나이브 베이지안 학습법에 기초한 북마크 분류 에이전트

최정민, 김인철
경기대학교 전자계산학과

Bookmark Classification Agent Based on Naive Bayesian Learning Method

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

요 약

최근 인터넷의 발전으로 많은 정보와 지식을 우리는 인터넷에서 제공받을 수 있게 되었다. 인터넷에 존재하는 정보는 수많은 웹서버에 산재되어 있으며, 정보의 위치는 주소(URL)를 가지고 존재하게 되는데 사용자는 자신이 관심있는 정보의 주소를 저장하기 위하여 웹브라우저 북마크(Bookmark)기능을 사용한다. 그러나, 북마크 기능은 웹문서의 주소 저장에 일차적인 목적를 두고 있으며, 이후 북마크의 개수가 증가하면, 사용자는 북마크 관리가 어렵게 되므로 사용자 북 마크 파일을 자동으로 분류하여 관리 할 수 있는 에이전트 기술을 사용 하고자 한다. 대표적인 분류 에이전트 시스템으로는 전자우편 분류 에이전트인 Maxims, 뉴스 기사 분류 에이전트인 NewT, 엔터테인먼트(Entertainment) 선별 에이전트인 Ringo 등이 있다. 이러한 시스템들은 분류할 대상에 따라 조금씩 다른 모습의 에이전트 기능을 보이고 있으며, 본 논문은 기계학습 이론 중 교사학습 알고리즘인 나이브 베이지안 학습방법(Naive bayesian learning method)을 사용하여 사용자가 분류하지 못한 북마크를 자동으로 분류하는 단일 에이전트 기반 북마크 분류기를 설계, 구현하고자 한다.

1. 서론

최근 인터넷의 발전은 급속도로 빠르게 발전해 나가고 있다. 평균 매일 20억 이상의 웹문서(Web-Page)가 증가하고 있으며, 다양한 정보를 우리는 인터넷상에서 접할수 있게 되었다. 그러나 이러한 정보의 웹문서들은 수많은 웹서버에 산재되어 있으므로, 사용자는 원하는 정보의 위치를 정확히 찾아내는 것도 어렵다. 그러므로 사용자는 자신이 관심있는 정보의 주소를 저장하기 위하여 웹브라우저의 북마크 기능을 사용한다. 그러나, 북마크 기능은 웹문서의 주소 저장에 일차적인 목적를 두고 있으며, 이후 북마크의 개수가 증가하면, 사용자는 북마크 관리가 어렵게 되므로 사용자 북마크를 자동으로 분류하여 관리 할 수 있는 에이전트 기술을 사용 하고자 한다. 북마크의 사용자 분류는 사용자가 웹문서의 클래스(class)를 파악하여 직접 분류를 하게 된다. 그러나 이러한 분류는 북마크 되어진 웹문서의 내용에 충실한 단어를 사용하여 북마크를 분류하는 것이 쉬운 작업이 아니다. 또한, 분류되지 않고 저장된 북마크의 개수가 많아지게 되면, 사용자는 북마크를 관리함에 있어 매우 어렵게 된다. 따라서, 본 논문에서는 웹브라우저 전체 북마크 파일에서 분류되지 않고 저장된 북마크를 자동으로 분류하여 북마크 관리를 사용자 대신 할 수 있는 에이전트 시스템을 설계, 구현 하였다. 북마크 분류는

인터넷상의 웹문서 주소를 북마크가 저장하고 있는 것이므로, 북마크가 가지고 있는 주소의 웹문서 문서를 분석하여 클래스를 결정하면, 웹문서의 주소를 나타내는 북마크를 분류할 수 있다. 따라서, 북마크 분류는 문서분류 방법을 사용하여 북마크를 분류하는 것이다. 본 시스템에서 사용된 북마크 분류를 위한 학습방법은 문서분류 기계학습 이론 중에 교사학습 알고리즘(Supervisor Learning Algorithm)을 사용하여 분류하는 기능을 가진 북마크 분류 에이전트(Bookmark Classification Agent)이다.

2. 관련연구

에이전트를 분류 시스템에 응용하여 복잡한 분류작업을 자동으로 수행하는 사례는 여러 가지가 있는데, 그중 대표적인 분류 대상으로는 전자우편, 뉴스 기사, 엔터테인먼트 선별 등이 있다. 다음은 분류대상에 따른 분류 에이전트 시스템에 대한 사례이다.

2.1. Maxims

에이전트 기반 분류 시스템에서 분류대상으로 전자우편을 분류 하는 시스템은 Maxims(Lashkari, Metral, and Maes 1994)이다.[2] 이 시스템은 메모리 기반 학습 알고리즘을 사용하여 개발 되었으며, 사용자 전자우편에 대하여 순위결정, 삭제, 정렬, 기록을 수행하여 전자우편을 관리하는 에이전트 이다. 이 에이전트는 평소 사용자가 전자우편을 받았을 때 행동하는 모습을 학습의 자료로 사용하기 위해 메모리에 저장하였다가, 비슷한 상황이 발생하였을 때 사용자의 행동에 조언을 하는 에이전트이다.

2.2 NewT

인터넷의 발전으로 수많은 정보가 네트웍으로 들어오는 가운데 뉴스분야의 정보는 계속적인 스트림(stream)의 형태로 네트웍상으로 유입된다. 이러한 뉴스의 스트림 가운데 사용자가 원하는 기사의 선택을 위한 에이전트 시스템으로 NewT가 있다.[2] NewT는 뉴스의 기사를 정치, 경제, 컴퓨터, 스포츠 4가지 클래스로 필터링 한다. 뉴스기사 문서 분석은 벡터-공간 모델을 사용한 풀-텍스트 분석으로 이루어지며, NewT 에이전트의 특징은 에이전트 협동 부분인데, 사용자는 충분히 학습된 에이전트를 복사하여 다른 사용자에게 제공할수 있도록 유닉스환경의 c++로 구현되었다.

2.3 Ringo

에이전트 기반 분류 시스템의 응용프로그램 중에서 엔터테인먼트 선별을 위한 시스템으로 Ringo가 있다.[2] 엔터테인먼트 선별은 어떤 다른 분류 대상 보다 앞으로 가장 핵심이 되는 대상이 될 가능성이 높은 에이전트 시스템이다. Ringo는 개인을 대상으로 구성된 음악추천시스템 (Standanand and Maes 1995)으로 유닉스 환경에서 perl언어로 구현되었다.

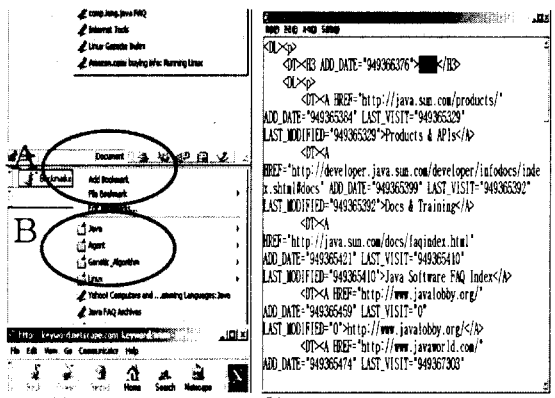
3. 북마크 파일 분류 에이전트

3.1 기본 가정

본 논문에서는 북마크 파일 분류를 위해서 필요한 기본 가정은 세 가지가 있다. 첫번째, 웹브라우저 북마크 파일에서 북마크가 나타내는 주소는 웹사이트 전체를 나타내는 것이 아니라 주소가 가르키는 웹문서를 나타낸다. 두번째, 북마크 파일 분류는 각각의 북마크에 저장되어 있는 웹문서를 분석하여 북마크를 분류한다. 세번째, 교사학습을 위한 훈련예제는 북마크 파일에서 분류된 부분의 클래스와 클래스에 따른 소속된 북마크에 저장된 주소의 문서들과 디렉토리 서비스를 제공하는 야후 사이트의 14가지 클래스와 각각의 클래스에 따른 소속 문서들이다.

3.2 북마크 파일의 구성

웹브라우저 북마크는 사용자가 평소에 인터넷에서 관심있는 웹문서의 주소를 저장한 모습을 나타낸다.

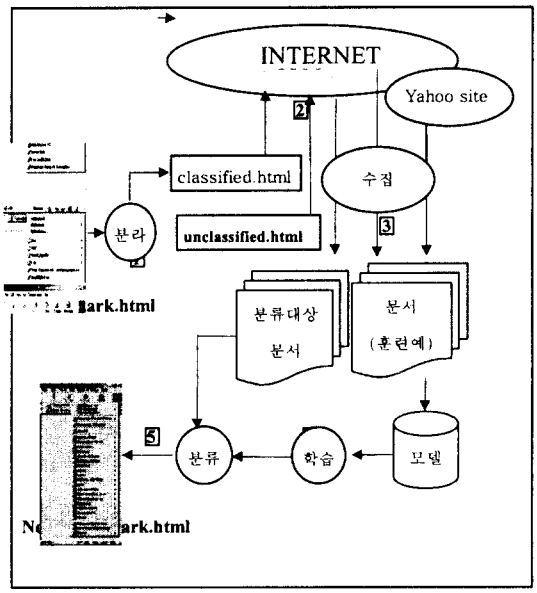


[그림1] 분류전의 북마크 파일

[그림1]은 분류전의 상용 프로그램인 네스케이프 네비게이터 웹브라우저 북마크 모습(a)과 북마크 파일의 소스코드(b)를 나타내고 있다. 분류전의 북마크는 저장은 되어 있지만 북마크의 분류는 충분히 이루어지지 못하고, 단순히 사용자의 관점에 따른 임의의 분류가 이루어진 부분(A)과, 분류가 이루어 지지 않은 부분(B)으로 구분 되어 있다.

3.3 처리 과정

본 논문에서 구현하고자 하는 북 마크 파일 분류 시스템은 에이전트 기반 분류 시스템의 분야로서 분류 대상인 웹브라우저 북마크 파일에서 분류되지 않은 부분의 분류를 에이전트가 사용자 대신 수행하여 북마크의 관리를 보다 효과적으로 수행하기 위한 시스템이다.



[그림 2] 시스템의 전체 구성도

[그림2]는 시스템의 전체 구성을 보여주고 있으며, 다음과 같은 순서에 의해 수행된다.

1 북마크 파일의 분리 : 웹브라우저 북마크 파일에서 분류된 북마크(classified.html)와, 분류되지 않은 북마크(unclassfied.html)를 분리한다.

2 문서수집 : 분리된 두개의 파일에서 각각의 북마크 주소에 해당되는 인터넷상의 웹문서를 분리하여 수집하고, 디렉토리 서비스를 제공하는 사이트의 문서분류 클래스와 클래스에 따른 웹문서를 같이 수집한다.

3 훈련예제 : 분류된 북마크의 사용자가 만든 클래스와 웹문서 그리고, 디렉토리 서비스를 제공하는 사이트의 클래스와 웹문서를 통합하여 학습의 훈련예제로 사용한다.

4 모델화 및 학습 : 훈련예로 사용될 클래스와 문서의 모델화 작업을 수행하고 학습을 한다.

5 분류 및 북마크 파일의 생성 : 학습된 자료를 사용하여 분류대상 문서를 분류하고 문서에 해당되는 북마크를 분류한다. 그리고, 새로운 북마크 파일 (New_bookmark.html)을 생성한다.

3.4 문서수집

본 논문의 에이전트는 북마크 파일 분류를 위하여 교사학습에 사용될 훈련예제를 분류전의 북마크 파일에서 분류된 부분(classification.html)의 클래스에 소속된 북마크의 웹문서를 클래스당 50개씩 너비우선 탐색에 의하여 수집한다. 그리고, 보다 정확한 분류를 위하여 디렉토리 서비스를 제공하는 야후 검색사이트의 최상위 14개의 클래스 체계와 클래스에 소속된 웹문서를 또한, 클래스당 50개씩 훈련예제로 추가하여 사용한다. 분류대상 북마크의 문서수집은 북마크에 저장된 주소의 해당 웹문서를 사용한다.

3.5 모델 및 학습

본 논문에서는 에이전트 북마크 파일 분류를 위한 기계학습 이론중 교사학습 알고리즘의 대표적인 나이브 베이지안 학습기법을 통하여 북마크의 분류가 이루어진다. 이 학습방법은 모든 문서에서 특정단어의 출현으로 구별되는 이진 속성 벡터(vector of binary attributes)로 표현된 모델로 문서를 정형화 하는데, 모델은 다형성 베르누이 사건 모델(multi-variate Bernoulli event model)을 기초로 하여 각 클래스의 문서 마다 다르게 모델을 만들게 된다. 나이브 베이지안 가설은 문서들의 모든 속성은 주어진 전체 클래스의 다른 문서의 전후관계에 대해서 독립적이다.

$$P(d_i|c_j) = \prod_{t=1}^{|V|} (B_{it}P(w_t \mid c_j) + (1 - B_{it})(1 - P(w_t \mid c_j)))$$

[식-1]

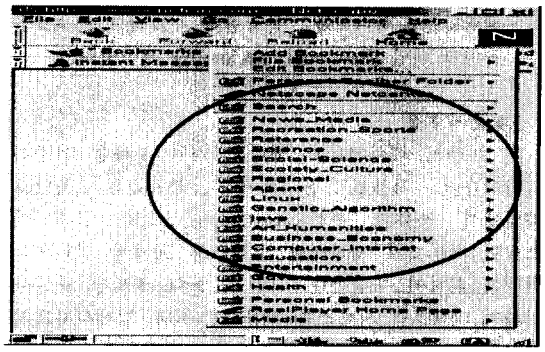
- B_{it} : 문서 d_i를 위한 벡터의 값 ($d_i = 0,1$)
- $p(d_i \mid c_j)$: 클래스 c_j에 문서 d_i가 나올 확률
- $P(w_t \mid c_j)$: 클래스 c_j에 단어 w_t가 나올 확률

[식-1]은 모델화 작업으로 만들어진 문서의 모델을 사용하여 각각의 클래스에 따른 문서의 확률값을 구하고, 구해진 확률값중 가장 높은 확률값을 가진 클래스에 문서를 분류하게 된다.

3.6 북마크 파일의 생성

북마크 파일 분류 에이전트 시스템의 북마크 분류작업이 모두 이루어지면 다음과 같은 새로운 북

마크 파일이 생성된다.



[그림 3] 분류후의 북마크

4. 실험

4-1 실험 목표

본 논문에서 구현한 북마크 파일 분류 에이전트의 성능실험은 나이브베이지안 기계학습법을 사용한 본 시스템의 성능과 다른 문서 분류 기계학습법인 k-NN(k-Nearest Neighbor) 방법과 비교하여 각각의 알고리즘을 적용했을 때 문서를 분류하는 정확도를 측정하고, 실험 데이터 개수를 다르게 했을 때의 정확도 값을 측정하였다.

4-2 실험 방법

분류 에이전트의 성능실험을 위한 데이터는 디렉토리 서비스를 제공하는 검색 사이트의 분류체계에 따른 웹문서를 받아서 사용하였다. 전체 클래스는 14개이며, 각각의 클래스에서 50개씩, 총700개의 웹문서를 실험에 사용하게 된다. 실험은 50개씩 문서를 가지고있는 14개의 클래스에서 임의로 10개의 문서를 분류대상 북마크로 발췌한다. 그리고 나머지 문서들 중에서 훈련예제의 개수를 10개, 20개, 30개로 증가시키면서 두가지 기계학습법을 비교하고, 정확도 값을 구하게 된다.

4-3 결과분석

(정확도 %)

학습법 / 훈련예(t)	Naive Baysian	k-NN
T=10	82.14	55.00
T=20	79.00	47.55
T=30	81.14	58.57

(분류대상 북마크 개수(m)=10)

[표 1] 분류기의 성능 비교표

[표 1]의 나이브 베이지안과 k-NN의 분류성능을 비교해 보면, 본 논문의 북 마크 분류 에이전트에서 사용한 나이브 베이지안 방법이 k-NN 방법보다 우수한 문서분류 성능을 보여주고 있다는 것을 알 수 있다.

5. 결론 및 향후 연구과제

본 논문에서는 분류되지 않은 웹브라우저 북마크 파일을 기계학습 방법 중 나이브 베이지안 학습방법을 이용하여 북마크 파일을 분류하는 에이전트 시스템을 구현하였다. 웹브라우저 북마크 파일의 기존의 분류는 사용자의 관점에서 분류작업이 이루어졌는데 북마크의 개수가 증가하면 분류작업에 쉽게 이루어지지 않는다. 이러한 문제의 해결을 위하여 에이전트를 이용한 분류 시스템을 구현하게 되었다. 에이전트 분류 시스템의 성능과 정확도을 높이기 위해서 앞으로 시행되어야 할 향후 연구과제로서 요구되는 부분은 충분한 분류 학습기반지식의 확보를 위한 정확한 분류 학습능력 배양을 위한 문서가 요구되며, 현재 교사학습 방법이 아닌 비교사 학습방법으로 분류할 수 있도록 하기위한 연구가 진행 중이며, 향후 연구해야 할 과제이다.

[참고문헌]

[1] 임윤택, 윤충화 '고정 분할 평균법에 기반한 점진적 알고리즘' 정보처리학회 가을 학술발표논문집 제6권 제1호, pp.559 1999

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

[3] McCallum, A. Nigam, K. 1998 "A Comparison of Event Models for Naïve Bayes Text Classification" In AAAI-98 Workshop on Learning for Text Categorization. http://www.cs.cmu.edu/~mccallum.