어휘를 제거하는 것으로, 관사, 전치사, 접속사, 대명사 및 그 밖에 자주 사용되는 동사 등으로 구성된 제거어휘리스트(stop list)를 사용한다. 어미처리과정은 동사, 형용사, 명사 등의 어미(ending) 변화를 감지하여 기본형으로 변형하는 과정이다. 이를 위하여 널리 사용되고 있는 어휘사전(lexicon)인 워드넷(WordNet) 2.0을 사용하였다[Miller, 1993].

전처리 과정이 끝나면 추출된 어휘들을 이용하여 <그림 2>와 같이 문서집합의 색인을 생성한다. 그림에서 각 어휘는 발생 위치가 저장된 연결 리스트를 참조한다. 어휘의 발생위치는 문서의 시작위치를 1로 했을 때의 상대위치로 구성되어 있으며, 해당 문서와 연결되어 있다. 색인구조를 통해 얻어지는 정보를 이용하여 어휘의 발생빈도 계산 및 어휘간의 관련성 계산이 수행된다. 이러한 방식으로 총 2011개의 어휘로 구성된 색인을 생성하였으며, 이를 이용하여 TFIDF 방식의 문서-어휘 행렬과 개념 네트워크를 적용한 문서-어휘 행렬을 각각 구성하였다.



<그림 2> 색인구조

4.2 검색 성능의 비교

검색 성능의 측정은 가장 널리 사용되는 recall과 precision의 두 가지 척도를 사용하였다. recall은 검색 방식이 얼마나 많은 관련문서를 추출해 낼 수 있는 가를 측정하는 것으로, 질의어를 통해 검색되어야 할 전체 문서(관련문서)중 검색결과에 나타난 문서의 비율로 계산된다. 반면, precision은 검색결과의 정확성을 의미하는 것으로 총 검색결과에 나타난 문서 중 실제 해당 질의어와 관련이 있는 문서의 비율로 계산된다 [Faloutsos, 1994].

<표 5>는 10개의 어휘 중 'neural network'를 사용하여 검색을 수행한 결과를 나타낸다. Google 방식의 검색 결과는 Google에서 제공하는 Google Desktop 검색 엔진을 사용하여 직접 검색을 수행하고 그 결과를 관련성 순서로 정렬한 것이다.[1] 성능 측정은 각 순위까지 검색된 문서를 대

1) Google의 경우 Google desktop 프로그램의 특성으로 인해 특정 순위 이하의 검색 결과를 얻을 수 없었다.

<표 5> 검색결과 및 성능 측정(어휘 "neural network")

순위	Google	Tfidf	C-Tfidf	Google		Tfidf		C-Tfidf	
				Recall	Precision	Recall	Precision	Recall	Precision
1	**683**	**683**	**978**	0.125	100.0%	12.5%	100.0%	12.5%	100.0%
2	957	**183**	**120**	0.125	50.0%	25.0%	100.0%	25.0%	100.0%
3	**448**	957	**448**	0.25	66.7%	25.0%	66.7%	37.5%	100.0%
4	**183**	**448**	**340**	0.375	75.0%	37.5%	75.0%	50.0%	100.0%
5	**978**	195	**211**	0.5	80.0%	37.5%	60.0%	62.5%	100.0%
6	195	**211**	697	0.5	66.7%	50.0%	66.7%	62.5%	83.3%
7	**211**	**340**	**183**	0.625	71.4%	62.5%	71.4%	75.0%	85.7%
8	**340**	697	195	0.75	75.0%	62.5%	62.5%	75.0%	75.0%
9	**120**	**978**	64	0.875	77.8%	75.0%	66.7%	75.0%	66.7%
10	-	391	872	-	-	75.0%	60.0%	75.0%	60.0%
11	-	**120**	699	-	-	87.5%	63.6%	75.0%	54.5%
12	-	...	284	-	-	87.5%	38.9%	87.5%	38.9%
13	-	...	329	-	-	87.5%	36.8%	87.5%	36.8%

상으로 하여 매 순위마다 precision과 recall을 계산하는 방식으로 하였다. recall과 precision의 특성상 검색된 문서 수가 증가할수록 recall은 증가하며, precision은 감소하는 현상을 보인다.[2] 이와 같은 방식으로 총 10개의 검색어를 대상으로 각 검색결과의 성능을 측정하였다. 세 검색 방식의 성능을 비교하기 위해서, <표 6>에 보는 바와 같이 10개의 검색결과를 대상으로 recall의 구간 별로 측정된 precision의 평균값을 구하였다. 이를 통해 동일한 recall 구간을 기준으로 precision을 비교할 수 있다.

<표 6>에서 보는 바와 같이 본 연구에서 제시한 방식은 기존의 TFIDF를 이용한 방식에 비해 모든 구간에서 우수한 검색성능을 제시하였다. 특히 recall이 30%~60%의 경우에는 precision이 평균 10% 이상 높게 나타나는 결과를 보였다.

2) 문서의 개수를 1개씩 증가시키며 recall과 precision을 측정하였기 때문에, 특정 부분에서 한시적으로 recall과 precision이 동시에 증가하는 현상이 있다.

<표 6> 실험 결과

Recall	Precision (평균치, %)		
	Google	Tfidf	C-Tfidf
0~10%	97.5	81.7	84.4
10~20%	87.6	70.2	72.9
20~30%	87.2	64.3	65.1
30~40%	87.4	53.6	64.6
40~50%	86.6	52.3	62.7
50~60%	84.9	50.8	59.2
60~70%	84.8	46.7	52.1
70~80%	83.2	42.4	46.4
80~90%	81.4	39.4	40.0

주목할 만한 것은 세 가지 검색방식 중 전문가와 가장 유사한 검색 결과를 제시하는 것은 Google의 검색방식이었다. Google의 Pagerank 알고리듬은 문서간의 관련성을 정량화하고 이를 바탕

<표 7> 실험 결과(검색어 'information retrieval'과 'concurrency control'의 비교)

Recall	precision(평균치), 'information retrieval'			precision(평균치), 'concurrency control'		
	Google	Tfidf	C-Tfidf	Google	Tfidf	C-Tfidf
0~10%	100.0	61.1	100.0	100.0	100.0	100.0
10~20%	83.3	41.1	64.2	70.8	80.6	100.0
20~30%	75.0	40.2	35.1	81.7	81.7	87.8
30~40%	80.0	37.0	23.5	79.5	69.9	79.5
40~50%	83.3	34.5	21.3	80.9	61.9	75.9
50~60%	85.7	34.4	21.5	84.0	63.6	74.4
60~70%	82.6	35.0	-	85.7	50.4	51.7
70~80%	59.5	38.1	-	-	-	-
80~90%	-	36.2	-	-	-	-

으로 검색을 수행한다. 이러한 방식은 특히 문서간에 하이퍼링크를 통해 참조가 일어날 경우 매우 우수한 성능을 보이는 것으로 알려져 있다[Brin, 1999]. 실험에서는 기존 방식과 개선된 방식의 성능을 비교하는 데 초점을 맞추었으나 상용 검색시스템과 검색 성능이 이와 같이 차이나는 이유는, 검색 성능이 문서의 표현방식 이외에, 형태소 분석의 최적화, 문서간 참조관계의 활용, 어휘의 사전적 의미 및 품사의 활용 등 여러 가지 추가적인 요소에 의해 영향을 받기 때문으로 보인다.

실험을 통해 본 연구에서 제시한 어휘간의 거리 관계를 활용한 검색방식이 기존의 방식에 비해 우수한 성능을 보임을 개괄적으로 증명할 수 있었으나, 여러 문서에서 발생 빈도가 높은 일부 어휘들에서 검색 성능이 뒤처지는 것을 발견하였다. <표 7>은 검색어 'information retrieval'의 검색결과와 'concurrency control'의 검색결과를 비교한 것이다. 이와 같은 현상이 발생하는 원인으로는 제안된 방식이 'information'과 같이 자주 등장하는 어휘의 경우, 많은 다른 어휘와의 높은 연관관계로 인해 IDF의 영향을 감쇄하는 역할을 하기 때문으로 짐작된다. 즉, 자주 출현하는 어휘일수록 개념 네트워크상에서 관련 어휘들이 많아지고, 그렇게 될수록 해당 어휘가 특정 문서에 가지는 독점성이 감쇄되기 때문이다. 이러한 현상은, 'information', 'database' 등과 같이 DF가 높은 어휘들을 검색어로 사용했을 때 공통적으로 나타나는 현상이었다. 반면 'concurrency', 'security' 등 특수한 의미에 사용되는 어휘들을 검색어로 사용하였을 때 상대적으로 우수한 성능을 보였다. 추후 연구를 통해 이러한 부분의 원인을 논리적으로 규명하고, 이를 해소할 수 있는 방안이 필요하다고 판단된다.

V. 결론 및 토의

본 연구에서는 구조화되지 않은 문서집합에 대하여 내용 기반의 자동화된 검색 방법론을 제시하였다. 본 연구가 가지는 중요한 의미는 어휘들의 발생횟수 뿐만 아니라 어휘 간의 물리적 위치관계를 동시에 고려하는 새로운 문서 - 어휘 표현방식을 제시하였다는 점이다. 또한 실험결과를

통해 제시된 방법이 기존의 방식보다 우수한 성능을 가질 수 있음을 보였다. 더욱이 대부분의 기존 연구들이 어휘 자체의 사전적 의미와 어휘 사전에서 제공하는 의미의 연관성을 사용하였으며, 어휘의 물리적 거리를 활용한 방식은 거의 찾아보기 힘든 상황에서 본 연구가 가지는 의미는 크다고 하겠다.

다만, 본 실험을 위해 전문가가 직접 문서의 내용을 참고하여 검색을 수행한 결과를 여러 가지 자동화된 검색 방식과 비교하는 방식을 취하였기 때문에 보다 많은 수의 문서로 실험을 수행하기에는 제약이 있었다. 또한, 개발된 검색시스템은 문서가 증가할수록 어휘의 숫자도 증가하게 되는데 어휘가 많아지면 성능이 급격하게 감소되는 현상을 보였다. 실험에 사용된 문서 집합의 크기가 제시된 방식의 우월성을 통계적으로 완벽히 입증하기에는 다소 작은 수이지만, 전반적인 성능의 향상 및 효과성을 보여주는 데는 무리가 없었다고 판단된다. 차후 문서의 분류 결과가 알려진 문서집합을 대상으로 한 추가적인 실험과 검색 시스템의 최적화를 통해 보다 명확한 통계적 검증이 필요할 것이다.

본 연구는 문서집합의 내용기반 색인 방식에 대하여 초점을 맞추었으나, 궁극적으로는 효과적인 지식관리 시스템을 설계 및 적용하기 위해서는 질의와 문서의 표현, 관련성 평가, 결과 제시, 사용자 feed-back의 처리 등의 항목에 대하여 다음과 같은 추가 연구들이 수행되어야 할 것이다. <표 4>는 효과적인 검색 시스템이 갖추어야 할 요구사항을 나열하였다[Kalt, 1996; Chen, H., 1994; Bartell, 1992].

<표 8> 지식관리 시스템의 기반기술

항목	고려 요소	필요 기술
질의/문서 표현	자연어 질의, 다의어 지원, 고유 명사 검색, 어휘 사전 사용, 다개국어 지원 등	자연어처리,
연관성 평가	벡터 모델 및 확률 모델의 적용, 어휘의 의미를 사용한 색인 기법의 적용, 구조적 문서 표현 모델 개발	통계이론, 확률이론
검색 결과 제시	검색 결과의 체계적인 우선순위 부여, 내용에 따른 계층적 분류, 사용자 기호(Preference)의 반영	의사결정나무, 신경회로망, 정보가시화
사용자 feed-back	Feed-back에 따른 질의어 확장, 사용자와의 상호적인(Interactive) 검색. 학습을 통한 지속적 검색 성능의 향상	에이전트 기술, 질의 최적화

〈참 고 문 헌〉

[1] Alavi, M. and Leidner, D., "Knowledge Management Systems: Emerging Views and Practices from the Field," In IEEE Proceedings of the 32nd Hawaii International Conference on System Sciences, pp.1-11, 1999.

[2] Awad, E.M. and Ghaziri, H. M., Knowledge Management, Prentice Hall, Upper Saddle River, NJ 07458, 2004

[3] Tiwana, A., "Knowledge Management Toolkit: Orchestrating IT, Strategy, and Knowledge Plafrom," Prentice Hall, Upper Saddle River, NJ 07458, 2002.

[4] Bartell, B.T. et al., "Optimizing Parameters in a Ranked Retrieval System Using Multi-Query Relevance Feedback," In the Proceeding of 3rd Annual Symposium on Document Analysis and Information Ret-

rieval, 1992.

[5] Chen, H., "Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms," MIS Department, College of Business and Public Administration, University of Arizona, July 1994.

[6] Chen, S.M., Wang, J.Y., "Fuzzy Query Processing for Document Retrieval Based on Extended Fuzzy Concept Network," Transactions on Systems, Man, and Cybernetics, Vol. 29, No. 1, February 1999.

[7] Combarro, E.F., Montanes, E., Diaz, I., Ranilla, J, and Mones, R., "Introducing a Family of Linear Measures for Feature Selection in Text Categorization," IEEE Transaction on Knowledge and Data Engineering, Vol. 17, No. 9, 2005.

[8] Faloutsos, C., Orad, D., "A Survey of Information Retrieval and Filtering Methods," University of Maryland, College Park, MD20742, 1994.

[9] Fraeks. W. and Baeza-Yates, R., Information Retrieval: Data Structures and Algorithms, Prentice-Hall, 1992.

[10] Gopal, B. and Manber, U., Integrating Content-Based Access Mechanisms with Hierarchical File Systems, In ACM OSDI, pp. 265-278, 1999.

[11] Grossman, D.A. and Frieder, O., "Information Retrieval: Algorithms and Heuristics," Kluwer Academic Publishers, 1998.

[12] Gudivada, V.N. and Raghavan, V.V., Design and evaluation of algorithms for image retrieval by spatial similarity, ACM Transaction on Information Systems, Vol. 13, No. 2, April, pp. 115-144, 1995.

[13] Gudivada, V. N., Reaghavan, V. V., Grosky, W. I., and Kasanagottu, R., "Information Retrieval On The World Wide Web," IEEE Internet Computing, September. October 1997.

[14] Kalt, T., "A New Probabilistic Model of Text Classification and Retrieval", University of Massachusetts, 1996.

[15] Kim, Y. H. et al., "InfoFlow: A Web- based Workflow Management System," Proceedings of International Conference CALS/ EC Korea'99, 1999.

[16] Lewis, D.D., "Representation and Learning in Information Retrieval," University of Massachusetts, 1992.

[17] Luhn, H., Keyword in Context Index for Technical Literature, American Documentation, XI (4), 1960.

[18] Miller, G. A., Beckwith, R., "Introduction to Word-Net: An On-line Lexical Database," Princeton University,1993.

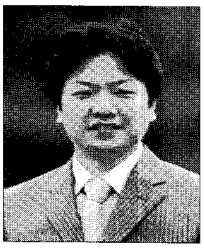
[19] Singhal, A., Salton, G., Mitra, M., and Buckley, C. "Document length normalization," Information Processing and Management, Vol. 32, No. 5, pp 619-633, 1996.

[20] Sebastiani, F., "Machine Learning in Automated Text Categorization," ACM Computing Surveys, Vol. 34, No. 1, pp. 1-47, Mar. 2002.

[21] Brin, S. and Page, L., "The anatomy of a large-scale hypertextual Web search engine," Proceedings of the seventh international conference on World Wide Web 7, pp. 107-117, 1998.

[22] Zimmermann, H. J., Fuzzy Set Theory - and Its Applications, Kluwer Academic Publishers, 1991.

◆ 저자소개 ◆



허원창 (Hur, Wonchang)
2005년부터 인하대학교 경영학부에 재직하고 있다. 서울대학교 산업공학과에서 학사 및 석사학위를 받은 후 2004년 2월 동 대학원에서 박사학위를 받았으며, 2004년~2005년까지는 (주)CyberMed에서 연구소장을 역임하기도 하였다. 주요 관심분야는 비즈니스 프로세스 관리, Knowledge Engineering, 기술경영 등이다.

이상진 (Lee, Sang Jin)
1998년 서울대학교 산업공학과 학사 졸업하였고, 2000년 서울대학교 산업공학과 석사 졸업하였다. 2000년~2005년 (주)핸디소프트 BPM 개발팀에서 근무하였으며, 2005년~2006년 (주)삼성네트웍스 기업솔루션 사업팀에서 근무하였다. 주요 관심분야는 BPM, RFID/USN, Web2.0, UCC 등이다.

◈ 이 논문은 2006년 02월 03일 접수하여 1차 수정을 거쳐 2006년 12월 14일 게재확정되었습니다.