카이제곱 통계량을 이용한 문서분류 자질 자동추출 방법

박종현*, 박소영**, 장준호**, 길태숙**
*상명대학교 컴퓨터과학과,**상명대학교 디지털미디어학부

Text Categorization Features Automatic Extraction Method Using Chi-squared Statistic

Jong-hyun Park*, So-young Park**, Juno Chang**, Tae-suk Kihl**
*Dept. of Computer Science, Sang-Myung University
**Dept. of Digital Media, Sang-Myung University
E-mail : ssoya@smu.ac.kr

요 약

문서에 포함되는 어휘는 문서 분류의 정보를 가지므로 문서를 분석하여 유용한 단어를 추출하는 것은 다양한 서비스와 연계되어 사용될 수 있어 매우 유용한 일이다. 문서 자동 분류에서는 분류자질 선정 방식에 따라 분류정확도가 서로 달라질 수 있으며, 문서에서 추출되는 유용한 단어에 따라 인지되는 분야가 달라질 수 있다. 이에 본 논문에서는 각 문서에 포함되는 단어에 대한 카이제곱 통계량 점수를 사용하여 단어별 문서 분류에 대한 단어의 자질을 평가하고 문서의 분류별 유용한 단어를 자동 추출하는 방법을 제안하고 개발한다.

키워드

카이제곱통계량, 문서분류, 자동추출

Ⅰ. 서 론

컴퓨터 과학이 발달하면서 웹(www)과 같은 매일 대량의 정보들이 만들어지고 있는 현대의 문서들에 대한 검색에서 각 문서분류에서 유용한 단어를 파악하는 것은 매우 유용한 일이다. 문서의 분류란 정해진 분류체계 하에서 분류하고자 하는 각 문헌들을 가장 적합한 카테고리에 배정함으로써 문헌을 집단화하는 작업이다[1]. 분류된 문서에 포함된 어휘들은 문서 분류의 정보를 가지고 있으며, 문서가 가진 정보를 표현한다.

문서 자동 분류는 문서의 검색 및 문서에서 데이터 추출 등의 다양한 서비스에 적용되어 사용될 수 있다[2].

문서 자동 분류에서는 분류 자질 선정방식에 따라 분류정확도가 서로 달라질 수 있으며, 대부분의 문서 자동 분류에서는 학습이나 통계 혹은 확률 자료에 의한 것과 키워드들간의 관계를 벡터수치로 표현하고 적용하여 문서를 분류하고 있다[3]. 각 문서의 분류에서 사용되는 키워드는 문서 분류에 중요한 단초를 제시한다.

본 논문에서는 문서에서 추출된 단어의 출현여부와 관련이 없는 분류에서의 단어 출현여부를 측정하여 가중치 점수를 부여하여, 각 단어의 해당 분류에 대한 기대치와 관측치를 측정하는 카이 제곱 통계량을 이용한 유용한 단어를 자동으로 추출하는 방법을 제안하고, 그를 위한 모듈을 개발한다.

2장에서는 문서의 범주화와 관련된 자질로 단어를 선택하는 관련 연구에 대해 논의하고, 3장에서는 카이스퀘어제곱 통계량에 의한 자동 추출 방법을 제안하고 성능 향상을 위한 방안을 모색한다. 마지막으로 4장에서 결론 및 향후 연구 과제를 정리한다.

Ⅱ. 관련연구

기존에 문서 관리를 위한 자동 문서분류에 대한 연구는 많이 이루어져 왔으며, 각 문서 분류에 대한 단어의 자질을 평가하는 연구 또한 많이 이루어져 왔다.

[5]는 자질 선택에 있어 자료 내의 특정 문서에서 어떤 단어의 중요도를 평가하기 위해 사용되는 통계적인 수치를 이용하는 TF-IDF 방법을 설명하고 있다. 단어의 중요도는 문서 내에서 해당 단어가 많이 나타날수록 증가하며, 전체 자료 내에서 해당단어가 많이 나타날수록 감소한다.

상호정보량을 이용하여 문서에서 특정 단어의 출현이 문서의 범주에 포함이 되는지 여부를 예측하는데 제공한다[6]. 특정 범주에서 특정 단어가 많이 출현할수록 범주에서의 단어 정보량이 크다. 하지만 단어의 전체 출현 빈도가 낮을수록 상호정보량이 높아지는 단점이 있다[7].

정보획득량은 문서에서의 출현 빈도뿐 아니라 출현하지 않은 빈도까지 고려해서 각 범주에서의 용어 정보량을 계산한다. 모든 범주의 평균값으로 특정 문서에서 나타나는 모든 용어들의 정보 획득량을 계산하여 일정 임계값 이상의 값을 가지는 용어들만을 자질로 선택하게 된다[7]. 연관성을 측정하는 유사성 함수로서의 기본 성질을 만족하지 못하며 지나치게 저빈도인 경우를 선호하는 것이 제한점이 된다[8].

Ⅲ. 카이 제곱 통계량

카이 제곱 통계량은 관측된 값들과 이론에 의해 예측된 값을 비교하여 적합도를 검정하는 것이다. 통계적 자료에서 범주적 자료의 경우에 범주별로 빈도를 측정하게 되는데 이때, 카이 제곱 통계량을 일반적으로 많이 사용하게 된다.

카이 제곱 통계량은 특정 단어 t와 범주 c간의 비독립성을 측정하는 것으로, 자유도가 1인 카이 제곱 분포와 비교될 수 있으며, 다음 식 1과 같이 계산되어진다.[9]

$$\chi^2(t,c) \approx \frac{(A+B+C+D)(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (1)$$

표 2. 식의 표현

		문서가 분류 c와	
		관련있다	관련없다
단어 t가 문서에	포함	A	B
	미포함	C	D

계산된 카이 제곱 통계량은 특정 단어 t에 대해 각 분류에 대한 자질 점수를 의미한다. 본 논문에서는 (식 1)에 의해 계산된 특정단어에 대해 각 문서 분류별로 자질 점수를 순위화한 후 상위 순위의 자질을 선택한다.

Ⅳ. 문서분류 자질 자동추출 모듈

본 논문에서 제시하는 문서분류별 단어의 자질을 자동으로 추출하는 것은 문서의 검색 및 문서에서 유용한 데이터의 획득 등의 다양한 서비스에 적용될 수 있다[3]. 타 서비스와의 원활한 협업이나 연계를 위하여 XML 형태로 데이터를 주고 받을 수 있게 설계되었으며, 이식성을 고려하여 Java 기반으로 제작되었다.

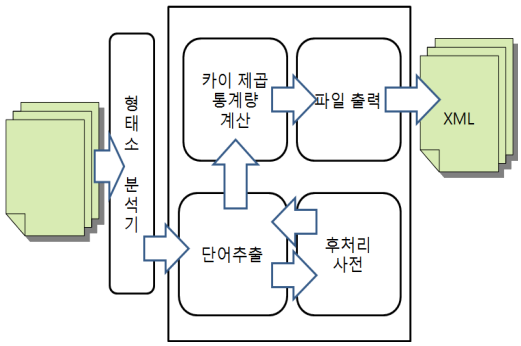

그림 1. 문서분류 자질 자동추출 모듈구성

제작된 모듈의 데이터 이동과 절차는 다음과 같다. 첫째, 분류된 문서의 정보와 설명을 형태소 분석기[10]를 이용하여 단어를 자동으로 추출한다. 최근 신조어나 새로운 상품명, 상호를 견고하게 분석할 수 있도록 후처리 사전을 두어 형태소 분석기의 결과에서 분석을 하지 못한 형태소에 대해 원활히 분석하도록 한다.

둘째, 추출된 단어에 대해 카이제곱 통계량을 계산한다. 계산법은 (식 1)에서 설명한 것과 같이 단어에 대해 문서 분류의 포함여부와, 문서에서의 출현 여부로 판단하여 계산하였다.

모듈의 테스트를 위하여 스마트폰 애플리케이션의 앱스토어에서 관련 분류와 설명을 이용했으며, 문서분류 32개, 애플리케이션 105개에 대한 설명에서 유용한 단어를 자동 추출하였다. 아래의 [표 2]은 모듈의 실행결과로 나온 카이 제곱 통계량 점수 중 “쇼핑” 분류에서 일부 발췌한 것이다.

분류에 유용한 품사인 일반명사, 고유명사, 동사, 형용사 등의 품사는 포함하고, 의존명사, 대명사, 수사, 보조용언 등의 품사는 제외하여 2590개의 단어가 추출되었다. [표 2]의 결과는 각 문서 분류별 카이 제곱 통계량의 점수별로 내림차순으로 정렬하여 상위 10개를 추출한 것이다.

표 2. 카이 제곱 통계량 결과

순위	단 어	카이제곱통계량
1	쿠폰/NNG	50.24079
2	마켓/NNG	37.444855
3	몰/NNG	26.823725
4	쇼핑/NNG	26.823725
5	쇼핑몰/NNG	26.823725
6	11번가/NNG	24.720873
7	가맹점/NNG	24.720873
8	디앤샵/NF	24.720873
9	옥션/NF	24.720873
10	의류/NNG	24.720873

본 실험 결과에서 [표 2]의 “쇼핑” 문서 분류에 포함된 애플리케이션의 수는 8개 였으며, 결과를 보면 알수 있듯이 점수가 같은 단어가 많이 있음을 알 수 있다. 실험에 사용된 문서 분류 중에서 문서가 6~7개 이상이 포함된 분류에서는 신뢰적인 결과가 나왔으나, 그 이하의 개수를 포함한 문서 분류에서는 카이 제곱 통계량의 점수에 따른 유용한 단어의 추출 성능이 급격히 감소함을 볼 수 있었다.

Ⅴ. 결 론

본 논문에서는 카이 제곱 통계량을 이용하여 분류된 문서에서 각 단어별 자질을 계산하여 문서 분류별 유용한 단어를 자동을 추출하는 방법을 제안하고 모듈을 개발하였다. 카이 제곱 통계량을 이용한 방법은 포함된 문서와 관계없는 분류에 대한 관계도 포함되어 계산하였다. 그러나 실험에 사용된 문서의 개수가 결과를 평가하기엔 그 문서의 수가 부족했다.

본 논문에서 제안한 방법을 기초로 한 모듈이 다른 시스템과 연계되어 사용될 수 있도록 현실적인 방법을 모색하고, 더 많은 문서의 실험으로 유용성을 평가하여 발전시켜나갈 계획이다.

참고문헌

[1] 조광제, 김준태, “역카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동 분류”, 한국정보과학회 학술 발표논문집, 1997.

[2] David D. Lewis, "Feature Selection and Feature Extraction for Text Categorization", Proc. Speech and Natural Language Workshop, 212-217, Morgan Kaufmann, 1992.

[3] 최동시, 정경택, “카테고리와 키워드의 밀접성 정보에 의한 문서 자동 분류 시스템 설계 및 구현”, 한국정보과학회 학술발표논문집, 1995.

[4] David D. Lewis, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task", ACM SIGIR conference on Research and development in information retrieval, p.37-50, 1992.

[5] Stephen Robertson, "Understanding inverse document frequency : on theoretical arguments for IDF", Jounal of Documentation, Vol. 60 lss : 5, pp.503-520, 2004

[6] Yang. Y, J. O. Pederson, "A comparative study on feature selection in text categorization." Proceeding of the 14th International Conference on Machine learning, 1997.

[7] 고영중, 서정연, “문서관리를 위한 자동문서 범주화에 대한 이론 및 기법”,정보관리 연구, vol. 33, no. 2, 2002.

[8] 이재윤, “상호정보량의 정규화에 대한 연구”, 한국문헌정보학회지, 2003.

[9] 김경선, 서정연, “정보과학회 논문지”, 소프트웨어 및 응용, 2003.

[10] 류원호, 이상주, 임해창, “통계적 결정 그래프 학습 방법을 이용한 한국어 품사 부착 오류 수정”, 한국정보과학회 학술발표 논문집, 2001.