인터넷 뉴스 기사에 대한 자동 분류 정보 시스템에 관한 연구

백용규[a], 서용무[b]

[a]고려대학교 대학원 경영학과
136-701, 서울시 성북구 안암동5가 1
Tel: +82-2-3290-1369, Fax: +82-2-922-7220, E-mail:ykbeak@korea.ac.kr

[b]고려대학교 경영대학
136-701, 서울시 성북구 안암동5가 1
Tel: +82-2-3290-1945, Fax: +82-2-922-7220, E-mail:ymsuh@korea.ac.kr

요약

방송 및 신문 산업 분야에서 인터넷 기술의 활용은 확산되고 있다. 특히, 인터넷 뉴스 기사 서비스는 뉴스 기사를 인터넷, 핸드폰 또는 무선 단말기를 통하여 실시간으로 제공하여야 하기 때문에 뉴스 기사를 빠르고 정확하게 분류하는 정보 시스템이 필요하다.

본 논문에서는 한글 뉴스 기사를 인터넷으로 빠르게 분류, 제공하는 정보시스템의 설계와 구현에 대하여 연구하였다. 입력 데이터는 전자 신문사의 인터넷 기사를 사용하였고, 문서의 분류는 k-NN, 의사결정 트리, 베이지언 네트워크, 신경망 등 이미 존재하는 다양한 분류자 중에서 실험을 통하여 가장 정확하고 빠른 분류를 하는 분류자를 선택하여 사용하였다.

이 시스템은 실시간으로 인터넷 뉴스 서비스를 제공하여야 하는 방송 및 신문 산업에서 활용 할 수 있다.

키워드:

분류 시스템, 정보 관리, 정보 시스템, 데이터/텍스트 마이닝

I. 서론

방송 산업에서 정보 기술과 통신 기술의 사용은 뉴스 기사 서비스를 위한 다양한 기회들을 제공한다. 정보 기술은 뉴스 데이터의 저장, 조회 및 분류를 가능하게 하며 통신 기술은 원격의 뉴스 데이터를 통합, 저장 및 의사 소통을 위한 채널을 제공한다. 뉴스 기사의 빠른 수집과 보도를 가능하게 하는 이러한 기술들을 통하여 뉴스 제공자는 신속하게 뉴스 기사를 고객에게 제공 할 수 있다.

현재 국내외에서 인터넷 기술을 활용한 뉴스 기사 서비스를 제공하고 있다. 인터넷 뉴스 기사 서비스는 기존 뉴스 보도와는 다른 가치를 고객에게 제공한다. 뉴스를 제공 받는 고객은 인터넷을 이용하여 실시간으로 뉴스 기사를 조회 할 수 있으므로, 뉴스 방송 시간 외에 긴급한 뉴스 속보를 접할 수 있다.

대부분의 뉴스 정보 시스템은 뉴스 데이터의 수집, 기사 선택 및 편집에 대한 기능을 하는데, 이러한 뉴스 정보 시스템은 다음과 같은 문제점이 있다. 모든 뉴스 기사는 분야 전문가가 기사 내용을 확인한 후 사전 정의된 뉴스 범주 중 하나에 분류한다. 따라서, 이러한 수작업 처리는 신속한 보도에 장애요소가 될 뿐만 아니라, 분류를 위한 전문인력 자원의 투입으로 추가 인건비가 소요된다.

또한, 뉴스 데이터는 비 구조적이기 때문에 기존의 구조적 데이터와는 다르게, 기사에서 중요 단어를 추출하는 작업과 분류를 위한 기타 다른 데이터를 생성하는 사전 작업을 요하는 등 많은 차이점을 보인다.

본 논문은 인터넷 뉴스 기사의 자동 분류를 위한 정보 시스템 설계 및 구현에 대한 방법을 제시한다. 내부 여러 조직에 분산된 데이터를 통합하여 사례 베이스를 구축하는 방법과 사례 베이스 구축 후 새로운 뉴스 기사에 대한 자동 분류를 수행하는 정보 시스템을 구축하는 방법을 제시한다. 이러한 목적을 수행하는 시스템을 구축하기 위하여 정보 검색과 텍스트 마이닝, 인공 지능의 기계 학습을 이용하였다.

본 논문의 구성은 다음과 같다. 제2장은 인터넷 뉴스 정보 시스템에 대한 개요 및 특징에 대하여 살펴본다. 제3장은 인터넷 뉴스 기사 분류 시스템에 대한 접근 방법 및 처리 단계를 살펴본다. 제4장은 자동 뉴스 기사 분류 시스템에서 구현한 데이터 집합, 전처리기 및 기계 학습에 대하여 살펴본다. 제5장은 본 논문에서 작성된 시스템에 대한 개발 도구, 전처리기, 뉴스 기사 분류기 및 실험과 평가에 대하여 살펴본다.

II. 인터넷 뉴스 정보 시스템

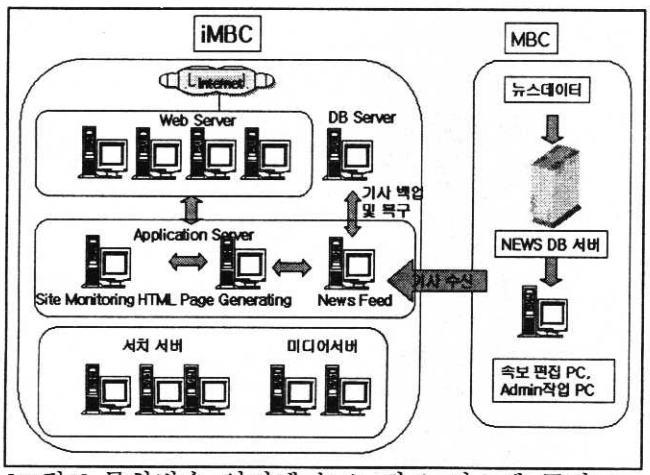
인터넷 기술을 활용한 뉴스 정보 시스템은 국내외

많은 기존 방송 산업에서 만들고 활용되고 있다. 인터넷 뉴스 정보 시스템은 기존의 뉴스 보도 정보 시스템을 확장하여 인터넷 서비스를 위한 페이지 자동 생성기 및 동영상 데이터 처리를 위한 기능을 포함한다. 본 절은 인터넷 뉴스 정보 시스템의 구성 요소 및 특징을 살펴본다.

2.1 인터넷 뉴스 정보 시스템 구성

인터넷 뉴스 정보 시스템의 가장 큰 특징은 뉴스 정보의 요청이 불특정 다수에 의하여 수시로 발생된다는 것이다. 기존 뉴스 정보 시스템은 내부 방송 관계자만 사용하였다. 즉, 기자가 뉴스를 작성하여 본사의 뉴스 데스크에게 전송하면 뉴스를 선별하여 뉴스 보도를 하였다. 하지만 인터넷 뉴스 정보 시스템은 인터넷을 통하여 많은 사용자에 의하여 접근이 가능하다. 또한, 새로운 뉴스 기사에 대한 실시간 서비스를 위한 자동 분류 기능을 요구한다. [그림1]은 문화 방송사의 인터넷 뉴스 정보 시스템의 시스템 구성도를 보여 준다. 문화 방송사는 기존 뉴스 정보 시스템을 확장하여 현재 인터넷을 이용한 뉴스 기사 서비스를 제공하고 있다.

그러나, 현재 인터넷 뉴스 정보 시스템은 뉴스 기사에 대한 수집 후 내용에 맞는 분류 범주에 전문가가 할당한다. 이러한 이유로 기사의 작성 후 실시간 서비스를 제공하지 못 한다. 인터넷 뉴스 서비스를 제공하고 있지만 기존의 인터넷 정보 시스템의 기능을 그대로 사용하고 있다.



[그림1] 문화방송 인터넷 뉴스 정보 시스템 구성도

2.2 인터넷 뉴스 정보 시스템 구성 요소

문화 방송사 구성도를 통하여 인터넷 뉴스 정보 시스템 포함하는 다양한 구성 요소를 살펴보겠다. 문화 방송사의 구성도는 기존 뉴스 정보 시스템(MBC)에서 인터넷 뉴스 정보시스템(iMBC)로 확장된 시스템 구성도이다.

인터넷 뉴스 서비스를 제공하는 정보 시스템의 구성 요소는 다양한 정보 기술의 통합을 통하여 이루어진다. 웹 서버(web server)은 인터넷 서비스를 제공한다. 즉, 사용자의 요청에 따라 뉴스 기사를 생성하여 서비스를 한다.

응용 서버(application server)은 세 가지의 중요한 서버들로 구성된다. 사이트 모니터 서버는 인터넷 뉴스 서비스의 진행을 감시하고 보고하는 기능을 제공한다. 각 웹 서버의 시스템 부하와 응답 시간 등 성능 관련 요소들을 웹 기반으로 한곳에서 확인한다. 웹 문서 생성 서버는 데이터베이스에 저장된 뉴스 기사를 자동으로 웹 문서로 변환하는 기능을 제공한다. 뉴스 피드 서버로부터 데이터를 받아 XML 데이터를 생성하고 HTML 파일을 생성 후 각 서버에 전송한다. 마지막으로 뉴스 피드 서버는 기존 뉴스 정보 시스템에서 뉴스 자원인 기사를 제공 받는 기능을 제공한다. 기존 뉴스 송출 프로그램에서 데이터를 입력 받아 데이터베이스와 페이지 생성 서버에 데이터를 보낸다.

인터넷 뉴스 정보 시스템은 기존 뉴스 정보 시스템과 통합되어 사용된다. 문화 방송의 뉴스 정보 시스템도 기존의 뉴스 데이터베이스 시스템을 활용하며, 기사 편집 및 속보 작업 기능을 유지하면서 사용된다.

이외 동영상 서비스를 위하여 별도의 동영상 미디어 서버를 설치 운영할 수 있다. 동영상 미디어 서버는 방송 뉴스에 대한 동영상 서비스를 제공한다. 하지만, 본 논문에서는 데이터의 범위는 텍스트 형식의 뉴스 기사에 한정 한다. 즉, 현재 서비스되는 뉴스 동영상 서비스의 데이터는 제외하였다.

III. 인터넷 뉴스 기사 분류 시스템

인터넷 뉴스 기사 자동 분류를 위한 정보시스템을 만들기 위해서는 어떤 접근 방법과 절차가 있는지에 대하여 살펴보도록 하겠다.

3.1 인터넷 뉴스 기사 분류 시스템 개요

인터넷 뉴스 기사 자동 분류 시스템은 입력된 새로운 기사를 사전에 정의된 기사 범주(class)에 자동으로 저장한다.[16] 이러한 자동 분류에 대한 연구는 데이터 마이닝과 정보 이론 및 인공 지능 분야에서 활발히 연구되어 왔다. 하지만 기존의 연구들이 구조적인 데이터 형식을 근간으로 사용한다.[6,7,8] 즉, 일정 형식을 가진 데이터 자원을 이용하고 있다. 예를 들어서 고객번호, 고객 성별, 구매상품명 등의 구분된 형식에 데이터를 가지고 있다. 하지만, 본 연구는 뉴스 기사의 기사 내용인 텍스트 데이터를 기반으로 처리한다.

또한, 대부분의 연구가 영어권에서 진행되어 한글 뉴스 기사에 대한 연구 및 기존 정보 시스템에 대한 통합 방법에 대한 연구는 적다.[12,13,16,17]

그리고, 뉴스 기사는 자연어로 작성이 되어

비구조적으로 구성된다. 이러한 비구조적 데이터를 처리하기 위한 연구는 텍스트 마이닝(text minig) 분야에서 활발히 연구되고 있다.

Feldman은 수많은 정보들이 문서 형식으로 저장되고 있으며 이들의 대부분은 비구조적 이라고 했다. 또한, 인터넷 기술의 발전에 따라 이러한 비구조적 데이터의 양은 계속적으로 증가한다고 했다. Merk은 수많은 정보들이 텍스트 형식으로 저장되기 때문에 문서들 간의 내포된 관계를 찾는 것이 중요하며 이를 통하여 새로운 정보 즉 지식을 찾을 수 있다고 했다. 임희석은 최근 인터넷이 폭넓게 보급되어 온라인 상에서 얻을 수 있는 텍스트 양이 증가하고 있으며 이를 처리 하기 위한 연구의 필요성에 대하여 언급하였다.([4,6])

3.2 인터넷 뉴스 기사 분류 절차

인터넷 뉴스 기사 분류를 처리하는 방법은 기존 구조적 절차와는 다른 사전 작업이 필요하다. 이는 처리 대상이 되는 데이터가 비구조적으로 구성되어 있기 때문이다. 또한, 처리되는 데이터의 양이 많다. 즉, 뉴스 기사가 포함하는 단어들 모두를 처리 대상으로 한다.

3.2.1 문서 추출

문서 추출은 인터넷 뉴스 기사에 대한 자동 추출을 하는 단계이다. 이 단계는 지정된 인터넷 주소를 입력 받아 웹 서버의 문서를 읽어 저장하게 된다. 저장된 뉴스 기사는 문서 표현을 위한 변환 작업이 필요하다.

3.2.2 전처리

불필요한 데이터의 삭제 및 데이터 형식의 통일 단계이다. 추출된 웹 문서는 HTML 및 스크립트 언어를 포함한다. 전처리 단계에서는 이러한 웹 데이터를 삭제하고 순수 기사의 내용을 얻게 된다. 이 단계에서 자연어 처리 기법을 이용하여 저장된 뉴스 기사의 단어를 자동으로 추출하여 사용한다.

본 연구에서는 문서의 표현을 위하여 벡터 모델을 사용한다. 채용한 벡터 모델인 TFIDF 기법[Baeza-Yates and Ribeiro-Neto, 1999]에 대해서 간략히 소개한다. 벡터 모델은 부울리안 모델의 0 또는 1의 가중치의 한계를 극복하고 질의어와 검색문서 간의 부분일치(Partial Matching)를 가능하게 한다. 즉, 문서의 색인어들에 연속형 수치의 가중치를 부여하고, 이 가중치들을 이용하여 유사도를 계산한 후, 상위의 유사도를 갖는 문서들을 검색해오는 방법인데, 부울리안 모델에 의한 방법보다 정확하게 사용자의 정보요구사항에 부합하는 문서들을 검색할 수 있다는 장점 때문에 현재 널리 쓰이고 있다.([5,11])

벡터 모델은 하나의 문서를 t개의 정규화된 단어로 구성된 t-차원의 벡터로 표현한다.

문서 =(단어1, 단어2, ... , 단어n)

TF(Term Frequency) 즉 단어빈도는 문서에서 단어가 나타나는 빈도를 의미하며 (식 1)과 같이 계산된다.

$$TF_{i,j} = \frac{f_{i,j}}{\max f_{i,j}} \qquad \text{(식1)}$$

$f_{i,j}$ 검색문서 j 내의 단어 i의 출현 횟수를 나타내며, $\max f_{i,j}$ 은 검색문서 j에서 가장 많이 출현한 단어의 출현 횟수이다.

그리고, DF(Document Frequency), 즉 문서빈도는 보유한 전체 문서 중 해당 단어를 갖고있는 문서의 비율을 나타내는 것으로서 (식2)와 같이 계산된다.

$$DF_i = \frac{n_i}{N} \qquad \text{(식2)}$$

n_i 단어 i를 가진 문서의 개수 이며, N 은 보유한 전체 문서의 개수이다.

IDF(Inverse Document Frequency), 즉 역문서빈도는 (식 3)과 같이 계산되는데, 적은 개수의 문서에 걸쳐 나타난 색인어의 가중치는 높이고, 많은 개수의 문서에 걸쳐 나타난 색인어의 가중치는 낮추는 효과를 준다.

$$IDF_i = \log \frac{N}{n_i} \qquad \text{(식3)}$$

위의 (식1)와 (식3)을 사용하여, 문서 j 내의 색인어 i의 가중치, 즉 TFIDF는 (식4)과 같이 계산된다.

$$w_{i,j} = TF_{i,j} \times IDF_i \qquad \text{(식4)}$$

이렇게 얻어진 문서 벡터는 가중치의 값으로 표현된다. 이때 문서의 벡터 통일을 위하여 문서에서 발생한 모든 단어를 이용하여 인덱스 벡터를 만들고, 문서가 가진 단어의 가중치를 저장하게 된다.

3.2.3 단어 필터링

뉴스 기사는 수많은 단어들의 데이터로 구성된다. 이러한 이유로 단어 인덱스 벡터의 크기는 커지게 된다. 단어 필터링 단계에서는 단어의 차원을 축소한다. 이를 통하여 가중치가 낮은 단어를 분류 단계에서 계산에서 제외하게 된다.

차원을 축소하는 방법은 여러 가지

속성선택(feature selection) 기법을 적용할 수 있다. 다양한 차원 축소 기법 중 본 연구에서는 정보획득량(information gain)을 이용하여 차원을 축소하였다. 정보 획득량은 특정 문서 내에서의 특정 단어의 출현 여부로 얻은 정보의 비트 수를 이용하여 특정 문서가 포함되어야 할 범주를 예측한다. 이러한 알고리즘은 기계학습분야에서 단어의 유용성 측정 기준이 된다.

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (식5)$$

(식5)는 정보 획득량에 대한 식이다. S은 데이터 집합이며, A는 S은 정보 획득에서 계산된 단어이다. Sv는 단어 A가 v값을 가지는 S의 부분 집합을 나타낸다.

임계치(entropy)는 다음과 같은 (식6)로 계산된다.

$$Entropy(S) \equiv \sum_{i=1}^{c} -p_i \log_2 p_i \quad (식6)$$

위의 식을 이용하여 단어 차원을 수를 줄일 수 있다.

3.2.3 문서분류

필터링된 단어 데이터는 분류자(classifier)에 맞는 자료구조 형태로 변경되며 분류자를 이용하여 사전에 정의된 문서 범주에 분류되게 된다. 뉴스 기사에서 사용되는 대부분의 분류자는 데이터 마이닝에서 이용되는 분류자를 사용하게 된다.

즉, k-NN, 의사결정 트리, 베이지언 네트워크, 신경망을 사용한다. 최근에서 SVM에 대한 사용이 증가하는 추세이지만 본 논문에서는 세 가지의 분류자만을 이용한다.

IV. 자동 뉴스 기사 분류

인터넷 뉴스 가사의 자동 분류를 위해서는 데이터 집합(dataset)이 필요하다. 본 절은 자동 뉴스 기사 분류를 위한 데이터 집합의 생성과 전처리 과정을 설명한다.

5.1 데이터 집합

자동 뉴스 기사 분류를 위해서는 데이터 자원인 데이터 집합이 필요하다. 국외의 연구자들은 사전에 정의된 여러 가지 문서 집합을 이용하여 실험을 진행하다. 예를 들어 로이터 뉴스 기사 집합이나 인터넷 뉴스 그룹 데이터를 사용한다. 이들은 사전에 데이터 전문가에 의하여 분류가 되어 있으며 실험자가 쉽게 사용할 수 있다. 하지만, 이들의 데이터가 모두 영어로 작성되어 본 논문은 직접 데이터 수집을 하였다.

그러나, 국내에서는 문서 집합에 대한 표준적인 데이터 집합이 없다. 이러한 이유로 인터넷 뉴스 기사를 자동적으로 추출한다. 뉴스 기사의 자동 분류는 사전에 분류에 대한 정의 및 데이터 집합을 가지고 있음을 전제로 한다.

본 논문에서 사용한 뉴스 기사는 전자 신문사의 2003년도 3월에서 4월의 인터넷 뉴스 기사를 자동 추출하여 데이터 집합을 만들었다. 추출된 뉴스 기사의 수는 805개이며 17개의 사전 범주(class)를 가지고 있다. 분류별 문서의 수는 [표1]와 같다. 문서 집합은 "대분류명>소분류명" 으로 구성되어 있다.

[표1] 인터넷 뉴스 기사 문서 집합

문서분류	문서 수
IT/인터넷 > 정보통신	82
IT/인터넷 > 정보보호	158
e 비지니스_SI > E 비즈	56
e 비지니스_SI > 정보화	76
솔루션/시스템 > 솔루션	27
솔루션/시스템 > 시스템	13
국제 > 국제뉴스	90
국제 > E 월드	40
국제 > IT 마켓뷰	2
문화산업 > 게임	23
벤처/증권 > 증권	68
벤처/증권 > 전국	43
벤처/증권 > 벤처	18
반도체/산업전자 > 과학기술	28
반도체/산업전자 > 반도체	12
정보가전 > 디지털가전	27
정보가전 > 정보기기	42

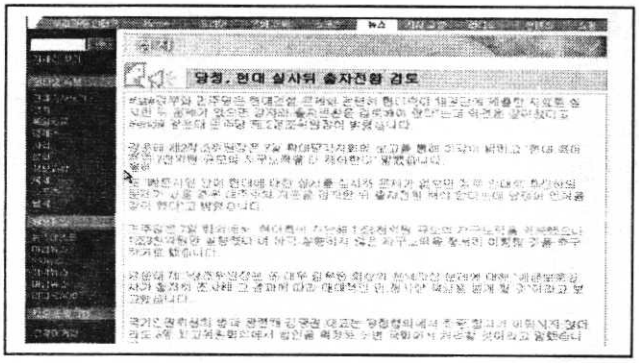
5.2 전처리기

인터넷에서 추출된 뉴스 기사 집합은 HTML 및 스크립트 언어를 포함하고 있다. 전처리 단계에서 전처리기는 문서 분류에 필요 없는 데이터를 삭제하게 된다. 이를 통하여 기사에 대한 데이터만을 얻을 수 있다. 이때 제거되는 대부분의 데이터는 태그(tag) 데이터가 된다.

또한, 문서 분류에 영향을 주지 않는 관사나 조사, 대명사의 데이터는 삭제를 한다. 한글에 대한 전처리기 작업은 국민대학교 자연어 처리 연구실의 HAM을 이용하여 작업을 하였다. HAM은 국내에서 작성된 쉐어웨어 프로그램으로 한글 자연어 처리에서 사용되고 있다.

[그림2]은 전처리기 추출 이전의 데이터이다. 즉, 인터넷 태그 데이터를 포함한 뉴스 기사이다. [그림3]은 전처리기를 이용하여 불필요한 데이터를

삭제한 후의 그림이다. 즉, 인터넷 태그를 제거한 뉴스 기사이다. 이러한 기사는 파일로 저장된다. 이때 뉴스 기사는 로컬 시스템에 디렉토리별로 저장된다. 폴더의 이름은 "대분류.소분류명"으로 작성된다.



[그림2] 전처리 추출 이전 데이터

개인용 디지털녹화기(PVR)가 홈네트워크 가전의 핵심으로 급부상하면서 업계도 시장선점 경쟁에 돌입하는 등 시장개화를 위한 경쟁이 치열해지고 있다. 3일 관련업계에 따르면 소니, 필립스 등 외국 가전사들이 지난해말 이미 홈네트워킹의 허브역할을 할 PVR 응용상품을 내놓고 시장공략에 나선 가운데 그동안 관망하던 삼성·LG전자 등이 잇따라 신제품을 출시하면서 시장을 달구고 있다. 이에 따라 관련업계는 올해를 콤보형 PVR를 비롯한 PVR

[그림3] 전처리 추출 기사 데이터

VI. 자동 뉴스 기사 분류 시스템 구현

본 절은 한글 인터넷 뉴스 기사 분류에 대한 실험을 위하여 개발된 NICS(news information classification systems) 시스템에 대한 구현 설명을 한다.

6.1 개발 도구

NICS 시스템은 한글 뉴스 기사 데이터를 처리하기 위한 목적으로 개발 되었다. 한글 뉴스 기사의 처리를 위하여 유니코드를 지원하는 자바(java)를 사용하여 시스템을 개발하였다. 한글 뉴스 기사에서 단어(term) 처리를 위하여 국민대학교 자연어 처리 연구실의 HAM 모듈을 사용하였다. HAM 모듈은 한글 데이터에서 사전에 정의된 단어 인덱스를 사용하여 단어를 자동으로 추출한다.[9] 하지만 작성된 소스는 C 언어로 작성된 관계로 실행 파일을 호출하여 수집된 뉴스 기사의 단어를 추출한다.

문서 분류기는 오픈 소스를 지원하는 WEAK를 사용하여 분류처리를 하였다.[19] WEAK는 입력 파일의 형식을 ARFF 형식을 요구하기 때문에 이에 대한 데이터 형식을 만들기 위한 모듈을 개발하여 사용하였다. HAM에서 추출된 문서 단어들을 [그림4]와 같이 ARFF 파일형식으로 변경하여 WEAK에 보낸다.

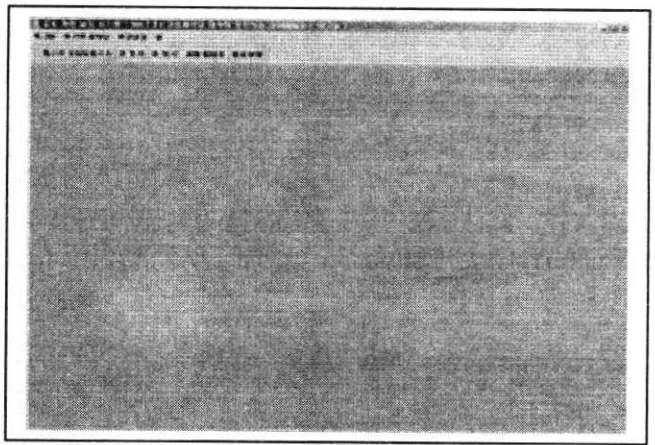
```
@relation doc

@attribute 기술 {0,1}
@attribute 연구 {0,1}
@attribute 범주 {경영.경영정보,경영.인사}

@data
0,0,경영.경영정보
1,1,경영.인사
0,1,경영.경영정보
```

[그림4] doc 뉴스 기사의 ARFF 파일 형식 예

ARFF 파일은 기존 문서 편집 프로그램을 이용하여 작성될 수도 있다. ARFF 파일의 태그를 이용하여 관계명, 속성과 데이터로 구성된다. @relation 태그는 문서의 명이다. @attribute은 문서가 가지는 단어들과 분류 범주를 나타낸다. @data는 @attribute에 대응하는 데이터이다. 이러한 형식의 파일은 자동 생성을 위하여 구현되었다. 즉, 문서에 따른 단어들을 입력 받아 ARFF파일을 자동으로 생성한다.

[그림5]는 NICS 시스템의 메인 화면이다. 화면은 자바의 스윙(swing)을 이용하여 사용자 인터페이스를 구성하였다. 작성된 시스템은 자바로 개발되어 어떤 운영체제 환경에서도 쉽게 접근이 가능하며 기존 뉴스 정보 시스템과도 데이터 교환이 자유롭다.

사용자 인터페이스는 데이터 처리 절차에 따라서 선택하여 사용할 수 있도록 하였다. 즉, 데이터 선택, 전처리, 문서분류 등으로 메뉴를 구성하였다.



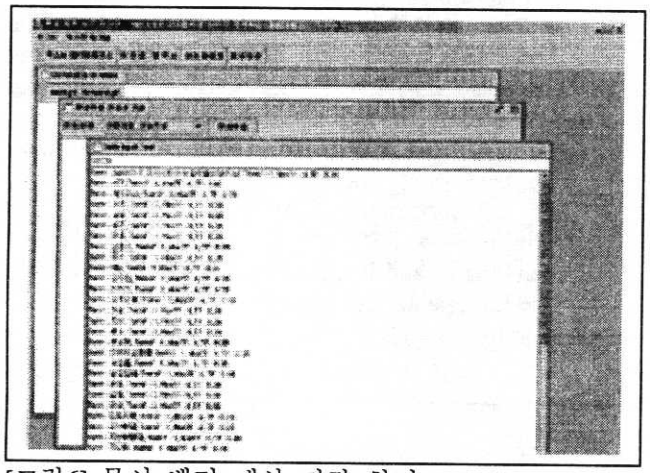
[그림5] NICS 시스템 메인 화면

6.2 전처리기

인터넷 뉴스 기사는 웹에 게시된 뉴스 기사를 이용하기 때문에 NICS 시스템은 뉴스 기사 추출을 위한 모듈을 개발하여 사용했다. 기사 추출 모듈은 웹 페이지에 대한 정보를 입력 받아 자동으로 로컬 시스템에 텍스트 파일의 형식으로 기사 데이터를 저장한다.

저장된 데이터는 문서별 단어들을 추출하며 가중치를 값을 계산하여 데이터 베이스에 저장된다. 가중치의 계산은 앞장에서 소개된 벡터 모델을 사용하여 계산된다.

벡터 모델에 대한 계산은 포함되는 문서와 단어의 수가 매우 큰 관계로 데이터 베이스의 내장 언어를 이용하여 개발하였다. 이는 데이터 내장 언어를 이용한 연산 속도가 개발 언어를 이용한 메모리 계산보다 빠르기 때문이다.[17]



[그림6] 문서 벡터 계산 저장 화면

6.3 뉴스 기사 분류기

벡터 모델을 이용한 가중치 계산이 완료되고 문서에 따른 가중치 할당이 완료되면 기사 분류 작업을 진행하다. 실험에서 사용된 분류자(classifier)는 WEAK를 이용하여 구현되었다. WEAK는 다양한 분류자를 제공한다. 실험에 사용한 세 가지의 분류자는 다음과 같다.

· 베이지언 네트워크(weak.classifier.NaiveBayes)
· 의사결정 트리(weak.classifier.j48.J48)
· K-NN(weak.classifier.IBk)
· 인공신경망

분류자 모두는 자바 패키지의 클래스로 제공이 된다. NICS 사전 처리된 문서 단어들을 세 가지의 분류자에 전달하게 된다.

단어의 차원을 축소하기 위하여 InfoGainAttributeEval과 Ranker 알고리즘을 사용하였다. 실험에서는 단어 차원별 분류자의 분류율을 발결하기 위하여 100개 단위로 차원을 축소하였다.

6.4 실험과 평가

실험은 한글 뉴스 데이터 집합을 이용하여 진행하였다. 차원 축소 알고리즘이 문서 분류에 어떤 영향을 주는지를 확인하기 위하여 문서 내의 단어 수를 다르게 하면서 뉴스 기사의 분류를 하였다. 전체 데이터 집합에서 훈련용 데이터와 테스트용 데이터 집합으로 나누었다. 그 비율은 70%와 30%로 하였다. 훈련용 데이터 집합은 분류자를 훈련시키며 테스트 데이터 집합은 분류자의 정확도를 확인하기 위하여 사용하였다.

[표2] 문서 분류 실험결과

단어수	베이지언 네트워크	의사결정 트리	신경망	K-NN
100	55%	53%	55%	44%
200	62%	59%		
400	66%	63%		
800	67%	65%		
1600	69%	66%		

두 개의 데이터 집합을 이용한 실험 결과는 [표2]와 같다. 실험에서 단어 수를 100, 200, 400, 800, 1600개로 변경을 하면서 분류자를 실험하였다. 이때, 신경망과 K-NN 분류자는 100이상의 단어에 대하여 입력이 되지 않음을 확인 할 수 있었다. 때문에 100개 이상에 대해서는 베이지언 네트워크와 의사결정 트리를 이용하여 실험을 하였다.

실험 결과 분류자에 입력되는 단어의 수가 증가하면서 분류율도 증가하는 것을 알 수 가 있었다. 또한, 의사결정 분류자 보다는 베이지언 네트워크 분류자의 분류율이 더욱 높아 짐을 확인 할 수 있었다. 하지만 전체 적인 분류율의 차이는 1%에서 3%사이의 차이를 보였다. 이를 통하여 두 개의 분류자를 이용하는 것이 큰 차이를 가지지 않는다는 것을 알 수 있었다.

기존의 뉴스 정보 시스템에 개발된 NICS 시스템을 통하여 분류를 할 수 있었다. 이때 신규로 입력된 뉴스 기사에 대해서는 피드 서버를 통하여 입력받아 자동 분류 하였다. NICS는 자바로 개발되어 쉽게 통합을 할 수 있었다.

VII. 결론

인터넷 뉴스 기사에 대한 자동 분류에 대한 시도는 국내외에서 많은 연구와 성과를 얻었다. 또한, 자동 분류에 대한 활발한 연구 교류 및 공동 프로젝트가 진행이 되고 있다.

본 논문은 한글 인터넷 뉴스 기사를 이용하여 한글

문서 분류 정보 시스템의 구현 및 절차에 대하여 연구해 보았다. 이를 통하여 국내에서 현재 활발히 시도되고 있는 비구조적 데이터에 대한 접근 방법과 정보 시스템으로 반영하기 위한 방법을 연구하였다.

한글 뉴스 기사에 맞는 분류자를 찾기 위하여 베이지언 네트워크, 의사결정 트리, 신경망, k-NN 분류자를 이용하여 실험을 하였다. 이때 분류자에 입력되는 단어 수가 커지면서 분류율이 높아지는 것을 알 수 있었다.

연구 초기 많은 데이터의 수집을 목적으로 하였지만 인터넷 뉴스 기사 서버가 한 달 이상의 데이터를 자동으로 삭제하는 관계로 많은 데이터의 수집이 되지 않은 점이 아쉽다. 또한, 뉴스 기사 분류자가 100개 이상의 데이터를 입력 받지 않아 이에 대한 한글 처리에 대한 문제점을 남겨 두었다.

끝으로 한글 뉴스 기사 분류에 대한 많은 연구의 활용을 기대 한다.

감사의 글

연구 논문 작성에서 문화방송 방송 정보 시스템에 대한 많은 조언을 주신 ㈜컴텍코리아의 방송 기술 연구소 팀원에게 감사를 전합니다.

참고문헌

[1] 김시천, Memory-Based Reasoning을 이용한 HTML 문서분류 시스템의 설계 및 구축, 아주대학교 경영정보학과 석사학위 논문, 1999.

[2] 김진상, 신양규(2000),"베이지안 학습을 이용한 문서의 자동분류", 한국데이터정보과학회

[3] 고수정, 이정현(2001),"Apriori 알고리즘에 의한 연관 단어 지식 베이스에 기반한 가중치 부여된 베이지안 자동 문서 분류", 멀티미디어 학회지 논문지 ,제4권 제2호

[4] 이상진, 데이터 활용에 대한 연구, 고려대학교 통계학과 석사 학위 논문, 1998

[5] 이재식, 이종운(2002),"사례기반 추론을 이용한 한글 문서분류 시스템", 경영정보학연구, 제12권 제2호

[6] 임희선과 남기춘(2002),"경험적 정보를 이용한 kNN 기반 한국어 문서 분류기의 개선", 한국컴퓨터교육학회지, 제5권, pp.37-44

[7] 조성준, 데이터 마이닝을 이용한 의사결정 시스템, 고려대학교 전산학과 석사학위 논문, 1999

[8] 최승경, 데이터 마이닝을 이용한 웹사이트 분석, 고려대학교 경영학과 석사학위 논문, 2001

[9] 한글공학 연구소, 한국어 분석 라이브러리 HAM 사용 설명서, 국민대학교, 2003

[10] 허원창, 신경회로망을 사용한 WWW 문서의 자동분류, 서울대학교 산업공학과 석사학위 논문, 1999

[11] Baeza-Yates, R. and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999

[12] Caldas, Carlos H., and Soibelman, Lucio (2003), "Automating hierachical document classification for construction management information systems", *Journal of Automation in Construction*, pp. 1-12

[13] Dash,M.,and Liu, H.(1997),"Feature Selection for Classification", *Intelligent Data Analysis*, pp. 131-156

[14] Klosgen, Willi and Zytkow, Jan M. (2002), *Hand of Data mining and Knowledge discovery*, Vol. 1, pp. 749-757

[15] Ko,Youngjoog., Park, Jinwoo., and Seo, Jungyun. (2002),"Improving text categorization using the importance of sentence", *Information Processing and Management*, Vol. 38, pp.231-240

[16] Sebastiani, Fabrizio (2002), "Machine Learing in Automated Text Categorization", *ACM Computing Surveys*, Vol. 34, pp.1-47

[17] S. T. Dumais, J. Platt, D. Heckerman and M. Sahami (1998),"Inductive learning algorithms and representations for text categorization", *Proceedings of ACM-CIKM98*, Nov. 1998, pp. 148-155

[18] Tan,Chade-Meng, Wang., Yuan-Fang, Wang., and Lee, Chan-Do.(2002)."The Use of Bigrams to Enhance Text Categorization" *Information Proceeding and Mangement*, pp.1-30.

[19] Witten, Ian H. and Frank, Elbe. (2000). *Data Mining.* Morgan Kaufmann Publishers.