웹 검색 환경에서 범주의 동적인 분류
(Dynamic Classification of Categories in Web Search Environment)

최 범 기 † 이 주 홍 †† 박 선 †
(Bumghi Choi) (Ju-Hong Lee) (Sun Park)

요 약 분류검색 방법은 색인검색 방법과 함께 중요한 요소로서 웹 검색 엔진에서 지원되고 있다. 사용자가 분류나 색인검색 방법 중 하나를 이용하여 원하는 검색결과를 찾지 못하면 다른 검색방법을 이용하여 찾을 수 있도록 대부분의 검색엔진에서는 두 가지 방법 모두 지원하고 있다. 색인검색 방법에서는 검색결과의 재현율이 높지만 검색결과가 너무 많이 나오기 때문에 원하는 검색결과를 찾아내는 것이 어렵다는 단점이 있다. 분류검색 방법은 찾고자 하는 문서의 해당 분류가 애매모호하거나 명확하게 알지 못할 때에는 문서를 찾지 못하는 경우가 빈번히 발생한다. 즉, 검색결과의 정확도는 높으나 재현율이 떨어지는 단점이 있다. 본 논문은 이러한 문제점을 해결하기 위해서 분류와 검색어간의 관계를 퍼지논리를 이용하여 정량적으로 계산하고 이를 바탕으로 범주간의 함의관계를 유도함으로써 동적인 범주체계를 구성하는 새로운 방법을 제시한다. 이 방법의 장점은 범주간의 함의관계를 유사한 하위범주로 간주함으로써 분류검색 결과의 재현율을 높일 수 있다는 것이다.

키워드 : 웹 검색엔진, 분류검색, 퍼지집합, 퍼지 관계 곱

Abstract Directory searching and index searching methods are two main methods in web search engines. Both of the methods are applied to most of the well-known Internet search engines, which enable users to choose the other method if they are not satisfied with results shown by one method. That is, Index searching tends to come up with too many search results, while directory searching has a difficulty in selecting proper categories, frequently mislead to false ones. In this paper, we propose a novel method in which a category hierarchy is dynamically constructed. To do this, a category is regarded as a fuzzy set which includes keywords. Similarly extensible subcategories of a category can be found using fuzzy relational products. The merit of this method is to enhance the recall rate of directory search by expanding subcategories on the basis of similarity.

Key words : web search engine, directory search, fuzzy set, fuzzy relational products

1. 서 론

웹 검색엔진에서, 문서가 속한 그룹을 장르 등의 기준에 따라서 계층적으로 구분하는 것을 분류 또는 디렉토리라 하고, 분류를 나타내는 단어를 주제 또는 분류어라고 한다. 그리고 웹 검색 엔진에서 사용자가 검색하기 위해서 입력하는 단어나 해당 문서를 검색할 수 있도록 색인된 문서의 분류어나 문서 안의 단어를 검색어라고 한다.

검색엔진에서 문서를 찾는 방법은 크게 색인검색과 분류검색으로 나뉘어 진다. 색인검색에서는 문서들을 그 안의 단어들로 색인하여 데이타베이스에 저장하고, 입력된 검색어와 일치하는 단어들을 데이타베이스에서 검색하여 검색어를 포함한 문서들을 중요도가 큰 순서대로 나열한다. 분류검색에서는, 각 분류는 문서들의 요약정보와 링크정보 또는 여러 개의 하위 분류를 포함한다. 찾고자 하는 문서가 해당 분류 없으면 하위 분류로 범위를 축소시켜 주제와 일치하는 분류 경로(path)로 찾아가 원하는 정보를 얻는다.

색인검색은 검색어를 입력하여 색인된 모든 문서를 신속하게 찾을 수 있는 장점이 있다. 그러나 단일 검색

· 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 육성·지원사업의 연구결과로 수행되었으며, 이 논문은 2006년도 두뇌한국21사업에 의하여 지원되었음
† 학생회원 : 인하대학교 컴퓨터공학과
neural@inha.ac.kr
sunpark@datamining.inha.ac.kr
†† 종신회원 : 인하대학교 컴퓨터공학 교수
juhong@inha.ac.kr
논문접수 : 2005년 6월 13일
심사완료 : 2006년 5월 23일

어나 검색어들의 조합이 찾고자 하는 문서들의 조건을 충분히 만족하지 못하고, 광범위한 의미로 확대되거나, 검색어가 동철이음이의어(heteronym), 동음이의어(homonym) 이거나, 문서의 내용이 검색어들로 적절히 표현되지 못할 때에는 불필요한 문서들을 너무 많이 찾거나, 아무것도 찾지 못하는 치명적인 단점이 있다. 색인검색의 이와 같은 단점들을 보완하기 위하여 자동분류 방법이나 질의어 확장 등의 방법들이 연구되고 있다. 자동분류 방법은 인터넷에서 문서자료와 해당링크를 수집과 동시에 분류하거나 검색된 결과를 군집화 하여 분류한다. 질의어 확장 방법은 검색어와 문서간에 다양한 관계를 설정하여 정확한 결과를 찾을 수 있도록 질의어를 확장한다.

색인검색의 장점과 단점에 대한 보완에도 불구하고, 분류검색 방법은 사용자가 정확한 분류를 알고 있으면 하위분류로 범위를 축소해 나갈 수 있어서 빠르게 검색할 수 있고 자주 검색되는 중요한 정보들이 잘 정리되어 있어서 색인검색 방법의 보완적인 방법으로서 많이 사용된다. 그러나 분류 검색 방법도 사용자가 찾고자 하는 문서의 해당 분류를 정확하게 알지 못하거나, 문서들이 정확하게 분류되어 있지 않을 때는 만족스러운 결과를 얻지 못하는 단점이 있다. 즉 찾고자 하는 문서를 어느 한 분류에서 찾지 못한 경우에는 다른 분류에서 다시 검색하여야 하는 불편한 경우가 자주 발생한다.

분류검색 방법에서의 이와 같은 문제점은 기존 검색엔진의 범주체계가 다른 범주에 속한 하위범주들 간의 관계를 분석하여 자동으로 설정하여 주는 적절한 방법이 없어서 상위범주 아래에 그와 관련된 좀더 세분화된 주제의 하위범주를 수동으로 구성하는 고정계층구조로 되어 있기 때문이다.

그림 1은 검색 엔진에서 문서 검색에 관련된 3가지 객체인 분류, 검색어, 문서의 관계를 보여준다. 분류 검색이 문서와 분류간의 관계를 이용한 검색이라고 한다면, 색인 검색은 검색어와 문서의 관계를 이용하는 검색이다. 문서의 자동 색인 기법은 검색어와 문서의 관계에 관련이 있으며, 문서의 수동 분류나 자동 분류는 문서와 분류와의 관계에 관련이 있다. 분류와 검색어는 각 검색방법에서 각각 중요한 역할을 하고 있다. 따라서 분류검색 방법을 개선하여 검색결과의 효율을 높이기 위해서는 검색어와 분류사이의 관계를 규정하고 좀더 유연한 분류간의 관계를 설정하여 이를 검색에 활용할 수 있는 방법이 고려되어야 한다.

본 논문은 위와 같은 동기에서, 검색어와 분류 간의 관계를 규정하고, 분류들 간의 상호 관계를 규명함으로써 분류검색의 범주체계를 동적인 체계로 재구성함으로써 검색효율을 높일 수 있는 새로운 방법을 제안한다.

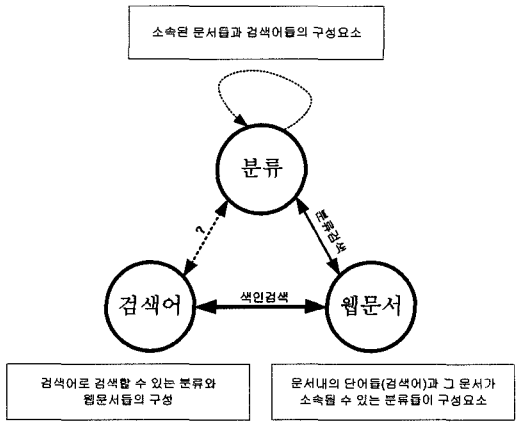


그림 1 검색 객체들 간의 관계

분류와 검색어간의 관계는 문서에서의 검색어의 중요도와 문서의 분류에서의 중요도 등의 관계를 구하여 설정할 수 있다. 이러한 관계는 분류를 검색어로 구성된 퍼지 집합으로 간주할 수 있게 한다. 두 범주 간의 관계는 유사도를 계산함으로써 규정할 수 있는데, 유사도는 한 퍼지 집합이 다른 퍼지 집합을 포함하는 정도로써 계산할 수 있다. 이것을 이용하면 서로 다른 범주의 유사관계를 동적으로 생성할 수 있다. 본 논문에서 제안한 방법은 다음과 같은 장점을 가진다. 첫째 범주의 공유성 및 범주 레벨의 유동성을 제공하여 고정된 계층형 범주 체계가 아닌 유연한 동적인 범주 체계를 생성한다. 둘째, 선택한 범주와 가장 유사한 하위범주들을 퍼지논리로 자동 생성하여 분류검색의 범위를 적절한 범위 내에서 확대함으로써 재현율을 올릴 수 있다. 본 논문의 구성은 다음과 같이 구성되어 있다. 2절에서는 관련 연구를 보인다. 3절은 본 논문에서 사용되는 퍼지이론에 대하여 알아본다. 4절은 동적인 범주 체계를 구성하는 방법을 설명한다. 5절에서는 실험 결과를 보인다. 6절에서 결론을 맺는다.

2. 관련연구

검색엔진의 검색 효율을 높이기 위해서, 웹문서분류 방법, 문맥기반 방법, 퍼지기반 방법들이 연구되었다.

웹문서분류 방법으로는, 문서를 수작업으로 분류하는 AskJeeves, 문서를 자동으로 분류하는 NorthernLight, 문서를 자동으로 군집하는 Vivisimo, 검색결과를 이용하여 문서를 동적으로 군집하는 Grouper가 있다. AskJeeves는 미리 전문가들이 수작업으로 분야별로 분류된 문서를 구축하고, 사용자의 질의로부터 분야를 추정하여 해당 분야의 문서를 제공한다. 사용자가 원하는 검색결과를 정확히 찾을 수 있으나, 문서의 분류를 수작업으로 해야 하고 사용자의 정보요구가 변화함에 따라 문서를

재분류 해야 하는 단점이 있다. NorthernLight는 문서를 수집하는 동시에 자동적으로 분류한다. 분류가 계층적으로 구성되어 있어 신속히 검색할 수 있는 장점이 있으나, 광범위하고 모호한 분류에 대한 재분류 작업이 필요하다. Vivisimo는 문서간 유사도에 의해 군집하여 분류한다. 검색결과를 그룹별 계층적 트리 형태로 자동 생성하여 빠르게 검색할 수 있고, 전처리 과정 없이 실시간으로 자동 분류를 제공하여 분류에 대한 유지보수가 필요 없으나, 색인검색 결과를 자동으로 군집화하기 때문에 시스템의 부하가 증가하고, 분류의 적합도가 떨어지는 단점이 있다. Grouper는 메타검색엔진인 Husky-Search의 결과를 동적으로 군집화 한다. Grouper는 사용자의 컴퓨터에서 실행되기 때문에 분산 정보검색 시스템에 적합하며, 검색 후 군집화 하기 때문에 검색된 결과들이 잘 정리가 되나 사용자의 시스템에 설치가 되지 않으면 사용할 수 없는 단점이 있다. Peng등은 전자상거래 검색엔진을 위한 유사한 제품을 군집하여 지원하는 방법을 지원하였다. Si와 Callan은 연합검색(federated search)를 위한 새로운 알고리즘을 제안하였다. Pandey등은 검색결과를 향상시키기 위해서 검색순위의 부분 부작위성에 기반을 둔 새로운 순위 정책을 제안하였다[11-15].

문맥기반 방법은 명시된 문맥정보를 지원하는 Inquirus 2와 자동적으로 문맥정보를 추리하는 방식인 IntelliZap가 있다. Inquirus 2는 검색효율이 좋으나 검색시 이용되는 분류가 한정되어 있다. IntelliZap는 재현율은 높으나 검색결과가 너무 많이 나열된다[5,6].

퍼지기반 방법에는 퍼지 개념 네트워크를 이용한 방법이 있다. 퍼지 개념 네트워크를 이용하여 문서와 개념에 대한 관련 정도를 나타내고, 질의와 가장 관련이 높은 문서를 선택한다. 그러나 퍼지 개념 네트워크가 결과를 검색하기 위해서는 많은 시간이 걸리기 때문에 계산은 퍼지 개념 행렬을 이용한다[3,7,9,10].

3. 퍼지 이론

본 절에서는 이 논문에서 사용되는 퍼지 이론에 관하여 간략하게 소개한다. 퍼지 개념은 하나의 대상이 하나의 값으로 정의되는 것이 아니라 여러 값을 통해 단계적으로 정의되기 때문에 집합의 개념을 사용하여 표시한다. 이러한 집합은 일반집합(crisp sets)의 표현 및 특성과 서로 비교하여 몇 가지 다른 점을 가지고 있기 때문에 일반집합과 구분하기 위하여 퍼지집합(fuzzy sets)이라 부른다[8]. 퍼지집합의 정의는 다음과 같다.

정의 1. X를 전체집합이라고 하자. X에서 정의되는 퍼지집합 A는 다음과 같은 순서쌍의 집합이다.

$$A = \{ (x, \mu_A(x)) \mid x \in X \}$$

여기서 $\mu_A : X \rightarrow [0,1]$는 A의 멤버쉽 함수(membership function)로서 A에서의 x의 소속 정도를 나타낸다. 즉, 단위 구간 [0,1] 사이의 실수 값을 멤버쉽 정도로 취하는 원소들로 구성되는 집합이다. □

일반집합은 퍼지 집합의 특수한 경우로 볼 수 있으며 일반집합의 특성함수(characteristic function)는 멤버쉽 함수의 특수한 경우로 볼 수 있다. 특성함수는 주어진 집합에 대해 어떤 원소의 소속 여부를 나타낸다.

정의 2. Aa로 표시되는 퍼지집합 A의 a-cut은 지정된 값보다 크거나 같은 소속 정도를 가지는 X의 모든 원소들의 보통집합으로 다음과 같이 표현된다.

$$Aa = \{ x \in X \mid \mu_A(x) \geq a \}$$ □

소속 함수의 값을 기준으로 일정한 값 이상의 요소를 취해야할 필요가 있다. 이럴 경우에 사용되는 것이 a-cut이다. a는 일정한 값을 의미한다. 퍼지 집합에 a-cut를 적용하면 일반 집합이 얻어 진다. 이것은 일정수준 이상을 소속정도 '1'로 취하겠다는 의미이다.

퍼지 함의 연산자(Fuzzy Implication Operator)는 일반 함의 연산자(Crisp Implication Operator)를 확장하여 퍼지에 적용한 것으로서, 일반 함의 연산자는 $\{0,1\}\times\{0,1\}\rightarrow\{0,1\}$로 정의되는데 반해, 퍼지 함의 연산자는 $[0,1]\times[0,1]\rightarrow[0,1]$로서 단위 구간의 다치 논리로 확장된 것이다. 퍼지 함의 연산자의 종류는 무수히 많으며 대표적인 Kleene-Diense 퍼지함의 연산자의 예는 다음과 같다[2].

$$a \rightarrow b = (1-a) \vee b = \max(1-a, b),$$
$$a = 0 \sim 1,\ b = 0 \sim 1 \qquad (1)$$

표 1은 Kleene-Diense 퍼지 함의 연산자의 구체적인 예를 보여준다.

표 1 Kleene-Diense 시스템의 퍼지 함의 연산자

a → b	0	0.3	0.6	1
0	1	1	1	1
0.4	0.6	0.6	0.6	1
0.7	0.3	0.3	0.6	1
1	0	0.3	0.6	1

집합이론에서 "$A \subseteq B$"는 "$\forall x,\ x \in A \rightarrow xB$"와 같고 "$A \in \wp(B)$"와도 같다. 여기서 $\wp(B)$는 B의 멱집합(power set)이다. 따라서 퍼지 집합에서의 "$A \in B$인 정도"는 $A \in \wp(B)$인 정도이므로 $\mu_{\wp(B)}A$로서 나타낼 수 있으며 다음과 같이 정의된다.

정의 3. 퍼지 함의 연산자 $\rightarrow$와 일반 전체집합 U의 퍼지 집합 B가 주어진 상태에서 B의 퍼지 멱집합의 멤버쉽 함수 $\mu\wp B$는 다음과 같이 주어진다.

$$\mu_{\wp B}A = \bigwedge_{x \in U} (\mu_A x \rightarrow \mu_B x)$$ □

정의 4. U_1, U_2, U_3는 유한한 전체 집합이라 하고, R은 U_1에서 U_2로의 퍼지관계이고, S는 U_2에서 U_3로의 퍼지관계이다. 즉, R은 $U_1 \times U_2$의 퍼지 부분집합이고 S는 $U_2 \times U_3$의 퍼지 부분집합이다. 퍼지 관계곱은 $a \in U_1$이고 $c \in U_3$일 때, a가 c에 관련되어 있는 정도를 나타낼 사용되는 퍼지연산이다. U_1에서 U_2로의 퍼지관계인 삼각논리곱 ,◁, 는 다음과 같이 정의된다.

$$(R \triangle S)_{ik} = \frac{1}{N_j} \sum_j (R_{ij} \rightarrow S_{jk})$$

이것을 퍼지 관계곱(Fuzzy Relational Products)이라 한다. □

정의 5. 퍼지 함의 연산자는 주어진 문제의 범주에 따라 달라진다. $a \in U_1$에 대한 후위집합(afterset) aR는 a와 연관된 $y \in U_2$로 구성된 U_2의 퍼지 부분집합이며 그 멤버쉽 함수는 $\mu_{aR}(y) = \mu_R(a,y)$로 주어진다. $c \in U_3$에 대한 전위집합(foreset) Sc는 c에 연관된 $y \in U_2$로 구성된 U_2의 퍼지 부분집합이며 그 멤버쉽 함수는 $s_c(y) = s(y,c)$로 주어진다. aR이 Sc의 부분집합인 평균정도는 $y \in aR$의 멤버쉽 정도가 $y \in Sc$의 멤버쉽 정도를 함의하는 평균정도로서 다음과 같이 정의된다.

$$\pi_m(aR \subseteq Sc) = \frac{1}{N_{U_2}} \sum_{y \in U_2} (\mu_{aR}(y) \rightarrow \mu_{Sc}(y)) \qquad (2)$$

여기서 π_m은 평균 정도를 나타내는 함수이다.

위의 평균 정도는 $R \triangleleft S$에 의해서 a가 c에 관련되는 정도를 나타낼 수 있다[3].

4. 동적인 분류 체계

본 논문에서 검색어와 분류간의 관계는 검색어와 문서 간의 관계 및 문서와 분류간의 관계에 의해서 결정된다. 검색어와 문서간의 관계는 기존의 연구에 의해 많이 연구되었는데[1], 문서를 검색어로 구성된 퍼지 집합이라고 간주할 수 있고, 마찬가지로 분류를 소속한 문서들의 검색어들로 구성된 퍼지집합으로 간주할 수 있다. 즉, 본 논문에서는 검색어와 문서의 관계를 구하기 위하여, 먼저 검색어가 문서에 몇 번이나 나타났는가 하는 빈도테이블을 구성한다. 구성된 빈도테이블을 이용하여 검색어와 문서간의 코사인유사도[1]를 이용하여 유사도를 계산한다. 계산된 유사도는 검색어와 문서 간에 얼마나 관계를 가지고 있는가 하는 유사도를 나타낸다. 본 논문에서는 유사도를 검색어가 문서를 얼마나 포함하는가하는 퍼지 값으로 간주한다. 그런 다음, 두 범주간의 관계는 생성된 두 범주의 퍼지 집합의 함의 정도를 계산하여 결정할 수 있다. 두 퍼지집합의 함의 정도는 퍼지함의 연산자를 사용하여 한 퍼지집합이 다른 퍼지집합에 포함되는 정도를 계산하여 구할 수 있고, 이를 이용하여 서로 다른 두 범주의 유사관계를 동적으로 생성할 수 있다.

퍼지 함의 연산자는 각 응용의 필요성에 맞게 제시되어야 하는데 본 논문에서는 위의 식 (1)의 Kleen-Diense 퍼지 함의 연산자를 사용한다. 퍼지 함의 연산자를 식 (2)의 퍼지관계곱을 적용하여 분류들 간의 퍼지함의관계, $C_i \rightarrow C_j$를 유도할 수 있다. 그러나 $C_i \rightarrow C_j$은 $C_i \subseteq C_j$의 정도를 나타내는 척도로서 약간 문제가 있다. 즉 C_i에 멤버쉽 값($\mu_{Ci}(x)$)이 작은 원소 x가 많으면, $C_i \subseteq C_j$의 포함여부와 관계없이 항상 1에 가까운 값이 나오는 문제점이 있다. 따라서 변형된 a-cut 퍼지관계곱을 다음과 같이 정의하여 두 범주 퍼지 집합의 함의 관계, $C_i \xrightarrow[\alpha]{} C_j$,를 계산한다.

$$C_i \xrightarrow[\alpha]{} C_j = (R_\alpha^T \triangle R)_{ij} = \frac{1}{|C_{i\alpha}|} \sum_{K_k \in C_{i\alpha}} (R_{ik}^T \rightarrow R_{kj}) \qquad (3)$$

여기서, K_k는 k번째 검색어이고, C_i와 C_j는 i번째와 j번째 범주이며, $C_{i\alpha}$는 C_i의 α-cut이고, $|C_{i\alpha}|$ 는 $C_{i\alpha}$의 원소의 갯수이다. R는 $m \times n$ 행렬로서 R_{ij}는 $\mu_{C_j}(K_i)$, 즉, $K_i \in C_j$인 정도이다. R^T는 행렬 R의 전치 행렬로서 $R_{ij} = R^T_{ji}$이며, R_α^T의 $(R_\alpha^T)_{ij}$는 다음과 같다.

$$(R_\alpha^T) = \begin{cases} 0 & \text{if } R_{ij}^T < \alpha \\ R_{ij}^T & \text{if } R_{ij}^T \geq \alpha \end{cases}$$

다음은 식 (3)을 적용한 예이다.

예 1. a = 0.9 일 때 $C_2 \xrightarrow[\alpha]{} C_3$ 는 $(R_\alpha^T \triangle R)_{23}$ = 0.94 이고 $C_1 \xrightarrow[\alpha]{} C_3$는 $(R_\alpha^T \triangle R)_{13}$ = 0.7 이다. 각 분류간 함의 관계는 a-cut 퍼지 관계곱(3)에 의해 표 2의 (a), (b) 및 (c), (d)와 같이 설정될 수 있다.

표 2 범주와 검색어의 α-cut 퍼지 관계곱

	K_1	K_2	K_3	K_4	K_5
C_1	0.9	1	1	1	1
C_2	0.1	1	0.1	0	1
C_3	1	0.8	0	1	1
C_4	0	0.2	1	0	0.1
C_5	0.1	1	1	0.8	1

(a) R^T

	C_1	C_2	C_3	C_4	C_5
K_1	0.9	0.1	1	0	0.1
K_2	1	1	0.8	0.2	1
K_3	1	0.1	0	1	1
K_4	1	0	1	0	0.8
K_5	1	1	1	0.1	1

(b) R

	K_1	K_2	K_3	K_4	K_5
C_1	0.9	1	1	1	1
C_2	0	1	0	0	1
C_3	1	0	0	1	1
C_4	0	0	1	0	0
C_5	0	1	1	0	1

(c) R_α^T

	C_1	C_2	C_3	C_4	C_5
C_1	0.98	0.44	0.76	0.28	0.78
C_2	0.98	0.96	0.94	0.64	0.98
C_3	0.98	0.62	0.96	0.26	0.78
C_4	1	0.82	0.76	0.94	1
C_5	0.98	0.64	0.76	0.48	0.94

(d) $(R_\alpha^T \triangle R)$

다음에 $(R_\alpha^T \triangle R)$ 를 a' 으로 a-cut 하여 크리스프 값으로 바꾼다. 표 3의 (a)는 $(R_\alpha^T \triangle R)$를 $a' = 0.94$ 로 a-cut 한 최종 결과이다. 즉 0.94 미만의 값은 0이 되고 원래 0.94 이상인 값은 1이 된다. 표 3의 (b)는 $a' = 0.76$ 로 a-cut한 최종결과이다.

표 3 $(R_\alpha^T \triangle R)$를 a' 으로 a-cut 한 최종결과

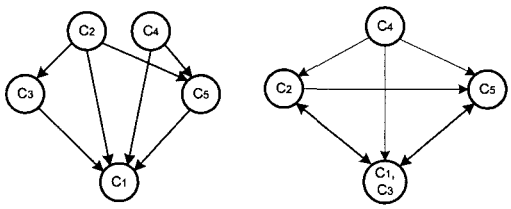
	C_1	C_2	C_3	C_4	C_5
C_1	1	0	0	0	0
C_2	1	1	1	0	1
C_3	1	0	1	0	0
C_4	1	0	0	1	1
C_5	1	0	0	0	1

(a) a-cut for $a' = 0.94$

	C_1	C_2	C_3	C_4	C_5
C_1	1	0	1	0	1
C_2	1	1	1	0	1
C_3	1	0	1	0	1
C_4	1	1	1	1	1
C_5	1	1	1	0	1

(b) a-cut for $a' = 0.76$

그림 2는 표 3에 의하여 얻어진 최종 결과로서 각 범주간 관계를 보여준다. $a' = 0.94$일 때는 (a)와 같으며, $a' = 0.76$ 일 때는 (b)와 같다. 그림 2의 (a)에서 $a' = 0.94$일 때 범주간의 함의관계를 살펴보면, C_1 범주 항목은 모든 범주 항목의 하위범주이고, C_3과 C_5 각각은 C_2, C_4의 하위범주이다. (b)에서는 $a' = 0.76$ 일 때 범주간의 함의관계를 살펴보면, C_4가 최상위 범주에 위치하며, C_1, C_3는 최하위 범주에 위치한다. $a' = 0.94$ 일 때의 범주관계를 모두 포함하면서 확장된 것을 알 수 있다.



(a) $a' = 0.94$ 일 때의 범주관계 (b) $a' = 0.75$ 일 때의 범주관계

그림 2 최종결과의 범주 관계도

그림 2와 같은 동적인 범주관계를 생성하면, 검색시 원하는 대상이 없을 때는 유사한 하위범주로 확장하여 검색할 수 있다.

본 논문에서 제안한 동적인 범주체계는 일반적인 검색엔진의 고정 범주체계와는 다른 다음과 같은 특이성을 가지고 있다.

(1) 범주의 공유성 : 그림 2의 (a)에서 C_1은 C_2, C_3, C_4, C_5에 대해서 동시에 하위범주로 구성된다. 이것은 일반 고정범주체계의 배타적 개념 대신 공유개념과 다중계승의 개념이 도입된 것이다. 즉 범주 C_1에 의해 분류되는 사이트들은 상위 검색범주로서 C_2, C_3, C_4, C_5를 공유하게 된다.

(2) 범주 레벨의 유동성 : 그림 2의 (a)에서 C_1은 여러 개의 범주계층에 위치할 수 있는 유동성을 가지고 있다. 검색 경로가 $C_2 \rightarrow C_3 \rightarrow C_1$ 인 경우는 3번째 범주 계층에 속하지만 검색 경로가 $C_2 \rightarrow C_1$인 경우는 2번째 범주 계층에 속하게 된다.

(3) 하나의 의미에 대한 둘 이상의 표현 : 그림 2의 (b)에서 또 다른 특이성을 발견할 수 있다. 그림에서 보는 바와 같이 C_1, C_3는 같은 의미의 범주로서 다르게 표현될 뿐이다.

(4) 범주의 호환성 : 그림 2의 (b)에서 보여지는 또 다른 특이성은 C_2과 C_5는 서로가 하위 개념이면서 동시에 상위 개념이기도 한다. 그러나 C_2, C_5는 C_1과 C_3의 관계처럼 같은 의미를 갖는 표현들은 아니다. C_5는 C_1, C_3의 하위 범주이지만 C_2는 그렇지 않다. 이는 C_1, C_3와 C_5의 관계에서도 비슷하게 설명된다.

(5) 범주 체계의 주관성 : 그림 2의 (a) 및 (b)에서 보는 바와 같이 a-cut의 값 , a', 에 따라 범주 체계는 달라질 수 있다. 검색 환경과 검색 주체에 따라서 a-cut 의 값을 연동시킬 수 있다.

4.1 검색방법

본 논문에서 제시한 방법을 검색엔진에 다음과 같이 적용할 수 있다.

(1) 문서의 주제를 알고 있을 때는 범주체계에서 해당하는 주제의 범주를 선택한다.

(2) 해당 범주에서 찾고자 하는 내용이 없을 때는 유사한 하위범주로 확장하여 찾는다.

(3) 범주가 모호한 경우는 검색어를 입력하여 그 검색어와 연관된 범주들 중 중요도가 높은 순으로 열거된 범주 중에서 찾고, 열거된 범주에서 찾지 못할 경우 마찬가지로 유사한 하위 범주로 확장하여 찾는다.

다음 그림 3은 위 검색방법이 처리되는 절차를 나타낸 것이다. 사용자가 검색할 범주를 입력하면, 검색정보

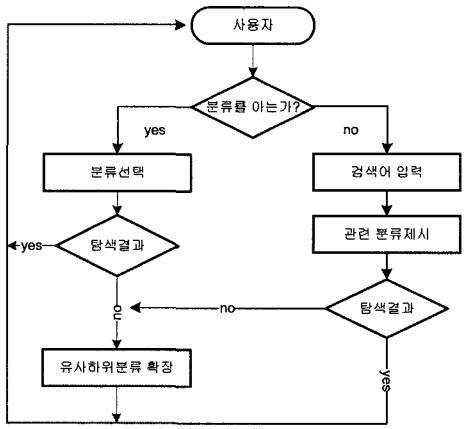


그림 3 퍼지 관계곱을 이용한 검색 과정

데이타베이스로부터 하위범주 항목들을 추출하고, 검색어를 입력하면 그 검색어의 중요도가 높은 범주들을 추출한다. 그리고 a-cut 퍼지관계곱에 따라 상위범주에 대하여 하위범주 항목을 확장하여 유사 범주를 구한다. 각각의 범주에 속한 문서들을 대상으로 검색처리의 결과를 사용자에게 보내준다.

다음의 예는 기존 검색엔진과 제안된 검색방법에서 검색어 은을 이용하여 검색하는 방법을 설명한 것이다. 사용자가 원하는 검색결과는 귀금속 은제품의 인터넷 판매 회사이다.

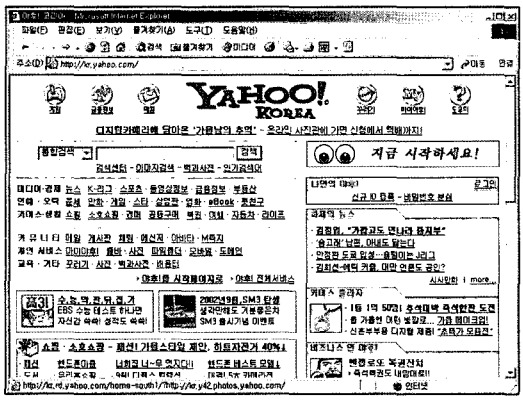
예 2. 기존 검색엔진을 이용한 검색방법

분류검색과 색인검색 방법을 따로 사용한다. 분류검색에서는 정확한 범주를 알지 못하면 찾기 힘들다. 만약 정확한 범주를 알고 있더라도 여러 단계를 거쳐야 원하는 결과를 찾을 수 있다. 분류검색 결과인 그림 4의 (a)는 은의 인터넷 판매회사를 찾기 위해서 하위분류로 7단계나 이동하였다. 검색어를 사용한 색인검색에서 은은 중국의 '은'나라와 귀금속 '은'등의 여러 가지 다른 의미(동음이의어, 동철이음이의어)들이 혼합되어 있기 때문에 많은 결과들이 검색된다. 그림 4의 (b)와 같이 이 결과들은 범주별로 모아져 있지 않다. 그래서 사용자가 원하는 결과를 찾기 위해서는 다시 한번 필터링해야 하므로 많은 시간이 소요된다.

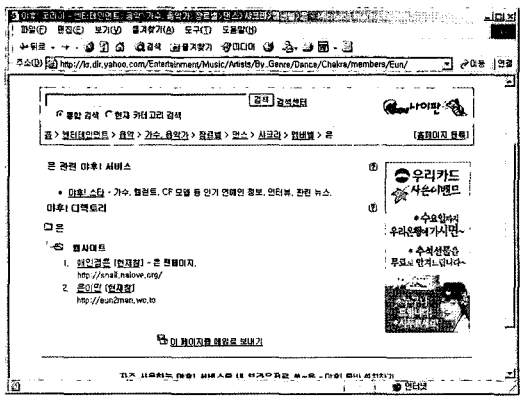
예 3. 제안된 검색방법

기존 검색엔진과 마찬가지로 정확한 범주를 알지 못하면, 검색어를 이용한 색인검색으로 검색결과들과 그것들의 범주를 찾을 수 있다. 그림 5의 (a)와 같이 검색어를 입력하면 곧 관련 범주가 왼쪽 분류트리에 나타난다. 여기서 온라인 쇼핑 하나만 선택하여 한번 만에 원하는 결과를 찾을 수 있다.

만약 하위범주에 원하는 결과가 없다면, (b)와 같이 확장된 범주를 선택하여 원하는 결과를 선택할 수 있다. 범주 트리에서 전개되는 범주들은 '은'의 여러 의미와 관련된 범주임을 알 수 있다. 이와 같이 범주를 유사 하위범주로 확장하면 다른 주제의 범주에 속하면서 검색어와 관련 있는 범주로도 검색할 수 있다.

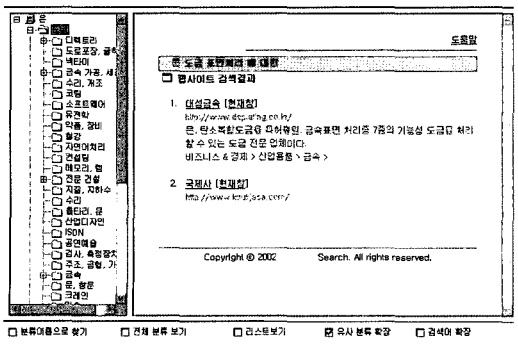


(a) 분류검색 결과

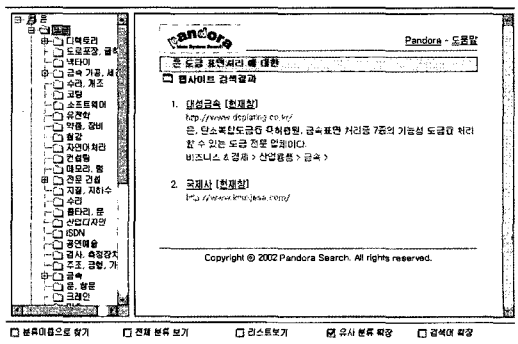


(b) 색인검색결과

그림 4 검색엔진 '야후 코리아'의 검색 결과



(a) 기본 분류 결과



(b) 확장 유사 분류 결과

그림 5 제안된 검색방법의 검색결과

5. 실험결과

본 논문에서는 제안방법이 분류검색방법의 문제점을 개선하였는가에 대한 평가를 하기 위한 두 가지의 실험을 하였다. 정보검색에서 가장 많이 사용되는 평가방법은 재현율과 정확률 이다. 그러나 웹과 같이 빠르게 변화하는 큰 자료에 대하서 재현율과 정확률에 필요한 문서의 개수를 결정하는 것은 불가능하다. 따라서 본 논문에서는, 성능평가를 하기 위하여 기존의 검색 엔진에 비하여 분류의 재현율이 얼마나 향상되었는가와, 사용자가 원하는 정보를 얼마나 효율적으로 검색하는지를 비교하기 위하여 원하는 사이트를 검색하는 접근속도를 계산하였다. 실험 1 에서 분류의 재현율을 알기 위해서는 주어진 주제의 적합한 사이트의 전체 수를 알아야 하지만 현실적으로 그것을 알 수 있는 방법이 없다. 따라서 같은 조건 하에서 각각 찾은 결과로부터 주제에 적합한 사이트의 수를 센다. 그리고 하위분류들 중에서 적합한 사이트를 포함하고 있는 적합한 하위분류의 수를 세어 비교하였다. 실험의 비교 대상은 '야후 코리아'이며, 본 실험을 위하여 α-cut 퍼지 관계곱을 사용한 검색 엔진을 구현하였다. 같은 조건에서 실험하기 위하여 구현한 검색 엔진의 데이타는 '야후 코리아'의 자료를 가져와 데이타베이스를 구축하였다.

실험 1. 실험에 사용된 범주는 다음 표 4와 같다. 하위 범주의 수가 적은 범주와 많은 범주를 고르게 선택하였다. 각 범주의 하위범주에서 찾아지는 적합한 사이트 수를 세고, 하위범주들 중에서 주제에 적합한 사이트를 포함하고 있는 적합한 하위범주의 수를 세었다.

실험에서 주제는 범주가 속하는 범주의 이름이며, 범위는 범주에 포함된 사이트와 주제간의 관계를 나타낸다. 카니발의 주제를 검색시 야후에서는 기아자동차와 소개/평가 등의 2개의 범주가 나타났으나 구현한 검색엔진에서는 기아자동차, 메이커, 노동조합 등의 3개의 기본 분류와 이를 확장한 유사 하위범주로서 중고자동차 매매 등과 같은 관련된 모든 범주를 찾을 수 있었다. '야후 코리아'에서는 고래의 범주가 존재하지 않아서 검색할 수 없었다. 그러나 제시된 검색방법은 고래라는 단어를 범주로 간주하여 유사한 하위 범주로 확장할 수 있었으므로 관련 사이트를 찾을 수 있었다. 그림 6은 '야후 코리아'와 구현된 검색 엔진에서 각 분류의 포함된 하위분류의 수를 보여준다. 실험 결과로서 검색된 범주에 포함된 하위범주의 수는 '야후 코리아'에서 90개이며 구현된 검색엔진에서 116개이다. 그리고 전체 범주의 주제와 일치하는 사이트의 수는 '야후 코리아'가 6,960개, 구현된 검색엔진은 7,934이다. 따라서 기존의 검색엔진에 비해 범주의 재현율은 28.9%가 향상되었고 검색된 웹사이트의 재현율은 14%정도 향상되었다. 즉, 분류검색에서 범주의 범위(α = 0.5)를 확대함으로써 재현율을 올렸다.

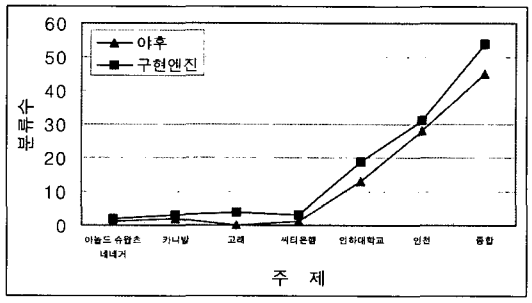


그림 6 '야후 코리아'와 구현된 검색엔진의 적합한 하위분류의 수

실험 2. '야후 코리아'와 구현된 검색엔진에서 분류검색만 이용하여 검색할 때, 원하는 검색 결과를 얻기까지 내려가는 하위범주 단계의 수를 비교하였다. 이때 하위범주의 단계의 수가 작으면 검색을 보다 빨리 할 수 있는 것으로 간주한다. 실험에 참여한 대상은 대학교 1학년 학생 151명으로, 컴퓨터 사용능력은 초보자들이다. 이들은 먼저 5개의 주제를 정하고 그 주제에 가장 적합한 범주 5개를 선택하여 그 곳에서 검색을 시작한다.

그림 7은 각 실험자별로 '야후 코리아'와 구현된 검색엔진에서 최종 검색되는 하위범주 단계의 평균을 보여준다. 실험 결과로서 '야후 코리아'에서는 평균 6.22 단

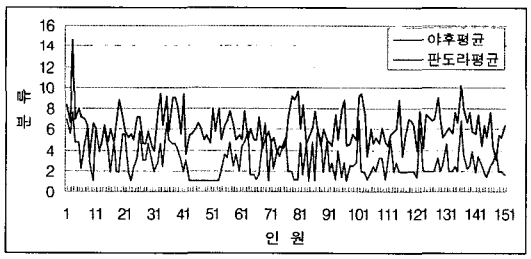


그림 7 '야후 코리아'와 구현된 검색엔진의 분류단계의 비교

표 4 실험 1에 사용된 범주

분 류	C_1	C_2	C_3	C_4	C_5	C_6
주 제	아놀드 슈왑츠제네거	고래	카니발	씨티은행	인하대학교	인천
범 위	영화배우	동물	자동차	은행	학교정보	종합정보

계 만에 검색되었으며, 구현된 검색 엔진에서는 평균 4.3 단계 만에 검색되었다. '야후 코리아'에서는 범주가 모호한 최악의 경우에는 최고 15단계나 거쳐서 찾았다. 반면, 구현된 검색엔진에서는 최악의 경우에도 7회 만에 찾았으며 이중 45% 정도는 2.6회 이하에서 찾았다. 기존의 방법에 비해 접근속도를 평균 30.9% 향상 시켰다. 즉, 이는 야후의 분류검색 방법이 정확한 범주 경로를 알지 못하면 찾기 어려운데 비해, 제안된 검색방법은 범주를 동적으로 재구성 하여 범주와 유사한 주제의 하위 범주를 분류에 포함시킴으로써 검색의 효율을 높였다.

6. 결 론

이 논문에서 우리는 분류 검색에 있어서 α-cut 퍼지 관계곱을 이용하는 새로운 방법을 제안하였다. 본 논문에서 제시한 방법은 α-cut 퍼지 관계곱을 이용하여 각 분류의 유사한 하위범주를 찾아냄으로써 분류 검색의 재현율을 향상시킨다. 실제로 시스템을 구현하였고, 실험을 통하여 다음의 잇점을 확인 하였다.

(1) 범주가 모호한 검색어에 대하여 유사한 하위범주로의 확장하여 검색을 용이하게 한다.
(2) 하위분류가 여러개의 상위 범주에 속할 수 있는 범주의 공유성과 범주 레벨의 유동성을 제공하여 검색 범주체계를 동적으로 관리할 수 있다.
(3) 에러의 한계를 규정하는 α-cut 값인 α와 α'를 다양하게 설정함으로써 범주 체계를 다양하게 변동시킬 수 있다.
(4) 하나의 검색 범주 항목이 다른 검색 범주에 대하여 하위 개념이면서 동시에 상위 개념이 되는 검색 범주간의 상하위 개념에서의 특이성을 통해 범주간의 호환성을 제공한다.

본 논문에서 제시한 방법은 검색 엔진 뿐만 아니라 기업문서 관리 및 검색시스템, 도서 관리시스템, 상품 및 부품 관리시스템 등 지능적 분류 방식을 필요로 하는 다양한 분야에 적용할 수 있다.

향후의 연구과제로는 동적 분류 체계에서 확장된 질의어를 처리할 수 있는 방법에 대한 연구와 유사한 하위분류를 생성할 때 소요되는 시간을 단축시키기 위한 연구가 있다.

사 사

본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 육성·지원사업의 연구결과로 수행되었으며, 이 논문은 2006년도 두뇌한국21사업에 의하여 지원되었음

참 고 문 헌

[1] R. Baeza-Yates and B. Ribeiro-Neto., "Modern Information Retrieval. Addison Wesley," 1999.
[2] W. Bandler and L. Kohout., "Fuzzy Power Sets and Fuzzy Implication Operations," Fuzzy Set and Systems, Vol.4, No.1, pp. 13-30, 1980.
[3] W. Bandler and L. Kohout., "Semantics of Implication Operators and Fuzzy Relational Products," International Journal of Man-Machine Studies, Vol. 12, pp.89-116, 1980.
[4] S. Chen and Y. Horng., "Fuzzy Query Processing for Document Retrieval Based on Extended Fuzzy Concept Networks," IEEE Transaction on Systems, Man and Cybernetics, Part B, Vol. 29, Issue 1, pp.96-104, 1999.
[5] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin., "Placing Search in Context : The Concept Revisited," In Proceedings of the 10th International Conference on World Wide Web, pp.406-414, HongKong, Chine, May 2001.
[6] E. J. Glover, S. Lawrence, W. P. Brimingham, and C. L. Giles., "Architecture of a Metasearch Engine that Support User Information Needs," In Proceedings of the 8th International Conference on Information and Knowledge Management, pp. 210-216, Kansas City, Missouri, 1999.
[7] K. Kim and S. Cho., "A Personalized Web Search Engine Using Fuzzy Concept Network with Link Structure," IFSA, pp.81-88, July 2001.
[8] K.H. Lee and G.L. Oh., "Fuzzy Theory and Application Volume I : Theory," HongReung Science Publishing Co., 1991.
[9] M. NikRavesh., "Fuzzy Conceptual-Based Search Engine using Conceptual Semantic Indexing," NAFIPS-FLINT 2002, pp.146-151, New Orleans, LA, June 2002.
[10] T. Takagi and M. Tajima., "Query Expansion Using Conceptual Fuzzy Sets for Search Engine," In Proceedings of the 10th IEEE International Conference on Fuzzy Systems, pp. 1303-1308, 2001.
[11] J. Wen, J. Nie, and H. Zhang., "Clustering User Queries of a Search Engine," In Proceedings of the 10th International Conference on World Wide Web, pp. 162-168, Hong Kong, China, 2001.
[12] O. Zamir and O. Etzioni., "Grouper : A Dynamic Clustering Interface to Web Search Results," In Proceedings of the 8th International Conference on World Wide Web, Toronto, Canada, 1999.
[13] S. Pandey, S. Roy, C. Olston,. "Shuffling a Stacked Deck: The Case for Partially Randomized Ranking of Search Engine Results," In Proceedings of the 31st VLDB Conference, pp. 781-792, Trondheim, Norway, 2005.

[14] Q. Peng, W. Meng, H. He, C. Yu., "WISE-Cluster: Clustering E-Commerce Search Engines Automatically," In Proceedings of the 6th annual ACM international workshop on web information and data management, pp. 104-111, Washington DC, USA, 2004.
[15] L. Si and J. Callan,. "Modeling Search Engine Effectiveness for Federated Search," In Proceedings of the 28th SIGIR Conference, pp. 83-90, Salvador, Brazil, 2005.

최 범 기

1986년 서울대학교 수학과 졸업(학사) 1995년 Florida State University 대학원 Computer Science 졸업(석사). 2006년 인하대학교 대학원 컴퓨터 정보공학과 겸임교수, 박사과정. 관심분야는 데이타베이스, 데이타마이닝, 정보검색, 신경망

이 주 홍

1983년 서울대학교 컴퓨터공학과 졸업(학사). 1985년 서울대학교 대학원 컴퓨터공학과 졸업(석사). 2001년 한국과학기술원 컴퓨터공학과 졸업(박사). 2001년~현재 인하대학교 컴퓨터공학부 부교수 관심분야는 데이타마이닝, 데이타베이스, 정보검색, 소프트컴퓨팅, 신경망

박 선

1996년 전주대학교 전자계산학과 졸업(학사). 2001년 한남대학교 정보산업대학원 정보통신학과 졸업(석사). 2002년~현재 인하대학교 컴퓨터 정보공학과 박사과정 관심분야는 데이타마이닝, 정보검색