자동 문서 분류를 위한 분류 주제어의 자동 증식 방법

정호석∪ 임종태 나혜숙 민철호
공주대학교 전산학과
hsjeong@mail.kribb.re.kr jtlim@kcs.kongju.ac.kr hsna@hsna.kict.re.kr churo@mail.kribb.re.kr

A Method of an Automatic Increment of Class Representatives for an Automatic Document Classification

Ho-Seok Jeong∪ Jong-Tae Lim Hei-Suk Na Chul-Ho Min
Dept. of Computer Science, KongJu University

요 약

현재의 자동 문서 분류 시스템에서는 문서분류는 지식베이스를 구축하고 전문가가 클래스의 분류 주제어를 수동 입력함으로써 이루어진다. 이것은 대단히 어렵고 번거로운 일이며 많은 시간과 노력이 소요되고 지속적으로 이루어지기 힘들다. 본 논문에서는 지식베이스와 문서의 구조적 정보, 통계적 정보, 키워드 간의 응집도를 이용하여 자동 문서 분류를 위한 분류 주제어의 자동 증식 방법을 제안한다.

1. 서론

현재 대부분의 자동화된 인터넷 문서 분류 시스템에서는 전문가가 각 클래스의 분류 주제어를 수동 입력함으로써 지식베이스를 구성한다. 이것은 많은 시간과 노력을 필요로 하며 지속적으로 이루어지기도 힘든 일이다.

본 논문에서는 문서 분류 시스템의 핵심인 분류 주제어를 자동으로 증식할 수 있는 방법을 제안한다. 하나의 문서를 키워드의 묶음으로 보고, 여러 문서들을 그 성격에 따라 각 클래스로 분류하고, 각 문서에서 분류 주제어에 포함되지 않는 단어들에 대해서는 후보 분류 주제어로 관리한다. 후보 분류 주제어들에 대해서 클래스 포함도를 계산하여 클래스와 관련성이 높은 후보 분류 주제어를 그 클래스의 분류 주제어에 추가한다.

본 논문의 구성은 다음과 같다. 먼저 제 2 장에서는 자동 문서 분류 및 분류 주제어의 자동증식에 관련된 연구에 대해 기술한다. 제 3 장에서는 시스템의 개요와 구조, 구현 기술에 대해 설명하고 자동 증식 시스템을 설계한다. 제 4장에서는 이 시스템에 대한 평가를 기술한다. 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

자동 문서 분류는 시스템에 등록된 문서를 특정한 기준에 따라 사람의 손을 거치지 않고 자동으로 여러 분류로 나누는 것이다. 자동 문서 분류 시스템은 통계적 방법 또는 문맥 정보 등을 이용하여 자동으로 문서들을 분류한다. 따라서, 단시간에 방대한 양의 자료를 처리할 수 있으며 관리 비용을 절감할 수 있다.

문서 분류기는 계층의 각 노드에서 분류 주제어를 사용해서 필터링 효과를 갖는다. 이것이 작동되는 기본 원리는 각 클래스를 계층적으로 하향 비교하면서 각 문서의 키워드들을 각 클래스의 분류 주제어와 비교해서 유사하다고 판단되면 그 클래스의 문서로 인정하는 것이다.

클래스는 동일한 계통의 정보(문서)의 대표 명칭이고, 분류 주제어(class representative)는 클래스의 키워드와 동의어, 시소러스를 포함해서 그 클래스의 성격을 규정짓는 단어 및 구를 말한다. 이 분류 주제어는 특정한 클래스에 속한다고 이미 알려진 문서들에서 발생하는 공통적인 용어들에서 추출한다[1].

문서를 자동 분류하기 위해서는 클래스와 분류 요청 문서의 유사도(similarity)를 계산해야 한다. Salton과 Buckley는 벡터 스페이스 모델을 이용한 코사인(cosine) 유사도 계산법을 개발하였고 Rijsbergen은 Dice 지수(Dice's coefficient)를 개발하였다[2, 3]. 이러한 통계적 방법의 단점을 Sparck Jones는 역문헌빈도를 제안하여 보완하였다[4]. 또, Rada는 동일한 문장이나 문서의 색인어 사이의 응집도(Cohesion)를 계산하였는데, 이것으로 문장이나 문서의 의미를 파악하고 식별할 수 있다[5].

3. 문서 분류 및 증식 시스템의 설계

3.1 개요

본 논문에서는 지식베이스를 이용하여 분류 주제어의 자동증식이 가능한 웹 문서의 문서 분류 시스템을 제안한다.

먼저 전문가가 분류 테이블을 만들고 분류할 문서들을 각 클래스에 대해서 수동 분류하고 클래스 별로 분류된 문서들에 대해서 키워드들을 수동으로 추출하고 해당 클

래스에 분류 주제어로 등록한다.

전문가의 작업이 끝나고 문서의 분류 요청을 받으면 시스템은 문서수집기에 의해서 수집된 문서를 토큰 단위로 분리하고 명사사전에 등록된 토큰에 한하여 문서 색인으로 등록한다. 문서와 각 클래스에 대한 유사도를 구하고 이를 근거로 문서를 클래스별로 분류한다.

분류된 문서 중 분류 주제어가 아닌 색인은 후보 분류 주제어로 등록하고 이 후보 분류 주제어 중에서 클래스에의 포함도를 평가해서 일정한 값 이상이면 분류 주제어로 포함시킨다.

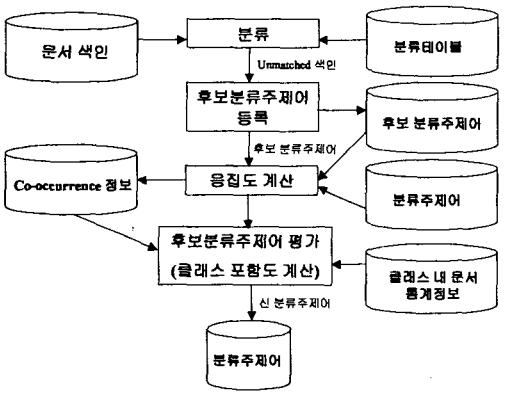

그림 1 : 분류 주제어 자동증식기 구조

3.2 분류

분류 요청된 문서가 있을 때 이 문서를 분류하는 것은 클래스와 문서와의 유사도를 계산함으로서 행해진다. 본 논문에서는 분류에 필요한 유사도 계산방법으로 Salton과 Buckley의 코사인(Cosine) 유사도 계산법을 이용한다[2].

유사도를 계산하여 일정한 임계값(threshold) 이상일 경우 새로운 클래스 멤버 문서로 인정한다. 문서를 분류할 때 계층적 성질이 많이 이용된다.

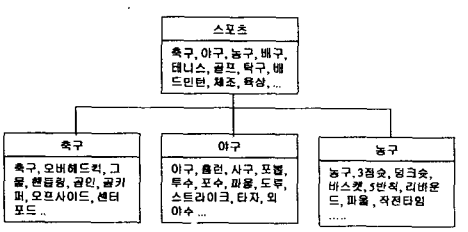

그림 2 : 2단계 분류 계층의 클래스와 분류 주제어

분류의 간단한 예로서 그림 2의 가상적인 2단계 분류 계층을 보자. 여기에서 스포츠, 축구, 야구, 농구는 클래스들이고 야구 클래스의 홈런, 포볼, 투수 . . . 등은 분류 주제어들이다. 상위 레벨은 스포츠 클래스이고 하위 레벨은 축구, 야구, 농구 3개이다. 먼저 스포츠 클래스와 문서를 비교해서 문서가 유효한 유사도를 갖는 것으로 판정되면 그것은 다음 단계로 3개의 하위 클래스와 비교될 것이다.

3.3 분류 주제어의 자동 증식

3.3.1 문서 내의 구조적 정보의 활용

각 문서는 독자적인 내부적인 문서구조를 가지고 있다. 특히 웹 문서의 구조적인 정보는 문서의 자동 분류의 성능을 향상시키기 위한 휴리스틱으로 활용 가능하다. 본 논문에서 이 성질을 이용해서 문서 내 분류 주제어에 대한 가중치를 다음과 같이 정의한다.

$$TF_{ij} = (1 + \omega T_{ij}) \times tf_{ij}$$

TF_{ij} : j번째 클래스의 i번째 분류 주제어의 가중치

tf_{ij} : j번째 클래스의 i번째 분류 주제어의 발생횟수

T_{ij} : j번째 클래스의 i번째 분류 주제어가 클래스 내의 문서 전체에서 TITLE, META 또는 IMG 태그 내에 나타난 횟수

ω : T_{ij}에 대한 가중치

3.3.2 Co-occurrence 정보의 이용

본 논문에서는 분류 주제어와 후보 분류주제어 간의 Co-occurrence 정보를 이용하여 분류 주제어를 증식하는데 이용한다. 이것은 동일 클래스 내의 문서들에서 같은 문서나 문장에서 분류 주제어와 동시에 자주 발생하는 후보 분류 주제어는 그 분류 주제어와 관련성이 높고 또한 클래스와도 관련성이 높다는 것을 의미하기 때문이다. 이를 계산하기 위하여 분류 주제어와 후보 분류주제어 간의 응집도(cohesion)를 구하여 증식에 이용한다[5].

3.3.3 역문헌빈도의 활용

본 논문에서는 Sparck Jones의 역문헌빈도(inverse document frequency)를 이용한다[4]. 클래스 C의 i번째 분류 주제어의 역문헌 빈도는 다음과 같다.

$$IDF_i = \log \frac{N}{n_i} + 1$$

n_i : 클래스 C의 i번째 분류 주제어가 나타난 문서의 수

N : 클래스 C의 전체 문서들의 수

클래스의 후보 분류주제어가 빈번하게 나타나는 문서일수록 높은 유사도를 나타낸다. 그러나, 여러 문서에서 많이 나타나는 단어는 대부분 중요하지 않은 단어인 경우가 많다.

3.3.4 클래스 포함도 계산

분류 작업이 끝난 후에 후보 분류 주제어들 중 클래스 포함도가 일정한 값 이상인 것은 새로운 분류 주제어로 인정하고 이를 클래스에 추가한다. 기본 방안은 클래스에 등록된 문서들에 지속적으로 자주 나타나며 기존 분류 주제어와 관련이 많은 후보 분류주제어는 그 클래스

의 분류 주제어로 포함시킬 수 있다는 것이다. 따라서, 후보 분류 주제어의 총 발생횟수와 후보 분류 주제어와 클래스 분류 주제어들 사이의 응집도가 높을 때 새로운 분류 주제어로 인정할 수 있다.

$$CI_i = \omega_1 \times ((1+\omega_3 T_i) \times tf_i \times IDF_i) + \omega_2 \times \frac{\sum_{j=1}^{m} COHESION(CW_i, SW_j)}{m}$$

CI_i : 클래스 C에 대한 후보 분류 주제어 i의 대한 클래스 포함도

T_i : 클래스 C의 문서들에서 발생한 후보 분류 주제어 i의 TITLE, META, IMG 태그 내 총 발생횟수

tf_i : 클래스 C의 문서들에서 발생한 후보 분류 주제어 i의 총 발생횟수

m : 클래스 C의 분류 주제어의 총 개수

IDF_i : 클래스 C의 문서들에서 발생한 후보 분류 주제어 i의 역문헌 빈도

ω_1 : 클래스 C의 문서들에서 발생한 후보 분류 주제어 i의 총 발생횟수에 대한 가중치

ω_2 : COHESION 값에 대한 가중치

ω_3 : T_i에 대한 가중치

$$\text{COHESION(ti, tj)} = \frac{co-occurrence\text{발생횟수}}{\sqrt{frequency(ti) \times frequency(tj)}}$$

그림 3 : 클래스 포함도의 계산

4. 평가

분류 주제어의 증식의 평가는 특정한 분야들의 문서들에 대하여 각 분야의 전문가가 수작업으로 증식한 분류 주제어에 대해서 시스템이 증식한 분류 주제어와 비교하여 재현율과 정확률로 성능을 평가한다.

$$\text{분류주제어 증식 재현율} = \frac{\sum_{i=1}^{n} \frac{\text{증식된 적합분류주제어의 수}}{\text{적합분류주제어의 수}}}{n}$$

$$\text{분류주제어 증식 정확률} = \frac{\sum_{i=1}^{n} \frac{\text{증식된 적합분류주제어의 수}}{\text{증식된 분류주제어의 수}}}{n}$$

n : 분류한 클래스의 수

적합분류주제어: 각 분야의 전문가가 수작업으로 증식한 분류 주제어

그림 4 : 분류 주제어의 증식의 평가 방법

본 논문의 증식 방법을 평가하기 위하여 PowerBuilder로 자동 분류 및 증식 시스템을 개발하였고, 야구 분야의 100개의 문서에 대하여 ω_1, ω_2, ω_3를 각각 1로 설정하고 클래스 포함도 10 이상의 후보 분류 주제어를 증식한 결과, 표 1의 결과를 얻었다.

표 1을 보면 대체로 재현율은 높지만 정확률이 낮다는 것을 알 수가 있다. 즉, 본 시스템은 전문가가 수작업으로 증식한 분류 주제어 중 절반 이상을 증식하지만, 부정확하게 증식된 분류주제어를 절반 이상 발생시킨다.

재현율	정확률
0.53	0.43

표 1 : 분류 주제어 증식의 평가 결과

자동 문서 분류 시스템의 성능은 분류 주제어의 정확성과 밀접한 관계가 있다. 따라서 부정확한 분류 주제어는 잘못된 분류를 일으킨다. 이것은 분류와 증식이 주로 통계적인 방법에 근거하여서 일상적인 용어가 적정 빈도의 용어로 채택되기가 쉽기 때문이다. 이러한 약점을 보완하기 위해서 의미적인(semantic) 문맥 정보가 반영되는 방법의 보충이 필요하다.

5. 결론

현재의 문서 분류 시스템에서 수동적으로 정의된 분류 주제어의 사용은 문맥 인식 필터링을 수행하고 정확도를 높인다. 그러나, 수동적인 분류 주제어의 입력은 많은 노력과 시간, 비용을 요구되고 지속적인 유지보수가 어렵다.

이에 비해 본 논문에서 제안된 분류기의 자동 증식 기능은 분류 시스템의 유지와 관리를 훨씬 쉽게 함으로써 비용 절감의 효과를 가져온다. 이것은 통계적인 메타데이타를 사용하고 웹 문서의 구조적 정보를 사용함으로써 가능하다. 또한 용어 정합(matching) 방식을 사용함으로써 형태소 분석의 시간을 줄이고 정확한 명사를 찾아온다. 본 시스템은 지식베이스를 이용하여 정보들을 특정 조직이나 단체의 관심 분야에 대해 독자적인 데이터베이스를 구축할 수 있게 한다.

앞으로, 사용자 인터페이스의 성향을 데이터로 관리하고 이를 분석하여 분류 주제어를 증식하는 데 이용하는 연구를 하고자 한다.

참고 문헌

[1] C. Jenkins, M. Jackson, P. Burden, and J. Wallis, "Automatic Classification of Web resources using Java and Dewey Decimal Classification", http://www.scit.wlv.ac.uk/~ex1253/classifier/, 1995

[2] G. Salton and C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval", Information Processing and Management, 24(5), 1988

[3] R. C. J. van Rijsbergen, Information Retrieval: Second Edition, Chapter 3, Butterworths, 1981

[4] K. Sparck Jones, "A Specification Interpretation of Term Specificity and Its Application in Retrieval", J. Documentation, 28(1), 1972

[5] R. Forsyth and R. Rada, Machine Learning-Applications in Export Systems and Information Retrieval, Ellis Horwood Series in Artificial Intelligence, England, 1986