제19회 한국정보처리학회 춘계학술발표대회 논문집 제10권 제1호 (2003.5)

효율적인 바이그램을 이용한 자동 문서 범주화

최준영, 이찬도
대전대학교 정보통신공학과
e-mail: lovejun02@hotmail.com, cdlee@dju.ac.kr

Automated Text Categorization using high quality Bigrams

Joon-Young Choi and Chan-Do Lee
Dept of Information & Communications Eng., Daejeon Univ.

요 약

본 연구는 바이그램을 이용하여 자동문서범주화 성능을 향상시키는 알고리즘의 개발을 목표로 한다. 기존의 문서 범주화 알고리즘의 장단점을 비교하여 개선된 바이그램 추출 알고리즘을 구현하고, 이 알고리즘을 실험한 결과 Reuters-21578 data set은 개별 단어를 사용하여 실험한 결과보다 단어+바이그램을 사용하였을 경우 BEP은 2.07%, F1은 1.40% 향상률을 보였고, Korea-web data set은 BEP의 8.12%, F1의 6.25% 향상을 보였다. 이와 같은 실험결과는 단어를 사용한 경우보다 단어+바이그램을 사용한 자동 문서 범주화 시스템이 더 효율적이라는 것을 보여준다.

1. 서 론

자동문서범주화(Automated text categorization)는 문서의 내용에 기초하여 미리 정해진 범주에 컴퓨터가 자동으로 분류해 넣는 기술을 말한다[1]. 정보화 시대의 물결에 따라 웹 문서가 기하급수적으로 늘고 있으며 효율적인 정보 관리 및 검색을 위해서는 내용별 분류작업이 필요하다. 이를 수작업으로 처리할 경우 막대한 비용이 들게 되고 일관성 또한 결여될 수 있는 가능성이 높아지므로 컴퓨터가 자동으로 분류 관리하는 자동문서 범주화 시스템이 필요하다.

자동문서범주화 기술의 응용분야는 문서검색을 위한 indexing, 텍스트에서 특정 내용 추출, 웹페이지 분류, 메일 필터링 등 다양하게 사용되어진다. 과거 수년간의 연구결과는 괄목할 만한데, 예를 들면 Apté, Damarau, & Weiss[2]는 Reuters-21578 데이터에서 87.8%의 Precision/recall break-even point를 보고하고 있다.

[그림 1]에서 보는 바와 같이 자동문서 범주화 과정은 크게 자질 추출 과정, 범주화 학습 과정, 범주할당 과정으로 나누어진다. 자질 추출 과정에서는 단어 중 기계학습에 필요한 자질만을 추출하는데 이를 위해 term frequency, document frequency, information gain 수치가 사용된다. 범주화 학습 과정에서는 이러한 자질을 이용하여 이미 분류된 문서를 범주화하는 분류기(classifier)를 훈련하는데, Naïve Bayes, maximum entropy, neural network 등의 방법을 사용한다. 이렇게 해서 얻어진 문서범주화 시스템을 이용하여 새로운 문서를 범주화하는데 사용한다. 자동문서범주화의 세 단계 중 본 연구는 자질 추출 과정에 중점을 두고, 어떻게 자질을 추출했을 경우 시스템 성능을 향상시킬 수 있는가 연구하였다.

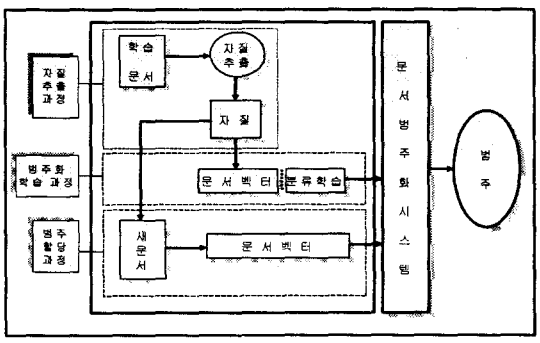

그림 1. 자동문서범주화 과정

2. 구(phrases)를 사용한 문서 범주화

Lewis[3]은 구를 사용한 문서범주화를 행하고 심도 깊은 분석을 통하여 구가 단어보다 더 나쁜 결과를 가져왔음을 보고하고 있다. 이와 같은 이유는 탐색공간의 확장, 빈도의 축소, 동의성의 증가 등 부정적인 측면이 더 강하게 작용하기 때문이다. 그렇지만 구를 사용하면 의미의 애매성은 줄일 수 있다. (예를 들어 "computer"나 "science"등 단어를 핵심어로 사용하는 것보다 "computer+science" 와 같이 구를 사용하는 경우 그 문서가 전산학 관련 문서임을 쉽게 결정할 수 있다). 여러 연구자들이 위의 문제점들을 해결하기 위하여 노력해오고 있는데, n-gram(단어가 n개 모인 자질)을 BOW에 추가할 때 향상된 성능을 가져왔음을 보여주는 희망적인 연구 결과들이 보고되고 있다.

Mladenic and Grobelnik[4]는 term frequency에 따라 단어 열을 5개까지 연속적으로 증가시키며 실험한 결과 단어 군에 3개까지의 단어 열을 추가했을 때 성능향상이 있었음을 보여주고 있으며, Fürnkranz[5], Schapire et. al[6], Schutze et. al[7]도 비슷한 결과를 보고하고 있다.

국내에서도 자동문서범주화 연구가 진행되어지고 있는데 한글문서를 자동으로 범주화하는 데에는 한글이 내포하고 있는 고유의 문제점들이 있다. 예를 들면, 핵심어를 자질로 추출하기 위해서는 문서의 내용과 관련도가 적은 조사, 수사, 어미 등의 기능어(function word) 처리가 선행되어져야 한다.

본 논문은 위에서 말한 관련 연구들과의 몇 가지 면에서 독창성을 가지고 있다. 첫째, 개별 단어를 바이그램으로 대치하는 것이 아니라 개별 단어에 바이그램을 추가함으로써 바이그램이 가지고 있는 의미 애매성 해소 능력을 활용한다. 둘째, 탐색 공간을 줄이기 위해 추가하는 바이그램 수를 전체 단어수의 2% 내로 선정한다. 셋째, 바이그램 선정기준으로서는 document frequency, term frequency 뿐만 아니라 information gain을 사용한다. 넷째, 한글 문서에 대한 범주화를 한다.

3. 바이그램 추출 알고리즘

이 알고리즘은 문서에 일정수(df_seed * 문서의 수) 나타나는 단어를 seed로 하여 이들 단어가 한번이라도 나타나는 바이그램을 추출한 후 빈도수가 일정기준 (df_thresh * 범주에 속하는 문서의 수, tf_thresh * 모든 문서의 수)를 초과하고, 또한 information gain이 기준 (ig_thresh)보다 높은 바이그램을 찾아내어 자질에 추가하여 범주화를 수행한다. [그림 2]는 바이그램 추출 알고리즘의 의사코드이다. pilot study를 통하여 df_seed는 0.01, df_thresh는 0.005, tf_thresh 3, 그리고 ig_thresh는 단어수의 1%에 해당하는 단어가 갖는 information gain수로 정하였다.

```
Find S = { 문서에 일정 수 나타나는 단어들}
Set B = {}.
For each 훈련문서집합
{
      모든 단어들을 미리 분석하여 필요 없는 기능어를 제거
      For each 인접 단어(w1, w2)
         if(w1 ∈ S of w2 ∈ S)  // w1,w2가 집합 S의 원소이면
         add bigram "w1+w2" to B. //w1과 w2를 더하여(바이그램으로 인식하여)
                                  집합 B에 넣는다.
}
For each b in B
{
     For each 카테고리 c
        if (b의 수 < 카테고리에 C에 속하는 문서의 수 * df_thresh)
        OR (b의 수 < 총 문서의 수 * tf_thresh )
        B의 집합에서 b를 제거한다.
     if (제거되지 않은 b 의 infogain < ig_thresh)
        //ig_thresh(단어수의 1%에 해당하는 단어가 갖는 information gain)
        B의 집합에서 b를 제거한다.
}
B를 출력. //바이그램을 이용하여 자질에 추가 범주화 실행
```

그림2. 바이그램 추출 알고리즘

4. 실험

본 실험은 Reuters-21578 corpus 와 Korea-web corpus를 사용하여 실험을 하였다.

Reuters-21578 corpus는 1987년에 나타난 Reuters 통신회사의 문서들을 모아 분류 해놓은 corpus인데 총 135 범주에 23460문서로 되어있다. 그러나 이 실험에서는 일부 테스트 문서가 없는 범주를 제외한 나머지 93 범주 13343 문서 (9592 학습문서, 3751 테스트문서)로 실험을 하였다. 한글문서 실험으로는 고려대, 호서대, 한남대에서 함께 만든 문서분류학습 및 테스트용 문서집합인 Korea-web corpus를 이용하였다. 이 corpus는 소분류(79), 중분류(19), 대분류(8)로 나누어져 있으며, 전체 문서 수는 4726개로서 3150 학습문서와 1576 테스트문서를 가지고 각 분류별로 실험을 진행하였다.

이 실험에서 보면 우리가 사용한 알고리즘은 매우 성공적이었다. Reuters-21578 corpus, Korea-web corpus를 가지고 실험한 결과, "computer+science" 등과 같이 명확한 개념을 나타내는 바이그램들을 성공적으로 추출하였고, 바이그램의 수는 단어 수에 비해 적었지만 information gain 수치에 따라 나열한 결과 많은 수가 높은 순위를 차지하였다. 추출된 바이그램들을 분석한 결과 information gain을 중

가 시켰다. 즉, 전체 자질들의 질을 향상시킨다는 것을 알 수 있었다.

바이그램을 자질에 추가하여 Nïave Bayes 분류기를 학습시킨 결과 [표1]과 같이 단어만 사용한 경우보다 recall값이 증가하였다.

표 1. Recall and Precision

	Recall		Precision	
	단어	단어+바이그램	단어	단어+바이그램
Reuters	0.810	0.844	0.715	0.708
Korea-web	0.710	0.718	0.813	0.809

Break-even point 와 F1 계산 결과 역시 많은 문서에 대하여 단어와 바이그램을 함께 사용할 때 성능이 향상하였다.

표 2. Break-even point and F1 measure

	F1 measure		
	단어	단어+바이그램	향상률(%)
Reuters	0.759	0.770	1.408
Korea-web	0.758	0.761	0.359

	Break-even point		
	단어	단어+바이그램	향상률(%)
Reuters	78.35	79.97	2.07
Korea-web	73.96	74.28	0.43

[표1], [표2]의 내용을 그래프로 보면 [그림 3]과 같이 나타내어진다.

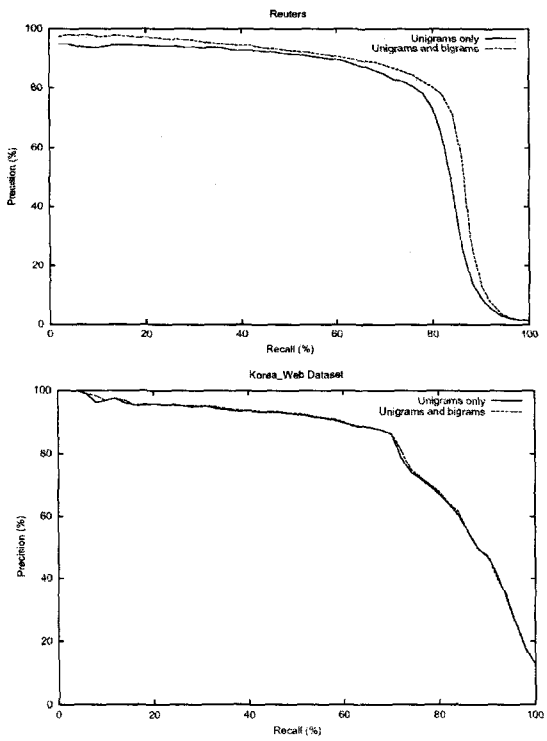

그림 3. Precision-recall graphs

[표1], [표2]와 [그림 3] 에서 보여진 값은 Korea-web 대분류 데이터 와 Reuters의 각각의 범주에서 나온 데이터 값들을 합한 전체 데이터의 값으로 나타내어진 것이다. [그림 3] 에서 점선은 바이그램+단어의 recall-precision을 나타낸 것이고, 실선은 단어의 recall-precision을 나타낸 것인데 점선이 더 효율적인 결과 값을 얻은 것을 볼 수 있다.

[표3]은 Korea-web data 대분류 각 범주에 대한 실험결과인데, 전반적으로 향상되었음을 보여준다.

표 3. Korea-web (대분류 데이터)
F1 and BEP

	단어		단어+바이그램		F1 (%) 향상률	BEP (%) 향상률
	BEP	F1	BEP	F1		
의학	84.49	0.843	84.49	0.847	0.434	0
공학	69.37	0.691	69.37	0.696	0.645	0
스포츠	76.31	0.751	76.31	0.750	-0.094	0
사회	70.19	0.713	69.63	0.717	0.529	-0.79
경제	72.28	0.744	71.03	0.747	0.463	-1.72
교육	55.91	0.616	57.14	0.647	4.949	2.19
문화	75.28	0.760	75.28	0.755	-0.726	0
컴퓨터	83.99	0.811	83.99	0.811	0	0
macro	73.96	0.758	74.28	0.761	0.359	0.43
micro	73.47	0.741	73.40	0.746	0.775	-0.04

Korea-web에 대한 중분류와 소분류 실험결과는 [표4], [그림 4]와 같다.

표 4. Korea-web (중분류, 소분류 데이터)
F1 and BEP

	단어		단어+바이그램		F1(%) 향상률	BEP(%) 향상률
	BEP	F1	BEP	F1		
중분류	53.68	0.562	51.19	0.566	0.75	0.95
소분류	12.56	0.175	13.58	0.186	6.25	8.12

대분류와 마찬가지로 중분류, 소분류 역시 바이그램을 이용한 값이 전반적으로 향상한 것을 볼 수 있다.

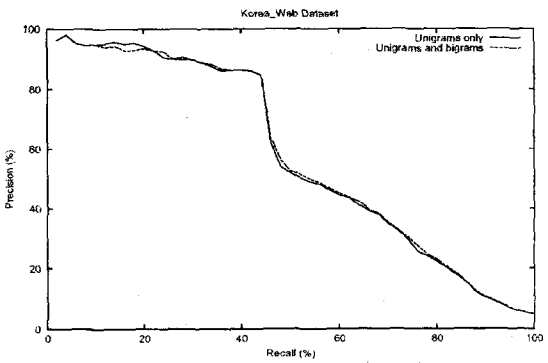

그림 4. Korea-web (중분류)

대분류에서 소분류로 즉, 큰 주제(예: 문화)를 갖는 범주에서 전문적인 주제(예: 문화->(신문, 방송))을 갖는 범주로 진행되어 가면서 F1, BEP의 더 많은 향상을 보였다. 그러나 전반적인 정확도는 감소하였다.

이와 같은 이유는 문서를 분류하기 전에 사전 인식 즉, 학습 과정에서의 데이터 수가 적어짐에 따라 학습 과정의 바이그램 또한 적어지기 때문에 정확도는 감소하였다.

그러나 좀더 전문적인 주제를 갖는 범주로 실험을 할 경우 정확도는 감소하였지만 향상률은 증가한다. 이와 같은 이유는 주제의 폭이 좀더 전문적으로 되어짐에 따라 발생하는 전문지식의 전문용어가 증가하기 때문이다. 즉, 전문용어 같은 경우는 일반적인 단어보다 바이그램의 구로 나타냄으로서 그 뜻을 정확히 알 수 있기 때문에 의미의 애매성이 감소하여 향상률은 증가하게 된다.

5. 결론

본 연구에서는 단어+바이그램을 사용한 자동문서 범주화시스템을 사용하여 단어를 사용한 시스템보다 정확도를 향상시켰으며, 지금까지 진행되어진 영어 문서 자동 문서범주화 시스템에서 한글 문서 범주화를 할 수 있는 방법을 획득하였다.

그러나 한글 문서의 실험 시 정확도가 감소하는 이유는 실험 데이터 문서수가 대분류에서 소분류로 갈수록 작아지므로 정확도 감소의 원인이 된다. Reuters-21578 같은 경우도 학습 문서의 수가 작으면 정확도가 대체적으로 떨어지는 것을 볼 수 있다.

문서 범주화 연구의 발전을 위해서는 일반성을 가질 수 있고 좀더 정확한 범주를 갖는 많은 학습 문서 및 테스트 문서를 토대로 한 corpus가 있어야 될 것이다.

향후연구로는 한글 문서 자동 분류 시스템의 구축과 좀더 효율적인 알고리즘의 개발로 자동문서 분류 시스템의 성능 향상에 기여 하고자 한다.

참고문헌

[1] Sebastiani, F. (2002). Machine learning in automated text categorization, ACM Computing Surveys, 34(1):1-47.

[2] Apté, C., Damerau, F., and Weiss, S. (1994). Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, 12(3), 233-251.

[3] Lewis, D. (1992). *Representation and learning in information retrieval.* Technical Report UM-CS-1991-093. Department of Computer Science, University of Massachusetts, Amherst, MA.

[4] Mladenic, D. and Grobelnik, M. (1998). Word sequences as features in text learning. In *Proceedings of the 17th Electrotechnical and Computer Science Conference* (ERK-98) (pp.145-148), Ljubljana, Slovenia (pp. 81-93).

[5] Fürnkranz, J. (1998). *A study using n-gram features for text categorization.* Technical Report OEFAI-TR-98-30, Austrian Research Institute for Artificial Intelligence, Vienna, Austria.

[6] Schapire, R, Singer, Y., and Singhal, A. (1998). Boosting and Rocchio applied to text filtering. In Croft et. al. (Ed.), Proceedings of SIGIR-98, 21st ACM *International Conference on Research and Development in Information Retrieval* (pp.215-223). New York: ACM Press.

[7] Schütze, H., Hull, D., and Pederson, J. (1995). A comparison of classifiers and document representations for the routing problem. In Croft et. al. (Ed.), Proceedings of SIGIR-95, *15th ACM International Conference on Research and Development in Information Retrieval* (pp.229-237). New York: ACM Press.