소셜 북마킹 시스템의 스패머 탐지를 위한 기계학습 기술의 성능 비교

(Comparative Study of Machine Learning Techniques for Spammer Detection in Social Bookmarking Systems)

김 찬 주 † 황 규 백 ††
(Chanju Kim) (Kyu-Baek Hwang)

요 약 소셜 북마킹(social bookmarking) 시스템은 사용자가 북마크를 저장하고 공유할 수 있는 플랫폼을 제공하는 웹 기반(web-based) 시스템으로 폭소노미(folksonomy)를 이용한 대표적인 웹2.0 서비스이다. 소셜 북마킹 시스템에서의 스패머(spammer)란 자신들의 이익을 위해서 시스템을 고의적으로 악용하는 사람을 말한다. 스패머는 많은 양의 잘못된 정보를 시스템에 포스팅(posting)하기 때문에 전체 소셜 북마킹 시스템의 리소스(resource)를 쓸모없게 만들어 버린다. 따라서, 스패머를 빠른 시간 안에 탐지하고 그들의 접근을 차단하는 것은 시스템의 붕괴를 방지하기 위해 중요하다. 본 논문에서는 사용자가 사용한 태그에 대한 데이터를 추출하여, 사용자가 스패머인지 아닌지를 예측하는 모델을 기계학습의 다양한 방법을 적용하여 생성한 후 그 성능을 비교해 보았다. 구체적으로, 결정테이블(decision table, DT), 결정트리(decision tree, ID3), 나이브 베이즈 분류기(naïve Bayes classifier), TAN(tree-augmented naïve Bayes) 분류기, 인공신경망(artificial neural network)의 방법을 비교하였다. 그 결과 AUC(area under the ROC curve)와 모델 생성시간을 고려하였을 때 나이브 베이즈 분류기가 가장 만족할 만한 성능을 보였다. 나이브 베이즈 분류기의 분류 결과가 가장 좋았던 이유는 성능을 비교하는 데 사용된 AUC가 결정트리 계열의 방법(ID3 등)보다 나이브 베이즈 분류기에서 일반적으로 높게 나오는 경향이 있다는 것과, 스패머 탐지 문제가 선형으로 분리 가능한 경우(linearly separable)와 유사할 가능성이 높기 때문으로 여겨진다.

키워드 : 웹2.0, 소셜 북마킹, 스패머 탐지, 기계학습, 나이브 베이즈 분류기

Abstract Social bookmarking systems are a typical web 2.0 service based on folksonomy, providing the platform for storing and sharing bookmarking information. Spammers in social bookmarking systems denote the users who abuse the system for their own interests in an improper way. They can make the entire resources in social bookmarking systems useless by posting lots of wrong information. Hence, it is important to detect spammers as early as possible and protect social bookmarking systems from their attack. In this paper, we applied a diverse set of machine learning approaches, i.e., decision tables, decision trees (ID3), naïve Bayes classifiers, TAN (tree-augment naïve Bayes) classifiers, and artificial neural networks to this task. In our experiments, naïve Bayes classifiers performed significantly better than other methods with respect to the AUC (area under the ROC curve) score as well as the model building time. Plausible explanations for this result are as follows. First, naïve Bayes classifiers are known to usually perform better than decision trees in terms of the AUC score. Second, the spammer detection problem in our experiments is likely to be linearly separable.

Key words : web 2.0, social bookmarking, spammer detection, machine learning, naïve Bayes classifiers

1. 서 론

소셜 북마킹(social bookmarking) 시스템은 폭소노미(folksonomy)를 이용한 대표적인 웹2.0 서비스로 시스템의 사용자와 데이터가 급격히 증가하는 등 많은 발전을 하고 있다[1]. 폭소노미를 사용한 시스템은 리소스(resource)를 기계가 자동적으로 분류하는 것이 아니라 리소스의 내용을 이해하고 있는 사람이 직접 태그(tag)를 리소스에 붙여서 분류하기 때문에 기존의 택소노미(taxanomy)로는 수행할 수 없었던 일을 가능하게 해주는 장점이 있다. 하지만 반대로 태그에 대한 표준과 제한이 없기 때문에 잘못된 사용으로 인해서 시스템 전체가 쓸모없게 될 수 있다는 약점을 가지고 있다[2]. 그렇기 때문에 시스템을 악용하는 사용자를 빠른 시간 안에

· 본 논문은 숭실대학교 교내연구비, 서울시정개발연구원의 서울시 기술기반 구축사업(GS070167C093111) 및 2007년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원(KRF-2007-331-D00414)을 받아 연구되었음
· 이 논문은 제35회 추계학술대회에서 '소셜 북마킹 시스템의 스패머 자동 탐지를 위한 기계학습 기술의 성능 비교'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 숭실대학교 컴퓨터학부
cjkim@ml.ssu.ac.kr
†† 정 회 원 : 숭실대학교 컴퓨터학부 교수
kbhwang@ssu.ac.kr

논문접수 : 2008년 12월 19일
심사완료 : 2009년 3월 1일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제15권 제5호(2009.5)

미리 탐지하고 접근을 차단하는 것은 시스템 전체의 붕괴를 막기 위해 중요하다. 본 논문에서는 시스템을 악용하는 사용자-스패머(spammer)-를 탐지하기 위한 모델을 다양한 기계학습 기법들을 이용해 생성한 뒤, 각각의 성능을 비교 및 분석하여 이 작업에 가장 적합한 방법을 결정하고 그 이유를 분석하였다.

논문의 구성은 다음과 같다. 2장에서는 소셜 북마킹의 정의와 특징에 대해서 살펴보고, 3장에서는 스패머 탐지 문제와 사용한 데이터에 대해 자세히 설명한다. 4장에서는 각각의 기계학습 방법을 적용한 결과를 비교 및 분석하고, 마지막으로 결론과 향후 연구 방향에 대해서 논의한다.

2. 소셜 북마킹(Social Bookmarking)

2.1 소셜 북마킹 시스템의 정의

소셜 리소스 공유 시스템(social resources sharing system)은 사용자들이 자신들의 리소스를 업로드(upload)하고 태그라고 불리는 임의의 색인어로 리소스에 이름을 붙이는 형태의 웹 기반(web-based) 시스템이다. 이러한 시스템은 어떠한 종류의 리소스를 다루는가에 따라서 구분될 수 있다. 예를 들면, 플리커(Flickr)[1]는 사진을 공유하며, 43things[2]는 개인적인 목표를 공유한다. 소셜 북마킹 시스템은 북마크(bookmark)를 공유하는 시스템으로 그 대표적인 예로는 딜리셔스(del.icio.us)[3], 빕소노미(Bibsonomy)[4], 마가린(mar.gar.in)[5] 등이 있다[3].

다시 말해, 소셜 북마킹 시스템은 사용자가 자신의 북마크를 온라인에 저장하고 태그를 이용해 분류할 수 있는 시스템이다. 또한 사용자 간의 북마크 정보 공유를 통해서 북마킹의 분류와 검색까지 효과적으로 행하고 관리할 수 있는 도구를 제공한다. 이들 시스템의 중요한 특징은 사용자가 리소스를 시스템에 더한 후 직접 태그를 붙이고 그것을 공유한다는 것이다. 이렇게 여러 사용자의 집단지성(collective intelligence)이 모여서 폭소노미(folksonomy)를 이룬다.

2.2 폭소노미(Folksonomy)

폭소노미는 택소노미(taxanomy)와 포크(folk)의 합성어로 사람에 의해 만들어진 개념의 구조를 뜻한다[3]. 택소노미는 전문가가 모든 자료를 일정한 기준에 따라 계층적 구조로 분류해 둔 것을 말한다. 택소노미는 크게 두 가지의 한계를 가지고 있는데, 하나의 리소스가 두 가지 이상의 카테고리에 속할 때 분류하기가 힘들다는 점과 계층구조 전체를 파악하고 있지 않으면 리소스를 찾기가 힘들다는 점이다. 기존의 택소노미와 다르게 집단지성을 이용한 폭소노미는 특정한 리소스에 그와 연관된 태그-색인어-를 꼬리표처럼 붙여서 택소노미로 수행하기 어려운 작업들을 가능하게 해준다[4]. 예를 들어 두 가지 이상의 카테고리에 속하는 자료를 쉽게 관리 할 수 있으며, 집단 지성을 이용한 분류와 검색이 가능하다.

2.3 폭소노미를 이용한 소셜 북마킹 시스템의 특징

폭소노미를 이용한 소셜 북마킹 시스템은 다음과 같은 특징이 있다. 첫째, 북마크가 기계에 의해서 자동으로 분류되지 않고 북마크의 내용을 이해하는 사람에 의해서 분류되기 때문에 북마크가 의미를 가지고 있으며 유용하다. 둘째, 사용자들은 아직 유명하지 않거나 기존의 웹 검색엔진에 등록되지 않은 웹 페이지를 북마크하는 경향이 있기 때문에 새로운 웹 페이지를 찾을 수 있다. 셋째, 사용자에 의해서 특정 링크가 얼마나 많이 북마크 되었는지를 알 수 있기 때문에 이를 가지고 북마크의 순위를 매길 수 있다.

반면에 덧붙일 수 있는 태그 구조의 표준이 없기 때문에 잘못되거나 불명확한 태그가 만들어져 시스템이 잘못된 결과를 낼 수 있고, 시스템을 고의적으로 악용하는 사용자에 의해서 시스템 내용의 상당 부분이 쓸모없어질 수 있다는 약점을 가지고 있다.

2.4 소셜 북마킹 시스템에서의 스패머(Spammer)

소셜 북마킹 시스템에서의 스패머란 의미 없거나 잘못된 태그를 등록하는 등 시스템을 고의적으로 악용하는 사람을 말하며, 스패머가 시스템을 악용하는 이유에는 크게 다음과 같은 두 가지가 있다. 첫째, 시스템에 특정 사이트의 링크를 두어서 이목을 끌어 자신들의 사이트를 광고한다. 둘째, 유명한 웹2.0 사이트에 그들의 사이트로 향하는 링크를 최대한 많이 두어서 자신들의 사이트의 페이지랭크(PageRank)를 높여 구글과 같은 검색엔진에서의 노출을 확대한다. 이러한 목적을 가지고 시스템을 남용하는 사람을 스패머라고 부른다.

3. 소셜 북마킹 시스템에서의 스패머 탐지

3.1 문제 설명

소셜 북마킹 시스템의 인기가 상승하면서 스패머들은 자신들의 사이트를 광고하기 위해서 시스템에서 적극적으로 활동하기 시작하였다. 이러한 스패머는 시스템의 성능에 큰 손실을 주고, 시스템 전체를 쓸모 없게 만들 수 있기 때문에 최대한 빨리 활동을 하지 못하도록 막아야 한다. 자동가입방지(captcha)[6]라는 방법이 있지만, 스패머들은 자동가입방지를 침투할 수 있는 프로그램이

1) http://www.flickr.com
2) http://www.43things.com
3) http://del.icio.us
4) http://bibsonomy.org
5) http://mar.gar.in
6) Completely Automated Public Turing test to tell Computers and Humans Apart

나 값싼 노동력을 이용하여 활동하기 때문에 이것만으로는 스패머에 의한 시스템의 악용을 효과적으로 차단할 수 없다. 스패머의 활동을 최대한 저지하기 위해서는 사용자가 스패머인지 아닌지를 잘 판단하는 효과적인 모델을 데이터로부터 학습하는 것이 효율적이며 효과적인 방법이 될 수 있다.

3.2 실험에 사용한 데이터

실험에 사용한 데이터는 ECML & PKDD 2008의 Discovery Challenge[7]에서 제공한 것이다. 데이터는 북마크와 빕텍스(bibtex) 정보를 공유하는 빕소노미(Bib-Sonomy)에서 수집된 데이터로 2,467명의 액티브 유저(active user 혹은 non-spammer)와 29,248명의 스패머에 대한 것이며, 이는 수작업으로 분류되어 제공되었다[8]. 데이터는 총 7개의 데이터베이스 테이블로 구성되어 있으며, 그 각각의 내용과 크기는 표 1과 같다.

표 1 실험에 사용한 데이터의 구성 및 내용

테이블 명	테이블 크기 (라인 수)	테이블 설명
tas	816,197	액티브 유저 중에 누가 어떤 리소스에 어떠한 태그를 붙였는지에 대한 테이블
tas_spam	13,258,759	스패머 중에 누가 어떤 리소스에 어떠한 태그를 붙였는지에 대한 테이블
bookmark	181,833	액티브 유저에 의해 시스템에 더해진 북마크 정보에 대한 테이블
bookmark _spam	2,059,991	스패머에 의해 시스템에 더해진 북마크 정보에 대한 테이블
bibtex	219,417	액티브 유저에 의해 시스템에 더해진 빕텍스(bibtex) 정보에 대한 테이블
bibtex _spam	716	스패머에 의해 시스템에 더해진 빕텍스(bibtex) 정보에 대한 테이블
user	31,715	각 유저가 스패머인지 스패머가 아닌지 표시되어 있는 테이블

3.3 특성(Feature) 추출 및 학습데이터 생성 과정

데이터에 포함된 여러 가지 특성(feature) 중에서 사용자가 사용한 태그를 모델 학습에 사용하였다. 주어진 데이터에서 사용된 약 400,000개의 태그 중에서 유용한 태그를 선택하기 위해서 태그의 상호정보량(mutual information)[5]을 이용하였다. 각 태그의 상호정보량을 계산한 후에 그 값이 높은 태그들에 대하여 유저가 그 태그를 사용했는지 사용하지 않았는지를 속성으로 하여 학습데이터를 생성하였다. 높은 상호정보량을 가진 태그로 학습데이터를 생성하는 과정은 그림 1과 같으며 아래는 각 단계에 대한 자세한 설명이다.

7) http://www.kde.cs.uni-kassel.de/ws/rsdc08/
8) http://www.kde.cs.uni-kassel.de/ws/rsdc08/dataset.html

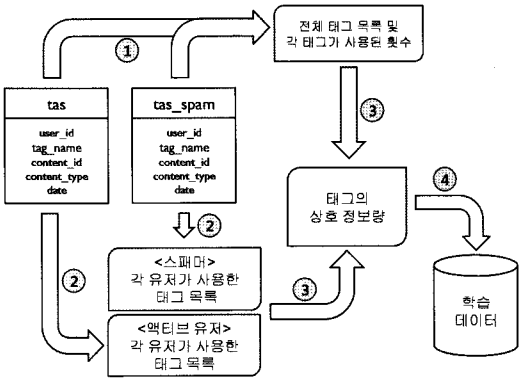


그림 1 학습데이터 생성 과정

1) tas, tas_spam 테이블에서 사용된 태그와 각 태그가 사용된 횟수를 추출한다.
2) tas, tas_spam 테이블에서 스패머와 액티브유저[non-spammer]를 나누어서 각 유저가 사용한 태그를 추출한다.
3) 1)과 2)에서 추출한 데이터를 이용하여 태그의 상호정보량을 식 (1)과 같이 계산한다. 아래 수식에서 Tag_i는 이진 변수로 i번째 태그가 사용되었는지 아닌지를 나타내고, $Target$은 목표 변수로 스패머와 액티브 유저의 두 가지 값을 가진다. $\hat{P}$는 주어진 데이터에서 추정된 확률 값을 나타낸다.

$$I(Tag_i; Target) = \sum_{tag_i, target} \hat{P}(Tag_i, Target) \log \frac{\hat{P}(Tag_i, Target)}{\hat{P}(Tag_i)\hat{P}(Target)} \quad (1)$$

4) 상호정보량이 높은 태그들을 가지고 학습데이터를 생성한다. 그 개수에 대해서는 다양한 경우를 실험하였다.

4. 실험 및 결과

4.1 학습데이터 및 테스트데이터

주어진 데이터는 1989년 1월부터 2008년 3월까지의 데이터로 시간이 지남에 따라서 데이터의 양이 증가하는 특징을 가지고 있다. 학습데이터로는 2008년 1월까지의 데이터를 사용하고, 테스트데이터로는 2008년 2월과 3월의 데이터를 사용하였다. 데이터에 대한 통계는 다음과 같다.

표 2 포스팅 수

	액티브유저	스패머	합계
학습데이터	260,271	1,264,539	1,524,810
테스트데이터	8,421	362,266	370,687
합계	268,692	1,626,805	1,895,497

표 3 유저 수

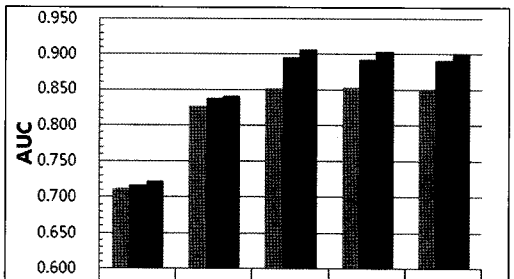
	액티브유저	스패머	합계
학습데이터	2,466	29,248	31,714
테스트데이터	656	10,610	11,266

4.2 적용된 기계학습 기법 및 결과 비교

결정테이블(decision table, DT), 결정트리(decision tree, ID3), 나이브 베이즈 분류기(naïve Bayes classifier), TAN(tree-augmented naïve Bayes) 분류기, 인공신경망(artificial neural network)의 방법으로 태그를 100개, 300개, 500개로 다르게 하여 성능을 비교하였다. 추가로 각 실험에서 모델 생성시간을 비교하였고, 실험은 웨카(Weka) 패키지[9]를 이용하였다.

그림 2는 기계학습 기법들의 실험결과를 AUC(area under the ROC curve)를 가지고 비교한 것이다[6]. 모든 방법에 대해서 사용한 태그가 많을수록 성능이 향상되었다. 각 기법들의 성능을 비교해 보았을 때, 나이브 베이즈 분류기, TAN 분류기, 인공신경망이 85%~90%로 비교적 높은 결과를 보였고, 결정테이블과 결정트리가 70%~85%로 비교적 낮은 결과를 보였다. 표 4는 각 알고리즘의 모델 생성시간을 비교하고 있다.[10]

모델 생성시간을 비교해 볼 때, 나이브 베이즈 분류기의 모델 생성시간이 압도적으로 빠른 것을 알 수 있다. 이는 나이브 베이즈 분류기의 학습이 다른 방법과 달리 탐색(search)을 필요로 하지 않기 때문이다. 나이브 베이즈 분류기, TAN, 인공신경망이 비록 비슷한 성능을 보인다고 하여도, 모델 생성시간에 비추어 볼 때 나이브



	DT	ID3	NB	TAN	ANN
100	0.713	0.826	0.851	0.852	0.848
300	0.716	0.838	0.895	0.893	0.891
500	0.722	0.840	0.906	0.903	0.900

그림 2 기계학습 기법들의 실험결과(DT: 결정테이블, ID3: 결정트리, NB: 나이브 베이즈 분류기, TAN: TAN 분류기, ANN: 인공신경망)

9) http://www.cs.waikato.ac.nz/ml/weka/

10) 실험환경: Intel Core2 Quad CPU 2.4GHz, 2GB RAM, Windows XP

표 4 모델 생성시간 비교(단위: 초) (DT: 결정테이블, ID3: 결정트리, NB: 나이브 베이즈 분류기, TAN: TAN 분류기, ANN: 인공신경망)

태그개수	DT	ID3	NB	TAN	ANN
100	71.55	24.45	0.11	43.53	237.92
300	764.99	131.91	0.31	427.38	2416.14
500	2122.39	345.44	0.64	1254.89	7252.39

베이즈 분류기가 태그를 이용한 스패머 탐지 문제에는 가장 적절한 방법임을 알 수 있다.

4.3 결과 분석

실험 결과 나이브 베이즈 분류기가 가장 좋은 성능을 보였고, TAN 분류기와 인공신경망은 그에 준하는 성능을 보였다. 반면에 결정트리 계열(decision table 및 ID3)의 기계학습 방법의 성능이 상당히 좋지 않았다. 결정트리 계열의 기계학습 방법의 성능이 낮게 나온 원인이 실험에 사용한 데이터의 액티브유저와 스패머의 비율이 지나치게 편향되어 있기 때문이라는 가설을 세우고, 이를 검증하기 위해서 추가적인 실험을 진행하였다. 데이터의 편향을 제거하기 위해서 표본추출[under-sampling]을 통해 학습데이터와 테스트데이터의 스패머와 액티브유저의 비율을 1:1로 조정한 후 추가 실험을 진행하였다.

표 5 데이터 불균형 해소 후 실험 결과, 수치는 3번의 실험을 통한 AUC값의 평균값과 표준편차를 나타냄 (DT: 결정테이블, ID3: 결정트리, NB: 나이브 베이즈 분류기, TAN: TAN 분류기, ANN: 인공신경망)

태그개수	DT	ID3	NB	TAN	ANN
100	0.656 ±0.010	0.837 ±0.002	0.853 ±0.002	0.851 ±0.003	0.851 ±0.001
300	0.655 ±0.003	0.841 ±0.002	0.880 ±0.005	0.869 ±0.001	0.876 ±0.003
500	0.680 ±0.020	0.848 ±0.003	0.910 ±0.005	0.900 ±0.004	0.893 ±0.006

표 5의 실험 결과를 보면, 결정트리 계열인 ID3의 성능이 조금 상승하였으나 여전히 다른 기법들에 비해서 낮은 결과를 보여주었다. 이를 통해서 데이터의 불균형이 실험 결과에 의미있는 영향을 주지는 않았음을 확인하였다.

나이브 베이즈 분류기가 다른 방법에 비해 우수한 성능을 보이는 것에 대해 다음과 같은 이유를 찾을 수 있다. 첫째, 본 논문에서 기계학습 방법의 성능을 비교하는 데 사용한 평가 기준인 AUC는 나이브 베이즈 분류기가 결정트리 계열의 방법(ID3, C4.5 등)보다 높게 나

오는 경향이 있다[8]. 그 원인은 AUC의 특성과 나이브 베이즈 분류기와 결정트리의 사후 확률을 표현하는 능력의 차이와 관련이 있다. AUC 점수는 알고리즘이 얼마나 표본의 사후 확률의 값을 정확하게 추정하여 순위를 매기는가에 따라서 결정된다[6]. 한편, 결정트리 알고리즘은 일반적으로 작은 결정트리를 만드는 것을 목표로 한다[7]. 그 결과 적은 수의 말단노드(leaf node)가 만들어지게 되고 한 말단노드안에 더 많은 표본이 들어가게 된다. 같은 말단노드의 표본들은 같은 사후 확률을 가지게 되므로 이들은 무작위로 순위가 매겨진다. 이런 이유로 결정트리는 각 표본의 사후 확률을 다르게 표현하는 능력이 약하다[8]. 반면, 나이브 베이즈 분류기는 표본의 사후 확률 $P(c|e)$의 값을 표본의 속성들이 서로 조건부독립적이라는 가정 아래 $P(a_1, a_2, a_3, \cdots, a_n|c)$의 값을 $P(a_i|c)$의 곱에 기반하여 추정한다(단, c는 클래스, e는 표본, a_i는 속성 A_i의 값, n은 속성의 개수를 나타냄)[7]. 구체적으로, 나이브 베이즈 분류기는 $2n+1$개의 파라미터만을 가지고 있지만 최대 2^n 개의 서로 다른 사후 확률의 값을 나타낼 수 있다[8]. 그러므로 나이브 베이즈 분류기가 사후 확률을 표현하는 능력[granularity]에 있어서 결정트리보다 앞선다고 할 수 있다. 이러한 이유로 AUC로 비교하였을 때 나이브 베이즈 분류기의 성능이 결정트리 계열의 성능보다 일반적으로 우수하다[8].

둘째 이유로는 스패머 탐지 문제가 선형으로 분리 가능한 경우(linearly separable)와 유사할 가능성을 생각할 수 있다. 문제가 선형으로 분리 가능한 경우에 복잡한 학습 모델은 단순한 학습 모델에 비해 과대적합(overfitting) 문제에 빠질 가능성이 크기 때문에 일반적으로 성능이 떨어진다. 본 논문의 실험에서 사용한 학습 데이터처럼 속성이 모두 이진인 경우에는 나이브 베이즈 분류기는 선형분류기와 동일하기 때문에 그 구조가 TAN 분류기나 인공신경망의 구조보다 단순하다[9]. 그렇기 때문에 나이브 베이즈 분류기가 선형으로 분리 가능할 가능성이 높은 '태그를 이용한 스패머 탐지 문제'에서 TAN 분류기 및 인공신경망보다 좋은 성능을 발휘하였다고 추측할 수 있다.

5. 결 론

소셜 북마킹 시스템에서의 스패머 탐지는 시스템의 장점을 활용하고 단점을 보완하는 데 매우 중요한 요소이다. 스패머를 탐지하는 모델을 만들기 위해서 주어진 실험 데이터에서 상호정보량이 높은 태그들을 추출하였고 이를 속성으로 이용하여 기계학습의 여러 알고리즘을 적용하여 성능을 비교해 보았다. 여러 방법 중에서 나이브 베이즈 분류기 방법이 성능과 모델 생성시간에서 만족할 만한 결과를 주었으며 이는 문제의 특성 및 그 평가 기준에 기인하는 것으로 여겨진다.

폭소노미를 이용하여 리소스를 분류하는 웹2.0 사이트가 점점 더 증가하고 있기 때문에 태그를 이용한 스패머 탐지 모델은 비단 소셜 북마킹 시스템뿐만 아니라, 다양한 종류의 웹2.0 사이트에도 적용될 수 있을 것으로 예상된다.

참 고 문 헌

[1] Heymann, P., Koutrika, G., and Garcia-Molina, H., Can social bookmarking improve web search?, *Proceedings of the First ACM International Conference on Web Search and Data mining*, 2008.

[2] Mathes, A., Folksonomies - cooperative classification and communication through shared metadata, unpublished paper, http://www.adam-mathes.com/academic/computermediatedcommunication/folksonomies.html, 2004.

[3] Hotho, A., Jaschke, R., Schmitz, C., and Stumme, G., BibSonomy: a social bookmark and publication sharing system, *Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, pp. 87-102, 2006.

[4] Hotho, A., Jaschke, R., Schmitz, C., and Stumme, G., Information retrieval in folksonomies: search and ranking, *Proceedings of the Third European Semantic Web Conference*, pp. 411-426, 2006.

[5] Cover, T.M. and Thomas, J.A., *Elements of Information Theory*, Wiley-Interscience, 1991.

[6] Fawcett, T., An introduction to ROC analysis, *Pattern Recognition Letters*, Vol.27, pp. 861-874, 2006.

[7] Mitchel, T.M., *Machine Learning*, McGraw-Hil, 1997.

[8] Huang, J., Lu, J., and Ling, C.X., Comparing naïve Bayes, decision trees, and SVM with AUC and accuracy, *Proceedings of the Third IEEE International Conference on Data Mining*, pp. 553-556, 2003.

[9] Ling, C.X. and Zhang, H., The representational power of discrete Bayesian networks, *Journal of Machine Learning Research*, Vol.3, No.Dec., pp. 709-721, 2002.