웹 검색 분류어의 동적인 분류

최범기* 박선° 이주홍**
*쿼크(주), °**인하대학교 컴퓨터공학과
{bgchoi, sunpark°}@datamining.inha.ac.kr, juhong@inha.ac.kr

Dynamic Classification of Web Search Categories

Bumghi Choi* Sun Park° Ju-Hong Lee**
*Quark Co., Ltd., °**School of Computer Science and Engineering, Inha University

요 약

본 논문은 웹 탐색중 디렉토리 검색엔진의 분류검색에 대한 문제점을 해결하기 위해서 분류와 검색어간의 관계를 퍼지논리를 이용하여 계산하고 분류간의 함의관계를 유도함으로써 동적인 분류체계를 구성하는 새로운 방법을 제시한다. 이 방법의 장점은 분류간의 함의관계를 유사한 하위분류로서 간주함으로써 분류검색 결과의 재현율을 높일 수 있다는 것이다

1. 서론

웹 탐색에는 기본적으로 세 가지 형태가 있다. 첫 번째는 전문가에 의해 선별된 웹 문서와 이를 주제별로 분류하여 디렉토리로 검색하는 분류검색방법이다. 두 번째는 자동으로 수집된 웹 문서의 한부분에 색인을 달아 검색하는 색인검색방법이다. 세 번째는 하이퍼링크 구조를 이용하여 웹을 탐색하는 방법이 있으나, 현재 성능의 제약과 상용제품 부족 등으로 넓게 사용되지 못하고 있다.

검색엔진의 효율을 높이기 위해 자동분류방법을 사용하나, 대부분이 색인검색의 보완방법으로 연구되고 있다. 대표적인 연구로는, 답변문서를 분류하는 AskJeeves, 문서를 자동으로 분류하는 NorthernLight, 문서를 자동으로 군집하는 Vivisimo, 검색결과를 이용하여 문서를 동적으로 군집하는 Grouper가 있다[1,5,6].

분류검색에서의 분류는 Yahoo!와 같이 수작업으로 분류하는 방법이외에 자동분류에 대한 연구는 미흡한 실정이다[1].

분류검색 방법은 사용자가 정확한 분류를 알고 있으면 빠르게 검색할 수 있고 검색 정보들이 잘 정리되어 있어서 컴퓨터 사용에 익숙하지 않은 사람들이 널리 사용하기 때문에 색인검색 방법의 보완적인 방법으로서 검색엔진에서 반드시 필요한 기능이다. 그러나 사용자가 찾고자 하는 문서의 해당 분류를 정확하게 알지 못하거나 문서들이 정확하게 분류되어 있지 않을 때는 만족스러운 결과를 얻지 못하는 단점이 있다. 분류검색 방법에서의 이와 같은 문제점은 기존 검색엔진의 분류체계가 고정계층구조로 되어 있기 때문이다.

검색 엔진에서 문서 검색에 관련된 3 가지 객체는 분류, 검색어, 문서이다. 분류 검색이 문서와 분류간의 관계를 이용한 검색이라고 한다면, 색인 검색은 검색어와 문서의 관계를 이용하는 검색이다. 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 분류체계를 자동으로 동적인 체계로 재구성함으로써 검색효율을 높일 수 있는 α-cut 퍼지 관계곱을 이용하는 새로운 방법을 제안한다.

2. 퍼지 이론

본 절에서는 이 논문에서 사용되는 퍼지 이론에 관하여 간략하게 소개한다[4]. 퍼지 함의 연산자 (Fuzzy Implication Operator) 는 $[0,1]\times[0,1]\rightarrow[0,1]$ 로서 단위 구간의 다치 논리로 확장된 것이다. 퍼지 함의 연산자의 종류는 무수히 많으며 대표적인 Kleene-Diense 퍼지함의 연산자는 다음과 같다[2].

$$a\rightarrow b = (1 - a)\vee b = \max(1 - a, b), \quad a = 0\sim1,\ b = 0\sim1 \qquad (1)$$

(정의) 퍼지 함의 연산자는 주어진 문제의 범주에 따라

달라진다. $a \in U_1$ 에 대한 후위집합 (afterset) aR 는 a와 연관된 $y \in U_2$ 로 구성된 U_2 의 퍼지 부분집합이며 그 멤버쉽 함수는 $\mu_{aR}(y) = \mu_R(a,y)$ 로 주어진다. $c \in U_3$ 에 대한 전위집합 (foreset) Sc 는 c 에 연관된 $y \in U_2$ 로 구성된 U_2 의 퍼지 부분집합이며 그 멤버쉽 함수는 $\mu_{Sc}(y) = \mu_S(y,c)$ 로 주어진다. aR 이 Sc 의 부분집합인 평균정도는 $y \in aR$ 의 멤버쉽 정도가 $y \in Sc$ 의 멤버쉽 정도를 함의하는 평균정도로서 다음과 같이 정의된다.

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

여기서 π_m은 평균 정도를 나타내는 함수이다[3].

3. 동적인 분류체계

검색어와 분류간의 관계를 도출해 내는 방식은 여러 가지가 있을 수 있다. 본 논문에서는 분류에 속한 각 문서들에 대한 검색어들과의 관계를[1] 종합하여 만들어진다. 이렇게 생성된 분류들은 검색어들을 요소로 하는 퍼지 집합이 된다. 두 분류간의 관계는 생성된 두 분류의 퍼지 집합의 함의 정도를 계산하여 결정할 수 있다. 이를 이용하여 임의의 두 분류의 유사관계를 동적으로 생성할 수 있는 자동화된 시스템을 구성할 수 있다.

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 에 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 α-cut 퍼지관계곱을 다음과 같이 정의하여 두 분류 퍼지 집합의 함의 관계, $C_i \xrightarrow[\alpha]{} C_j$, 를 계산한다.

$$C_i \xrightarrow[\alpha]{} C_j = \frac{1}{|C_{i\alpha}|} \sum_{K_k \in C_{i\alpha}} (R_{jk}^T \rightarrow R_{kj}) \quad (3)$$

여기서, K_k 는 k번째 검색어이고, C_i , C_j 는 i번째와 j번째 분류이며, $C_{i\alpha}$는 C_i의 α-cut이고 $|C_{i\alpha}|$ 는 $C_{i\alpha}$ 의 원소의 갯수이다. R 는 $m \times n$ 행렬로서 R_{ij} 는 $\mu_{Cj}(K_i)$,즉, $K_i \in C_j$ 인 정도이다. R^T 는 행렬 R 의 전치 행렬로서 $R_{ij} = R^T_{ji}$ 이다.

다음은 식(3)을 적용한 예이다.

예) α = 0.9 일때 $C_2 \xrightarrow[\alpha]{} C_3$ 는 0.94 이고 $C_1 \xrightarrow[\alpha]{} C_3$ 는 0.7 이다. 각 분류간 함의 관계는 α-cut 퍼지 관계곱 (3)에 의해 표1의 (a), (b) 와 같이 설정될 수 있다

표1 분류와 검색어의 α-cut 퍼지 관계곱

	C_1	C_2	C_3	C_4	C_5
K_1	0.1	0.1	1.0	0.0	0.1
K_2	1.0	1.0	0.8	0.2	1.0
K_3	1.0	0.1	0.0	1.0	1.0
K_4	1.0	0.0	1.0	0.0	0.8
K_5	1.0	1.0	1.0	0.1	1.0

(a) R

	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	1.00	0.82	0.76	0.94	1.00
C_5	0.98	0.64	0.76	0.48	0.94

(b) $C_i \xrightarrow[\alpha]{} C_j$

다음에 $C_i \xrightarrow[\alpha]{} C_j$ 를 α' 으로 α-cut 하여 크리스프 값으로 바꾼다. 표2의 (a)는 $C_i \xrightarrow[\alpha]{} C_j$ 를 α' = 0.94 로 , (b)는 α' = 0.76 로 α-cut 한 최종 결과이다.

표2 $C_i \xrightarrow[\alpha]{} C_j$ 를 α'으로 α-cut 한 최종결과

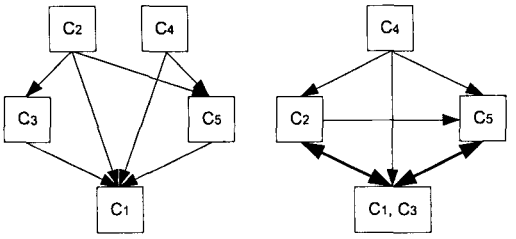
	C_1	C_2	C_3	C_4	C_5
C_1	1	0	0	0	0
C_2	1	1	1	0	1
C_3	1	0	1	0	0
C_4	1	0	0	1	1
C_5	1	0	0	0	1

(a) α-cut for α'=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	0	1	0	1

(b) α-cut for α'=0.76

α' = 0.94 일 때 분류간의 함의관계를 살펴보면, C_1 분류 항목은 모든 분류 항목의 하위분류이고, C_3 과 C_5 각각은 C_2, C_4의 하위분류이다. α' = 0.76 일 때는, C_4가 최상위 분류에 위치하며, C_1, C_3는 최하위 분류에 위치한다. α' = 0.94 일 때의 분류관계를 모두 포함하면서 확장된 것을 알 수 있다. 그림1와 같은 동적인 분류관계를 생성하면, 검색시 원하는 대상이 없을 때는 유사한 하위분류로 확장하여 검색할 수 있다.



(a) α'=0.94의 분류관계 (b) α'=0.75의 분류관계

그림1 최종 결과의 분류 관계도

4. 실험

실험의 비교 대상은 '야후 코리아'이며, 본 실험을 위해 α-cut 퍼지 관계곱을 사용하여 검색엔진을 구현하였다. 같은 조건에서 실험하기 위해 구현한 검색엔진의 데이터는 '야후 코리아'로부터 추출하였다. 실험은 150명의 비전산계열 대학생 1학년이 하였다. 실험 내용은 개인별 임의의 주제 5개를 선택하였고, 주제와 관련된 사이트의 개수와 관련 사이트가 나올 때까지의 하위 디렉토리의 단계 수로 재현율과 검색효율을 비교하였다.

전체 주제에 대한 적합한 사이트 수는 '야후 코리아'가 6,960개, 구현된 검색엔진이 7,934이었다. 최종 검색에 대한 하위 디렉토리의 평균단계는 '야후 코리아'가 6.22단계, 구현된 검색엔진은 4.3단계에 검색되었다. 제안한 방법이 검색 사이트의 재현율을 14%정도 향상시켰으며, 하위 디렉토리 접근 속도는 30.9% 향상 시켰다.

5. 결론

이 논문에서 우리는 분류 검색에 있어서 α-cut 퍼지 관계곱을 이용하여 각 분류의 유사한 하위분류를 찾아냄으로써 분류 검색의 재현율을 향상시키는 새로운 방법을 제안하였다. 실제로 시스템을 구현하였고 다음의 이점을 확인 하였다.

(1) 분류가 모호한 검색어에 대하여 유사한 하위분류로의 확장을 제공하여 검색을 용이하게 한다.

(2) 하위분류가 여러개의 상위 분류에 속할 수 있는 분류의 공유성과 분류 레벨의 유동성을 제공하여 검색 분류체계를 동적으로 관리할 수 있다.

(3) 에러의 한계를 규정하는 α-cut 값인 α 와 α' 을 다양하게 설정함으로써 분류 체계를 다양하게 변동시킬 수 있다.

본 논문에서 제시한 방법은 기업문서 관리 및 지식 포털, 도서 관리시스템, 컨텐트 관리 시스템 등 지능적 분류 방식을 필요로 하는 다양한 분야에 적용할 수 있다. 향후의 연구과제로는 동적 분류 체계에서 질의어 확장과 유사한 하위분류의 생성시간 단축에 대한 연구가 있다.

참고 문헌

[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] K.H. Lee and G.L. Oh. Fuzzy Theory and Application Volume I : Theory. HongReung Science Publishing Co., 1991.

[5] 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.

[6] 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.