주제어와 미분류 문서들을 이용한 문서의 자동 분류 방법

이강일° 이창환
동국대학교 정보통신공학과
{leeki816°, chlee}@dongguk.ac.kr

Automatic Text Classification Method Using Keywords and Unlabeled Text

Kang-Il Lee° Chang-Hwan Lee
Information and Communication Engineering, Dongguk Univ

요 약

문서를 분류하기 위해서는 분류주제에 맞춰 미리 분류가 된 자료(labeled data)가 필요하다. 하지만 미리 분류가 된 자료를 만들기 위해서는 사람이 직접 그 문서의 의미를 해석하고 일일이 분류를 해야 하기 때문에 시간이 많이 소모가 된다. 본 논문에서는 비록 사람이 직접 분류한 자료를 이용하는 것에 비해서 분류 정확도는 조금 떨어지지만, 대신 주제어와 미분류 문서(unlabeled data)를 이용해서 문서를 분류하는 방법을 제시하려고 한다. 이와 같은 주제어와 미분류 문서의 경우에는 구하기가 쉽고, 사람이 일일이 분류하는 작업이 필요로 하지 않기 때문에 비용과 시간이 크게 절약이 된다는 장점이 있다.

1. 서 론

이제 웹은 단순히 정보를 찾기 위한 수단이 아니며, 점차 개개인의 삶의 일부로 자리를 잡아가고 있는 실정이다. 따라서 인터넷 서비스를 제공하는 입장에서는 서비스 이용자가 원하는 어떠한 문서를 정확히 찾아 주는 것이 가장 중요한 주제가 되어가고 있다. 하지만 기존의 검색엔진의 한계가 나타남에 따라 점차 문서분류에 대한 관심이 높아져 가고 있다. 분류가 정확하게 되어 있는 상황에서 검색의 적중률은 그렇지 않은 경우에 비해 상대적으로 월등히 높아지게 되고 검색하는데 소요되는 시간 역시 현저하게 줄어든다. 일련의 연구 결과에 의하면 분류체계를 도입함으로써 검색엔진만 사용하는 검색에 비해 소요시간이 약 50%정도 감소되었다고 보고되고 있다[1]. 그러나 이러한 분류를 할 때 문제점은 사람이 분류를 하는 행위는 시간과 노력의 소모가 많다는 것이다.

따라서 본 논문에서는 미분류 문서(unlabeled data)들만을 가지고 특정한 분류 주제에 따라 문서를 자동으로 분류하고자 한다. 이 때, 각각의 문서를 분류하는 기준은 입력된 분류 주제에 따라 가중치를 부여하고, 그것을 이용해서 주제에 대한 핵심단어를 확장한 다음 다시 각각의 문서를 분류주제에 따라 가중치를 부여하는 방식을 반복 수행하게 된다. 그리하여 나온 각각의 주제에 대한 문서의 가중치 값을 비교하여, 다른 주제에 비해 상대적으로 높은 가중치 값을 가지는 주제에 문서가 속하게 하는 방법으로 문서를 분류 하고자 한다. 이러한 실험을 통해 본 논문에서는 미분류 문서와 사용자가 제공하는 주제어만으로서 과연 우리가 원하는 정도의 문서 분류 효과를 거둘 수 있을지를 구할 수 있을지에 대해서 알아보고자 한다.

2. 관 련 연 구

사전에 분류가 문서(labeled Data)를 이용하는 문서 분류에 대한 연구는 Text Classifier[2]나 Information Extraction[3]을 통해 이미 높은 정확도의 분류를 나타내고 있다. 하지만 이와 같은 감독자 학습(Supervised Learning)에서는 충분한 학습 자료가 주어져야 하며 또한 이러한 자료를 생성하는데 있어서 사람이 직접 문서를 찾아서 분류해야 하는 번거로운 노력이 필요하게 된다. 하지만 이미 그 이전에 미분류 문서가 유용하다는 것은 증명이 되어있다[4]. 미분류 문서는 분류를 하는데 있어서 필요한 정보를 분류된 문서에 비해 그 양은 적지만 어느 정도는 포함을 하고 있을 뿐만 아니라, 완벽하게 자동으로 분류를 할 수 있게 하는 정보도 가지는 경우도 종종 발생한다. 또한 적은 양의 분류 된 문서와 대량의 미분류 문서를 이용해 문서를 분류하는 준 감독자 학습(semi-supervised learning) 역시 사람이 만족할 만한 수준의 분류 수준을 나타낸다고 한다[5]. 미분류 문서는 간단한 Perl Script를 이용하면 원하는 문서를 웹을 통해서 손쉽게 구하는 것이 장점이다. 본 논문에서는 사람이 직접 분류하는 작업은 하지 않고, 오직 미분류 문서가 가지고 있는 문서내의 단어정보와 주제어만을 가지고 문서를 자동으로 분류하는 방법을 이야기하려 한다.

3. 수 행 방 법

본 논문에서는 문서를 분류하는 방식은 하나의 문서에 정해진 여러 개의 주제에 대한 가중치를 부여하고, 상대적으로 높은 값을 가지는 주제에 대해 해당 문서가 속하게 되는 방식으로 분류하게 된다. 이 때 가중치 연산은 단어의 가중치와 각각 문서의 가중치를 동시에 고려한다. 이 절에서는 가중치의 연산 수행 순서와 방법을 이야기하고자 한다.

3.1 단어의 가중치

우선 단어의 가중치를 구하기 위해서는 상대 빈도와 연관성이라는 2가지 요소를 고려하게 된다. 상대 빈도는 어떠한 단어가 얼마나 자주 문서에 등장하는 가를 나타낸다. 상대빈도가 높다는 것은 여러 문서에 자주 나타나

는 단어라는 의미로서 그 값이 큰 문서일수록, 그 단어는 어떠한 특정한 뜻을 포함하는 것이 아니라, 포괄적인 의미를 나타내는 단어라고 의미하게 된다. 따라서 이는 실제 어떠한 문서를 대표한다고 나타내기에 보기 어려운 단어라고 볼 수라고 있다. 따라서 특정한 단어의 상대빈도 g_i는 inversed document frequency[6]를 이용해서 과 같이 나타내게 된다.

$$g_i = \log(\frac{D}{df_i})$$

이 때, D는 전체 문서의 수를 의미하며, df_i는 어떠한 단어 i가 1회 이상 등장하는 문서의 수이다.

단어의 가중치를 구하는데 필요한 또 다른 요소인 연관성은 어떠한 단어가 특정 키워드와 얼마나 함께 문서에 등장했는가를 나타낸다. 상호 연관성이 높은 단어들은 한 문서에 함께 등장할 확률이 높기 때문이다. 여기서는 주제어로 정의된 단어가 등장하는 문서 내에서 얼마나 자주 등장하는 가를 측정하게 된다. df_{ik}는 특정 단어 i가 주제어 k와 함께 등장하는 문서의 수를 의미한다. 또한 df_k는 주제어 k가 등장하는 문서의 수를 나타내는데 이는 전체 문서 내에서 가중치 값이 0보다 큰 문서의 개수를 의미하게 된다. 따라서 특정한 단어 i의 연관성 a_i는 다음과 같이 나타낼 수 있다.

$$a_i = \frac{df_{ik}}{df_k}$$

따라서, 위 2가지 요소를 이용하여 어떠한 단어 i의 가중치 w_i를 나타내면 $w_i = g_i * a_i$가 된다.

3.2 문서의 가중치

문서의 가중치의 경우에는 현재 주제어들이 가지고 있는 가중치에 대한 0~1 사이의 정규화 된 값에 각각의 단어의 등장 빈도수를 곱한 것에 대한 합으로 표현한다. 또한 하나의 문서 내에서 자주 등장하는 단어는 그 문서를 대표하는 성질이 강하기 때문에 빈도수를 곱해서 문서의 가중치를 나타내었다. 따라서 특정한 문서 j의 가중치 d_j에 대한 표현은 다음과 같다.

$$d_j = \sum_{i=1}^{n} k_i * wf_i$$

이 때, n은 전체 주제어의 수를 나타내며, k_i는 현재 문서 j에 등장하는 주제어 i가 가지는 가중치 값이다. wf_i는 주제어 i가 문서 j에서 등장하는 횟수를 의미한다.

3.3 문서의 가중치를 구한 이후의 단어의 가중치

문서의 가중치를 이후에 단어의 가중치를 구하는 방식이 바뀌게 되는 데, 단어의 상대빈도는 문서의 가중치에 포함이 되었으므로 더 이상 고려하지 않으며, 연관성을 구할 때는 단순히 문서의 수로만 으로 계산 되었던 것을 문서의 가중치로 바뀌게 된다. 따라서 연관성을 다음과 같이 다시 변환하게 된다.

$$w_i = \frac{1}{df_k} * (\sum_{j=1}^{D} d_j)$$

이렇게 구한 새로운 단어의 가중치가 높은 단어는 주제어 집합에 포함이 된다. 그 이후에는 다시 전체 주제어에 대한 가중치를 정규화 시키고 앞에서 설명한 문서의 가중치를 구하는데 사용이 된다. 그림 1은 이와 같은 수행 순서를 간단한 슈도코드를 통해 나타내었다.

```
Preprocessing() // Initialize for this job
COMPUTE idf of each word
FOR i = 0 to loop_size
  COMPUTE association & weight of each word
  COMPUTE weight of each text
  COMPUTE category of each text
END FOR
```

그림 1 알고리즘 진행방식

4. 실 험

실험을 하는데 사용된 문서는 'Reuter21578'이다. 독일의 Reuter 신문의 웹 기사 21578개를 모아 놓고 그것을 5개의 카테고리(PLACE, PEOPLE, ORGANIZATION, EXCHANGE, 기타)별로 분류가 되어있는 문서이다. 이 때 각각의 카테고리에 속하는 문서의 수는 PLACE가 16,880개, PEOPLE이 997개, ORGANIZATION이 879개, EXCHANGE가 482개, 기타에 2540개 이다. 이는 신문기사라는 가장 보편적인 데이터이며 실제 사람이 분류한 결과가 제공되기 때문에 본 논문에서 수행하는 실험 결과와의 비교 및 분석이 용이하기 때문에 이 데이터를 선택하게 되었다. 실험을 하는데 있어서는 우선 Reuter Data 내에 있는 HTML 태그 및 카테고리 분류 정보 등을 모두 제거하여 순수하게 unlabeled data로 만든 후에 실험을 하였다. 그리고 문서의 분류는 4개의 카테고리 주제 단어를 제공하고 분류를 하였고, '기타' 카테고리는 어떠한 주제에 대해서도 가중치 값을 가지지 않는 경우에만 '기타' 분류에 속하도록 하였다.

또한, 간단한 단어에 대한 필터작업도 하였는데, 우선 78,123개의 단어가 들어있는 사전에서 나타나는 동사에 대한 일반적인 과거형 어미인 '-ed'와 '-d'가 사전에서 사용되지 않는 경우에는 이 어미를 제거하여 동사원형을 사용하였고, 명사의 일반적인 복수형에 대한 어미인 '-s'와 '-es' 또한 그 자체적으로 특별한 의미를 지니지 않는 경우에는 제거하였다. 또한 영어에서의 관사나 전치사 및 의미 없이 자주 쓰이는 단어 482개를 stopword list[7]로 지정하여 해당 단어에 대해서는 연산이 이루어지지 않게 하였다. 그리고 전체 문서의 1%이하의 문서에서 등장하는 단어의 경우에는 단어의 연관성에 비해 그 상대빈도가 너무 높게 나오므로 단어의 가중치를 구하는 과정에서 연산이 수행되지 않도록 하였다.

단어의 가중치를 구한 후에는 각각의 주제에 따라 주제어 수가 확장 하게 하였다. 그리하여 앞에서 설명한 단어와 문서에 대한 가중치 연산을 총 50회를 반복 수행하였다. 그리하여 매 회 수행결과에 따라 Reuter 자료에서 미리 정해놓은 분류에 얼마나 맞게 분류를 하였는가를 측정하였다.

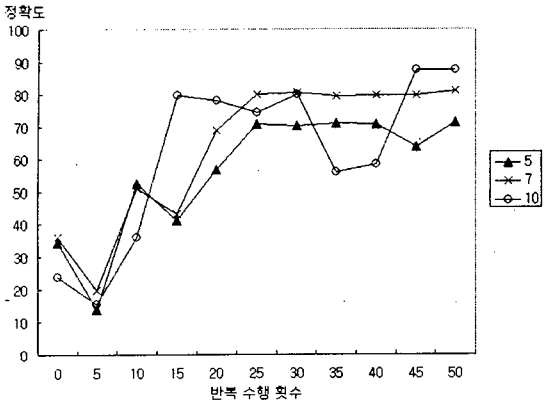

그림 2 반복 수행 횟수에 따른 문서 분류 정확도

그림 2는 전체 문서 21,578개에 대해 미리 정해진 분류에 대해 본 논문에서 수행한 분류 실험의 결과를 비교하였을 때, 얼마나 일치하는 가를 나타내고 있다. 또한, 각각의 선은 주제어를 확장 시키는데 있어 확장되는 주제어의 수 5개, 7개, 10개 일 때를 나타내고 있다. 주제어의 개수가 증가함에 따라 그 정확도 역시 증가하고 있음을 알 수 있다. 하지만 반복수행에 따른 변화의 폭이 크다는 단점이 존재하고 있다. 또한 이는 사람이 만족할 만한 수준의 분류인 72%[5]를 넘는 범위로서, 오직 미분류 자료와 사용자가 제공된 주제어만을 이용해도 그 안에 포함되어 있는 정보들이 자체적으로 분류를 향상시키는 데 크게 기여함을 알 수 있다. 또한, 지속적으로 확장되는 주제어의 수가 많을수록 정확도는 점점 더 높아지고 있는데 결국 주제어와 밀접한 단어들이 더 많이 선택 되어 가중치 값을 가짐에 따라 분류는 점점 더 정확해 진다고 볼 수 있다.

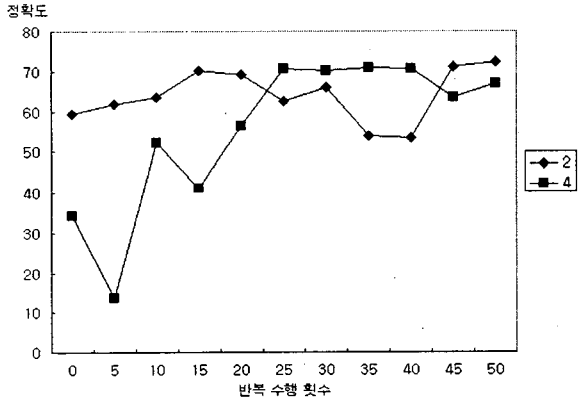

그림 3 카테고리 개수에 따른 정확도

그림 3은 Reuter 자료 중에서 자료의 수가 비슷한 PEOPLE과 ORGANIZATION 카테고리에 대한 자료만 선택해서 수행하는 경우와 전체 4개의 카테고리를 분류한 결과를 비교한 것이다. 이 때, 반복 수행에 따라 주제어를 5개씩 증가를 시켰다. 카테고리의 수가 적은 상황에서 확장되는 주제어의 수가 같은 경우 초반에는 상대적으로 많은 연관 단어를 주제어로 포함하여 정확도가 높게 나오지만, 점차 반복 수행함에 따라서 그 정확도는 거의 비슷해지고 있음을 알 수 있다.

5. 결 론

본 논문에서는 미분류 문서와 주제어만을 가지고 문서를 분류하는 방법을 시도해 보았다. 이러한 방식의 분류는 앞의 실험 결과에서 나타나는 것처럼 비록 사람이 직접 수집하여 분류한 문서를 이용하는 경우에 비해 정확도는 조금 떨어진다. 하지만, 어느 정도 사람이 만족할 만한 수준의 결과를 얻는데 큰 문제가 없고, 사람이 직접 자료를 구해서 분류 작업을 수행해야 하는 것에 비해 미분류 문서나 주제어는 손쉽게 구할 수 있다는 장점이 존재 한다. 이러한 장점은 문서를 분류하는 작업의 편리하게 해주고, 같은 시간에 많은 문서들을 분류할 수 있으므로 효율성을 증대시키게 된다.

6. 참 고 문 헌

[1] Chen, Hao and Dumais, Susan, Bringing Order to the Web: Automatically Categorizing Search Results, Proceedings of CHI2000, pp. 145-152, 2000

[2] Robert E. Schapire and Yoram Singer. Boostexter: A boosting-based system for text categorization. Machine Learning Journal, 1999

[3] M. E. Cali. Relational Learning Techniques for Natural Language Information Extraction. PhD thesis, Tech. Rept. AI98-276, Articial Intelligence Laboratory, The University of Texas at Austin, 1998

[4] V. Castelli and T. M. Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. IEEE Transactions on Information Theory, 42(6):2101, 1996

[5] Rosie Jones, Andrew McCallum, Kamal Nigam, Ellen Riloff, Bootstrapping for Text Learning Tasks : IJCAI-99 Workshop on Text Mining: Foundations, Techniques, and Applications, 1999

[6] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, and Sriram Raghavan, Searching the Web, ACM Transactions on Internet Technology, 2001

[7] http://www.lextek.com/manuals/onix/stopwords1.html