논문 13-38B-04-07

한국통신학회논문지 '13-04 Vol.38B No.04
http://dx.doi.org/10.7840/kics.2013.38B.4.291

# 메타 태그를 이용한 자동 웹페이지 분류 시스템

김 상 일◆, 김 화 성°

# An Automatic Web Page Classification System Using Meta-Tag

Sang-il Kim◆, Hwa-sung Kim°

요 약

최근 월드 와이드 웹(World Wide Web)의 사용이 폭발적으로 증가함에 따라 다양한 정보를 포함하고 있는 웹 페이지들의 양도 엄청나게 증가 하였다. 따라서 웹상에 존재 하고 있는 웹페이지들에 대한 접근을 용이하게 하고, 그룹화를 통한 검색을 가능하게 하기 위해 웹 페이지 분류의 필요성이 대두 되고 있다. 웹 페이지 분류는 기존의 웹 상에 산재 되어 있는 웹페이지들을 비슷한 문서 유형 또는 같은 키워드를 사용하는 문서들의 묶음으로 구분하는 작업을 의미하며, 웹 페이지 분류 기술은 웹페이지 검색, 그룹 검색, 메일 필터링 등의 분야에 응용될 수 있는 기술이다. 하지만 웹상에 존재하는 웹페이지들을 사람이 수동적으로 분류하는 방법으로는 현재 월드 와이드 웹에 존재하는 엄청난 양의 웹페이지들을 처리할 수 없으며, 자동적인 분류 방법 역시 서로 다른 형태로 작성된 웹페이지들을 정확하게 분류할 수 없다는 문제로 인해 한계를 보이고 있다. 본 논문에서는 서로 다른 형태로 작성된 웹 문서들에 대한 부정확한 분류 문제를 해결하기위해 웹페이지에 존재하는 메타 정보를 획득하여 자동적으로 분류하는 메타 태그기반의 자동화된 웹페이지 분류 시스템을 제안하였다.

**Key Words** : meta-tag, automatic classification, web page classification, weka naive bayes

ABSTRACT

Recently, the amount of web pages, which include various information, has been drastically increased according to the explosive increase of WWW usage. Therefore, the need for web page classification arose in order to make it easier to access web pages and to make it possible to search the web pages through the grouping. Web page classification means the classification of various web pages that are scattered on the web according to the similarity of documents or the keywords contained in the documents. Web page classification method can be applied to various areas such as web page searching, group searching and e-mail filtering. However, it is impossible to handle the tremendous amount of web pages on the web by using the manual classification. Also, the automatic web page classification has the accuracy problem in that it fails to distinguish the different web pages written in different forms without classification errors. In this paper, we propose the automatic web page classification system using meta-tag that can be obtained from the web pages in order to solve the inaccurate web page retrieval problem.

## Ⅰ. 서 론

최근 웹 서비스를 이용하는 이용자가 증가하면서 월드 와이드 웹 상에서 볼 수 있는 웹 문서의 양이 엄청나게 증가하고 있다. 이와 같은 대량의 웹 문서들은 사람이 수동적으로 직접 수집하고 정리 할 수

※ 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구 사업 지원을 받아 수행된 것임(2011-0025-226)
◆ 주저자 : 광운대학교 전자통신공학과, rlatkd234@kw.ac.kr, 학생회원
° 교신저자 : 광운대학교 전자통신공학과, hwkim@kw.ac.kr, 종신회원
논문번호 : KICS2012-04-212, 접수일자 : 2012년 4월 23일, 최종논문접수일자 : 2013년 3월 15일

없기 때문에 웹문서들을 알맞게 정해진 카테고리로 분류 하는 것을 도와주는 자동 웹페이지 분류 도구에 대한 필요성이 점차 커지고 있다. 웹 문서를 분류하는 목적은 특정 주제별로 중요한 문서들을 구분하여 사용자가 보다 빨리 필요한 문서를 검색할 수 있도록 하는 것과 사용자의 선호도를 바탕으로 개인화 또는 그룹화를 하려는 것으로 나눌 수 있다. 특히 웹의 효율적인 탐색을 위해 사용자가 관심 있어 할 웹 문서들을 자동으로 분류 하는 것은 개인화 추천 시스템 연구 및 데이터 마이닝 연구에서 중요시 되고 있다. 이와 같은 중요성으로 인해 검색엔진 최적화를 위한 Open Directory Project(ODP)가 진행되고 있으며 대표적으로 사람이 편집하는 최대 규모의 디렉토리 사이트인 Dmoz.org가 있다.

Dmoz.org와 같은 디렉토리 검색 방법은 사용자가 찾고자 하는 정보에 대한 지식이 분명하지 않거나 특정 분야에 대한 검색을 하고자 할 경우, 데이터가 인간의 판단에 의해 비교적 정확히 분류되어 있기 때문에 사용자가 원하는 결과를 얻기가 용이하며, 각각의 정보에 대한 평가가 가능해 신뢰성이 높다[1]. 하지만 데이터의 수집, 확인, 분류 등의 작업에 많은 시간과 인력을 요하므로 정보의 양이 상대적으로 적으며 분류 주제의 이해에 대한 개인차로 인해 검색이 어려워 질 수 있다. 또한 Dmoz와 같은 웹 문서들의 수동적인 분류는 현재 웹상에 존재 하는 웹 문서의 양을 수용 할 수 없기 때문에 사용자 참여를 유도 하고 있지만 무분별한 사용자 참여로 인해 분류의 정확도가 높지 않으며, 자동적인 분류 역시 서로 다른 형태로 작성된 웹 문서 획득 문제가 존재한다.

따라서 본 논문에서는 서로 다른 형태의 웹 문서의 본문이 아닌, 해당 웹 페이지에 포함된 메타 태그를 이용한 자동화된 웹 문서 분류 기법을 제안함으로써 웹페이지 형식에 구애받지 않는 자동화된 문서 분류를 가능하도록 하였다. 본 논문의 구성은 다음과 같다. 2장에서는 웹페이지 분류를 위해 사용가능한 메타 태그와 웹페이지 분류 알고리즘에 대해 살펴보고, 3장에서는 메타 태그를 획득하기 위한 자동화된 크롤러와 WEKA를 이용한 웹 페이지 자동 분류 시스템 성능평가를 진행한다. 마지막으로 4장에서는 결론 및 향후 연구 방향을 제시한다.

## Ⅱ. 관련 연구

### 2.1. 메타 태그

메타 태그는 HTML 문서의 맨 상단에 위치하는 태그로써, 브라우저와 검색엔진을 사용할 수 있도록 문서의 정보를 포함 하고 있는 태그를 말한다. 메타 태그의 구조는 다음과 같다.

```
<META NAME="Generator" CONTENT="페이지 제작툴">
<META NAME="Author" CONTENT="제작자">
<META NAME="Keywords" CONTENT="컴색 키워드">
<META NAME="Description" CONTENT="페이지의 요약설명">
<META NAME="Copyright" CONTENT="저작권 정보">
<META NAME="Subject" CONTENT="홈페이지주제 입력">
<META NAME="Title" CONTENT="홈피이름">
<META NAME="Publisher" CONTENT="만든단체 또는 회사">
<META NAME="Other Agent" CONTENT="홈피 책임자">
<META NAME="Classification" CONTENT="카테고리 분류">
<META NAME="Reply-To(Email)" CONTENT="메일주소">
<META NAME="FileNAME" CONTENT="파일이름 입력">
<META NAME="Author-Date(Date)" CONTENT="제작일">
<META NAME="Location" CONTENT="위치/대한민국인지등 나라이름기록">
<META NAME="Distribution" CONTENT="배포자">
<META NAME="Copyright" CONTENT="저작권">
```

그림 1. 메타 태그의 구조
Fig. 1. Structure of meta-tags

메타 태그에는 웹페이지를 설명하기 위한 정보들이 포함되어 있으며 웹페이지의 본문 내용 대신 메타 태그의 검색 키워드 및 페이지의 요약 설명을 이용하면 해당 웹페이지가 어떠한 내용의 웹페이지인지 알 수 있으며, 검색 엔진은 해당 정보를 수집하여 검색결과로 제공 한다. 메타 태그는 HTML4에 포함된 표준화된 웹 페이지 표기 방식이지만, 웹페이지의 제작자가 필요한 정보만을 입력하거나, 검색엔진이 정보를 수집하는 것을 피하기 위해 아예 정보를 입력하지 않는 문제점이 있다. 하지만 개인이 제작한 웹 페이지 이외에 사용자가 많이 접근하는 공공기관, 대형 포털 사이트, 카페, 블로그, 뉴스 사이트, 즉, 사용자가 검색하여 접근하는 웹 페이지들은 모두 검색 엔진이 메타 태그를 획득하여 제공되기 때문에 메타 태그가 존재하며, 웹페이지 분류에 충분히 활용될 수 있다.

### 2.2. 웹페이지 분류 알고리즘

웹페이지 분류 알고리즘들은 문서에 표현되는 단어로부터 문서 벡터를 생성하고 벡터화된 문서들을 Training Set 즉 기준이 되는 DB로 사용하여 학습함으로써 유사한 문서에 범주를 할당하게 된다. 문서 분류 알고리즘에는 대표적으로 최근접 이웃 분류기(k-Nearest Neighbors), 문서-범주벡터 관련도(Linear Text Classifiers), 결정트리(Decision Tree), 지지벡터기계(Support Vector Machines), 베이지안 확률(Bayesian classifier)등이 있으며 본 논문에서는 대표적인 웹페이지 분류 알고리즘인 Naive Bayes와 Naive Bayes의 단점을 보완한 Complement Naive

Bayes, 확률 기반이 아닌 의사 결정 기반의 알고리즘인 J48 알고리즘을 이용하여 메타 태그를 이용한 자동화 분류 시스템 성능평가를 진행 하였다[2-3].

#### 2.2.1. Naive Bayes 알고리즘

Naive Bayes 알고리즘은 주어진 학습 데이터를 이용하여 분류 대상이 포함될 확률이 가장 높은 클래스로 분류 하는 것으로 개별 속성 값들이 서로 독립적이라는 가정 하에 아래 식을 통해 계산 한다.

$$P(C_j \mid d) = \frac{P(C_j)P(d \mid C_j)}{P(d)} \qquad (1)$$

식(1)을 살펴보게 되면 $d$는 임의의 요소를 의미하고, $C_j$는 임의의 클래스를 의미한다. $P(d)$는 모든 클래스에 대하여 같은 값을 가지므로 확률을 계산하는데 있어 고려하지 않아도 된다. 따라서 $P(C_j)$와 $P(d|Cj)$만 이용하면 임의의 요소 $d$가 $C_j$에 할당될 확률을 계산 할 수 있다. $P(C_j)$는 모든 학습 문서들의 수와 $C_j$에 속하는 문서들의 수의 비율로 구할 수 있다.

#### 2.2.2. J48 알고리즘

J48알고리즘은 C4.5 알고리즘이라고도 불리며 트리기반의 대표적인 분류 알고리즘인 ID3의 단점을 보완한 알고리즘이다. 의사 결정 트리 알고리즘이기 때문에 분석하는 사람이 결과를 쉽게 이해하고 설명할 수 있지만 앞의 확률적인 분류 알고리즘과는 다르게 데이터를 분류 하는 기능만 있어 예측이 불가능 하다. C4.5 알고리즘이 보완하고자 한 ID3 알고리즘의 문제점들은 수치형 속성 취급, 무의미한 속성을 제외하는 문제, 나무의 깊이 문제, 결측치 처리, 비용고려, 효율성 문제가 있으며 C4.5에서는 바이너리 분할 방식을 통해 위와 같은 문제점을 다소 해결 하였다[4-5].

#### 2.2.3. WEKA 마이닝 툴

WEKA 마이닝 툴은 뉴질랜드 와이카토대학 컴퓨터 학과에서 자바 언어로 개발된 기계 학습 및 데이터마이닝을 위한 소프트웨어이다. 이 도구는 ARFF라는 단순 플랫 파일과 CSV파일도 지원을 하며 명령어 라인 또는 GUI 환경으로 모두 실행이 가능하도록 되어 있다. 또한 학습 알고리즘을 평가하거나 학습 알고리즘 비교, 응용 할 수 있으며 [그림 2]과 같이 모든 처리 결과를 시각적으로 보여주는 가시화 기능이 제공된다[6-9].

## Ⅲ. 메타 태그를 이용한 자동화된 웹페이지 분류 시스템

본 장에서는 본 논문에서 제안하는 메타 태그를 이용한 자동화된 웹페이지 분류 시스템에 대해 설명한다. 자동화된 웹페이지 분류 시스템은 크게 메타 태그를 자동적으로 수집하기 위한 메타 태그 크롤러와 컴퓨터 학습에 필요한 Training Set, Weka 기반의 classifier로 구성되어 있으며 시스템의 구성도는 다음과 같다.

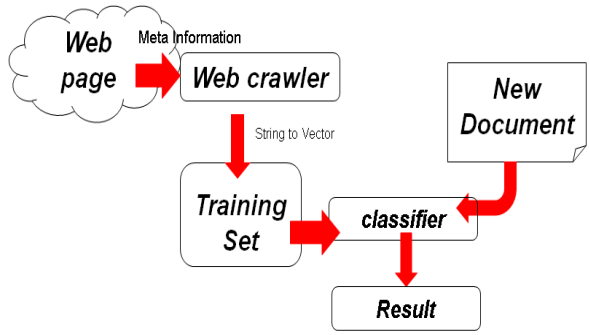


그림 2. 메타 태그를 이용한 웹페이지 분류 시스템 구성도
Fig. 2. Configuration of Web page classification system using meta-tag

### 3.1. 메타 태그 획득을 위한 메타 정보 획득 크롤링

2장에서 살펴본 메타 태그 구조를 통해 웹페이지의 내용을 알기 위해서는 메타 태그 항목 중 검색 키워드와 페이지 요약 설명 항목의 정보 획득이 필요 하다는 것을 알 수 있다. 아래의 그림은 오늘의 날씨를 알려주는 뉴스 기사의 메타 태그를 나타낸 그림이다. '<meta name="description"' 항목과 '<meta name="title"' 항목을 통해 전체의 뉴스 본문을 읽지 않아도 해당 기사가 날씨에 관련된 기사라는 것을 알 수 있다.


그림 3. 날씨 뉴스의 메타 태그 정보
Fig. 3. Meta tag information of weather news

따라서 본 논문에서 제안하는 크롤러는 <meta name="description"' , '<meta name="title"' 항목을 획득하는 형태의 크롤러로 구현 하였다. 실제로 실험을 위해서 구글 뉴스에서 제공하는 뉴스 100개를 분석한 결과 <meta name="description"' , '<meta name="title"' 항목은 100개의 구글 뉴스 전체에 존재 하였으며, 메타 정보를 크롤러를 통해 획득 하였다.


그림 4. 구글 뉴스에서 크롤링한 메타 태그 정보
Fig. 4. Meta tag information crawled to google news

### 3.2. Weka를 이용한 자동화된 웹페이지 분류

1절에서 구현한 크롤러를 통해 획득한 메타 정보는 Weka라는 마이닝 툴을 이용해 분류 하게 된다. Weka 마이닝 툴에는 Naive Bayes, J48 등 다양한 마이닝 알고리즘이 있으며 본 논문에서는 Complement, Naive Bayes, Naive Bayes, J48 알고리즘을 사용하여 웹페이지 문서 분류를 진행 하였다. 웹페이지 문서 분류는 다음과 같은 순서로 진행된다. 먼저 Training set을 구축하기위해 획득한 문서들을 구글 뉴스의 카테고리 기준으로 분류 한다. 이후 새롭게 문서 분류를 하고자 하는 크롤러로 크롤링 하지 않았던 새로운 뉴스 문서들을 Test set으로 활용하여 Weka를 통해 분류 한다. 새로운 뉴스 문서들의 카테고리 정보는 해당 카테고리에 정확하게 분류 되었는지 판단하기 위해 획득은 하였지만 마이닝 과정에는 포함시키지 않는다.

위 그림의 predicted 항목은 마이닝 알고리즘에 의해 예측된 Test set의 카테고리이며 획득한 카테고리 정보와 비교하여 정확하게 분류 되었는지 비교 하는 항목이다.

## Ⅳ. 메타 태그를 이용한 자동화된 웹페이지 분류 시스템

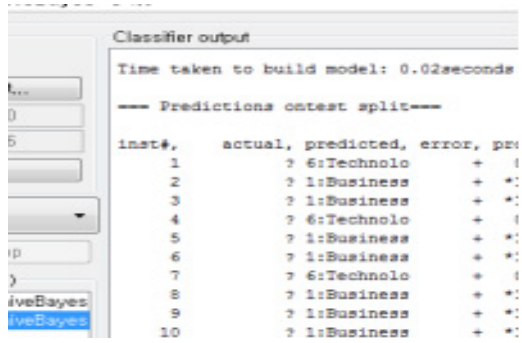


그림 5. WEKA의 데이터 마이닝 실행 화면
Fig. 5. Data mining execution screen of WEKA

본 논문에서는 메타 태그를 이용한 웹페이지 분류 기법의 효율성을 입증하기 위해 4가지 방법으로 실험을 진행하였다. 먼저 각 분류 알고리즘의 성능 분석을 위해 구글 뉴스 본문과 Dmoz에 분류된 웹페이지 본문을 Training set과 Test set을 이용하여 웹페이지 분류를 진행한다. 이후 획득한 메타 정보와 구글 뉴스 전문을 Training set과 Test set으로 번갈아 가며 분류 알고리즘을 적용하여 매칭 정확도를 비교한다.

### 4.1. 구글 뉴스본문과 Dmoz 웹페이지 본문

본 항에서는 알고리즘 별 정확도를 비교하기 위해 구글 뉴스 본문을 Training set으로 하여 Dmoz에 분류된 웹페이지를 자동적으로 분류 한다. 아래의 [그림 6]은 카테고리별 정확도를 나타내는 그림이다.

카테고리는 뉴스에서 가장 많이 사용 될 수 있는 카테고리를 기반으로 선정하였으며 구글 뉴스 본문 100개를 Training set으로 사용하였고 Dmoz 웹페이지 본문 100개를 Test set으로 사용하여 그 결과를 비교 하였다. 실험 결과 확률 기반의 Naive Bayes

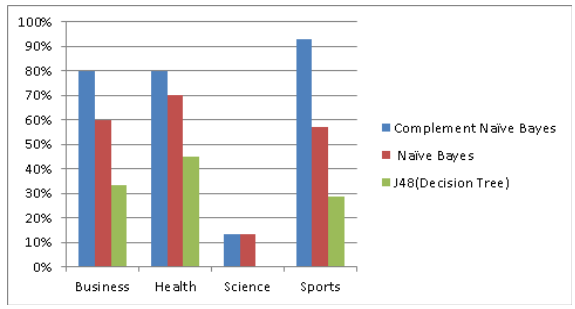


그림 6. 구글 뉴스 본문과 Dmoz 웹페이지 본문의 분류 알고리즘 정확도
Fig. 6. accuracy of the classification algorithm (Google news body text and Dmoz webpage body text)

알고리즘이 높은 성능을 보이고 그중에서도 Naive Bayes의 가중치 문제점을 해결한 Complement Naive Bayes 알고리즘이 가장 좋은 성능을 보이는 것을 알 수 있다. 알고리즘 별 평균 정확도는 Complement Naive Bayes 알고리즘이 67%, Naive Bayes 알고리즘이 50%, J48 알고리즘이 27%의 정확도를 보인다.

### 4.2. 구글 뉴스 본문과 구글 뉴스 본문

본 항에서는 기존의 분류 방식인 본문을 이용한 웹페이지 분류 기법의 정확도를 분석하기 위해 구글 뉴스 본문을 Training set과 Test set으로 사용하여 각 알고리즘 별 정확도를 비교 하였다. 아래 [그림 7]은 구글 뉴스 본문과 구글 뉴스 본문을 사용한 알고리즘별 정확도를 나타내는 그림이다.

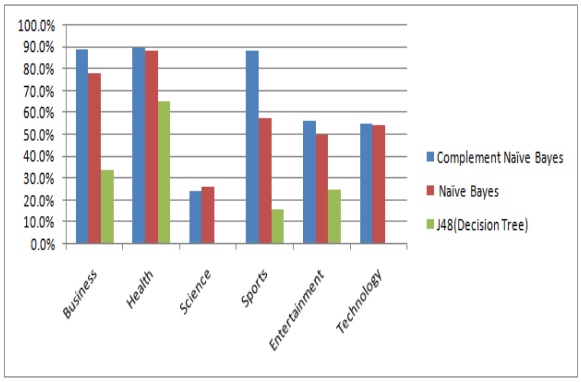


그림 7. 구글 뉴스 본문과 구글 뉴스 본문의 분류 알고리즘 정확도
Fig. 7. Accuracy of the classification algorithm (Google News body text and Google news body text)

실험 결과 Complement Naive Bayes의 경우는 Science 카테고리를 제외 하고 50% 이상의 문서 분류 정확도를 보이지만 Naive Bayes의 경우 Complement Naive Bayes와 비슷한 결과를 보였지만 정확도는 떨어졌다. 위의 알고리즘 성능평가와 같이 Complement Naive Bayes 알고리즘이 가장 높은 정확도를 보였으며 알고리즘 별 평균 정확도는 Complement Naive Bayes 67.1%, Naive Bayes 58.94%, J48 23.26%의 정확도를 보인다. 본 실험의 성능 평가 결과는 기존의 웹페이지 분류 방법인 본문을 이용한 웹페이지 분류 방법의 성능 평가 결과로써 메타 태그를 이용한 웹페이지 분류 성능 평과 결과와 비교 분석하여 메타 태그의 사용 가능 여부를 평가한다.

### 4.3 구글 뉴스 본문과 구글 뉴스 본문

본 항에서는 메타 태그의 Test set 활용 가능성을 확인 하기 위해 구글 뉴스 본문을 Training set으로 하고 구글 뉴스 메타 태그를 Test set으로 사용하여 각 알고리즘 별 분류 정확도를 비교 하였다. 아래의 [그림 8]은 구글 뉴스 본문과 구글 뉴스 메타 태그를 사용한 알고리즘별 정확도를 나타내는 그림이다.

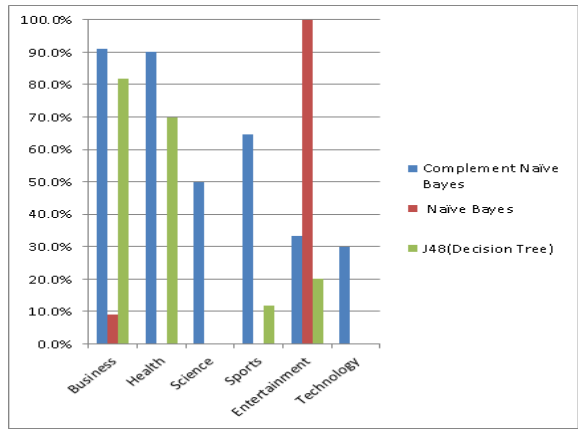


그림 8. 구글 뉴스 본문과 구글 뉴스 메타 태그의 분류 알고리즘 정확도
Fig. 8. Accuracy of the classification algorithm (Google News body text and Google news meta tags)

실험 결과 Complement Naive Bayes의 경우는 Technology, Entertainment 카테고리를 제외 하고 50% 이상의 문서 분류 정확도를 보이지만 Naive Bayes의 경우 다른 카테고리 항목을 다 분류 하지 못하고 Entertainment에 치중된 분류 결과를 보여주었다. 이와 같은 결과는 뉴스 본문을 Training set, 메타 태그를 Test set으로 사용하는 경우 표본이 되는 Training set에서 추출 되는 특정 단어의 빈도가 확률 기반의 문서 분류 알고리즘인 Naive Bayes에 영향을 주기 때문이다. Complement Naive Bayes는 위와 같은 Naive Bayes 알고리즘의 가중치 문제점을 해결하였기 때문에 본 실험에서는 비교적 안정적인 분서 분류 결과를 보여 준다. J48은 Science 항목과 Technology 항목을 제외한 나머지 항목에 있어서 가. 항에서 진행한 실험보다 더 높은 정확도를 보여주었다. 의사 결정 트리의 특성상 Science 내용과 Technology 항목은 유사 하거나 공통 적으로 사용하는 언어가 많기 때문에 분류 하지 못하는 문제점이 발생한다. 알고리즘 별 평균 정확도는 Complement Naive Bayes 59.8%, Naive Bayes 18.18%, J48 30.6%의 정확도를 보인다.

### 4.4. 구글 뉴스 메타태그와 구글 뉴스 메타태그

본 항에서는 메타 태그의 Training set과 Test set 활용 가능성을 보기위해 구글 뉴스 메타 태그를 Training set과 Test set으로 사용하여 각 알고리즘별 분류 정확도를 비교 하였다. 아래의 [그림 9]는 구글 뉴스의 메타 태그를 사용한 알고리즘별 정확도를 나타내는 그림이다.

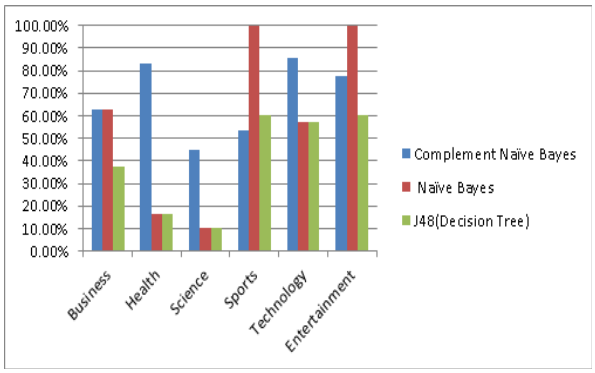


그림 9. 구글 뉴스 메타 태그의 분류 알고리즘 정확도
Fig. 9. Accuracy of the classification algorithm(Google news meta tags and Google News meta tags text)

실험 결과 Complement Naive Bayes의 경우 Science 카테고리를 제외하고 50%이상의 문서 분류 정확도를 보인다. Naive Bayes의 경우 가중치 문제로 인해 앞서 실험한 결과와 같이 편향된 결과를 보인다. 의사결정 트리는 Science, Health 항목을 제외한 나머지 항목에서 50%이상의 정확도를 보인다. 전체적인 알고리즘 별 정확도는 Complement Naive Bayes 67.87%, Naive Bayes 57.72%, J48 40.22%의 정확도를 보인다. 최종적으로 메타 태그만을 이용한 자동 분류 결과 Complement Naive Bayes를 기반으로 한 자동 분류가 가장 정확한 정확도를 보였으며 의사결정 트리인 J48알고리즘이 가장 낮은 정확도를 보였다. 자동 분류 실험 결과 Science 카테고리의 경우 다양한 분야가 존재하기 때문에 생명 공학 관련 뉴스기사의 경우 Health 또는 Sports로 분류 하는 문제점이 발생 하였다. 2절의 구글 뉴스 본문과, 구글 뉴스 본문 기반의 웹페이지 분류 정확도와 비교 했을 때 Complement Naive Bayes 알고리즘을 기준으로 웹페이지 본문을 전부다 사용하지 않고 메타 태그만 사용했을 때 더 비슷하거나 더 높은 결과를 보인다. 표1은 4가지 경우의 종합적인 결과를 표로 나타낸 것이다.

표 1. 실험 결과
Table 1. The experimental results

| Case | Complement Naive Bayes | Naive Bayes | J48 |
|---|---|---|---|
| Case1 | 67% | 50% | 27% |
| Case2 | 67.1% | 58.94% | 23.26% |
| Case3 | 59.8% | 18.18% | 30.6% |
| Case4 | 67.87% | 57.72% | 41.22% |

표1에서 Case1은 구글 뉴스 본문과 Dmoz 웹페이지 본문의 실험 결과이며, Case2는 구글 뉴스 본문과 구글 뉴스 본문의 실험결과 Case3는 구글 뉴스 메타 태그와 구글 뉴스 본문, Case4는 구글 뉴스 메타태그와 구글 뉴스 메타 태그를 실험한 결과를 나타낸다.

실험 결과 4가지 경우 모두 Complement Naive Bayes를 사용한 경우 가장 정확도가 높았으며, 4가지 경우 중 Case4가 가장 높은 정확도를 보였다.

## V. 결 론

본 논문에서는 메타 태그를 이용한 자동화된 웹페이지 분류 시스템을 제안하고, 제안한 시스템에서 웹페이지 분류에 대표적으로 사용되는 알고리즘들을 적용한 경우에 대해 성능 비교를 수행하였다. 실험 결과 Complement Naive Bayes 알고리즘을 이용한 자동화된 웹페이지 분류 방법이 가장 높은 정확도를 보였으며, 메타 태그만을 사용한 자동화된 웹페이지 분류기법이 본문을 사용한 자동화된 웹페이지 분류 기법보다 성능이 비슷하거나 높다는 것을 실험을 통해 증명하였다. 따라서 메타 태그를 이용하게 되면 서로 다른 형태로 작성된 웹페이지들을 일일이 사람의 손으로 정리할 필요가 없이 자동화된 크롤러를 이용하여 메타 태그정보만 획득한 후, 문서 분류 알고리즘을 이용하여 자동으로 웹페이지를 분류하는 것이 가능하다. 향후 연구로는 본 논문의 실험에 사용된 문서 분류 알고리즘보다 향상된 성능을 보이는 문서 분류 알고리즘을 도출하여 메타 태그 기반의 문서 분류 시스템의 정확도를 더욱 향상시킬 수 있는 연구가 필요하다.

## References

[1] dmoz web pages [open directory project],

from http://www.dmoz.org.

[2] J. D. M. Rennie and D. R. Karger, “Tackling the poor assumptions of naive bayes textclassifiers”, in *Proc. 20th Int. Conf. Mach. Learning*, pp. 616-623, Washington DC, U.S.A., Aug. 2003

[3] J.-U. Kim, H.-J. Kim, and S.-G. Lee, “A study on incremental learning model for naive bayes text classifier,” in *Proc. Int. Conf. Korea Database Soc.*, pp. 331-341, Seoul, Korea, June 2001.

[4] X. Qi and B. D. Davison. “Web page classification: features and algorithms,” J. ACM Computing Surveys, vol. 41, no. 2, Article No. 12, Feb. 2009.

[5] D. Shen, J.-T. Sun, Q. Yang, and Z. Chen, “A comparison of implicit and explicit links for web page classification”, in *Proc. 15th Int. Conf. World Wide Web (WWW 2006)*, pp. 643-650, Edinburgh, U.K., May 2006.

[6] I. Charalampopoulos, “A comparable study employing WEKA clustering/classification algorithms for web page classification”, in *Proc. 15th Panhellenic Conf. Inform. (PCI)*, pp. 235-239, Kastoria, Greece, Oct. 2011.

[7] weka web page, Weka 3: Data Mining Sofrware in Java [Online], from http://www.cs.waikato.ac.nz/ml/weka/.

[8] I. H. Witten and F. Eibe, *Data Mining: Practical Machine Learning Tools and Techniques,* 2nd Ed., Morgan Kaufmann, 2000

[9] Toby Segaran, *Programming collective intelligence*, O’Reilly Media, 2007

[10] X. Qi and B. D. Davison, “Classifiers without borders: incorporating fielded text from neighboring web pages,” in *Proc. 31st Annu. Int. ACM SIGIR Conf.,* pp. 643-650, Singapore, July 2008.

[11] G. Xu, C. Xiang, X. Zhao, and G. Yang, “Tibetan web page classification based on column navigator”, in *Proc. 2012 2nd Int. Conf. Intell. Syst. Design Eng. Applicat. (ISDEA)*, pp. 610-612, Hainan, China, Jan. 2012.

**김 상 일** (Sang-il Kim)

2010년 2월 서일대학교 정보통신공학과 졸업

2012년 9월 광운대학교 전자통신공학과 석사

2012년 9월~현재 광운대학교 전자통신공학과 박사과정

<관심분야> 시맨틱 웹서비스, 상황 인지 서비스

**김 화 성** (Hwa-sung Kim)

1981년 2월 고려대학교 전자공학과 졸업

1983년 2월 고려대학교 전자공학과(석사)

1996년 Lehigh Univ. 전산학(박사)

1984년 3월~2000년 2월 ETRI 책임 연구원

2000년 3월~현재 광운대학교 전자통신공학과 교수

<관심분야> 클라우드 컴퓨팅, 시맨틱 웹서비스, 이동 네트워크 프로토콜, 임베디드 소프트웨어