제 19 회 한국정보처리학회 춘계학술발표대회 논문집 제 10 권 제 1 호 (2003. 5)

# 동적분류체계를 사용한 웹 검색엔진의 설계 및 구현

박선*, 최범기**
*인하대학교 전자계산공학과
**쿼크㈜
e-mail : sunpark@datamining.inha.ac.kr, bumghichoi@yahoo.co.kr

# Design and Implementation of Web Search Engine Using Dynamic Category Hierarchy

Sun Park*, Bumgi-Choi**
*School of Computer Science and Engineering, Inha University
**Quark Co., Ltd.

요 약

분류검색 방법은 색인검색 방법과 함께 중요한 요소로서 웹 검색 엔진에서 지원되고 있다. 색인검색 방법에서는 검색결과의 재현율이 높지만 검색결과가 너무 많이 나오기 때문에 원하는 검색결과를 찾아내는 것이 어렵다는 단점이 있다. 또한 능숙한 컴퓨터 사용자는 색인검색을 자주 사용하지만, 컴퓨터에 익숙하지 않은 대부분의 사람들은 분류검색 방법을 사용한다. 이러한 이유 때문에 검색엔진에서 분류검색 방법이 반드시 필요하다. 그러나 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 분류검색에 대한 문제점을 해결하기 위해서 분류와 검색어간의 관계를 퍼지논리를 이용하여 정량적으로 계산하고 이를 바탕으로 분류간의 함의관계를 유도함으로써 동적인 분류체계를 구성하는 새로운 웹 검색엔진을 설계하고 구현하였다. 구현된 검색엔진은 분류간의 함의관계를 유사한 하위분류로서 간주함으로써 분류검색 결과의 재현율을 높일 수 있다.

## 1. 서론

검색엔진에서 문서를 찾는 방법은 기본적으로 세가지 형태가 있다. 첫번째는 자동으로 수집된 웹문서의 한 부분을 색인하여 데이터베이스에 저장하고, 입력된 검색어와 일치하는 단어들을 데이터베이스에서 검색하여 검색어를 포함한 중요도가 큰 순서대로 문서를 나열하는 색인검색방법이다. 두 번째는 전문가에 의해 선별된 웹 문서를 주제별로 분류하고, 각 분류의 페이지가 문서들의 요약정보와 링크정보를 보여주고 여러 개의 하위 분류를 보여주는데, 찾고자 하는 문서를 해당 분류에서 찾아보고 만약 없으면 하위 분류로 범위를 축소시켜 주제와 일치하는 분류 경로(path)로 찾아가 원하는 정보를 디렉토리로 검색하는 분류검색방법이다. 세 번째는 하이퍼링크 구조를 이용하여 웹을 탐색하는 방법이 있으나, 현재 성능의 제약과 상용제품 부족등으로 넓게 사용되지는 못하나, 활발히 연구되고 있다[1].

색인검색은 검색어를 입력하여 색인된 모든 문서를 신속하게 찾을 수 있다는 장점이 있다. 그러나 단일 검색어나 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 위

하여 자동분류 방법이나 질의어 확장 등의 방법들이 연구되고 있다. 자동분류 방법은 인터넷에서 문서자료와 해당링크를 수집과 동시에 분류하거나 검색된 결과를 군집화하여 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 하위 분류로 범위를 축소해 나갈 수 있어서 빠르게 검색할 수 있고 자주 검색되는 중요한 정보들이 잘 정리되어 있어서 컴퓨터 사용에 익숙하지 않은 사람들이 널리 사용하고 있기 때문에 색인검색 방법의 보완적인 방법으로서 검색엔진에서 반드시 있어야 하는 기능이라고 볼 수 있다. 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. 이것은 분류검색방법이 분류체계를 다른 분류에 속한 하위분류들 간의 관계로 분석하여 자동으로 설정하는 적절한 방법이 없기 때문이다. 즉, 기존의 분류검색방법은 상위분류 아래에 그와 관련된 좀더 세분화된 주제의 하위분류를 수동으로 구성하는 고정계층구조로 되어 있다.

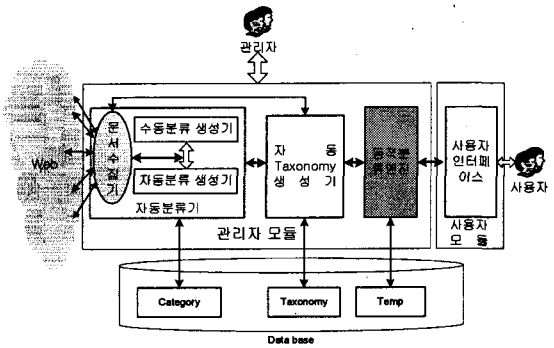
검색 엔진에서 문서 검색에 관련된 3 가지 객체는 분류, 검색어, 문서이다. 분류 검색이 문서와 분류간의 관계를 이용한 검색이라고 한다면, 색인 검색은 검색어와 문서의 관계를 이용하는 검색이다. 문서의 자동색인 기법은 검색어와 문서의 관계에 관련이 있으며, 문서의 수동 분류나 자동 분류는 문서와 분류와의 관계에 관련이 있다. 분류와 검색어는 각 검색 방법에서 각각 중요한 역할을 하고 있다. 따라서 분류검색 방법을 개선하여 검색결과의 효율을 높이기 위해서는 검색어와 분류사이의 관계를 규정하고 좀더 유연한 분류간의 관계를 설정하여 이를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

본 논문은 위와 같은 동기에서, 검색어와 분류 간의 관계를 규정하고, 분류들 간의 상호 관계를 규명함으로써 분류검색의 분류체계를 자동으로 동적인 체계로 재 구성함으로써 검색효율을 높일 수 있는 새로운 검색엔진을 설계 및 구현하였다.

본 논문의 구성은 다음과 같이 구성되어 있다. 2 장은 본 논문에서 제안한 검색 분류어의 동적인 분류를 위한 웹 검색엔진을의 설계를 보인다. 3 장에서는 이 검색엔진을 구현하고, 제 4 장에서는 결론과 향후 연구를 기술한다.

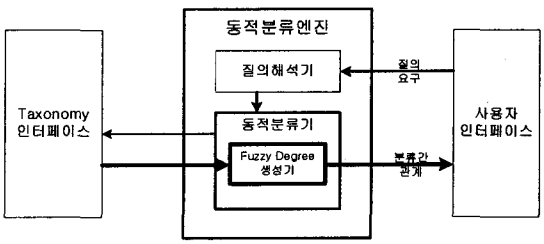
## 2. 검색 분류어의 동적인 분류를 위한 웹 검색엔진 설계

동적분류 검색엔진 시스템은 (그림 1)에서 보는 것과 같이 관리자 모듈과 사용모듈, 데이터베이스로 설계하였다. 관리자 모듈은 자동분류기, 자동 Taxonomy 생성기, 동적분류엔진으로 구성된다. 자동분류기는 문서수집기에서 수집된 문서들을 주제별로 분류하여 자동 또는 수동으로 데이터베이스를 구성한다. 자동 Taxonomy 생성기는 구축된 분류 데이터베이스로부터 Taxonomy 를 추출하다. 동적분류엔진은 사용자 인터페이스의 질의에 따라 fuzzy degree 를 생성하여 동적분류검색을 한다. 사용자 모듈은 사용자와 동적분류엔진과의 중계 역할을 한다. 데이터베이스는 자동분류기에서 생성되는 분류자료와 자동 Taxonomy 생성기에서 생성되는 Taxonomy 자료를 저장하며, Taxonomy 자료를 이용하여 동적분류엔진에서 동적으로 분류되는 임시 자료들은 Temp 데이터베이스에 저장된다.




(그림 1) 동적분류 검색엔진 시스템의 구조

동적분류엔진에서는 임의의 두 분류의 유사관계를 동적으로 생성하여 자동화 시스템을 지원하도록 설계한다. 동적분류엔진은 (그림 2)에서 보는 것과 같이 질의해석기와 동적분류기로 구성된다. 질의해석기는 사용자 인터페이스의 질의 요구사항을 해석하여 동적분류기에 전달한다. 동적분류기는 질의해석기의 사용자 질의에 따라 Taxonomy 인터페이스를 통해 해당 분류 자료를 추출한다. Fuzzy Degree 생성기는 추출한 분류자료에서 분류간 함의 정도를 계산하여 동적으로 분류간 관계를 임시 데이터 스키마로 구성한다. 계산된 분류관계는 사용자 인터페이스에 전달한다.




(그림 2) 동적 분류엔진

Fuzzy Degree 생성기에서 생성되는 두 분류간의 관계는 두 분류의 퍼지 집합의 함의 정도를 계산하여 결정한다. 두 퍼지집합의 함의 정도는 퍼지함의연산자를 사용하여 한 퍼지집합이 다른 퍼지집합에 포함되는 정도를 계산하고, 이를 이용하여 서로 다른 두 분류의 유사관계를 동적으로 생성한다. 퍼지 함의 연산

자는 각 응용의 필요성에 맞게 제시되어야 하는데 본 논문에서는 식(1)의 Kleen-Diense 퍼지 함의 연산자 [2,3,4]를 사용한다.

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

퍼지 함의 연산자를 식(2)에 적용하여 분류들 간의 퍼지함의관계, $C_i \rightarrow C_j$ 를 유도할 수 있다.

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

그러나 $C_i \rightarrow C_j$ 은 $C_i \subseteq C_j$ 의 정도를 나타내는 척도로서 약간 문제가 있다. 즉 $C_i$ 에 멤버쉽 값($\mu_{Ci}(x)$)이 작은 원소 $x$가 많으면, $C_i \subseteq C_j$ 의 포함여부와 관계없이 항상 1 에 가까운 값이 나오는 문제점이 있다. 따라서 식(2)를 식(3)과 같이 정의하여 두 분류 퍼지 집합의 함의 관계, $C_i \xrightarrow[\alpha]{} C_j$ ,를 계산한다.

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

여기서, $K_k$ 는 $k$번째 검색어이고, $C_i$, $C_j$ 는 $i$번째와 $j$ 번째 분류이며, $C_{i\alpha}$는 $C_i$ 의 $\alpha$-cut 이고 $|C_{i\alpha}|$ 는 $C_{i\alpha}$ 의 원소의 갯수이다. $R$ 는 $m \times n$ 행렬로서 $R_{ij}$ 는 $\mu_{C_j}(K_i)$ ,즉, $K_i \in C_j$ 인 정도이다. $R^T$ 는 행렬 $R$ 의 전치 행렬로서 $R_{ij} = R^T_{ji}$ 이다. 다음에 $C_i \xrightarrow[\alpha]{} C_j$ 를 $\alpha'$ 으로 $\alpha$-cut 하여 크리스프 값으로 바꾸어 동적인 분류 관계를 생성한다.

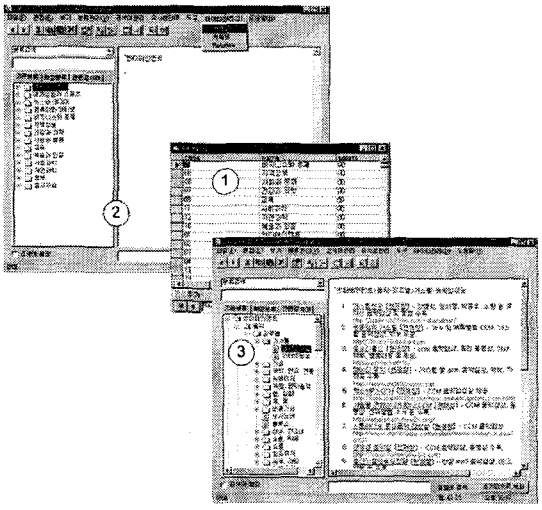
## 3. 동적분류 웹 검색엔진의 구현

본 장에서는 제안된 시스템을 구성하는 각 요소들의 구현내용을 설명한다. 웹 상의 문서를 수집하여 자동으로 분류를 생성하여 데이터베이스를 구축하는 자동분류프로그램, 구축된 분류 데이터베이스로부터 분류별 Taxonomy 데이터베이스를 구축하는 자동 Taxonomy 생성 프로그램, 분류별 Taxonomy 데이터베이스로부터 사용자의 질의에 따라 동적으로 분류검색을 하는 동적분류엔진 프로그램, 사용자 인터페이스 프로그램에 대해서 설명한다.

이러한 구현을 위하여 운영체제로 마이크로소프트사의 Windows 2000 Advanced Server, 데이터베이스는 마이크로소프트사의 SQL Server 2000, 동적분류엔진의 외부 인터페이스 지원을 위해 IIS 5.0 서버를 사용하였으며, 동적분류엔진은 마아크로소프트 Visual C++ 6.0 언어와 Java 를 사용하여 구현하였다.

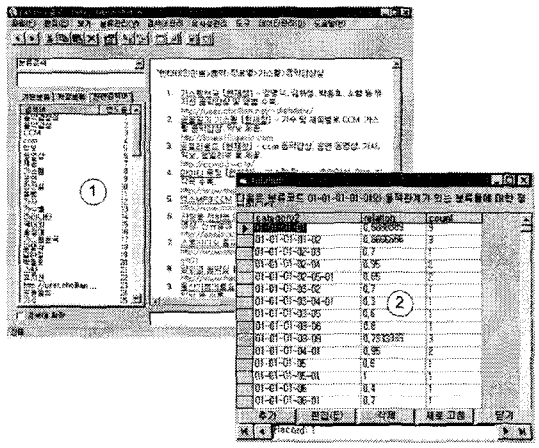
### 3.1 관리자 모듈의 구현

관리자 모듈은 사용자에게 검색서비스를 지원하기 위한 시스템을 구축한다. 관리자 모듈 프로그램은 (그림 4, 5, 6)과 같이 구성된다. 다음 (그림 4)는 자동분류기 모듈의 구현모습을 보여준다. (그림 4)의 ①은 최상위 디렉토리(대분류)를 수작업으로 입력하여 데이터베이스를 구성된 모습이다. 이와 같이 구성된 분류 데이터 베이스는 ②와 같은 인터페이스 형태로 최상위 디렉토리를 보여준다. 최상위 디렉토리와 관련된 하위 디렉토리(하위분류)는 자동으로 수집되어 데이터베이스를 구축하며, ③과 같은 형태로 구축되는 모습을 보여준다. 자동 하위 분류뿐만 아니라 ②의 인터페이스를 통하여 수작업으로 하위분류를 편집할 수 있다.




(그림 4) 자동 분류기

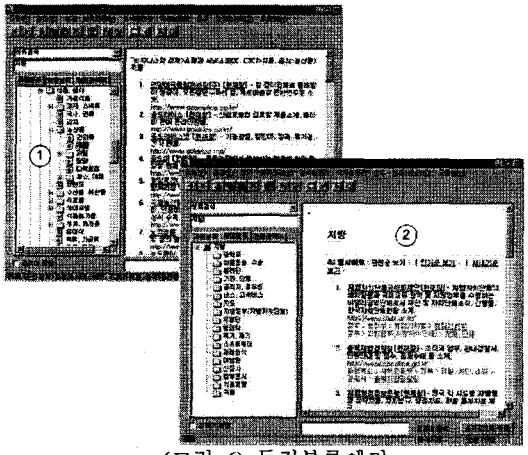
(그림 5)은 자동 Taxonomy 생성기와 Fuzzy Degree 생성기를 보여준다. (그램 5)의 ①은 (그림 4)의 ③에서 자동으로 생성된 분류 데이터베이스로부터 각각의 관련 사이트에 대한 Taxonomy 데이터베이스를 구성하는 모습을 보여준다. 이렇게 구성된 Taxonomy 데이터베이스는 사용자의 질의 요구에 따라 (그림 5)의 ②와 같은 분류간의 관계를 계산하여 보여준다.




(그림 5) 자동 검색어 생성기와 Fuzzy Degree 생성기

(그림 6)은 동적분류엔진을 보여준다. (그림 6)의

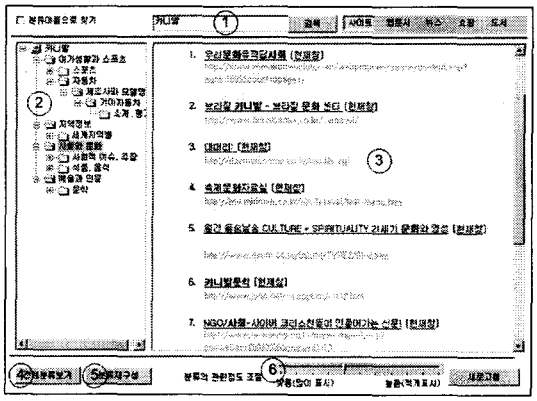
①은 지방에 대한 기본 디렉토리 검색결과를 보여준다. 기본 디렉토리 검색결과에 원하는 사이트가 없을 때 (그림 5)의 ②와 같이 Fuzzy Degree 생성기로부터 지방에 대한 분류간의 관계를 계산하여 동적분류한다. (그림 6)의 ②는 분류데이터베이스에서 지방과 관련된 분류들이 동적으로 분류되어 보여준다.




(그림 6) 동적분류엔진

### 3.2 사용자 인터페이스 모듈의 구현

사용자 인터페이스 모듈은 Visual C++로 구현하여 배포마법사에 의해 패키지로 만들었다. 처음 동적분류 엔진에 접속하는 사용자는 사용자 인터페이스 모듈을 다운받아 인터넷 브라우저에 프러그 인(Plug-in) 형태로 지원하도록 하였다.




(그림 7) 사용자 인터페이스

(그림 7)은 사용자 인터페이스를 보여준다. 사용자 인터페이스에서는 ①과 같이 검색단어를 입력 창에 입력하거나 ②의 전체 분류 TreeView 에서 분류를 선택함으로 사이트를 검색하는 것을 볼 수 있다. ③은 선택한 분류에 대한 관련 사이트를 보여준다. ④를 누르면 검색엔진의 전체 분류를 볼 수 있다. 만약, 색인 검색이나 분류 검색에서 원하는 검색결과가 나오지 않으면, 검색어나 분류를 선택 후 ⑤의 분류재구성 버튼을 눌러 관련 분류로 확장 할 수 있다. 관련 분류로 확장할 때 분류의 관련도를 ⑥의 선택 스크롤바로 조절할 수 이다. 스크롤바를 최고로 높게로 조정하면 분류어와 일치하는 사이트를 찾으며, 스크롤바를 낮음으로 조정하면 분류어와 관련이 있는 사이트들을 찾는다.

## 5. 결론

이 논문에서 우리는 분류검색의 효율을 높일 수 있는 검색엔진을 설계 및 구현하였다. 구현된 검색엔진은 새롭게 정의한 $\alpha$-cut 퍼지 관계곱을 이용하여 각 분류의 유사 하위분류를 찾아냄으로써 분류 검색의 재현율을 향상시켰으며, 다음과 같은 특징을 가진다. 첫째, 분류가 모호한 검색어에 대하여 유사한 하위분류로의 확장을 제공하여 검색을 용이하게 한다. 둘째, 하위분류가 여러개의 상위 분류에 속할 수 있는 분류의 공유성과 분류 레벨의 유동성을 제공하여 검색 분류체계를 동적으로 관리할 수 있다. 셋째, 분류의 관련도를 다양하게 설정함으로써 분류 체계를 다양하게 변동시킬 수 있다.

본 논문에서 구현한 검색엔진은 기업문서 관리 및 검색시스템, 도서 관리시스템, 상품 및 부품 관리시스템 등 지능적 분류 방식을 필요로 하는 다양한 분야에 적용할 수 있다.

향후 동적 분류 체계에서 확장된 질의어를 처리할 수 있는 방법에 대한 연구와 유사한 하위분류를 생성할 때 다양한 포함관계를 지원할 수 있는 퍼지함의 연산자에 대한 연구가 필요하다.

## 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.