제22회 한국정보처리학회 추계학술발표대회 논문집 제11권 제2호 (2004. 11)

웹서비스를 이용한 SVM 기반 분산 문서분류기 설계

김용수, 박용범
단국대학교 전자계산학과
e-mail : yongari@dankook.ac.kr

Design distributed document classifier based on SVM using Web Services

Yong-Soo Kim, Young B. Park
Dept. of Computer Science, Dan-Kook University

요 약

인터넷이 발달하면서 인터넷 상에서 공유 문서를 효율적으로 분류하기 위한 자동 분류의 필요성이 높아지고 있다. 또한 인터넷은 단순한 문서 제공의 한계를 넘어 어플리케이션간의 통합연동을 위한 기술이 대두되고 있다. 이러한 관점에서 본 논문은 새롭게 제시되고 있는 웹서비스를 이용하여 SVM 기반의 분류기를 분산 구성하여 설계하였고, 문서로부터 추출된 특성단어 벡터정보를 이용하여 SVM 학습 후 각각의 분류기를 통하여 분산 문서 분류를 수행한다. 특성단어 벡터는 TF*IDF에 기반한 특성 표현법을 사용하였으며, 분류 범주 별로 SVM 기반의 분류기 모델 데이터를 생성하기 위해 특성 단어 사전을 구축하여 분류 기준으로 구성하였다.

1. 서론

인터넷이 발달하면서 그 환경에 있는 사용자들은 자연스럽게 그들이 가지고 있는 문서를 서로 교환, 공유하고 있다. 이러한 인터넷 상에서 공유되고 있는 문서들 중에서 찾고자 하는 문서를 보다 빠르고 정확하게 찾아내기 위한 연구를 위해 많은 시간과 비용이 투자되었고, 자동으로 문서를 분류하는 연구가 진행되었다.[1]

또한 변화되는 인터넷 환경에서는 브라우저를 통한 정보제공의 개념을 벗어나, 어플리케이션의 정보제공까지 처리하고 있으며, 거대해진 인터넷 환경에서 효율적인 처리를 위한 분산처리 환경이 제시되고 있다. 이를 위하여 서버중심의 시스템에서 글로벌 컴퓨팅이 지원하는 네트워크 기반의 컴포넌트 시스템 환경을 통한 문서분류를 처리할 수 있는 구조를 제공하기 위해 웹서비스를 이용한 문서분류기의 활용을 제안하며, 일반적으로 자유롭게 사용할 수 있도록 인터페이스를 제공하는 분산 문서 분류 시스템을 설계하는데 중점을 두었다.

본 논문에서는 이러한 시스템을 구성하기 위한 환경과 분산 문서 분류를 위한 구조설계를 제안하고자 하며, 정해진 범위 내에서의 텍스트 문서를 기반으로 한 범주 별 다중 분류 테스트 실험결과를 보인다.

2. SVM 를 이용한 문서분류

문서 분류에 대한 방법은 NN(Neural Network)와 통계적 접근방법이 제시되고 있으며, 이러한 방법 중 하나가 SVM(Support Vector Machine)이다. SVM 은 기본적으로 두 개의 클래스를 분류하는 방법이다. 이 방법은 1979 년에 Vapnik 에 의해 발표되었다[2]. 문서분류에 SVM 을 적용하여 기존의 다른 방법에 비해 학습이 빠르고, 분류에 대한 높은 성능을 보여주고 있다.

SVM 을 이용하기 위해서는 아래의 <그림 1>과 같이 문서 데이터를 변환하는 표현법을 따른다.[3][4]

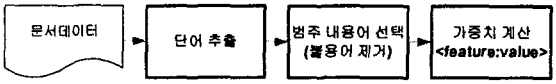
<line> ::= <target> <feature>:<value> <feature>:<value> ... <feature>:<value> # <info>
<target> ::= +1 | -1 | 0 | <float>
<feature> ::= <integer> | "gid"
<value> ::= <float>
<info> ::= <string>

<그림 1> 문서 데이터 표현법

이렇게 변환된 데이터는 SVM 을 통해 학습과 분류를 수행한다.

문서 데이터 중에는 분류에 필요한 내용어(Content word)와 필요하지 않은 불용어(Stop word)가 존재하게 되는데 우선적으로 불용어를 제거해야만 처리되는 데이터의 양을 줄이고, 분류의 정확도를 증가시킬 수 있다.

이러한 문제를 해결하기 위해 다음 <그림 2>와 같은 전처리 단계가 요구된다.



<그림 2> 전처리 단계

범주(Class) 별로 내용어를 선택하기 위해 문서빈도(Document Frequency) [5], 상호정보 척도(Mutual Information), 정보 획득량(Information Gain)[6], 카이 제곱 통계량(x^2 statistics)의 방법이 사용된다. 이 중 정보 획득량과 카이 제곱 통계량이 좋은 성능을 보이며, 논문에서의 시스템 구성에서는 그 중 하나인 정보 획득 방법을 선택하여 구성하였다.

정보 획득량 방법은 기계학습에서 자주 사용되는 방법으로 문서 출현 빈도와 출현하지 않은 빈도를 같이 포함하여 정보량을 계산한다.

$\{c_1,c_2,\ldots,c_m\}$이 분류 범주의 집합이라고 할 때 단어 t 의 정보 획득량은 다음과 같이 계산된다.

$$\text{Gain(t)} = \text{Entropy}(S) - \text{Expected Entropy}(S_t)$$

$$= \{-\sum_{i=1}^{m} P(c_i)\log P(c_i)\} -$$

$$[P(t)\{-\sum_{i=1}^{m} P(c_i \mid t)\log P(c_i \mid t)\} +$$

$$P(\bar{t})\{-\sum_{i=1}^{m} P(c_i \mid \bar{t})\log P(c_i \mid \bar{t})\}]$$

범주별로 내용어를 처리 후 작업인 가중치 계산을 위한 방법으로는 이진 가중치(Boolean Weighting), 단어 빈도 가중치(Word Frequency Weighting), TF*IDF 가중치(TF*IDF Weighting), 엔트로피 가중치(Entropy Weight), 역범주 빈도 가중치가 사용된다. 이 중 가장 많이 사용되는 방법은 TF*IDF 가중치 방법이며, 본 논문에서도 이 방법을 채택하였다.

TF*IDF 가중치 방법은 문서 데이터에서 각 단어의 가중치는 단어의 빈도와 역문헌 빈도(IDF)의 곱으로 계산된다.

$$w_{ij} = tf_{ij} \times idf_i$$

$$idf_i = \log\frac{N}{n_i}$$

n_i 는 단어가 나타난 문서 수, N 은 전체 문서 수를 의미한다.

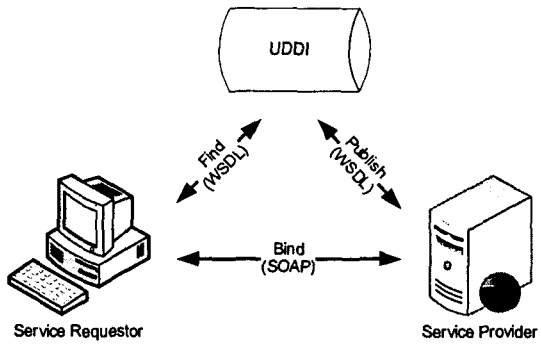
이와 같은 절차에 따라 SVM 에서 분류가 가능한 형태(feature:value)로 변환하여 학습/분류를 처리한다.[7]

3. 웹서비스(Web Services)를 이용한 분산 분류

웹서비스는 표준화된 XML 을 기반으로 한 인터페이스를 통하여 플랫폼 독립적이고 프로그램 언어에 중립적인 방법으로 네트워크 상에서 서비스를 통합하는 방법을 제시한다. 따라서 다양한 환경 안에서 하드웨어, 운영체제, 프로그래밍 언어에 종속되지 않고 상호 연계가 가능하게 된다.

이러한 방법을 사용하기 위해서는 SOAP(Simple Object Access Protocol), UDDI(Universal Discovery Description & Integration), WSDL(Web Services Description Language)과 같은 기술을 기반으로 W3C 등과 같은 표준기구가 제시하는 인터페이스로 <그림 3>과 같이 연동된다.[8]

- **SOAP** : XML 로 구성된 프로토콜로서 HTTP 를 기반으로 쉽게 어플리케이션들의 연동이 가능하도록 사용된다.
- **UDDI** : 웹서비스에 대한 등록 및 검색이 가능하도록 웹 기반의 글로벌 레지스트리 기능을 제공한다.
- **WSDL** : 웹서비스의 연동방법, 프로토콜, 데이터 포맷에 대한 정보를 XML 형태로 제공한다. WSDL 정보를 해석함으로써 어플리케이션이 대상 웹서비스를 사용할 수 있도록 한다.



<그림 3> 웹서비스 아키텍처

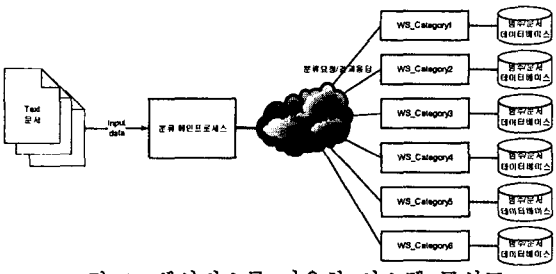
분산 분류를 처리하기 위해 범주 별 웹서비스를 구성해야 하며, 클라이언트는 모든 범주에 문서 데이터를 순차적으로 호출함으로써 독립적인 문서 분류가 가능하게 된다.

기존의 다중 분류를 처리하기 위한 시스템의 구성을 살펴보면 구축 시스템 이외의 이기종 시스템에서 분류기를 사용하고자 한다면 많은 비용과 시간을 투자하지만 웹서비스의 경우는 WSDL 에 기술된 정보로 이기종과 관계없이 통합연동이 가능하다.

범주 별로 처리되는 웹서비스는 독립적인 분류기의 역할을 수행하고, 모든 범주의 순차적 처리를 위한 웹서비스를 별도로 구성한다. 이와 같이 구성된 웹서비

스는 전체 분류기능 이외에 개별적 분류도 처리가 가능해진다.[9]

웹서비스를 이용한 시스템 구성은 <그림 4>와 같다.



<그림 4> 웹서비스를 이용한 시스템 구성도

4. 설계 모델

본 연구에서는 웹서비스를 이용한 SVM 기반 분산 문서 분류기의 모델을 구성하기 위해 웹서비스 구성, 문서 전처리, 분류기 구성으로 구분하여 설계하였다. 또한 실험적 구성을 고려하여 몇 가지 제약사항을 정의하여 그 범주 안에서 실험하였다.

웹서비스의 구성은 범주 별로 독립적으로 구성하여 분류 메인 프로세스에서 동기식으로 호출하여 각각의 분류결과를 얻어올 수 있도록 구성하였다. 이는 전체 범주의 분석 결과를 호출된 후 바로 얻기 위한 구성으로, 웹서비스의 구조상 비동기식 호출도 가능하다.[8]

문서 전처리를 위해서 다음과 같은 규칙을 적용하여 구성하였다.

첫째, 분류를 위한 특성 용어를 전문화 하기 위하여 범주 별로 특성 용어 사전을 구성하여 문서를 전처리하는 과정에서 사용한다.

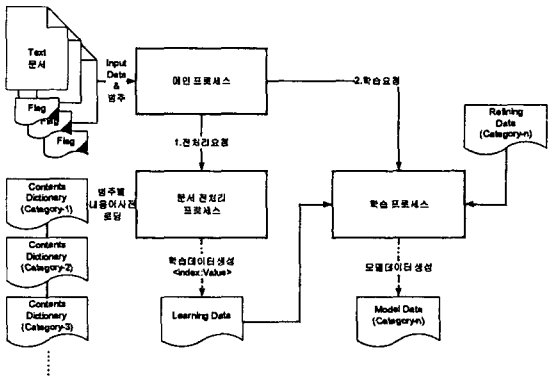
둘째, 특성 용어는 문서 내에 2 회 이상 출현한 단어 범위 내에서 카이 제곱 통계량 계산에서 산출된 수치(-1 < 산출값 < 1)의 단어를 선별하여 구성한다.

셋째, 학습 벡터는 TF*IDF 수식을 이용한 값을 기준으로 벡터를 구성한다.[10]

위와 같은 규칙이 적용된 데이터는 범주별로 구성된 특성 용어 사전에서 참조된 단어 인덱스로 벡터값이 계산되어 SVM 을 통한 학습/분류를 수행하게 된다. 특성 용어 사전은 학습 시에만 구성되며, 분류시는 학습에서 구성된 사전만을 사용하게 된다.

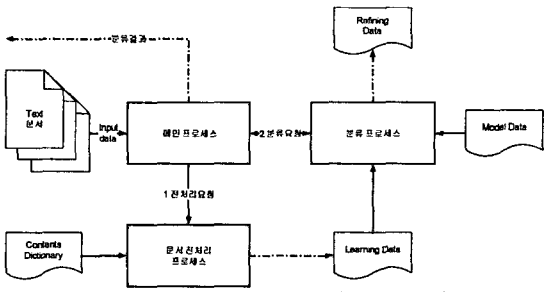
전처리 과정을 통하여 구성된 데이터로 SVM 을 이용하여 학습 및 분류를 수행한다.

실험 데이터는 인터넷 신문에서 기재되는 기사를 300 여개의 데이터를 사용하였으며, 범주는 정치, 경제, 사회, 문화, 국제, 스포츠로 구성하였다. 이는 범주별 연관성이 높은 범주와 낮은 범주를 같이 구성 함으로써 학습에 대한 결과 분석을 하기 위해서이다. 단, 학습을 위한 데이터는 그 중 범주별 특성이 높은 데이터만을 선별하여 구성하였다.



<그림 5> 학습기 시스템 구성도

<그림 5>는 문서를 기반으로 범주별 학습 모델 데이터와 특성 용어 사전을 구성하고 학습하기 위한 시스템 구성이며, <그림 6>은 학습 시스템에서 생성된 모델 데이터와 특성 용어 사전을 기준으로 분류하기 위한 시스템 구성이다.



<그림 6> 범주별 분류기 시스템 구성도

학습 모델 데이터와 특성 용어 사전을 구성한 후 전체 데이터를 기준으로 선별분류와의 정확도를 측정하도록 한다.

5. 실험결과 및 분석

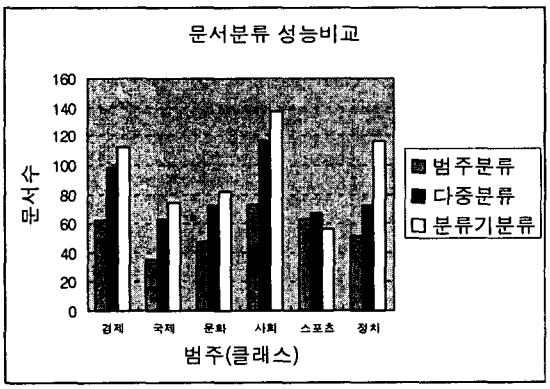
본 실험을 위해서는 6 개의 범주를 구성하기 위해 6 개의 SVM 이 분류기로 사용되었으며, 동일 개수의 특성 용어사전이 사용되었다. 초기 선별된 학습표본을 통해 학습모델을 수행하였으며, 전체 데이터를 각 범주 분류기를 통해 결과를 산출하였다. 본 실험에서는 독립적으로 분산된 분류기 구성에서 분류 성능과 기준으로 실험한 후 결과를 평가한다.

전체 실험에서 사용한 데이터 수와 학습에서 사용한 데이터 수는 <표 1>과 같다.

<표 1> 전체학습 데이터 수 및 특성용어사전 단어수

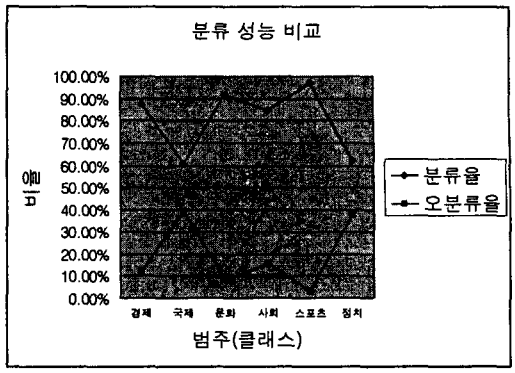
	전체 데이터	학습 데이터	특성용어수
경제	62	25	190
국제	35	23	192
문화	47	27	233
사회	73	22	97
스포츠	63	29	110
정치	51	25	172
계	331	151	-

독립적으로 구성된 분류기의 분류 성능을 분석하기 위해 전체 데이터를 각각의 분류기를 통해 결과를 분석한 결과 범주 분류의 경우 문서당 범주선택을 1 개로 제약하여 수작업으로 분류하였고, 분류기 분류의 경우 분산된 분류기를 통해 독립적으로 분류가 되었기 때문에 전체 데이터 보다 많은 문서수가 나왔다. 다중분류를 고려하여 수작업을 통한 다중 분류 수치를 비교한 결과 <그림 7>과 같이 특성 단어에 따른 다중 분류가 수행 되었음을 알 수 있다. 하지만 스포츠, 정치와 다르게 사회, 국제, 문화 범주와 같이 범주별 연관성이 높은 경우 범주별 분류가 낮게 이루어짐을 알 수 있다.



<그림 7> 문서분류 성능비교

<그림 8>은 실험 데이터를 기준으로 측정된 문서수를 기준으로 분류율을 계산하였다. 분류율은 다중분류시 측정된 문서 수를 기준으로 적중된 문서 수의 비율을 계산하였으며, 범주간 연동관계가 적을수록 분류율이 높은 것을 알 수 있었다.



<그림 8> 분류성능 비교

6. 결론

본 논문에서는 새롭게 대두되고 있는 웹서비스를 이용하여 SVM 기반의 분산 문서 분류기를 구성하기 위한 설계와 이에 대한 분류 처리를 기술하였다. 또한 별도의 SVM 을 구성함으로써 독립적인 다중분류가 처리됨을 보였다. 향후 분산구조를 통한 독립적 문서 분류를 위해서는 범주별 연관관계와 독립적으로 전문가적 분류가 가능한 효율적인 분류기의 학습방법에 대해 연구할 예정이며, 분류기가 처리하는 데이터와 에이전트를 이용한 자동 데이터 수집을 통한 시스템 성능 향상에 대해 연구할 예정이다.

분류 성능유지를 위한 재학습 방법론 및 에이전트를 이용한 자동 데이터 수집을 통한 시스템 성능 향상에 대해 연구할 예정이다.

참고문헌

[1] Joachims, T., "Text categorization with support vector machines : Learning with many relevant features", Proc. European Conference on Machine Learning (ECML), 1998

[2] Vapnik, V, "The Nature of Statistical Learning Theory", Springer, 1995

[3] Joachims, "Support Vector Machine(SVM-Light)", http://svmlight.joachims.org/, 2004

[4] Joachims, "Learning to Classify Text using Support Vector Machines: Methods, Theory and Algorithms, Kluwer Academic Publishers, 2001

[5] Yang. And Pederson, "A comparative study on feature selection in text categorization.", Proceedings of the 14 International Conference on Machine Learning, 1997.

[6] Mitchell, Tom, "Machine Learning", McCraw Hill, 1996

[7] Frakes, W. B. and R. B. Yates. "Information Retrieval Data Structures & Algorithm, Prentice-Hall, 1997

[8] Gustavo Alonso And Fabio Casati, Harumi Kuno, Vijay Machiraju, Web Services, Concepts: Architectures and Applications, Springer, 2004

[9] Damien, "Programming Microsoft .NET XML Web Services, Microsoft Press, 2003

[10] Salton, G. Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer, Addison-Wesley, 1989