색인어 선정을 위한 어휘결집력에 관한 연구
Study on the Vocabulary Synthesis for Index Term Selection

김 철(Chul Kim)[1] 정준민(Jun-Min Jeong)[2]

목 차

초 록

본 연구는 정보결집력을 응용한 자동 색인어 선정 기법에 관한 연구로 한 문장 내에 나타난 임의의 두 어휘가 그 문장을 표현하기 위한 의미 있는 집합이라는 가설 하에 어휘쌍 그래프를 통하여 색인어를 추출하였다. 특히, 그래프에 나타난 어휘 관계에서 각 어휘의 전체 어휘쌍 그래프에 대한 결집력을 분석하여 그 결집력을 색인어 선정의 우선 순위로 제안하였다.

가설을 검증하기 위하여 먼저 논문의 제목 및 초록에서 두 문장 이상에 동시 출현한 어휘쌍을 추출하였다. 다음으로 추출한 어휘쌍과 저자가 제시한 색인어 또는 주제명과 비교하였으며 그 결과 어휘쌍 그래프에 나타난 어휘가 대부분 색인어 또는 주제명에 사용되고 있음을 분석하였다. 그 중에서도 특히, 어휘쌍 그래프에서 어휘결집력이 높은 어휘일수록 그 논문의 내용을 전달하는 의미 있는 색인어로 채택될 가능성이 높음을 보여 주었다.

ABSTRACT

Under the hypothesis that any pair of terms in the sentence is meaningful to present the context of the paper, the Brillouin measure of term relatedness in automatic indexing is proposed. For the experiment, the pair of terms simultaneously appeared in two or more sentences of the paper are extracted from the title and abstract of the paper. Compared with the list of index terms or subject headings suggested by the author, the terms in term relatedness graph are highly matched with the terms in the list. Especially, it is revealed that the rank of terms by synthetic strength is useful in the selection of index terms.

* 광주교육대학교 전산교육과 조교수
** 전남대학교 문헌정보학과 교수
■ 논문 접수일 : 1996년 6월 17일

1. 서 론

1.1 연구의 목적 및 필요성

정보검색의 입장에서 색인은 개개의 정보 자료의 특성을 표현하는 데이터의 요소를 뽑아 각 정보 자료의 내용을 대표하도록 한 것이라고 할 수 있다. 이러한 색인은 지시와 선별의 기능을 갖는데 1차적으로는 특정한 정보를 필요로 하는 사람에게 그 정보의 위치를 지시해 주며 2차적으로 색인은 방대한 정보원으로부터 가장 유사한 내용의 정보 자료만을 선별하여 주는 선별 기능의 역할을 한다. 따라서 효과적인 정보검색은 적절한 색인의 사용을 전제로 하고 있으며 색인의 성능이 결국은 정보검색의 성능을 좌우하게 된다.

정보에 대한 끊임없는 연구와 불완전한 것에 대한 부단한 보완을 통해 색인 분야는 이용자의 정보 요구를 적절히 구현해 주기 위해 다양한 색인 기법의 개발과 그 미비점을 계속 고쳐 나가고 있다. 특히, 수작업 색인의 어려움을 극복하고자 자동 색인에 대한 연구가 활발히 이루어지고 있다.

수작업 색인은 용어의 추출이 색인자에 의해 이루어지는 색인이다. 이 색인은 색인자가 임의로 색인어를 부여하거나 통제 어휘집을 참고하여 색인어를 부여할 수도 있으며 정보 자료에서 추출한 용어 그대로를 색인어로 선택할 수도 있다. 수작업 색인은 색인자의 노력과 시간뿐만 아니라 색인 작성 비용도 많이 들게 되고 각 주제 영역에 있어 색인자는 고도의 전문적인 지식을 갖고 있어야 효과적인 색인어를 선택할 수 있으며 많은 인력을 필요로 하게 된다.

자동 색인은 색인 작성 시에 가장 많은 노력과 시간을 요하는 주제 분석 및 색인어의 선택을 컴퓨터가 대신하게 한 색인으로 최근 자동 색인 기법은 인간의 추론 기능과 주제 분석 능력을 갖춘 전문가 시스템이나 간단한 지식 베이스를 이용한 방법을 시도하고 있으나 자연 언어 처리 시 구문 분석의 의미론적 해석에는 어휘 개념 간의 다양한 관계를 제시하는 데 난점이 많아 경제적인 색인어 추출이 어렵다. 반면에 의미론적 해석과는 달리 문헌의 어휘 분산과 빈도에 의한 구조론적 해석은 경제적 이점과 아울러 분석의 용이함은 기대할 수 있으나 그 성능에 대해 많은 비판을 받아 온 것 또한 사실이다.

본 논문은 자동 색인어 선정 기법 중 구조론적 접근 방식에 어휘결집력을 응용하여 색인어 선정의 경제적 측면과 아울러 의미론적 접근의 문제를 해결하고자 하였다.

1.2 가설

문헌을 서지학적으로 분석하여 보면 내용과 형태로 나누어 볼 수 있으며 형태는 내용을 수록한 물리적 방법에서부터 저자, 서명, 수록 형태 등 내용과 관련지은 요소를 의미하며 내용 또한 저자에 의해 직접 만들어진 부분과 그것을 이해하거나 그것에 접근하기 용이하도록 가공되어진 부분으로 나누어 볼 수 있다. 초록이나 주제명, 분류 등이 여기에 속한다 할 수 있는데 이런 부분은 저자가 생산한 문헌의 내용과 무관하지 않다. 즉, 내용을 다른 식으로 표현한 것이라고도 볼 수 있는데 저자

와 이용자의 가교 역할을 할 뿐 아니라 필요에 따라서는 문헌의 대체 효과도 가질 수 있다.

문헌의 내용을 구조적으로 살펴보면 문헌의 내용을 전달하는 최소 단위가 어휘로 이루어져 있음을 알 수 있다. 경우에 따라 둘 이상의 어휘가 모여 한 뜻을 나타내는 복합어의 형태를 취하기도 하며 여러 가지 구두점을 이용하여 그 뜻을 보다 명확히 전달하고자 하고 있다. 그러나 가장 일반적인 형태는 마침표로 구분되어지는 하나의 문장과 하나 이상의 문장이 모여 만들어진 문단, 그리고 문단과 문단을 묶어 주는 다양한 계층의 단락을 볼 수 있다. 여기서 우리는 문헌의 내용을 전달하는 가장 최소 단위는 어휘이며 이들 어휘의 의미있는 최소 조합은 문장으로 구성되었다는 사실을 추론해 볼 수 있다.

문장을 사전적으로 보면 '하나의 내용을 갖춘 의미 있는 말'로서 이는 문장을 구성하는 어휘간에는 의미론적으로 상관관계가 있음을 뜻한다. 즉, 한 문장을 구성하는 어휘의 구조론적 분석을 통하여 문단이라는 의미론적 접근이 가능하게 된다. 이를 문헌 전체에 확장하여 보면 모든 의미 있는 문장의 단계별 의미론적 해석은 단위 문단의 어휘 조합의 구조론적 분석으로 가능하며 나아가 문장을 구성하는 어떤 어휘쌍도 전체적으로는 의미 있는 조합으로 볼 수 있다.

본 논문은 이와 같은 해석을 통하여 다음과 같은 가설을 수립하였다.

한 문단에 나타난 임의의 두 단어는 그 문단을 구성하기 위한 의미 있는 조합이며 이들 임의의 두 단어가 한 쌍으로 문헌 전체에 다른 단어쌍과 비교하여 상대적으로 높은 발생빈도를 보일 경우 그 두 단어는 그 문헌을 표현할 수 있는 의미있는 어휘쌍으로 볼 수 있다. 또, 이들 어휘쌍들이 나타내는 어휘쌍 그래프에서 각 어휘의 존재에 따른 그래프의 결집력을 분석하여 어휘결집력이 큰 어휘일수록 색인어로서의 가치가 크다.

2. 색인어 선정의 이론적 배경

2.1 자동 색인의 발달 과정

자동 색인 효시는 룬(Luhn)의 1957년과 1958년에 발표한 논문에서 찾아볼 수 있다(Luhn, 1957 & 1958). 그는 문헌에 출현한 단어들은 문헌의 내용 분석을 위해 사용될 수 있으며 단어의 출현 빈도가 문헌의 내용을 나타내는 주제어로서 중요성을 측정하는 기준이 된다고 하였다. 이와 같이 룬의 사상은 이후 개발된 다양한 자동 색인 기법의 기초가 되었다.

이러한 룬의 연구를 시초로 하여 발달된 자동 색인은 1958년 박센대일(Baxendale)이 발표한 연구 논문에서 자동 색인 발전의 두 가지 방향성을 제시하였다(Baxendale, 1958). 그가 제시한 방법을 보면 첫째는 기능어를 제외한 모든 단어를 색인어로 선택하거나 각 문단의 첫번째 문장과 마지막 문장에 출현한 단어를 색인어로 선택하는 방법으로 문헌의 구조적 특성을 이용하여 색인어를 선정하는 기법의 일종이다. 후에 통계적 방법이나 확률적

접근으로 발전하였으며 자동 색인 기법의 주류를 이루었다. 둘째는 문헌을 구성하는 전치사구로부터 색인어를 선택하는 방법으로 구문적 분석 기법의 일종으로 언어학적 기법 또는 의미론적 분석이라고 볼 수 있다.

자동 색인 연구에 있어 의미론적 분석의 잠재적 이점들이 인식되어 온 것은 사실이나 초기의 자동 색인 방법은 의미론적 접근보다 확률, 통계적 기법 등 구조론적 접근을 시도하였다. 그러나 1980년대 이후에는 의미론적 접근에 의한 자동 색인 연구가 많이 나타나게 되었다.

2.2 색인어 선정의 의미론적 접근

본 논문에서 말하는 의미론적이란 말은 구조론에 대한 상대적인 말로 어휘론 및 구문론을 포함한다. 의미론적 접근의 가장 단순한 어휘적 단계는 불용어 제거 기법을 대표로 들 수 있다. 다음으로 구문 분석 기법은 특정한 구문적 기능을 수행하는 단어나 단어 구가 문헌의 내용을 나타낸다는 가정 아래 이러한 구문 단위를 식별하는 작업을 의미한다. 이 방법은 구문 분석 수준에 따라 3단계로 구분할 수 있다. 간단한 방법으로는 구두점, 전치사, 접속사, 조사 등을 사용해서 문장을 분석하고 주제를 나타내는 단어 구를 식별하는 것이다. 두번째 단계는 어의적인 처리만 하지 않을 뿐, 구에서 절에 이르는 거의 완전한 문장 분석 단계이며 더욱 복잡하고 수준 높은 구문 분석은 컴퓨터에 내장된 문법과 어의적 사전을 이용해서 완전한 문장 분석을 행하는 것이다. 지금까지 구문론에 치우쳤던 의미론적 접근은 최근 어의 분석으로 영역을 넓히고 있는데 실제로는 이는 매우 복잡한 시스템을 필요로 하며 아직은 시작 단계이다.

2.3 색인어 선정의 구조론적 접근

구조론적 접근의 기본은 데이터의 통계적 처리라는 점이다. 통계적 처리란 단어의 출현 빈도를 근거로 하여 주제어로서의 중요도를 측정한 다음 색인어를 추출하는 방법을 말한다. 주제어로서 중요도를 측정하는 방법은 다양하며 통계적 처리의 기본적인 가설은 단어의 출현 빈도가 높을수록 그 단어가 문헌의 주제를 대표할 확률이 높다는 것이다.

1960년 마론(M.E. Maron)과 쿤(J.L. Kuhns)의 확률 색인 기법은 검색 문헌이 이용자를 충족시키는 확률에 따라 출력 문헌의 순위를 매기기 위해 색인 정보를 문헌 검색 시스템에 사용하도록 한 것이다 (Maron & Kuhns, 1960). 1965년에 다메로우(F.J. Damerau)는 단어 빈도를 사용할 경우 색인어를 선정하는 절대적 기준을 정하기가 어렵다는 데 착안하여 상대 빈도를 측정하여 색인어를 선정하는 방법을 제시하기도 하였다 (Damerau, 1965).

1970년대에 들어서면서 장서나 특정 문헌 내에서의 용어의 가치를 측정하여 단순히 색인의 목적에만 적용시키던 이론을 검색의 목적과 함께 고려하게 되었다. 다시 말해 문헌을 구분하는데 유용한 용어를 색인어로 선정함으로써 검색시 적합 항목을 구분할 수 있는 새로운 이론과 기법이 개발되었는데 그 대표적인 것이 스파크 존스(K. Sparck Jones)의

역문헌 빈도론 (Sparck Jones, 1972)과 샐튼 (G. Salton)의 문헌분리가 이론 (Salton, 1983)이다.

1975년 하터(S.P.Harter)는 동일 문헌에 함께 군집하여 출현하는 용어가 색인어로 유용하다는 가설 하에 포아송(Poisson)모델을 적용하여 용어의 출현 빈도에 대해 확률론적으로 모델화한 북스타인과 스완슨의 이론 (Bookstein & Swanson, 1975)을 발전시켜 2 - 포아송 분포 모델을 제시했다 (Harter, 1975).

3. 정보량 측정과 정보결집력

3.1 정보량 측정

정보란 발생 가능한 모든 경우로부터 하나의 경우를 선택할 때 부여되는 선택의 자유이며, 정보량은 부여된 선택의 자유를 정량화한 것으로 모든 경우의 집합으로부터 각각의 경우가 발생될 확률로 결정된다(Shannon & Weaver, 1959). 다시 말하면 어떤 특정한 경우의 정보량은 그 경우가 선택되어질 확률이 낮으면 낮을수록 많고 그 경우가 선택되어질 가능성이 높을수록 정보량은 작아지게 된다. 즉, 모든 경우의 발생확률이 동등한 경우, 각각의 경우의 정보량은 동일하며 이때 불확실성이 높다고 볼 수 있다.

$$M = \{M_1, M_2, M_3, \ldots M_i, \ldots\}$$

$$I(M_i) = \ln (1/p(M_i))$$

여기서 M은 경우의 집합을 나타내며 M_i는 각각의 경우이며 $p(M_i)$는 M_i가 발생할 확률을 의미한다. 결국 M_i에 대한 정보량 $I(M_i)$은 M_i이 일어날 확률과 반비례하게 된다.

섀논은 모든 경우에 대한 이론적 최대 정보량(H_m)을 구하고 실질 정보량 (H)과의 상대정보량(H/H_m)을 계산하여 이론상대치($H_m/H_m = 1$)에 대한 차이를 중복도(redundancy)라 하여 실제 상황에서 우리가 정보에 대한 결정력과 자유도를 계산하였다. 여기서 이론적 최대 정보량은 모든 경우의 발생확률이 동등하고 독립적일 때 일어나며 그 값은

$$H_m = \sum 1/N \cdot \ln N = \ln N$$

이다.

즉, 섀논이 설명한 상대 정보량이나 중복도의 계산은 정보에 대한 예측성에 대한 정량화이며 정보(그것이 어떤 형태로 표현되든 그것이 담고 있는 의미)에 대한 계량적 해석을 가능케 하는 척도이다.

정보란 섀논이 제시한 전달이나 표현의 개념에서 뿐 아니라 데이터의 분석, 즉 의미론과 구조적 해석에도 활용된다. 구조화란 데이터의 여러 요소로의 분해 및 의미 분석의 경로와 의미 선택을 최적화 시켜주는 행위이다 (Hayes, 1967).

주어진 데이터가 아무런 구조를 갖지 않을 경우, 이에 대한 해석은 많은 시간과 노력을 필요로 한다. 그러나 이 데이터가 구조적 틀을 갖으면 몇 개의 부분으로 요소화 시키며 이를 도식화하여 쉽게 의미 파악을 할 수 있다. 즉, 우리가 늘 접하는 문헌의 내용 전달

행위도 서명, 저자명 그리고 본문의 단어의 의미 있는 배열과 문장, 그리고 문단 등 그 구조적 틀 속에서 그 의미를 쉽게 파악하게 되는 것이다.

3.2 정보결집력

어떤 의미를 기호화하고 기호화된 정보의 의미론적 관계를 구조적으로 도식화할 경우, 우리는 의미구조 그래프(graph)를 생성할 수 있다. 이때 주어진 의미구조의 정점(node)과 그 정점을 잇는 의미관계(edge)에서 그 그래프를 해석할 수도 있을 뿐 아니라 그 그래프가 표현하고자 하는 의미를 분석하거나 또는 새로운 의미를 부여할 수 있다.

그 중에 하나가 정점과 그 정점을 잇는 의미관계에서 가장 핵심적인 정점 또는 그 의미관계를 찾아볼 수 있는데 이는 본래의 그래프에서 그 정점을 제거했을 경우 변화된 그래프와의 상대적 비교를 통하여 각 정점 및 의미관계의 중요성 정도를 측정할 수 있다. 즉, 그래프의 각 정점이 갖는 정보량의 변화를 통하여 그 정점의 그래프에 대한 기여도(결집력 ; synthetic point)을 측정할 수 있다 (Shaw, 1979).

쇼우 (Shaw, William M., jr.)가 제시한 결집력에 대한 분석은 그 원리를 섀논의 정보이론을 일반화한 브릴루인(Brillouin, L)의 정보측정 공식에서 찾아볼 수 있다. 섀논과 브릴루인은 정보 커뮤니케이션에 있어 각 정보가 갖는 정보량을 측정하고자 하였으며 쇼우는 이를 공저자 그래프를 통하여 각각의 저자가 전체 집단에 미치는 공헌도를 정보량으로 측정함으로써 그 정보량을 결집력(synthetic point)으로 평가하였다.

〈Brillouin의 정보량 측정 공식〉

$$I_k = K \cdot \ln \frac{N!}{n_1!\ n_2! \ldots n_s!}$$

I_k = 정점 k를 제거했을 당시의 정점 k의 정보량

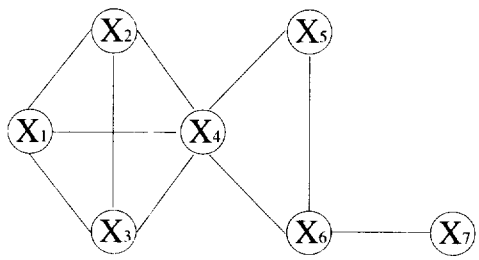
$K = 1 / \ln 2$

n_i = 어느 한 정점을 제거한 후 나뉘어진 sub-graph I의 정점 수

N = 원래 정점의 수 - 1

그래프에 변화를 주기 전의 정보량은 N을 전체 정점의 수로 계산을 하면 된다. 만일 하나의 연결 그래프일 경우의 정보량은 0이 됨을 알 수 있다. 이럴 경우 정점을 제거한 상태에서 그 정점에 대한 상대 정보량은 음의 값을 취하게 되며 그 절대값이 크면 클수록 그 정점의 그래프에 대한 결집력이 높다고 말할 수 있다.

한 예로 다음 그래프의 각각의 정점을 제거하여 그 정점이 갖는 정보량을 측정한다.



〈그림 3-1〉 임의의 그래프

〈그림 3-1〉에서 X_4에 대한 정보량을 계산해 보면

$$I_{x4} = K \cdot \ln \frac{N!}{n_1!\ n_2!}$$

$$= \frac{1}{\ln 2} \cdot \ln \frac{6!}{3! \cdot 3!}$$

$$= 4.3219$$

X_6에 대한 정보량

$$I_{x6} = K \cdot \ln \frac{N!}{n_1!\ n_2!}$$

$$= \frac{1}{\ln 2} \cdot \ln \frac{6!}{5!\ 1!}$$

$$= 2.5850$$

이들 각 정점의 정보량의 원래 그래프에 대한 상대 정보량은 원래 그래프가 하나로 연결되어 있으므로 그 정보량은 0이 되고 X_4의 상대 정보량은 -4.3219, X_6은 -2.5850이 된다.

X_4가 X_6보다 결집력이 높음을 알 수 있으며 만일 이 그래프가 정당의 인맥관계를 나타낸 그래프라면 정점 X_4가 그 정당에서 가장 인맥이 풍성할 뿐 아니라 그로부터 전체 정당의 인맥관계가 유지된다고 표현할 수 있다. 또 이 그래프의 정점을 저자(author)로, 의미관계를 공저자 활동으로 본다면 X_4가 그 집단의 여러 저자에 대한 사상적 연관성을 유지하고 전체를 하나의 주제군으로 형성시키는 데 크게 기여하는 저자라고 말할 수 있겠다.

위의 그래프 모형만을 가지고도 많은 적용과 해석이 가능한데 주제의 이합·집산의 개념, 핵심 참고문헌의 선정 및 인용과정의 전환점 등 그 응용 예는 이루 말할 수 없다. 본 논문은 부릴루인의 결집력을 한 문헌의 초록 및 서명에 나타난 의미를 문단과 문장의 구조적 개념에서 해석하여 정점을 어휘로 그 의미관계는 문장을 구성하기 위해 동시 출현한 임의의 두 어휘의 어휘쌍으로 보고 의미관계 그래프를 그려 그 결집력을 측정하여 보았다.

4. 어휘결집력 분석을 통한 색인어 선정 실험

4.1 데이터

본 가설을 검증하기 위하여 실험 데이터로는 'Fuzzy'와 '검색'이라는 두 개의 주제를 만족시키는 문헌 중 임의로 발췌한 5문헌을 사용하였다. 그 서지사항은 〈표 4-1〉과 같다.

본 실험을 위하여 선정된 5개 문헌의 서명과 초록만을 분석 대상으로 삼았다. 실험 데이터의 특성을 살펴보면 모든 문헌이 초록은 1개 문단에 약 4~9개의 문장으로 구성되어 있으며 불용어(관사, 전치사, 관계사 등 가장 기초적인 단어만을 불용어로 처리하였음)와 중복된 단어를 제외하면 사용한 단어의 수도 29~44개로 큰 차이를 보이지 않았다. 단지 저자가 선정한 키워드를 보면 거의 중복 없이

〈표 4-1〉 실험에 사용한 문헌리스트

	Title	Source
Set 1	General Asymmetric Neural Networks and Structure Design by Genetic Algorithms	*Neural Networks* 5(1992) : 327-334
Set 2	Knowledge engineering for a document retrieval system	*Fuzzy Sets and Systems* 38(1990) : 223-240
Set 3	A document retrieval system based on citation using fuzzy graphs	*Fuzzy Sets and Systems* 38(1990) : 207-222
Set 4	Information retrieval based on fuzzy associations	*Fuzzy Sets and Systems* 38(1990) : 191-205
Set 5	A fuzzy document retrieval method based on two-valued indexing	*Fuzzy Sets and Systems* 30(1989) : 103-120

다른 키워드를 사용하고 있음을 알 수 있으며 (부록 참조) 키워드를 유사어와 같은 개념으로 묶을 경우 16개로 묶을 수 있으며 〈표 4-2〉. 처음 검색에 사용하였던 'fuzzy' 와 'retrieval' 두 단어가 전체의 연결고리 역할을 하고 있음을 알 수 있었다.

〈표 4-2〉 수정 보완한 키워드 분포표

	Set 1	Set 2	Set 3	Set 4	Set 5
algorithms				X	
concept identification		X			
document retrieval					X
document relevance values		X			
fuzzy			X	X	X
genetic algorithm	X	X			
graph theory			X		
information retrieval		X	X	X	
knowledge engineering		X			
modal logic					X
neural network	X				
personal construct theory		X			
possible knowledge semantics					X
relations			X		
structure design	X				
thesaurus generation		X			X

특히, Set 1의 경우는 초록의 단어를 통하여 검색이 된 경우로 저자가 제시한 키워드만으로 검색을 한다는 것은 재현율에 문제가 될 수 있음을 보여주는 예라 할 수 있고 저자의 키워드 추출이 객관성이 결여되었음을 보여주는 반증이라 할 수 있다.

4.2 실험

4.2.1 실험과정

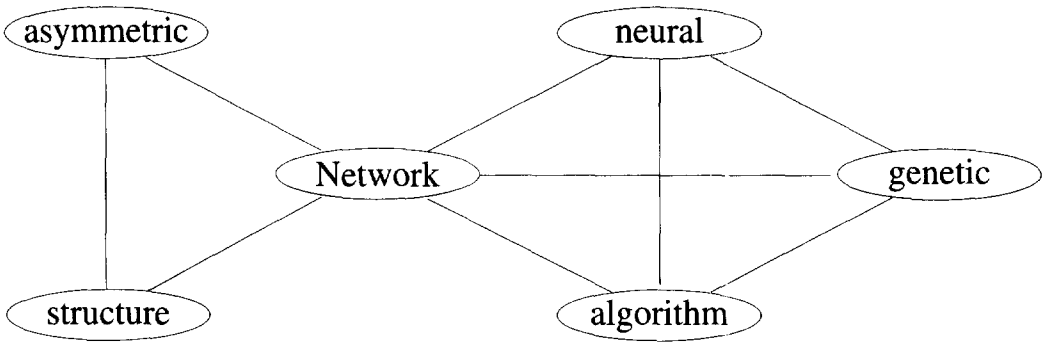
5개의 문헌에서 서명, 초록을 추출하여 문장별로 불용어(관사, 전치사, 관계사 등 가장 기초적인 단어만을 불용어로 처리하였음)를 제외한 모든 어휘(실험의 용이성을 위하여 복합어를 제외하였으며 단어의 변화형을 의도적으로 원형으로 통일하였다.)만으로 어휘쌍을 구성하였다. 서명도 하나의 문장으로 간주하여 둘 이상의 문장에 동시 출현한 어휘쌍민을 의미관계로 규정하였는데 그 이유는 한 문장에 나타난 어휘들의 어휘쌍은 완전 그래프를 형성하기 때문에 분석의 의미가 없고 최소한 관계가 의미를 갖기 위해서는 그 어휘쌍이 다른 문장에서도 출현하여야 다른 어휘쌍과의 차별화를 가질 수 있다고 보았다.

이미 추출된 의미 있는 어휘쌍을 중심으로 그래프를 그렸으며 그려진 결과를 토대로 그 그래프가 갖는 원래의 정보량을 측정하였으며 각 정점에 대한 정보량을 계산하여 상대 정보량 즉, 각각의 정점에 대한 결집력을 계산하였다.

각 정점에 대한 결집력을 기준으로 각 문헌의 색인어로 선정 가능한 어휘의 우선 순위를 매겼으며 그 순위에 따라 분석을 시도하였다.

4.2.2 데이터 처리

실험을 통해 얻어진 결과를 도식화하면 다음 그림과 식으로 표현할 수 있다. 그래프는 각 문헌에서 두 번 이상 나타난 어휘쌍에 대한 의미관계이며, I는 각 그래프의 고유 정보량을 뜻하며 I_i는 그래프 상에서 각 정점의 정보량을 그리고 S_i는 상대 정보량 즉, 결집성을 나타내며 결집성의 정도는 적으면 적을수록 결집력이 높음을 의미한다.



〈그림 4-1〉 Set 1의 어휘쌍 그래프

$I = 0 \qquad S_i = - |I_i|$

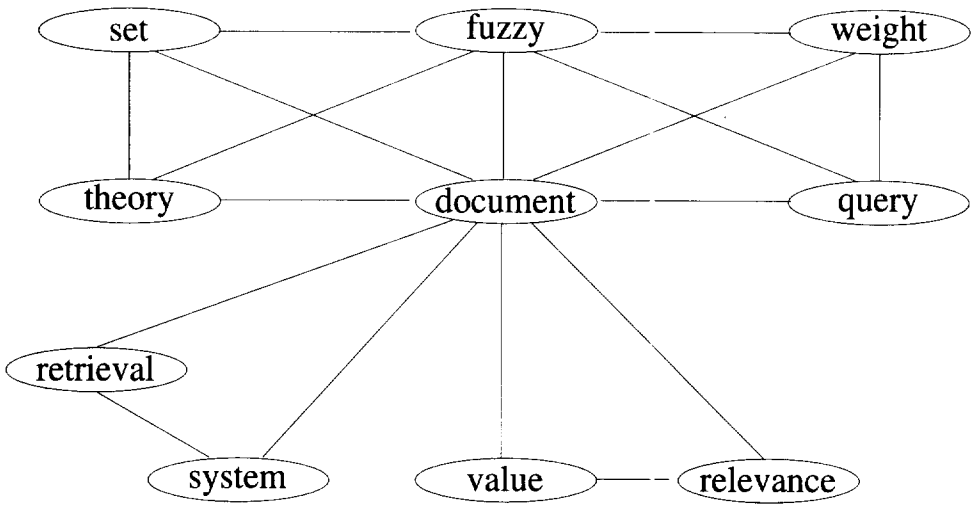
$I_i =$

$I_{network} = 1/l_n2 * \ln(5\,!\ /\ (3\,! * 2\,!)) = 1.4427 * 2.3026$

$= 3.3220$

$I_{asymmetric} = I_{structure} = I_{neural} = I_{genetic} = I_{algoritm}$

$= 1/l_n2 * l_n(5\,!\ /\ 5\,!\ ) = 0$



〈그림 4-2〉 Set 2의 어휘쌍

$I = 0 \qquad S_i = - |I_i|$

$I_i =$

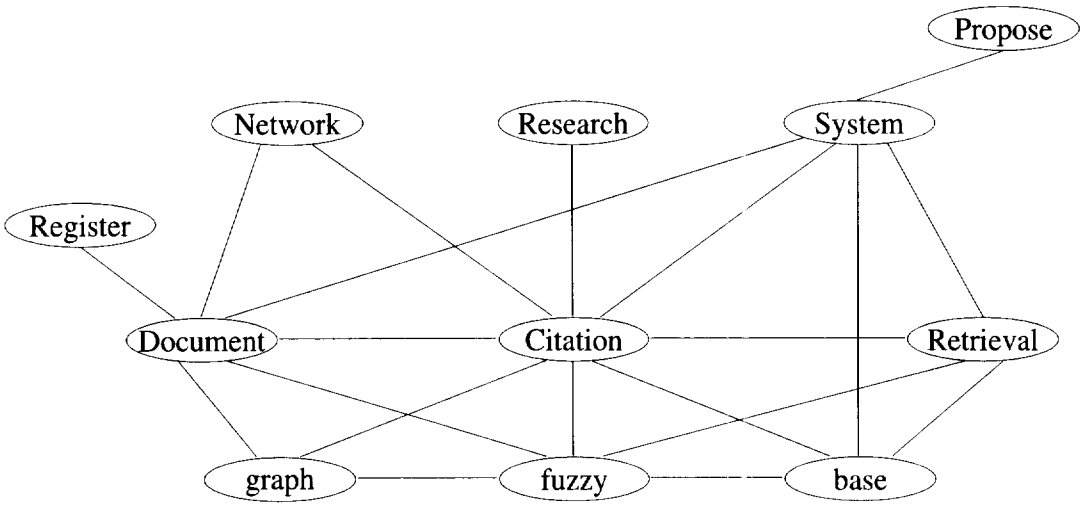
$I_{document} = 1/l_n2 * l_n(9\ !\ /\ (5\ !\ *\ 2\ !\ *\ 2\ !\ )) = 1.4427 * 6.6280$

$= 9.5622$

$I_{fuzzy} = I_{set} = I_{theory} = I_{weight} = I_{query}$

$= I_{retieval} = I_{system} = I_{value} = I_{relevance}$

$= 1/l_n2 * l_n(9\ !\ /\ 9\ !\ ) = 0$



〈그림 4-3〉 Set 3의 어휘쌍 그래프

$I = 0 \qquad S_i = -|I_i|$

$I_i =$

$I_{document} = I_{citation} = I_{system}$

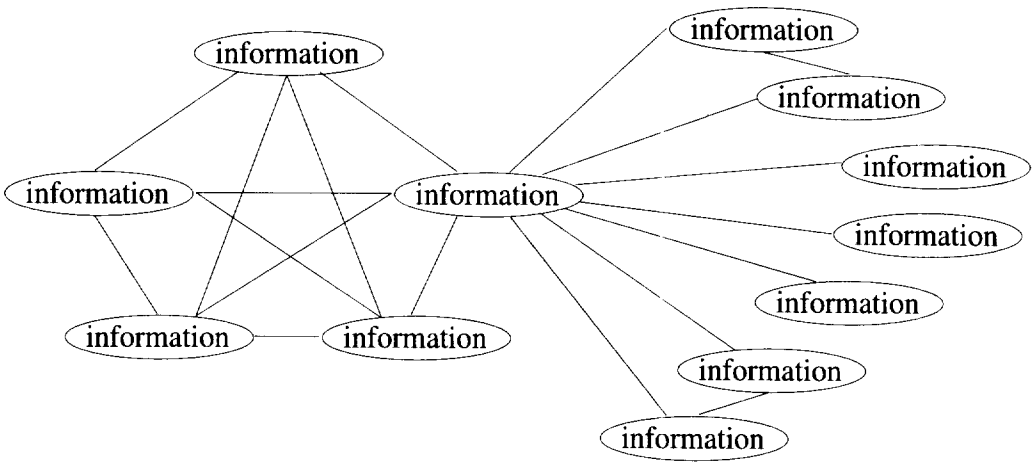
$= 1/\ln 2 * \ln(10\,!\,/\,(9\,!\,*\,1\,!\,)) = 1.4427 * 2.3026$

$= 3.3220$

$I_{register} = I_{graph} = I_{fuzzy} = I_{base} = I_{etrieval}$

$= I_{research} = I_{propose} = I_{network}$

$= 1/\ln 2 * \ln(10\,!\,*\,10\,!\,) = 0$



〈그림 4-4〉 Set 4의 어휘쌍 그래프

$I = 0 \quad S_i = - |I_i|$

$I_i =$

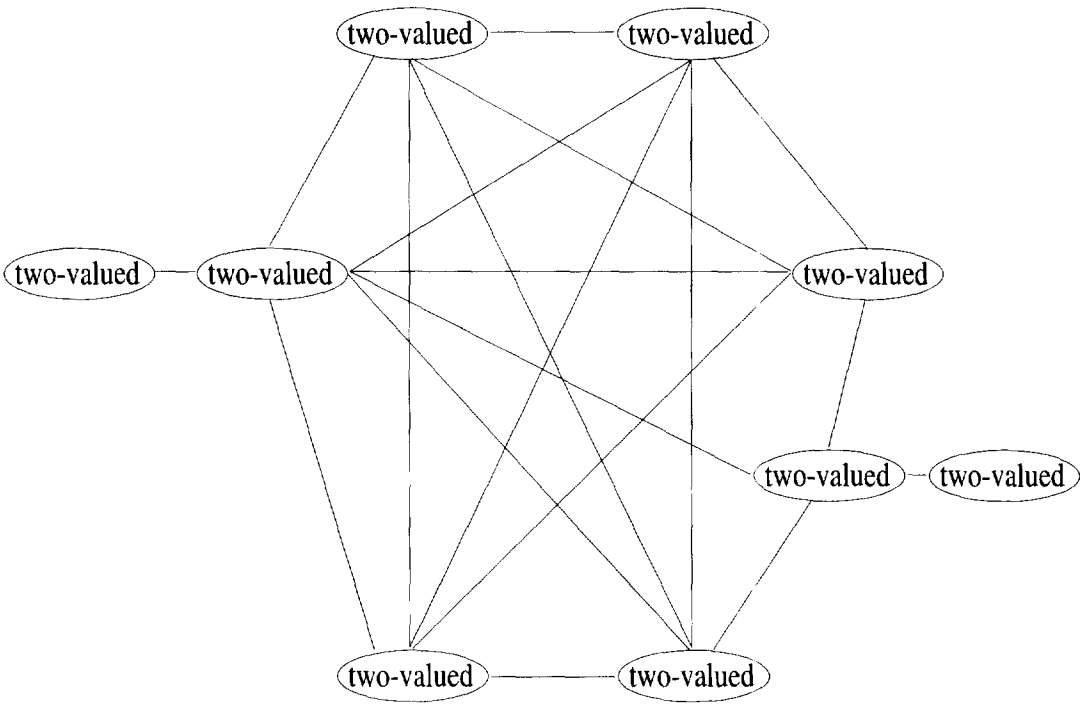
$I_{fuzzy} = 1/l_n2 * l_n(10\ !\ /\ (4\ !\ *\ 2\ !\ *\ 2\ !\ *\ 1\ !\ *\ 1\ !\ )) = 1.4227 * 10.5401$

$= 15.2062$

$I_{method} = I_{set} = I_{model} = I_{present} = I_{database}$

$= I_{bibliograph} = I_{association} = I_{retrieval} = I_{information}$

$= 1/l_n2 * l_n(10\ !\ /\ 10\ !\ ) = 0$



〈그림 4-5〉 Set 5의 어휘쌍 그래프

$I = 0 \qquad S_i = - |I_i|$

$I_i =$

$I_{retrieval} = I_{weight}$

$= 1/l_n2 * l_n(8\,!\ /(7\,!\ * 1\,!\,)) = 1.4227 * 2.0794$

$= 2.99995$

$I_{document} = I_{method} = I_{two\text{-}valued} = I_{indexing}$

$= I_{system} = I_{base} = I_{fuzzy}$

$= 1/l_n2 * l_n(8\,!\ / 8\,!\,) = 0$

4.3 결과

실험결과 특징만 우선 살펴보면 대부분의 그래프가 여러 개의 클리크(clique : 최소한의 완전연결그래프를 뜻함)를 갖고 있어 대부분의 정점이 유사한 값을 가졌으며 서명의 문장을 그대로 초록에 사용한 경우, 그래프 전체가 완전히 연결되는 현상을 보였다. 그러나 매 그래프마다 전체 그래프의 결집력에 중대한 역할을 하는 정점이 있었으며 이차적으로 비록 결집력의 값은 큰 차이를 보이지 않았으나 그 정점의 제거로 많은 의미관계(edge)가 떨어져 나가는 현상을 보였다.

특히 결집력이 높은 정점을 중심으로 만들어지는 클리크를 보았을 때 저자가 제시한 키워드와 비교하여 의미전달의 최소 집합으로서의 역할을 수행하고 있음을 보였다. 그러나 저자의 키워드는 경우에 따라 유추하여 제공된 경우도 있으며 아직 자연어 검색과 통제어 검색에 대한 확실한 검색평가가 이루어지지 않은 속에서 저자의 키워드가 옳다고 단정지을 수는 없지만 결집력과 동시발생 어휘쌍의 의미관계에서 저자의 의도를 충분히 객관화시킨 결과라고 보여진다.

4.3.1 결집력

Set 1의 경우, 정보량 측정에 있어 가장 이상적인 형태를 보이고 있다. network라는 어휘가 갖는 정보량의 변화량 즉, 결집력은 genetic algorithm과 asymmetric structure를 연결지을 수 있는 핵심어임을 나타내고 있다. 특히 두 개의 클리크를 연결지음으로써 문헌의 주제가 어떤 것인지를 쉽게 알 수 있다.

Set 2의 경우는 document라는 어휘가 중심적 역할을 수행하며 4개의 클리크를 형성하는 중앙에 위치하고 있다. 단지 그 단어를 제외하고는 모두가 같은 값을 갖는 무의미함도 보이지만 클리크의 특징을 볼 때 document에 대한 네 가지의 개념접근이 이루어지고 있음을 알 수 있었다.

Set 3은 구조가 다소 복잡한 양상을 보이고 있다. 결집력의 차이가 별로 없는 속에서 모든 클리크의 핵심적 역할을 하는 어휘가 제거되었을 경우 더 많은 의미관계를 상실함으로 비록 결집성의 차이는 보이지 않으나 주변적 해석으로 충분히 핵심 어휘임을 파악할 수 있다. 실제 내용에 있어서도 document의 citation 기반의 fuzzy graph로 citation의 중요성을 나타내고 있다.

Set 4의 분석은 하나의 커다란 클리크와 두 개의 작은 클리크 그리고 주변 어휘로 나타나는데 큰 클리크는 서명이 전부 초록에 다시 쓰임으로써 나타난 현상이다. 결과적으로 완전연결 그래프에서 fuzzy의 역할은 주변적 개념의 연결고리로서 충분히 의미를 갖는다.

마지막으로 Set 5는 클리크가 이중으로 겹쳐져 나타나면서 retrieval과 weight 두 어휘에 다소 높은 결집력을 보이고 있다. 그런데 서명과 초록의 내용을 분석해 보면 두 가지 값을 갖는 색인에 근거한 검색시스템에 관한 문헌으로 weight가 주어진 retrieval이란 개념이 도출될 수 있다. 비록 근소한 차이이고 전체 그래프가 많이 얽힌 양상을 보이면서도 결집성은 나름대로 의미 있는 값으로 표출되었다.

4.3.2 의미관계의 변화

결집력이 높은 어휘이든 그렇지 않든 그 어휘의 존재유무가 주변어휘와의 의미관계에 많은 영향을 미칠 때 그 단어는 꽤 의미 있는 어휘라 볼 수 있다. 특히, 소수의 어휘로 구성된 문장을 분석할 때에는 전체가 클리크에 가까운 형태들의 복합적 관계를 나타내기 쉬우므로 결집력과 아울러 의미관계의 변화에도 많은 관심을 가질 필요가 있겠다.

Set 2의 경우, 결집력이 상당히 높게 나타난 document의 경우 그 어휘의 부재로 인한 의미관계의 단절은 전 어휘집단에 미칠 정도로 지대하다. 물론 이 현상이 결집력으로 표현이 된다면 좋으나 Set 5의 weight처럼 결집력은 다른 어휘에 비해 상대적으로 높으나 자신의 부재가 주는 의미관계는 단지 4개의 관계만을 단절시킬 뿐 결집력이 약한 대부분의 어휘가 5개의 의미관계 단절을 보이는 것을 볼 때 결집력과 꼭 일치하지는 않는다. 그리고 결집력이 주는 의미, 즉 전체의 구조에 영향을 미친다는 가정에 비추어 볼 때 결집력과는 다른 어떤 의미 부여가 있어야 하지 않을까 한다.

4.3.3 주변어휘와의 관계

분석과정에 보이듯이 결집력이 높은 어휘를 중심으로 여러 개의 클리크가 형성되는 현상을 관찰할 수 있었다. 클리크는 그 의미가 갖듯이 연결 그래프에서 최대 완전연결 그래프를 형성하는 집합을 의미하는데 그래프에 있어 완전연결이라 함은 모든 정점이 똑같은 조건하에 모두 동등한 관계를 갖는다는 의미이며 관계에 있어 최적의 조화라 할 수 있다. 실험 데이터를 통해서도 알 수 있듯이 그래프 내에서 형성된 클리크는 두 문장 이상에 동시 발생한 경우로 그만큼 특정한 의미를 전달하는데 필수적인 어휘임을 알 수 있다. 아마 문장의 수가 많아지고 동시출현 빈도의 경계값이 높아지면(그 값의 설정은 그래프이론 중에서 hypergraph theory에 의해 규명지을 수 있으리라 본다.) 클리크의 단위도 줄어들 것이면 그 과정에서도 큰 덩어리의 클리크가 형성되면 그 집단은 특정 의미 전달에 굉장한 상관관계가 있으리라 예측할 수 있겠다.

그 경우 결집력이 높은 어휘를 중심으로 주변의 클리크에 대한 상대평가 등을 통해(일례로 fuzzy 기법을 이용한 간접 관계치를 계산하는 방법 등) 색인어의 확장을 꾀할 수 있으리라 본다.

4.4 결과의 활용 및 응용

본 실험에서 보여주고자 한 것은 어휘쌍 그래프의 의미결집력을 통한 색인어 선정의 가능성이다. 비록 서명과 초록만을 대상으로 시도하였으나 앞으로 원문 데이터베이스가 구축이 되고 원문에 대한 문자적 분석 및 접근이 가능하면 모든 원문을 대상으로 본 실험을 확대할 수 있으리라 본다. 즉, 한 문장에 나타난 임의의 어휘쌍이 문헌 전체에 분포된 정도에 따라 edge에 값을 부여하고 그 값의 크기에 따라(hypergraph theory 등을 원용하여) 으미밀집도를 계산하고 특정값을 기준으로 그래프를 형성할 수 있다. 물론 각 node는 단위어

휘(single term)일 수도 있으며 복합어(compound term)일 수도 있다.

일단 구축된 그래프를 통하여 결집력이 높은 어휘순으로 데이터를 정열하고 그 어휘를 기점으로 한 모든 어휘의 의미상대값을 추론하여(어휘쌍을 구축할 당시 가지고 있던 의미밀집도를 응용하여) 결집력이 높은 어휘와 이들 어휘와 어휘밀집도가 높은 확장된 어휘군을 선별할 수 있으며 이들 어휘를 그 문헌의 색인어 집단으로 응용할 수 있으리라 본다.

5. 결 론

5개의 임의 추출된 문헌집단을 통하여 서명과 초록에서 발췌한 어휘쌍을 통하여 자동색인의 구조론적 접근을 시도하여 보았다. 모든 의미는 그 의미를 전달하기 위한 구조를 가지며 그 의미구조의 정보량을 측정함으로써 의미전달의 최소한의 요소를 식별할 수 있다는 가설과 함께 실험결과, 의미 있는 결과를 도출하였다.

실험에 사용한 데이터가 비록 임의로 추출하였다고는 하나 전체에 대한 대표성을 갖지는 못한다. 본 실험의 의미는 주관적 행위의 객관화 작업이며 의미론적 표현을 구조적 틀 위에서 해결하는 데 있다. 본 실험은 현실을 대상으로 작업한 것이 아니라 실험을 위하여 상당히 가공되어지고 한정된 환경하에 이루어졌다. 본 실험의 의도는 구조적 방법이 저자나 또는 색인 전문가에 의해 작성된 색인을 경제적으로 대처할 수 있음을 보여준 것이며 문헌적으로 고찰해 봐도 저자에 의해 만들어진 서명, 초록, 키워드 그리고 나아가 본문을 구조적으로 객관화시킨다는 것은 통계적 기법 외에 별로 시행된 바가 없다. 그러나 통계가 갖는 문헌 집단에 대한 지속적인 재평가라든가 의미적 부분을 단순한 빈도 등의 숫자로만 표현해야 하는 논리전개의 비합리성 등을 최소한 정보량의 산출근거인 최소 의미단위와 의미의 구조화 등으로 이해하려 한 것 등을 본 논문의 성과로 볼 수 있겠다.

실험 결과에서도 보여주었듯이 최소 의미집단에 동시 출현한 어휘쌍의 결집력에 의한 색인어 선정은 앞으로 자연어 색인 시스템에서 통제어 색인 시스템으로 건너가는 계기가 되리라 본다. 본 실험에서는 제외시켰으나 복합어를 사전적으로 정리하여 불용어를 제외한 어휘분석이 아니라 사전적 방법에 의해 검출되어진 어휘만을 대상으로 색인어 평가 및 선정을 시도할 필요가 있다고 본다. 즉, 그동안 컴퓨터의 발전은 우리에게 자연어 검색에 많은 점수를 주어 왔으나 다양한 어휘의 사용은 어휘통제의 필요성을 제기하였으며 어휘에 의한 직접 접근보다는 통제된 어휘집단 또는 클리크와 같은 어휘집단으로 구축된 시소러스의 개발 등으로 일원화된 색인, 검색시스템의 개발이 있었으면 한다.

특히, 의미의 구조화에 사용된 정보량과 그를 결집력에 응용한 계량정보학적 분석을 다시 색인시스템에 활용하여 봄으로써 수년간 자동색인에 있어 통계적 방법에만 의존하던 구조론적 접근을 의미론적 또는 구문론적 해석과 접목시킴으로써 자연어 검색, 통제어 검색 등 그 성능 평가에 있어 많은 변수로 작용할 수 있으리라 본다. 또, 나아가 초록만이 아

넌 본문 전체를 대상으로 함으로써 저자의 의도를 최대한 반영시킨 객관적 시도가 있었으면 한다.

참 고 문 헌

김현희, 김용호. 1993. 계량정보학. 서울 : 구미무역.

김현희 1984. "An Investigation of Automatic Term Weighting Techniques". 정보관리학회지. 19.1 : 12-23

Baxendale, P. B. 1958. "Machine-Made Index for Technical Literature - An Experiment". IBM JRD 2 : 354-361.

Bert, Boyce and Martin, David. 1981. "The Brillouin Measure of an Author's Contribution to a Literature in Psychology". JASIS 32 : 73-76.

Bookstein, A. and Swanson, D. R. 1975. "A Decision Theoretic Foundation for Indexing". JASIS 26 : 45-50.

Boyce, B.R., Martin, D. 1981. "The Brillouin Measure of An Author's Contribution to A Literature in Psychology". JASIS 32 : 73-76.

Brillouin, L. Science and Information Theory. New York : Academic Press.

Damerau, F. J. 1965. "An Experiment in Automatic Indexing". American Documentation 6 : 283-289.

Doyle, L.B. 1962. "Indexing and abstracting by association". American Documentation 13 : 378-390.

Harter, S.P. 1975. "A Probabilistic Ap proach to Automatic Keyword Indexing : Part I. On the Distribution of Specialty Words in a Technical Literature". JASIS 26 : 197-206.

Harter, S.P. 1975. "A Probabilistic Approach to Automatic Keyword Indexing : Part II. An Algorithm for Probabilistic Indexing". JASIS 26 : 280-289.

Hayes, R.M. 1993. "Measurement of information". Information Processing and Management 29 : 1-11.

Hayes, R.M. 1965. "The measurement of information from a file.". In : M.E. Stevens et al. Statistical association methods for mechanized documentation, Proceedings of the Symposium, Washington, 1964 (pp. 161-162)

Luhn, H. P. 1958. "The Automatic Creation of Literature Abstracts". IBM JRD 2 : 159-165.

Luhn, H. P. 1957. "A Statistical Approach to Mechanized Encoding and Searching of Library Informa-tion". IBM JRD 4 : 309-317.

Maron, M. E. and Kuhns, J. L. 1960. "On Relevance, Probabilistic Indexing and Information Retrieval" Journal of ACM 7 : 216-244.

Miyamoto, S. and Nekayama, K. 1983. "A Technique of Two-stage clustering Applied to enviromental and Civil Engineering and Related Methods of Citation Analysis". JASIS 34 : 192-201.

Pao, M.L. 1982 "Collaboration in Computational Musicology". JASIS, 33 : 38-43.

Radecki, T. 1979. "Fuzzy set theoretical approach to document retrival". Information Processing and Management 15 : 247-259

Ramon Lopex de Mantaras, Ulises Cortes, Jaume Manerp, Enric Plaza 1990. "Knowledge engineering for a document retrieval systems". Fuzzy Sets and Systems 38 : 223-240.

Salton, G. 1968. Automatic Information Organization and Retrieval. New York : McGraw-Hill.

Salton, G. 1981. "A blueprint for automatic indexing". ACM SIGIR 16 : 22-38.

Salton, G. and McGill, M. J. 1983. Introduction to Modern Information Retrieval.. New York : McGraw-Hill.

Sparck Jones, K. 1972. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". Journal of Documentation 28 : 11-20.

Sparck Jones, K. 1971. Automatic Keyword Classification for Information Retrieval. London : Butterworth.

부 록 : 실험에 사용한 문헌 리스트

Set 1	
Title :	General Asymmetric Neural Networks and Structure Design by Genetic Algorithms
Author(s) :	Stefan Bornholdt ; Dirk Graudenz
Source :	*Neural Networks* 5 (1992) : 327-334
Abstracts :	A learning algorithm for neural networks based on genetic algorithm is proposed. The concept leads in a natural way to a model for the explanation of inherited behavior. Explicitly we study a simplified model for a brain with sensory and motor neurons. We use a general asymmetric network whose structure is solely determined by an evolutionary process. This system is simulated numerically. It turns out that the network obtained by the algorithm reaches a stable state after a small number of sweeps. Some results illustrating the learning capabilities are presented.
Keywords :	neural network ; genetic algorithm ; structure design

Set 2	
Title :	Knowledge engineering for a document retrieval system
Author(s) :	Ram n L pez de M ntaras ; Ulises Cort s ; Jaume Manero ; Enric Plaza
Source :	*Fuzzy Sets and Systems* 38 (1990) : 223-240
Abstracts :	A document retrieval system based on fuzzy set theory is described in this paper. Weights, as fuzzy characteristic functions, can be assigned to descriptors in the query expression and to index terms in the document description. Fuzzy set theory allows one to calculate a relevance value for each document from weights assigned to documents and queries. The relevance value for each document is calculated following different models. An elicitation mechanism is used to generate and to enhance the thesaurus structure. The thesaurus guides the retrieval operations.
Keywords :	information retrieval ; fuzzy sets theory ; knowledge engineering ; thesaurus generation ; document relevance values ; personal consuruct theory ; concept identification

Set 3	
Title :	A document retrieval system based on citation using fuzzy graphs
Author(s) :	K. Nomoto ; S. Wakayama ; T. Kirimoto ; Y. Ohashi ; M. Kondo
Source :	*Fuzzy Sets and Systems* 38 (1990) : 207-222
Abstracts :	A fuzzy retrieval system based on citation is proposed. A citation implies a trend of research or inheritance of knowledge. If we register a large number of documents and their citations, we will haase a network of citations. Using the network of citations, one can recognize background or development of a research of a document. A citation itself is Boolean, but we define graded relations among documents through fuzzy graph theory. The membership function resulting from the retrieval is given by this relation. We discuss mathematical properties and their meaning in practical retrieval. The proposed system has been implemented on a personal computer and more than 600 documents have been registered. We will show experimental results lastly.
Keywords :	information retrieval systems ; graph theory ; relations ; fuzzy set theory

Set 4	
Title :	Information retrieval based on fuzzy associations
Author(s) :	S. Miyamoto
Source :	*Fuzzy Sets and Systems* 38 (1990) : 191-205
Abstracts :	The aim of the present paper is to propose a fuzzy set model for information retrieval and to develop methods and algorithms for fuzzy information retrieval based on the fuzzy set model. A process of information retrieval is represented as a diagram that consists of three components. Each component has its inherent fuzziness. As typical example for describing the three components, we consider a fuzzy association as a generalization of a fuzzy thesaurus for the first component, a fuzzy inverted index for the second component, and a fuzzy filter for the third component. Efficient algorithms for fuzzy retrieval on large scale bibliographic databases are developed. The significance of the present method is that current techniques in researches of bibliographic databases without fuzzy sets are studied in the framework of fuzzy sets and their implications are made clear using the model herein.
Keywords :	information storage and retrieval ; fuzzy associations ; algorithms

Set 5	
Title :	A fuzzy document retrieval method based on two-valued indexing
Author(s) :	Tetsuya Murai ; Masaaki Miyakoshi ; Masaru Shimbo
Source :	*Fuzzy Sets and Systems* 30 (1989) : 103-120
Abstracts :	A fuzzy retrieval method that enables one to infer weights and ranking output is formulated in this paper, still based on two-valued indexing. It provides a weighted system of document retrieval : a keyword is represented by a set of keywords related to it as a document ; then, a weight is calculated by the degree of matching between the representation of a keyword and that of a document. This procedure can obviously be extended to Boolean queries by fuzzy logic, and hence to weighted Boolean system without assuming weighted indexing. An interpretation of this method from the point of view of the possible world semantics in modal logic is further discussed.
Keywords :	fuzzy sets ; document retrieval ; modal logic ; possible world semantics ; keyword representations