# 웹 페이지에서의 자질 선택과 분류

송무희° 임수연 박성배 강동진* 이상조
경북대학교 컴퓨터공학과, 경북대학교 정보전산원*
mhsong@knu.ac.kr° nadalsy@hotmail.com {parksb, djkang*, sjlee}@knu.ac.kr

# Feature Selection and Classification of Web Pages

Mu-Hee Song° Soo-Yeon Lim Sung-Bae Park Dong-Jin Kang* Sang-Jo Lee
{Dept. of Computer Engineering, Information Technology Services*} Kyungpook National University

요 약

본 논문에서는 웹 문서의 분류 성능을 향상시키기 위해 웹 페이지에서의 자질선택과 그에 따른 웹 문서 분류 방법을 제안 한다. 문서 분류에는 문서에 포함된 단어를 분류 자질로 사용하게 되며 이때 한 문서의 모든 단어를 분류 자질로 이용한다고 좋은 성능을 보인다고 보장할 수는 없다. 그러므로 문서에 필요한 단어만을 자동으로 추출하여 문서데이터의 자질을 축소하는 작업이 필요하다. 따라서 본 논문에서는 모집군 내의 자질벡터의 범위가 큰 것을 적은 수의 주요 성분으로 감소 시키기 위해 통계적 분석 기법중의 하나인 주성분분석 방법을 이용하여 자질감소와 그에 따른 문서분류의 성능 향상을 실험을 통하여 보인다. 야후 스포츠 뉴스 웹 페이지가 분류를 위해 사용되었으며, 분류기로는 Naïve Bayesian 분류 방법을 사용하였다. 실험 결과를 통해 본 논문에서 제안한 뉴스 웹 페이지 분류 방법이 스포츠 뉴스 데이터 군에서 만족할 만한 분류 정확도를 제공한다는 것을 알 수 있다.

## 1. 서 론

급속도로 발전하는 인터넷의 사용증가 추세에 맞추어 웹 상에서 볼 수 있는 전자문서의 양은 엄청나게 증가하고 있다. 이에 따라 문서를 알맞게 정해진 카테고리로 분류하는 것을 도와주는 도구에 대한 필요성이 점차 커지고 있다. 웹 문서를 분류하는 목적은 특정 주제별로 중요한 문서들을 구분하여 보다 빨리 사용자가 요구하는 문서를 검색하려는 것과 사용자의 선호도를 바탕으로 개인화를 하려는 것으로 나누어 볼 수 있다. 특히 웹의 효율적인 탐색을 위해 사용자가 관심 있어 할 웹 문서를 분류하는 것은 개인화 추천시스템을 연구하는데 중요하다.

본 논문에서는 효율적인 웹 문서의 분류를 위해 웹 페이지에서의 자질선택과 그에 따른 웹 문서 분류를 제안하고자 한다. 문서 분류에는 문서에 포함된 단어를 분류 자질로 사용하게 되며 이때 한 문서의 모든 단어를 분류 자질로 이용한다고 좋은 성능을 보인다고 보장할 수는 없으며 잡음정보를 배제하기 위한 일정한 기준이 필요하다[1]. 그러므로 문서에 필요한 단어만을 자동으로 추출하여 문서 데이터의 자질을 축소하는 작업이 필요하다. 분류문제에 있어서 자질 감소 (feature reduction) 또는 자질 집합 축소는 처리시간의 단축과 분류 성능의 향상이라는 두 가지 목적을 가지고 있으며, 다양한 자질 감소 방법이 연구되고 있다. 따라서 본 논문에서는 모집군 내의 자질벡터의 범위가 큰 것을 적은 수의 주요성분으로 감소시키기 위해 통계적 분석 기법중의 하나인 주성분분석(PCA: *principal component analysis*) 방법을 이용하여 자질감소와 그에 따른 문서분류의 성능 향상을 실험을 통하여 보인다. 모집군 내에 고유한 단어들의 수가 많기 때문에 주성분분석 방법이 분류를 위한 가장 적당한 자질을 선택하기 위해 사용되어 왔다[2]. 주성분분석으로부터 감소된 주요성분은 자질 벡터로서 쓰이게 될 것이다. 이러한 자질 벡터들은 그 뒤 분류를 위해 분류기의 입력으로 사용된다. 야후 스포츠 뉴스 웹 페이지가 분류를 위해 사용되었으며, 분류기로는 Naive Bayesian 분류 방법을 사용하였다. 실험 결과를 통해 본 논문에서 제안한 뉴스 웹 페이지 분류 방법이 스포츠 뉴스 데이터 군에서 만족할 만한 분류 정확도를 제공한다는 것을 알 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 문서분류에 대한 관련연구를 살펴보고, 3장에서는 본 논문에서 제안한 뉴스 웹 페이지 분류 방법에 대해 설명하고, 4장에서는 스포츠 뉴스 웹 페이지를 대상으로 한 실험과 다른 방법을 사용했을 때의 분류결과를 비교하며, 마지막으로 5장에서는 뉴스 웹 페이지분류 정확도에 대한 결론 및 향후 연구방향을 제시한다.

## 2. 관련연구

서로 다른 유형의 자질 벡터들을 가진 텍스트 문서들을 분류하기 위해 많은 연구자들이 광범위하게 문서 분류 방법들을 적용해왔다. 예를 들면 Lee et al.은 퍼지학습기법을 가진 언어 자질 선택에 기초한 문서분류를 위해 신경망 이용을 제안하였다[3]. Yalin Wang은 웹 페이지 내에서 테이블이 원래 고유한 테이블의 목적으로 사용된 경우와 그렇지 않은 경우로 분류하였는데 이를 위해 결정 트리 기법과 SVM(Support Vector Machine)을 사용하였다[4]. 그리고 Lam et al.[5]은 문서분류를 위한 자질 벡터들을 선택하기 위해 신경망을 이용한 용어빈도 방법을 제안하였다. 그러나 이러한 방법들은 데이터군의 한 특정계층에서 표현되는 문서의 수가 적다면 분류 정확도는 저하 될 것이다. 또한 특징 공간이 다차원으로 표현되며, 이러한 다차원 특징은 많은 학습 알고리즘에서 적용되기가 쉽지 않다 본 논문에서 제안한 웹 페이지 분류 방법은 적은 수의 문서로 표현되는 데이터 군에서 주성분분석 방법에 의한 자질 선택 접근법을 이용하여 웹 페이지의 범주 정확도를 개선시키는 데 그 목적을 두고 있다.

### 3. 웹 페이지 분류 방법

본 논문에서 제안한 방법을 이용한 웹 페이지 분류과정이 그림 1에 나타나 있다. 이것은 웹 뉴스 검색과정, 스테밍(stemming) 및 불용어(stopword) 처리 과정, 자질 감소와 선택 과정, 분류기를 이용한 웹 문서 분류 과정 등으로 구성된다. 스포츠 뉴스 웹 페이지 검색 과정은 가장 최신의 스포츠 뉴스 웹 페이지 범주가 웹으로부터 야후 웹 서버를 통해서 검색 될 것이다.

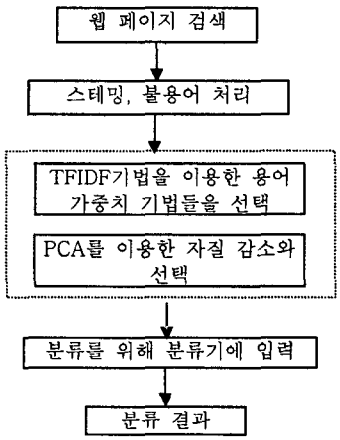


그림 1. 뉴스 웹 페이지 분류 과정

3.1 웹 페이지의 선처리

문서에서 단어를 추출하기 위한 방법으로는 전처리 단계로서 불용어 제거와 스테밍 처리, 그리고 정보검색 측정치 TFIDF(term frequency/inverted document frequency)에 기초를 두고 있다. 불용어 처리는 웹 페이지 문서에 존재하는 하는 가장 빈번한 단어들을 제거하는 과정이다. 이러한 단어들을 제거하면 문서 내용을 저장할 공간을 절약하고 검색과정에 소요되는 시간을 줄일 수 있다. 스테밍은 단어들을 가능한 어근 단어로 줄임으로써 웹 페이지 문서에서 각 단어를 추출하는 과정이다. 각 단어 $w_k$의 용어 가중치 $x_{jk}$의 계산은 Salton에 의해 사용된 방법을 사용하여 수행되며 수식 1에 주어진다[6].

$$X_{jk} = TF_{jk} \times idf_k \tag{1}$$

TFIDF는 역문헌 빈도수를 단어의 빈도수와 같이 적용함으로써 그 문서를 대표하는 단어들을 효율적으로 찾아주는 알고리즘이다. 역문헌빈도 $idf_k = \log(\frac{n}{df_k})$ 이며, $n$은 데이터베이스에 있는 문서의 총 개수이다.

뉴스 웹 페이지의 선처리 후에 뉴스 데이터베이스에 있는 모든 고유한 단어들을 포함하는 어휘집이 생성된다. 상이한 단어들의 수가 많으므로 어휘집에 있는 고유한 단어들의 수를 1000개로 제한하였다. 어휘집에 있는 각 단어는 하나의 자질 벡터를 나타낸다.

3.2 주성분분석을 이용한 자질 감소

아래와 같은 행렬 문서-용어 가중치인 A를 가정하자.

$$A = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2m} \\ x_{31} & x_{32} & \cdots & x_{3k} & \cdots & x_{3m} \\ M & M & \cdots & M & \cdots & M \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{nm} \end{pmatrix}$$

여기에서 $x_{jk}$는 문서집합에 존재하는 용어가중치이다. j는 뉴스 웹페이지에 존재하는 각각의 웹페이지의 문서($Doc_j$)의 수를 말하며 j=1,...,n이다. k는 하나의 단어($w_k$)가 문서($Doc_j$)에서 발생하는 횟수를 나타내며 k=1,...,m이다. 데이터 행렬 A의 주요 성분을 계산하기 위해서는 몇가지 단계를 거쳐야 한다[7][8]. 데이터행렬 A에 있는 m개 변수들의 평균은 다음과 같다.

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk} \tag{2}$$

다음에 공분산 행렬 $S = \{s_{jk}\}$가 계산된다. 분산 $s_{kk}^2$는 아래와 같이 주어진다.

$$s_{kk}^2 = \frac{1}{n}\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2 \tag{3}$$

공분산 $s_{ik}$는 아래와 같이 주어진다.

$$s_{ik} = \frac{1}{n}\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \tag{4}$$

여기에서 i=1,...,m이다. 그 뒤 우리는 공분산 행렬 S의 고유값과 고유벡터를 결정한다. 고유치 $\lambda$와 고유벡터e는 다음과 같다.

$$Se = \lambda e \tag{5}$$

고유벡터 e를 찾기 위해 특성방정식 $|S - \lambda I| = 0$을 풀어야 한다. S가 $m \times n$ 행렬이면 m개의 고유값들 $(\lambda_1, \lambda_2, ..., \lambda_m)$을 찾을 수 있다. 다음의 식을 이용하여 해당되는 모든 고유값들을 찾는다.

$$(S - \lambda_i I)e_i = 0 \tag{6}$$

고유값들과 이에 해당하는 고유벡터들은 $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$처럼 정렬될 것이다. 고유벡터 열들에서 정방행렬 $E = [e_1 e_2 e_3 ... e_m]$를 구성한다.

$$\wedge = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ M & M & M & \cdots & M \\ 0 & 0 & 0 & 0 & \cdots \lambda_m \end{pmatrix}$$

그리고 처음 나오는 $d \leq m$ 고유벡터들을 선택하는데 여기에서 d는 자질벡터로서 100, 200, 400등이다. 주요성분집합은 $n \times d$ 행렬M으로 다음과 같이 표현된다.

$$M = \begin{pmatrix} f_{11} & f_{12} & f_{13} & \cdots & f_{1d} \\ f_{21} & f_{22} & f_{23} & \cdots & f_{2d} \\ M & M & M & . & M \\ f_{n1} & f_{n2} & f_{n3} & \cdots & f_{nd} \end{pmatrix}$$

여기에서 $f_{ij}$는 $m \times n$ 데이터 크기에서 $n \times d$ 크기로 감소된 상태에서 추출된 자질벡터이다.

### 4. 실험 및 결과

본 논문에서는 야후 스포츠 뉴스의 웹 페이지 데이타군을 사용하였다. 데이터베이스의 뉴스 유형으로는 야구, 축구, 골프, 농구, 경마/바둑, 각종스포츠 등이다. 문서의 총 개수는 500개이다. 훈련을 위해 서로 다른 계층들에서 무작위로 100개의 문서를 선택하였다. 나머지 문서들은 시험 군으로 사용되었다.

주성분분석을 통해 원래 자질벡터들인 m=1000개를 보다 적은 수의 주요성분들로 줄였다. 우리의 경우에는 몇 가지의 d값들, 즉 100, 200, 300, 400, 500, 600을 선택하였다.

### 4.1 분류 실험

통계적 분류를 하기 위해 우리는 Naive Bayesian 분류기를 사용하였다[9]. Bayesian 분류기를 사용할 때 한 용어의 발생이 다른 용어의 발생과는 독립적이라고 가정하였다. 우리는 주어진 문서 *Doc*에 대해 가장 높은 조건부 확률을 제공하는 계층 *cs*를 찾으려 한다. $w^m{}_k = \{w_1, w_2, \Lambda, w_m\}$ 는 문서 *Doc*의 텍스트 내용을 나타내는 단어들의 집합이고 *k*는 용어 번호인데 *k=1,2,...,m*이다. 분류 점수는 아래의 수식4과 같이 측정된다.

$$J(cs) = \prod_{k=1}^{m} P(w_k \mid cs) P(cs) \ , \tag{7}$$

### 4.2 실험결과

표준 정보 검색 측정법인 정확도, 재현율, *F1*을 이용하여 TFIDF, 단순Bayesian과 본 논문에서 제안한 방법들이 평가되었다[10]. 평가식들은 다음과 같이 정의 된다.

$$precision = \frac{a}{a+b} \ , \tag{8}$$

$$recall = \frac{a}{a+c} \ , \tag{9}$$

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \ , \tag{10}$$

여기에서 *a, b, c* 값은 에서 표 1에 정의 되어있다. *F1* 측정법은 정확도와 재현율에서 일정의 평균치를 나타낸다. 본 논문에서 제안한 방법을 이용한 분류 결과가 표2에 나타나있다. 또한 F1측정을 사용하여 제안한 방법과 TFIDF와 단순Bayesian 방법과도 비교하였다.

표 1 *a,b,c* 파라미터들의 정의

| 값 | 의미 |
|---|---|
| *a* | 시스템과 전문가 모두 할당된 범주와 일치 |
| *b* | 시스템은 할당된 범주와 불일치하고 전문가는 일치 |
| *c* | 전문가는 할당된 범주와 불일치하고 시스템은 일치 |
| *d* | 시스템과 전문가 모두 할당된 범주와 불일치 |

표 2 웹 문서 분류 결과

| *class no* | Precision(%) | Recall(%) | F1(%) |
|---|---|---|---|
| 1. 야구 | 97.14 | 69.00 | 80.39 |
| 2. 축구 | 89.29 | 100.00 | 94.16 |
| 3. 골프 | 89.91 | 98.00 | 93.43 |
| 4. 농구 | 86.09 | 99.00 | 92.18 |
| 5. 경마/바둑 | 94.29 | 99.00 | 97.05 |
| 6. 각종스포츠 | 90.09 | 100.00 | 94.78 |
| **Average** | **91.14** | **94.17** | **92.00** |

표 3 *F1* 측정법을 사용한 분류 결과 비교

| class no | TFIDF(%) | 단순 Bayesian(%) | PCA-Bayesian(%) |
|---|---|---|---|
| 1. 야구 | 83.04 | 83.04 | 80.39 |
| 2. 축구 | 88.51 | 80.81 | 94.16 |
| 3. 골프 | 86.86 | 82.56 | 93.43 |
| 4. 농구 | 79.39 | 78.79 | 92.18 |
| 5. 경마/바둑 | 68.03 | 87.10 | 97.05 |
| 6. 각종스포츠 | 80.26 | 91.22 | 94.78 |
| **Average** | **81.02** | **83.92** | **92.00** |

표 3에 있는 것처럼 TFIDF, 단순 Bayesian에 의한 분류 정확도와 본 논문에서 제안한 방법인 PCA-Bayesian 각각에 대해 81.02% 83.92%, 92.00% 이다. 이것은 자질 벡터가 주의 깊게 선택된다면 본 논문에서 제안한 방법이 웹 스포츠 뉴스의 분류가 개선되며, 분류 정확도가 증가된다는 것을 나타낸다.

## 5. 결론 및 향후 과제

본 논문에서는 웹 페이지 내에서의 자질선택에 따른 웹 뉴스 분류 접근법을 제시하였다. TFIDF, 단순 Bayesian과 본 논문에서 제안한 주성분분석을 이용한 자질감소에 따른 웹 페이지분류 방법의 분류 정확도를 비교 실험을 통하여 제시하였다. 데이터로는 야후 스포츠 뉴스 웹 페이지가 분류를 위해 본 논문에서 제안한 웹 페이지 분류방법에 적용되었다. 자질선택에 따라 분류 알고리즘을 통한 실험평가에서 본 방법이 스포츠 뉴스 데이터 군에 대해 만족할 만한 분류 정확도를 제공한다는 것을 알 수 있다.

향후 과제로 본 논문에서 제안한 분류 방법을 사용하여 *F1* 분류 결과를 증가시키기 위해서는 각 계층에서 후보문서를 선택하기 위한 보다 개선된 문서 선택접근법이 사용되어야 한다. 또한 본 논문에서 제안한 분류정보를 정보검색에 활용하여 효율을 높이는 방안과 웹 문서에서부터 의미적 개념, 관계를 추출하는 방법을 계속 연구, 진행하고자 한다.

## 참 고 문 헌


[1] 정영미, 이재윤. "지식분류의 자동화를 위한 클러스터링 모형연구", 정보관리학회지, Vol.18, No.2, pp. 203-230, 2001.

[2] R.A.Calvo, M.Partridge, M.Jabri, "A comparatives study of principal components analysis techniques", In Proc. Ninth Australian Conference on Neural Networks, Brisbane, QLD, 1998, pp.276-281.

[3] H.M.Lee, C.M.Chen, C.W.Hwang, "A Neural network document classifier with linguistic feature selection", In Proc. The 13th International Conference on Industrial and engineering Applications of Artificial Intelligence and Expert Systems, New Orleans, Louisiana, USA, IEA/AIE 2000, 19-22 June 2000, pp. 555-560.

[4] Y.Wang and J.Hu, "A Machine Learning Based Apporach for Table Detection on The Web", In Proc. The 11th International World Wide Web Conference, Honolulu, Hawaii, USA, pp.242-250, May, 2002.

[5] W.Lam, M.E.Ruiz, P.Srinivasan, "Automatic text categorization and its applications to text retrieval", IEEE Transaction on Knowledge Data Engineering 11(6), 1999, 865-879.

[6] Salton&McGill, Introduction to modern information retrieval, New York, McGraw-Hill, USA, 1983.

[7] 김기영,전명식, "다변량 통계자료분석", 자유아카데미, 1997.

[8] Ali Selamat, Sigeru Omatu, "Web page feature selection and classification using neural networks", Information Sciences Vol.158, pp 69-88, 2004.

[9] T.Joachim, "Probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization", In Proc. International Conference on Machine Learning, Nashville, TN, USA, 1997, pp.143-151.

[10] D.D.Lewis, "Evauating and optimizing autonomous text classification", In:E.A.Fox,P.Ingwersen, R.Fidel(Ed.), SIGIR'95: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, 1995, pp.246-254.