유해어 필터링을 위한 자질어 추출 알고리즘에 관한 연구

정정훈[O] 이원휘 이신원 안동언 정성종
전북대학교 대학원 컴퓨터공학과
jhjeong@geonji.co.kr,{wony, swlee9237, duan, sjchung}@chonbuk.ac.kr

Study of Feature Extraction Algorithm for Harmful word Filtering

Junghoon Jeong[0], Wonhee Lee, Shinwon Lee, Dongun An, Sungjong Chung
Dept. of Computer Engineering, Chonbuk National University

요 약

유해 정보란 정보의 홍수 속에서 무차별적으로 제공되는 음란, 폭력 등의 내용을 담고 있는 정보를 말한다. 이러한 유해 정보들로부터 청소년 등 사회적으로 보호를 받아야 할 인터넷 이용자들을 보호하기위한 장치가 필요하다. 현재 다양한 방법이 제안되고 연구되고 있다. 본 연구에서는 유해 문서의 필터링을 기법 중 키워드 필터링에서 사용되는 유해어 사전을 위한 자질어 추출 알고리즘에 대해서 비교/연구하였다. 키워드 필터링에서 자질어는 필터링의 성능에 많은 영향을 미친다. 따라서 필터링의 성능을 높이기 위한 자질어 추출 알고리즘 선택은 매우 중요하다. 이에 본 논문에서는 다양한 알고리즘을 비교 분석하여 정확하고 효율적인 자질어 추출 알고리즘 조합을 찾고자 하였다. 그 결과 CHI/TF-IDF 조합이 높은 성능을 보였으며 92%의 정확도를 얻을 수 있었다.

1. 서 론

오늘날의 정보환경은 웹을 통한 정보가 홍수를 이루는 환경을 이루고 있다. 사람들은 원하는 정보를 웹을 통해 얻고, 제공하고자 하는 정보 또한 웹을 통해 제공하고 있다. 그러나 이러한 웹의 편리성은 그 이면에 무분별한 정보의 제공으로 인한 여러 가지 문제를 안고 있다. 너무 많은 정보의 제공으로 인한 정보 검색의 부담과 무분별한 유해 정보의 범람은 정보사회를 살고 있는 우리에게 커다란 문제가 되고 있다. 특히 음란, 폭력, 자살 등 유해 정보는 사회적으로 보호를 받아야 할 청소년들을 비롯한 판단력과 절제력이 부족한 인터넷 이용자들에게 심각한 사회적 문제를 야기하고 있다[7].

현재 이러한 문제를 해결하기 위한 제도 및 연구가 다양한 방법으로 이루어지고 있다. 게시자(publisher)의 자발적 등급 결정에 기반한 인터넷 내용 선택에 대한 플렛폼(PICS), 영상정보의 스킨컬러(skin color)에 기반한 연구, 유해한 단어나 구에 기반하여 필터링하는 키워드 필터링, 신경망 이론을 응용한 지능적 내용 분류 연구, 이미지 정보를 이용한 연구 등이 그것이다[5].

텍스트에 기반한 키워드 필터링의 경우, 적당한 자질어의 추출은 필터링 성능에 많은 영향을 미친다. 따라서 정확하고 효과적인 자질어의 추출을 위한 알고리즘 선택이 중요하다. 본 연구에서는 다양한 알고리즘에서 최적의 자질 추출 알고리즘과 인덱싱 알고리즘의 조합을 찾고자 한다. 이를 위해 2장에서 자질어 추출 알고리즘 및 인덱싱 기법에 대해서 알아보고 3장에서 각 알고리즘을 이용한 실험 및 평가를 한다. 4장에서 결론을 맺는다.

2. 관련 연구

2.1 자질어 추출 알고리즘

자질어 추출은 키워드 필터링에서 사용될 자질어의 목록을 생성에 필요한 부분이다. 자질어 추출은 학습대상 문서(웹 문서)를 파싱하고 형태소 분석의 전처리 단계를 거친 후의 데이터를 이용한다. 자질어 추출 알고리즘으로는 DF(Document Frequency), IG(Information Gain), MI(Mutual Information), CHI 등 다양한 알고리즘이 존재한다. 이들 각각의 알고리즘은 다음과 같다.

DF(Document Frequency)는 전체 문서 집합 중 특정 단어가 출현한 문서의 수를 의미한다. 본 연구에서는 이 DF를 이용하여 자질어를 추출하고 일정 임계치 이하의 문서에서 출현하는 용어를 제거한다. 이 때 "문서 빈도가 아주 작은 용어는 특정 주제 범주를 대표할 만한 충분한 정보가 되지 못하고 전체적인 성능에도 큰 영향을 미치지 못한다"는 기본 가정을 가지고 출발한다. 이 알고리즘은 매우 간단하고 계산량이 적다는 장점을 가지고 있는 반면 정보검색 분야에서 전통적으로 문서 빈도 값이 낮을수록 색인어로서의 가중치를 높게 할당하는 것과 대치되는 단점을 가지고 있다.

IG(Information Gain)는 특정 단어의 출현 여부가 문서 분류에 기여하는 정도를 계산하기 위하여 기여도가 높은 자질만을 선택하는 알고리즘으로 모든 용어들의 정보 획득량을 계산하여 일정 임계치 이상의 값을 갖는 용어들만을 자질로 선택하게 된다. 이 방법은 문서에서의 출현 빈도뿐만 아니라 출현하지 않은 빈도까지 고려하여 각 범주에서의 용어 정보량을 계산한다. 범주 집합이 {C1,C2, ..., Cn}일 때 IG의 알고리즘은 다음과 같다.

$$IG(t) = -\sum_{i=1}^{m} \Pr(C_i)\log\Pr(C_i) + \Pr(t)\sum_{i=1}^{m} \Pr(C_i|t)\log\Pr(C_i|t) + \Pr(\neg t)\sum_{i=1}^{m} \Pr(C_i|\neg t)\log\Pr(C_i|\neg t)$$

MI(Mutual Information)는 두 용어 중의 한 용어가 다른 용어에 대해 갖고 있는 정보량을 이용하는 방법으로 두 용어 중 한 용어가 출현했다는 사건이 다른 용어의 출현 여부를 예측하는데 기여하는 정도를 수치적으로 나타낸 값이 된다. 범주 c에서 용어 t가 많이 출현할수록 범주 c에서의 용어 t의 MI(t,c)의 값은 크다. A가 범주 c에 속한 문서 중 용어 t를 가진 문서의 개수라 하고, B가 범주 c에 속하지 않는 문서 중 용어 T를 가진 문서의 개수, C를 범주 c에 속한 문서 중 용어 t가 없는 문서의 개수라 할 때 MI값과 평균값, 최대값은 각각 다음과 같이 구해진다.

$$MI(t,c)=\log\frac{\Pr(t\wedge c)}{\Pr(t)*\Pr(c)}\approx\log\frac{A*N}{(A+C)*(A+B)}$$

$$MI_{avg}(t)=\sum_{i=1}^{m}\Pr(C_i)MI(t,c_i)$$

$$MI_{\max}(t)=Max_{i=1}^{m}\{MI(t,c_i)\}$$

CHI는 용어 t와 범주 c간의 의존성을 측정해 용어의 중요도를 구하는 방법으로 t와 c 두 값의 차가 클수록 용어 t가 자질로 선정될 확률이 높아진다. 또한 문서 빈도를 사용해 범부별 발생분포가 일반적인 단어들의 발생분포와 다른 정도를 계산하고, 그 차이가 특정 값 이상인 단어를 자질로 선정하게 된다. 최종 카이제곱 통계량을 구하기 위해서는 각 용어 및 범주에 대해 통계값을 계산한 후 각 용어마다 계산된 카이제곱 통계량의 평균이나 최대값을 구한다. 알고리즘은 범주 c에 포함된 문서에서 단어 t가 출현한 문서를 A, 출현하지 않은 문서를 C, 번주 c에 속하지 않은 문서에서 단어 t가 출현한 문서를 B, 출현하지 않은 문서를 D라 할 때

$$x^2(t,c)=\frac{N*(AC-CB)_2}{(A+C)*(B+D)*(A+B)*(C+D)}$$

단, $N=A+B+C+D$

$$x^2avg(t)=\sum_{i=1}^{m}\Pr(C_i)x^2(t,c_i)$$

$$x_2\max(t)=Max_{i=1}^{m}\{x^2(t,c_i)\}$$

이다. 이 때 N=A+B+C+D이다.

2.2 인덱싱

자질어 추출 과정에서 추출된 자질들이 문서에서 차지하는 비중을 계산하는 가중치 부여 작업이 필요하다. 즉, 한 문서의 특징을 표현하기 위해 가중치가 부여된 자질어를 이용하여 문서를 벡터화해야 한다. 인덱싱 및 가중치 작업에는 TF, TF-IDF, TF-ICF등의 알고리즘이 사용된다.

TF에 의한 가중치 계산 방법은 단순히 용어가 한 문서 내에서 나온 빈도수, 즉 Term Frequency를 사용한다. TF에 의한 가중치 계산에도 단순TF, 이진TF, 로그TF등이 있는데 본 연구에서는 tf가 너무 낮은 단어의 지나치게 낮은 영향력을 보완하고 너무 높은 tf의 영향력을 낮추기 위하여 로그TF를 사용하였다.

$$TF=1+\log(tf)$$

여기서 tf는 용어가 한 문서 내에서 나온 빈도수이며, TF는 적절히 변형된 빈도수를 의미한다.

TF만으로는 고빈도의 용어가 항상 문서를 대표하지 않고 대부분의 고빈도 용어는 기능어로써 많이 등장하지만, 그 문서의 내용을 나타내지 못하는 맹점을 가지고 있다. 그래서 TF-IDF(Inverse Document Frequency)와 TF-ICF(inverse Category Frequency)를 사용한다.

TF-IDF는 적은 수의 문서에 나타난 자질어에 대해 높은 가중치를 부여하는 방법으로 알고리즘은 다음과 같다.

$$IDF=\log N-\log DF_j+1$$
$$Weight=TF*IDF$$

여기서 DF_j는 자질어 w_i를 포함하는 문서의 개수이며 N은 총 문서의 개수가 된다.

ICF는 소수의 범주에 많이 나오는 용어에 높은 가중치를 부여하고, 여러 범주에 고르게 나오는 용어에는 낮은 가중치를 부여하는 방법으로 알고리즘은 다음과 같다.

$$ICF=\log M-\log CF_j+1$$
$$Weight=TF*ICF$$

여기서 M은 범주의 총 수이며, CF_j는 자질어 w_i를 포함하는 범주 수이다.

위 인덱싱 과정을 거친 가중치 값은 SVM의 성능을 높이기 위하여 값들을 정규화하는 과정이 필요하다. 정규화는 가중치값은 -1과 1사이의 값으로 정규화 하였다.

3. 실험 및 평가

실험 및 평가 환경은 다음과 같다.

CPU	Intel Pentium 4 Xeon 2.0 GHz (2 CPU Physical, 4 CPU Logical)
RAM	2GB
HDD	80GB
OS	RedHat Linux Fedora Core 3 (Kernel Version : 2.6.11)
Compiler	gcc 3.4.4

[표 1] 실험 및 평가 환경

사용된 데이터 셋은 ETRI에서 텍스트 기반 유해 문서 셋으로 구축한 EHDS-20000(ETRI Harmful Data Set)을 이용하였다. 위 문서 셋에서 실험 및 평가에 사용된 학습 데이터 및 테스트 데이터는 다음과 같다.

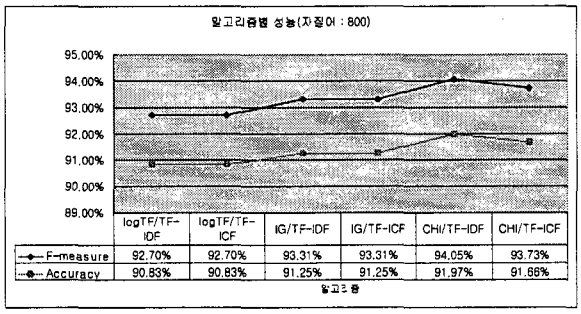
		수집데이터	
		무해	유해
문서	한글	2126	2572
	영문	3340	12250

[표 2] 데이터 셋

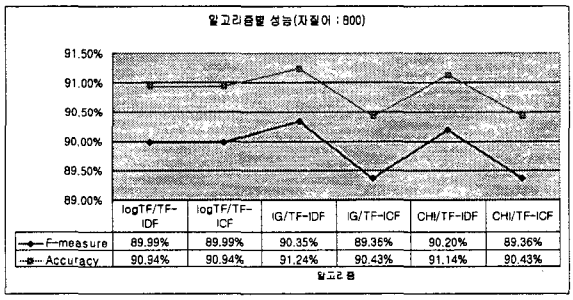
실험은 자질어 추출 알고리즘에 logTF, IG, CHI과 인덱싱 알고리즘에 TF-IDF, TF-ICF를 조합하여 자질어 수를 200과 800사이의 값을 100씩 조정하면서 실시하였다. 실험 결과 각각의 측정값은 다음 그림과 같이 나타났다. 성능은 정확도와 F-measure로 평가하였다.

$$F=\frac{2*P*r}{P+r}$$

한글과 영문으로 나누어 자질어 수 800의 경우 결과는 다음 [그림 1]과 [그림 2]과 같다.

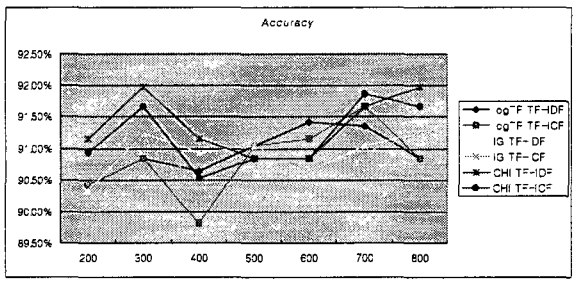


[그림 1] 한글 자질어 수 800의 경우 성능



[그림 2] 영어 자질어 수 800의 경우 성능

결과에서 자질어 수는 800에 CHI와 TF-IDF 조합이 가장 좋은 성능을 나타냄을 알 수 있다([그림 6] 참조).



[그림 3] 자질어 추출 알고리즘 비교

위 결과에서 보듯이 자질어의 추출은 알고리즘에 따라 많은 차이를 보이고 있다.

4. 결론

현재 인터넷 환경은 정보의 홍수와 함께 음란, 폭력, 자살 등의 유해 정보가 범람하는 환경으로 되고 있다. 이들을 해결하기 위하여 다양한 방법이 제안되고 연구되고 있다. 이들 방법 중 키워드 필터링은 비교적 높은 필터링 성능을 보여주고 있으나 과분류(over-blocking) 등의 문제를 안고 있다. 키워드 필터링의 성능은 자질어의 추출에 큰 영향을 받는다. 이에 본 연구에서는 자질어 추출을 위한 알고리즘들을 비교 분석함으로써 최선의 결과를 찾고자 하였다. 결과 평균 CHI/TF-IDF 조합이 가장 높은 성능을 보였고 92%의 정확율을 얻을 수 있었다.

향후 단순히 키워드 필터링만으로는 성능 개선에 한계가 있으므로 지능적 내용 분류 등의 기법과 병행한 연구가 필요하다.

Reference

[1] Mohamed Hammami, Youssef Chahir, and Liming Chen, "WebGuard:A Web filtering Engine Combining Textual, Structural, and Visual Content-Based Analysis", IEEE Transaction On Knowledge and Data Engineering, Vol. 18, No. 2, February 2006

[2] Huicheng Zheng, Hongmei Liu, Mohamed Daoudi, "Blocking Objectionable Image : Adult Images and Harmful Symbols", IEEE International Conference on Multimedia and Expo(ICME), pp1223-1226, Jun 2004

[3] M. Hammami, Y.Chahir, and L.Chen, "WebGuard:Web Based Adult Content Detection and Filtering System", IEEE WIC International Conference. Web Intelligence, pp. 574-578, 2003

[4] Dequan Zheng, Yi Hu, Tiejun Zhao, Hao Yu, Sheng Li, "Research of Machine Learning Method for Specific Information Recognition on the Internet", IEEE International Conference on Multimedia Interfaces(ICMI), pp, October 2002

[5] Christopher D. Hunter "Internet filter effectiveness : testing over and underinclusive blocking decisions of four popular filters", Proceedings of the tenth conference on Computers, freedom and privacy: challenging the assumptions, pp287-294, April 2000

[6] Wai Lam, "Automatic Text Categorization and Its Application to Text Retrieval", IEEE Transaction on Knowledge and Data Engineering, Vol.11, No.6, November/December 1999

[7] 김광현, 최정미, 이준호, "웹 문서 분석에 근거한 유해 웹 문서 검출", 한국정보처리학회 논문지 제12-D권 제 5호 683-688, 2005.10