BClassifier : 나이브 베이지안 학습법에 기초한 북마크 분류 에이전트

최정민, 김인철
경기대학교 전자계산학과
e-mail : {eclips, kic}@kuic.kyonggi.ac.kr

BClassifier : A Bookmark-Classification Agent Based on Naive Bayesian Learning Method

Jung-Min Choi, In-Cheol Kim
Dept. of Computer Science, Kyonggi University.

요 약

최근 고성능 PC의 보급과 네트워크의 발달로 인하여 인터넷의 가용 정보가 폭발적으로 증가하고 있다. 이러한 추세에 따라 우리는 인터넷을 사용하여 많은 정보를 얻고 있다. 그러나 인터넷에 존재하는 정보는 수많은 웹 서버에 주소(URL)를 가지고 존재하게 되는데 사용자는 자신이 관심 있는 정보의 사이트를 재방문하기 위하여 웹 브라우저 북 마크 기능을 사용한다. 그러나, 북 마크를 효율적으로 사용하기 위해서는 북 마크 분류, 수정, 편집, 정렬등의 북 마크 관리가 필수적이지만 이와 같은 북 마크 관리 작업이 전반적으로 수작업으로 이루어져야 하는 단점이 있다. 이러한 문제점을 해결하기 위한 한가지 방법으로 웹 문서 분류를 위한 기계 학습법을 적용하여 사용자의 북 마크를 카테고리별로 자동으로 분류, 재정렬해주는 북 마크 자동 분류 에이전트를 개발하고자 한다. 대표적인 분류 에이전트 시스템으로는 전자우편 분류 에이전트인 Maxims, 뉴스 기사 분류 에이전트인 NewT, 엔터테인먼트 선별 에이전트인 Ringo 등이 있으며, 이러한 시스템들은 분류 대상과 분류 방법, 기능 등에서 차이를 보이고 있다. 본 논문에서는 대표적인 교사학습 방법인 나이브 베이지안 학습법을 사용하여 북 마크를 자동으로 분류하는 북 마크 자동 분류 에이전트를 설계, 구현하였다.

1. 서론

최근 인터넷의 발전은 급속도로 빠르게 발전해 나가고 있다. 매일 평균 20 억 이상의 웹 문서(web document) 가 증가하고 있으며, 다양한 정보를 우리는 인터넷상에서 접할 수 있게 되었다. 그러나 이러한 정보의 웹 문서들은 수많은 웹 서버에 산재 되어 있으므로, 사용자는 원하는 정보의 위치를 정확히 찾아내는 것도 어렵다. 그러므로 사용자는 관심 있는 정보의 주소(URL)를 기록하여 두었다가 재방문 할 때 사용하고자 웹 브라우저의 북 마크(bookmark) 기능을 사용 한다. 그러나 현재 웹 브라우저의 북 마크 기능을 효율적으로 사용하는 데는 몇 가지 문제점이 있다. 먼저 북 마크의 개수가 늘어남에 따라 북 마크의 분류와 정렬 작업이 지속적으로 이루어져야 하는 점과 이와 같은 북 마크 관리 작업이 웹 브라우저의 북 마크 편집기능을 이용하여 모두 수작업으로 이루어져야 한다는 점이다. 일반적으로 북 마크는 인터넷상의 웹 문서에 대한 주소를 저장하고 있는 것이므로, 북 마크가 가지고 있는 주소의 웹 문서를 분석하면 각 북 마크의 카테고리를 결정할 수 있고 이것을 바탕으로 카테고리별로 북 마크들을 모아 재정렬 할 수 있다. 본 논문에서는 수작업으로 일관된 북 마크 분류 및 정렬을 에이전트 기술을 사용하여 자동으로 수행 할 수 있는 시스템을 설계하고 구현하였다.

2. 관련연구

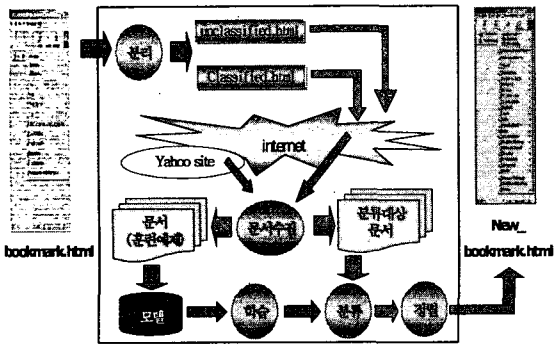
일반적으로 기계 학습법을 사용하여 복잡한 분류 작업을 자동으로 수행할 수 있는 자율적인 소프트웨어를 분류 에이전트라 한다. 문서분류를 위한 기계 학습법에는 대표적으로 통계적 확률을 이용한 나이브 베이지안기법(Naive Bayesian method)과 개체 기반 학습 기법인 k-NN 기법, 단어의 출현 빈도수를 이용한 TFIDF 기법 등이 있다.[3] 기계 학습법을 이용한 대표적인 에이전트 시스템을 알아보면 카네기 멜론 대학의 Personal WebWatcher 가 있다. 이 시스템은 사용자의 행동을 웹 브라우저를 통해 모니터링하여, 사용자의 관심영역을 학습한 뒤, 브라우징하는 웹 문서내의 링크들에 대해 사용자 관심영역에 속하는 것들과 그렇지 않은 것들을 분류하여 관심있는 링크들만을 제안 해주는 시스템이다. 또한, 앤더슨 컨설팅 연구실에서 만든 InfoFinder 역시 사용자의 관심 프로파일을 바탕으로 온라인 문서들에 대한 분류작업을 통해 사용자가 관심을 가질 문서들을 찾아주는 에이전트 시스템이다. 이외에도 MIT 대학에서 만든 전자우편을 분류하는 Maxims, 엔터테인먼트 선별 에이전트인 Ringo, 뉴스 기사 분류 에이전트인 NewT 등이 모두 문서 분류기법을 이용한 대표적인 에이전트 시스템이다.[4]

3. 시스템의 설계

3.1 기본 가정

본 시스템은 효과적인 북 마크 분류를 위하여 세 가지 기본 가정을 전제로 하고 있다. 첫 번째, 교사학습(supervised learning)을 위한 훈련예제는 웹 브라우저의 북 마크 파일에서 사용자가 이미 클래스별로 분류 해놓은 부분의 웹 문서와 인터넷에서 디렉토리 서비스를 제공하는 Yahoo 사이트의 최상위 14가지 분류 클래스에 따른 각 클래스 소속 웹 문서들이다. 두 번째, 사용자에 의해 분류 되어진 북 마크 클래스들과 Yahoo 사이트 14가지 클래스들은 서로 중복되지 않으며, 분류하고자 하는 북 마크는 반드시 이들 클래스중 하나에 속한다고 가정한다. 세 번째, 북 마크를 분류 할 때는 북 마크 된 주소의 웹 사이트 전체 문서를 고려하지 않고 주소가 가르키는 단일 웹 문서를 기초로 분류한다.

3.2 시스템 구조

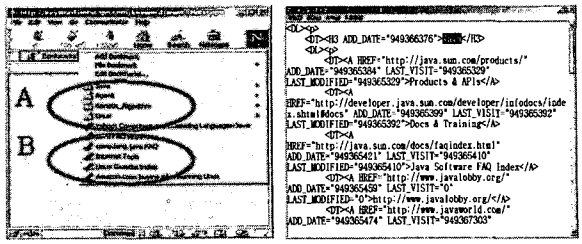


[그림 1] 시스템 구조

[그림1]는 전체 시스템의 구조를 보여주고 있으며, 수행 과정은 다음과 같다. 먼저 북 마크 파일(bookmark.html)에서 사용자에 의해 이미 분류된 북 마크들의 집합(classified.html)과, 분류되지 않은 북 마크들의 집합(unclassfied.html)을 분리한다. 분리된 두개의 파일에서 각각의 북 마크 주소에 해당되는 인터넷상의 웹 문서를 구분하여 수집하고, 디렉토리 서비스를 제공하는 Yahoo 사이트의 최상위 14가지 클래스에 대한 웹 문서를 인터넷에 접속하여 자동으로 수집한다. 이들 중에서 이미 분류된 북 마크 파일과, Yahoo 사이트에서 수집한 웹 문서들을 통합하여 모델화 과정을 거쳐 학습의 훈련예제로 사용하게 되며, 분류되지 않은 북 마크 파일의 웹 문서는 분류 대상 문서로 사용한다. 그리고 학습된 자료를 바탕으로 분류대상 문서의 소속 클래스를 결정하는 분류 작업을 거쳐 각 북 마크를 분류하게 된다. 그리고, 최종적으로 분류된 북 마크들을 재정렬하여 새로운 북마크 파일(New_bookmark.html)을 생성한다.

3.3 북마크 파일의 구성 및 분리

웹브라우저에서 하나의 북마크는 사용자가 평소에 인터넷에서 관심있는 웹문서의 주소를 나타낸다. [그림2]는 분류전의 상용 프로그램인 네스케이프 네비게이터 웹브라우저 북마크 모습(a)과 북마크 파일의 소스코드(b)를 나타내고 있다.



(a) (b)

[그림 2] 분류전의 북마크 파일

네스케이프의 북 마크 파일 소스 코드를 보면 [그림 2]의 (b)와 같이 html 문서 형식의 태그(<>)로 구성되어 있는 소스를 볼 수 있다. 북 마크가 분류된 부분(A)의 태그 구성(b)은 다음과 같다.

```
<DT><H3 ADD_DATE="949366376">Java</H3>
<DL><p>
<DT><A  HREF=http://www.javalobby.org/
ADD_DATE="949365459"  LAST_VISIT="0"
LAST_MODIFIED="0">http://www.javalobby.org/</A>
</DL><p>
```

위 소스코드를 보면 클래스의 이름이 "Java"라는 속에 http://www.javalobby.org의 주소가 소속되어 있는 것을 알 수 있다.

```
<DT><A  HREF=http://www-net.com/java/faq/
ADD_DATE="949367554" LAST_VISIT="949367592"
LAST_MODIFIED="949367523">Java FAQ Archives</A>
```

한편 분류 되지 않고 저장된 북 마크(B)는 클래스를 명시하지 않은 채 단지 http://www-net.com/java/faq/의 주소가 저장 되어 있는 형태로 존재 하게 된다. 따라서 북 마크 파일 분리는 소스 코드에서 </DL><P>를 구분자로 사용하여 분류된 부분과 분류되지 않은 부분을 분리 할 수 있다

3.4 웹 문서 수집

본 논문의 에이전트는 북 마크 파일 분류를 위하여 교사학습에 사용될 훈련예제를 북 마크 파일에서 이미 분류된 부분의 클래스에 소속된 북 마크의 웹 문서를 클래스 당 50개씩 너비우선 탐색(Breadth-First Search) 에 의하여 수집한다. 그리고, 인터넷 디렉토리 서비스를 제공하는 Yahoo 검색사이트의 최상위 14가지 클래스 체계와 클래스에 소속된 웹 문서를 클래스 당 50개씩 훈련예제로 추가하여 사용한다. 그리고 분류대상 문서들은 분류되지 않은 북 마크 파일에 저장된 주소를 이용하여 해당 웹 문서를 인터넷을 통하여 수집한다. 이러한 모든 수집 과정이 에이전트에 의해 자동으로 수행된다.

3.5 특징 추출 및 모델화

웹 문서의 정확한 키워드 추출과 학습의 신뢰성을 높이기 위해 전처리 과정을 먼저 거치게 된다. 전처리 과정은 html 형식으로 만들어진 웹 문서에서 태그(<>)를 제거하는 작업을 하여 보다 신속한 수행이 이루어 지도록 하며, 스테밍 처리와 불용어 처리를 통하여 키워드의 신뢰성을 높이게 된다. 전 처

리 과정 후 본 시스템에서는. 수집된 모든 웹 문서에 대하여 이진 속성 벡터(vector of binary attributes)로 모델화 한다. 수집된 웹 문서에 대해 이와 같은 이진 속성 벡터로 만들기 위해서는 먼저 웹 문서들로부터 각 문서를 표현하는데 사용할 단어들인 특징(feature)들을 추출하여야 한다. 본 시스템에서는 [식 1]과 같이 정보이론(Information Theory)에 입각해 엔트로피(entropy) 변화량을 기초로 특징 단어를 추출하는 방법인 정보 획득(Information Gain) 방법을 사용한다.[1]

$$InfoGain(w) = P(w)\sum_i P(c_i \mid w)\log\frac{P(c_i \mid w)}{P(c_i)} + P(\bar{w})\sum_i P(c_i \mid \bar{w})\log\frac{P(c_i \mid \bar{w})}{P(c_i)} \quad [식\ 1]$$

[식 1] 에 의해 [식 2] 와 같이 전체 .단어집합 V 에서 정보 획득량이 큰 L개의 단어를 추출한다.

$$K = \{w_1, w_2, w_3, ..., w_L\}, \quad K \subset V \quad [식\text{-}2]$$

그리고, 추출된 L개의 특징 단어들을 바탕으로 각 웹 문서에 대해 아래와 같은 모양의 이진 속성 벡터 모델을 만들게 된다.

$$d_i = (1,0,1,.... 1)$$

3.6 학습 및 분류

본 연구에서는 문서분류를 위한 대표적인 교사학습 알고리즘인 나이브 베이지안 학습기법을 통하여 북 마크의 분류가 이루어 진다. 그러나 이외에도 K -NN기법과 TFIDF 기법도 선택적으로 사용될 수 있도록 구현 되었다. C 를 [식 3]과 같이 전체 클래스들의 집합이라고 할 때

$$C = \{c_1, c_2, c_3, ... c_k\} \quad [식\ 3]$$

나이브 베이지안 분류법은 한 문서 d_i의 각 클래스 c_j 에 대한 조건부 확률들을 [식 4]와 같이 구해준다.

$$\Re(d_i) = \{P(d_i \mid c_1), P(d_i \mid c_2), P(d_i \mid c_3), ..., P(d_i \mid c_k)\} \quad [식\ 4]$$

본 시스템에서는 분류 대상 문서에 대하여 [식 5]와 [식 6]에 의해 가장 높은 확률값을 가지는 클래스로 분류하게 된다.

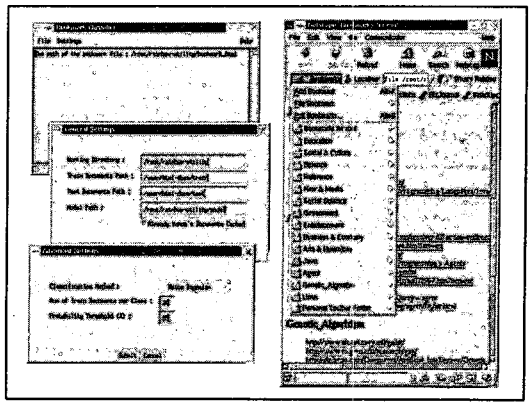
$$P_{\max}(d_i) = \max\{P(d_i \mid c_t)\}, \quad t = 1,...,k \quad [식\ 5]$$

$$c_{best}(d_i) = \begin{cases} c_j & \text{if } P_{\max}(d_i) = P(d_i \mid c_j) \geq T \\ c_{\text{unknown}} & \text{otherwise} \end{cases} \quad [식\ 6]$$

그러나, 클래스의 확률값이 일정한 임계값(threshold) T 이상 되지않으면 그만큼 분류에 대한 정확도와 신뢰도가 떨어진다, 따라서 이러한 경우에 분류가 자동으로 이루어지지 않고 사용자에게 분류 결정을 양도 하게 된다.

4. 시스템의 구현

본 시스템의 구현은 300Mhz 펜티엄 II 프로세서, 128M 주기억 장치와 리눅스 환경의 컴퓨터에서 인터페이스 구현 부분은 자바를 사용하였으며 내부수행은 펄을 사용하여 구현하였다. [그림 3] 은 구현한 북 마크 분류 에이전트 시스템의 실행화면을 보여준다. [그림 3]의 좌측 상단에 위치한 사용자 인터페이스의 주윈도우는 사용자 선택 메뉴들과 시스템의 실행상태를 보여준다. 그 아래에 위치한 두 개의 윈도우는 사용자의 북 마크 파일의 위치와 작업 폴더 위치, 모델 경로를 설정하는 " General Settings " 윈도우와 기계학습 기법 선택, 클래스 당 기계 학습 문서 개수, 임계값을 설정하는 " Advanced Settings " 윈도우이다. [그림 3]의 우측에는 에이전트가 북 마크 파일에 대한 분류 및 정렬 작업을 마친 후 그 결과를 네스케이프 웹 브라우저에서 보여주고 있다.



[그림 3] 시스템 수행 모습

5. 결론

본 논문에서는 분류되지 않은 웹 브라우저의 북 마크 파일을 문서분류 기계학습 방법을 사용하여 자동으로 분류하는 에이전트 시스템을 구현하였다. 본 에이전트 시스템의 성능과 정확도를 높이기 위해서 앞으로 시행되어야 할 향후 연구과제로는 분류 클래스들의 세분화, 분류 클래스들간의 계층관계 및 중복관계에 대한 해결 등이 있다. 또한, 분류 클래스별로 충분한 훈련예제를 제공 할 수 없는 경우를 위해 비교사 학습기법(unsupervised learning)에 대한 도입을 검토해볼 예정이다

[참고문헌]

[1]Dunja Mladenic and Marko Grobelnik " Feature selection for classification based on text hierarchy"

[2] McCallum, A. Nigam, K. 1998 " A Comparison of Event Models for Naive Bayes Text Classification " In AAAI -98 Workshop on Learning for Text Categorization.

[3]Tom M. Mitchell " Machine Learning " McGRAW -HILL international Edition.

[4] Jeffrey M. Bradshaw " Software Agent" AAAI Press/The MIT Press pp151-161

[5]백혜정, 박영택, 윤석환 ' 사용자 관심도를 이용한 웹에이전트' 정보처리학회지 1997. 9.