질의 분류를 이용한 합성적 웹 문서 검색 서비스의 구현

박제현° 김연정 최철희 양재영 최중민
한양대학교
{jhpark°, yjkim, chchoi, jyyang, jmchoi }@cse.hanyang.ac.kr

Implementation of Merge-able Web Document Search Service by using User-Query Classification

Jeahyun Park° Yeonjung Kim Cheolhee Choi Jaeyoung Yang Joongmin Choi
Hanyang University

요 약

웹 문서 검색은 인터넷을 통해 정보를 얻는 가장 손쉬운 방법이지만, 사용자의 질의 의도를 충분히 반영하지 못한다는 단점을 가지고 있기도 하다. 이 논문에서는 사용자 의도를 문서 분류라는 방법을 통해 질의로부터 얻어 내고 문서 검색 결과와 문서 분류에 의한 결과를 적절한 방법으로 합성하여 사용자에게 결과로서 제시하는 시스템을 구현하였다.

1. 서 론

인터넷 정보 검색의 가장 대표적인 응용 분야로서, 웹 문서 검색은 이미 웹을 사용하는 데 있어서 필수 불가결한 기능이 되었다. 그러나 전통적인 문서 검색은 주로 문서 내에 포함된 질의문을 기준으로 결과를 제시하며, 이 과정에서 사용자의 의도가 제대로 반영되지 못한다는 단점을 갖고 있다. 다시 말해서, 질의어의 외양만으로는 사용자가 갖고 있는 개념을 정확하게 표현할 수 없는 단점을 가지고 있다.

본 논문에서는 사용자의 의도를 문서 분류에서 사용하는 개념(class)과 연결하고, 이를 통하여 사용자의 의도에 좀더 근접한 결과를 제시할 수 있는 시스템을 구현하였다. 본 시스템은 전통적인 정보 검색 기능을 수행할 수 있으며, 그 기반 위에서 문서 분류 작업을 추가적으로 수행한다. 수행결과로 사용자의 질의에 대한 문서 검색 결과와 분류 결과를 동시에 내놓으며 이 두 결과를 적절하게 합성하는 방법을 포함하고 있다.

2. CNutch

2.1 시스템 개요

전체 시스템은 크게 문서 검색 기능을 수행하는 부분과 문서를 분류하기 위한 부분으로 구성되어 있다. 문서를 검색하는 기능과 문서 분류 기능은 각각 독립적으로 수행되며, 사용자 질의에 대한 결과도 서로 다른 데이터를 참조하게 된다. 그러나 최종적으로 제시되는 결과는 논문에서 제시하는 합성 과정을 거쳐서 통합된 결과로서 제시된다.

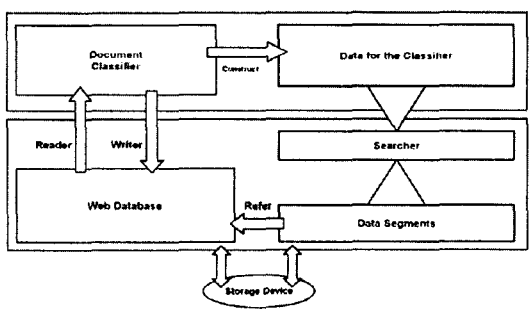

그림 1 CNutch 구조

2.1.1 문서 검색 기능

본 시스템에서는 문서 검색 기능을 수행하기 위하여 개방형 검색 엔진인 **Nutch**를 이용한다.[4] **Nutch**는 상업적인 검색 엔진으로 인한 정보의 독과점에 대한 대안으로 제시된 프로젝트이며 이 시스템에서는 기존의 **Nutch**에서 문서 수집과 인터페이스를 제외한 나머지 기능은 수정 없이 그대로 이용하였다.

2.1.2 문서 분류 기능

문서 분류 기능은 초기에 문서가 수집되는 단계에서 미리 정의된 개념 구조에 수집된 문서를 분류시키는 데

사용되고, 이후에 사용자 질의를 개념 구조에 대응시켜 단어 비교에서 얻을 수 없는 사용자 질의 의도를 파악하는 데 사용된다.

3. 시스템 구현

CNutch 시스템은 다음의 세 과정에 따라 작업을 수행한다.

3.1 학습 과정

3.1.1 특징 선택

문서 분류 기능을 수행하기 위하여 적절한 학습 데이터를 사용한 학습 과정이 필요하다. 학습 과정을 통해서 문서는 문서에 포함되어 있는 단어와 클래스 사이에서 존재하는 보이지 않는 관계를 수치적으로 계산한다. 일반적인 문서 검색이나 문서 분류에서 모든 단어를 고려치 않고 일정 수준 이상 관계가 있다고 판단되는 단어를 선택적으로 사용한다. 이 논문에서는 그 척도로서 χ^2 통계치를 사용하였다.[1]

$$\chi^2(t,c)=\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)}$$

위 식에서 **N**은 전체 문서의 수이며, **A**는 **t**를 포함하며 **c**에 속하는 문서의 수, **B**는 **t**를 포함하며 **c**에 속하지 않는 문서의 수, **C**는 **t**를 포함하지 않고 **c**에 속하는 문서의 수, **D**는 **t**를 포함하지 않고 **c**에 속하지 않는 문서의 수 이다.

3.1.2 분류 색인

CNutch에서는 분류를 위한 색인 구조를 별도로 유지한다.

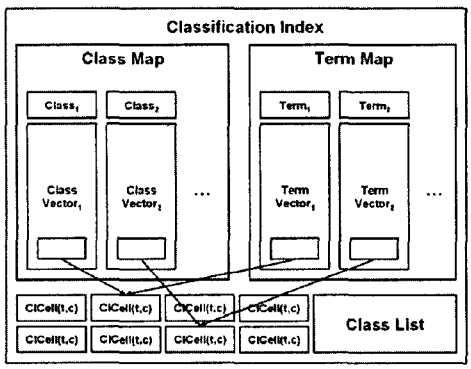

그림 2 Classification Index 구조

실제 데이터를 포함하는 **CICell**과 그에 대한 포인터를 통해 분류 과정에 필요한 수치를 찾을 수 있는 구조로서 설계되었다. 실제로 **CICell**에서는 학습 과정에서 선택된 특징들과 각 특징들이 갖는 기여도에 해당하는 수치들이 기록되어 있다.

3.2 분류 과정

3.2.1 베이지안 분류기

이 시스템에서 문서 분류는 베이지안 확률 모형을 사용하였다.[2] 문서 d_j를 임의의 클래스 c_i로 분류하는 기준은 d_j와 c_i가 단어 w_k의 모음으로 구성되어 있다고 할 때, 다음과 같은 조건부 확률의 형태로 나타난다.

$$P(c_i \mid d_j) \quad \begin{matrix} c_i=\{w_{i1},w_{i2},\Lambda,w_{i|T|}\} \\ \rho \\ d_j=<w_{j1},w_{j2},\Lambda,w_{j|T|}> \end{matrix} \qquad (1)$$

T: 모든 클래스에서 사용된 단어의 집합

이 시스템에서는 최상위 클래스를 그 문서가 해당하는 클래스로 간주하였다.

$$class(\vec{d}_j)=\arg\max_i \log_2 P(c_i \mid \vec{d}_j) \qquad (2)$$

3.2.2 분류 결과 데이터

문서 분류 과정은 문서를 수집하는 단계 이후에 일어난다. 분류된 문서는 질의 응답 과정에서 이용하기 위해서 어느 정도의 참여도(**membership**)를 가지고서 클래스에 분류되었는가를 기록해 두어야 한다. 모든 문서는 한 클래스에만 분류되며 그 때의 **membership**은 분류 시 가졌던 조건부 확률값을 그대로 이용하였다.

$$membership(\vec{d}_j,c_i)=\max_i P(c_i \mid \vec{d}_j) \qquad (3)$$

3.3 질의 응답 과정

3.3.1 질의 분류

전통적인 검색 방법의 문제점은 질의 의도를 단어 외양에 의존한다는 점이었다. 이 시스템에서는 사용자 질의 의도를 질의 문장이 속해 있는 개념에서 찾도록 하였다. 다시 말해서, 사용자 질의를 하나의 짧은 문서로 간주하고 이를 분류시킴으로써 검색 대상이 되는 문서를 만들어 냈던 의도 중 하나에 대응시킬 수 있도록 하였다.

식 (1)에서 문서 d_j를 단어들의 빈도수로 이루어진 벡터 형태로 모델링 하였듯이 사용자 질의 q 역시 같은 형태로 모델링한다.

$$\vec{q}=<w_{q1},w_{q2},\mathrm{L},w_{q|T|}> \qquad (4)$$

자연히, q가 속해있는 클래스도 같은 방법을 이용하여 구할 수 있다.

$$class(\vec{q})=\arg\max_i \log_2 P(c_i \mid \vec{q}) \qquad (5)$$

질의에 대한 분서 분류 결과는 질의가 속한 개념에 속한 문서 집합이 되며, 서열(**rank**)은 우선 각 문서가 해당하는 개념에 속하기 위해서 가진 참여도로서 주어진다.

3.3.2 결과 합성

정보 검색에서 도출된 웹 문서들과 문서 분류를 통해서 도출된 웹 문서들은 서로 합성되어 하나의 결과로서 제시되어야 한다. 합성은 결과가 갖는 순위의 재계산 과정이 된다.[3] 이 시스템에서는 두 가지 형태의 합성 과정을 사용하였다.

$$A = \{d \mid d \in \{R_R \cap R_C\}\}$$
$$B = \{d \mid d \in \{R_R - R_C\}\} \quad (6)$$
$$C = \{d \mid d \in \{R_C - R_R\}\}$$

여기서 R_R은 문서 검색에 의한 결과이며, R_C는 q가 속하는 분류에 속한 문서 집합이다. R_R에 속하는 문서 d는 초기 순위를 결정하기 위한 스코어를 가지고 있으며, R_C에 속하는 문서들은 클래스에 분류되며 가지는 참여도를 갖는다.

$$s_j : \text{score of } d_j \text{ in } R_R$$
$$m_j : \text{membership of } d_j \text{ in } R_C \quad (7)$$

이 때, B에 속한 문서의 m_j 값은 0이며, C에 속한 문서의 s_j 값도 0이다. 이제 R_R과 R_C에 속하는 문서들에 대한 클래스 참여도를 재계산한다.

$$m_j' = \frac{(m_j - m_{\min})(s_{\max} - s_{\min})}{(m_{\max} - m_{\min})} + s_{\min} \quad (8)$$
$$m_{\min} : \text{the minimum value of } m_j$$
$$m_{\max} : \text{the maximum value of } m_j$$
$$s_{\min} : \text{the minimum value of } s_j$$
$$s_{\max} : \text{the maximum value of } s_j$$

끝으로 사용자에게 제시되는 결과는 이 결과를 이용하여 두 결과를 합성한 결과가 제시된다.

3.3.2.1 이진 합성

이진 합성은 다음과 같은 합성 과정을 통해 이루어진다.

$$s_j' = \theta \cdot s_j + (1-\theta) \cdot m_j' \quad (9)$$
$$\theta : \text{user tuning factors, } 0 \le \theta \le 1$$

여기서 θ에 의해서 질의에 대한 응답이 문서 검색에 초점을 둔 것인지, 문서 분류에 초점을 둔 것인지 결정된다. 즉 θ가 1일 경우, 이 검색은 전적으로 문서 검색에 의존하며, 0일 경우 문서 분류에만 의존하게 된다.

3.3.2.2 수정 합성

수정 합성은 다음과 같은 합성 과정을 통해 이루어진다.

$$s_j'\big|_{d_j \in A} = \alpha(s_j + m_j')$$
$$s_j'\big|_{d_j \in B} = \beta \cdot s_j$$
$$s_j'\big|_{d_j \in C} = \gamma \cdot m_j'$$

이진 합성에서와 마찬가지로 α, β, γ는 사용자 인수이다. 이진 합성이 서로 상호 보완적인 특징을 갖는 반면에 수정 합성의 경우에는 다른 결과의 비중을 줄이지 않으면서 어느 한 결과에 대한 비중을 강조할 수 있는 특징을 갖고 있다.

4. 결론 및 향후 과제

이 논문에서는 문서 검색이라는 전통적인 정보 검색 기법과 문서 분류라는 지능형 문서 처리 기법을 접목시킴으로써 좀더 사용자의 의도에 부합하는 결과를 내놓을 수 있는 시스템을 구현하였다. 그리고, 다음과 같은 방향으로 좀더 나은 시스템으로 발전할 수 있는 여지를 갖고 있다. 첫 번째 개선 방안으로 CNutch에서는 단어 중심의 사용자 질의를 하나의 문서로서 간주하고 분류함으로써 사용자가 갖고 있는 질의 의도를 파악했지만, 사용자가 제시하는 단어는 단순히 단어의 외형만으로 판단하기에는 좀더 많은 의미를 포함하고 있다. 이런 숨어있는 의미를 추출할 수 있는 질의 분류법을 사용할 수 있다면 좀더 나은 검색 결과를 제시할 수 있을 것이다. 두 번째 방안으로 CNutch에서는 분류 체계 사이에 어떤 관계도 존재하지 않는다. 하지만, 온톨로지[5, 6]와 같이 계층 관계를 정의할 수 있는 체계 기술 구조를 이용하여 분류 체계를 정의할 수 있다면 지금보다 나은 검색 결과를 제시할 수 있을 것이다.

참고 문헌

[1] Yang, Y., Pedersen J.P. A Comparative Study on Feature Selection in Text Categorization, Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), pp412-420, 1997

[2] Fabrizio Sebastiani, Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol.34, No.1, pp.1-47, 2002

[3] Sang-Bum Kim, Hae-Chang Rim, Recomputation of Class Relevance Scores for Improving Text Classification, Lecture Notes in Computer Science(LNCS), 2004

[4] http://www.nutch.org

[5] http://www.w3.org/RDF

[6] http://www.w3.org/2004/OWL