# NewsML 기반의 뉴스 자동 분류 시스템에 관한 연구

이탁희*, 홍금원**
고려대학교 컴퓨터정보통신대학원 미디어공학과*
고려대학교 컴퓨터 전파통신공학과**
e-mail : thlee@joins.com*, gwhong@nlp.korea.ac.kr**

# Study on Automatic Classification System of News based on NewsML

Tak-Hee Lee*, Gumwon Hong**
Media Science & Engineering, Graduate School of Computer & Information Technology,
Korea University*
Dept of Computer and Radio Communications Engineering, Korea University**

**요 약**

뉴스 분류 체계는 각각의 기사에 정치, 경제, 사회 등 가장 적합한 주제별로 분류하는 것으로 언론사별 분류 체계는 통일성이 없이 전혀 다르게 구성되어 사용하고 있다. 이로 인해 방대한 콘텐트를 통합하는데 많은 어려움이 있으며, 그 만큼 시스템과 인력에 대해 중복 투자가 되고 있다. 이런 문제점을 개선하기 위해 국제 표준인 NewsML 에 기반한 뉴스 분류에 대해 제안한다. NewsML 은 XML 기반의 유연성과 확장성이 있는 구조적인 표준 형식으로 다양한 데이터 표현이 가능하여 자동 문서 범주화에 필요한 중요한 자질 선택이 가능하다. 본 논문에서는 NewsML 형식으로 되어 있는 뉴스와 그렇지 않은 뉴스를 구분하여 자동 분류에 대한 비교 실험을 한다. NewsML 의 구조화된 정보를 활용한 실험이 뉴스의 제목과 본문만으로 실험한 결과보다 좋은 성능을 보여 주었으며, 그 중에서 자질 공간이 아주 큰 경우에 유용하고 문서 분류에 효과가 뛰어난 지지 벡터 기계 모델이 가장 좋은 성능을 보였다.

## 1. 서론

뉴스 분류는 다매체, 매체간 융합, 온라인유통, 원소스 멀티유즈를 전략적으로 지향하는 미디어 기업이라면 간과할 수 없는 핵심적인 요소이다. 어떤 콘텐트를 생산할 것인가의 단계를 넘어 생산된 콘텐트를 어떻게 더 많이 활용할 것인가 즉, 이용자와의 접점을 극대화하는 쪽에 초점이 맞춰져 있다[2].

뉴스 분류 체계는 각각의 기사를 정치, 경제, 사회 등 가장 적합한 주제별로 분류하기 위해 만들었으며, 현재 언론사에서 사용하는 분류체계는 표 1 과 같이 통일성 없이 사용하고 있어 분류체계 표준화에 대한 개선작업이 필요하다.

<표 1> 언론사의 일반적인 분류방식

| 중앙일보 | 연합뉴스 | 동아일보 | 조선일보 |
|---|---|---|---|
| 정치 | 경제 | 정치 | 정치 |
| 사회 | 증권 | 사회 | 사회 |
| 국제 | 정치 | 국제 | 국제 |
| 경제 | 국제 | 경제 | 경제 |
| 스포츠 | 사회 | 스포츠 | 스포츠 |
| 건강/과학 | 전국 | 의학/과학 | 사설칼럼 |
| 문화 | 문화 | 문화/연예 | 연예 |
| 교육 | 스포츠 | 사설칼럼 | |
| 연예 | 연예 | | |
| 사설칼럼 | | | |

이로 인한 문제점은 방대한 콘텐트를 통합하는데 많은 어려움이 있으며, 그 만큼 시스템과 인력에 대해 중복 투자가 되고 있다. 또한, 비 표준 분류체계로 형식과 전송방식이 달라 유지보수의 어려움이 있으며, 컨버전스에 따른 다양한 디바이스 활용도 저하되어 콘텐트 경쟁력의 한계를 드러내고 있다. 국내에서도 이런 문제점을 개선하기 위해 국제 표준인 NewsML 의 분류체계를 도입하여 개선하고자 하는 노력이 지속되고 있다[2,3].

현재 국내 포털의 경우 많은 뉴스 콘텐트를 제공받아 서비스를 하고 있으나 분류 방법은 거의 수동 문서 분류에 의존하고 있다. 하지만 빠르게 생성되는 뉴스 속도에 비해 분류 방식은 따라가지 못해 인력과 비용이 많이 투자되고 있다[10]. 따라서 뉴스 전송방법을 표준화 하고 NewsML 이 갖고 있는 특성을 반영한 자동 문서 범주화에 대한 연구는 중요하다고 할 수 있다.

뉴스 기사의 특징은 일반적인 관심 내용과 전문적인 어려운 주제까지 다양하게 구성되며, 어휘가 어렵고, 문장이 다른 문서에 비해 짧아 중요한 핵심사항은 뉴스의 앞에 기술하는 특징이 있어 다양한 문제와 관점에 대해 색인 작업을 어렵게 한다[4]. 또한 전송방법도 제목, 본문, 카테고리, 기자, 출처 등 가장 기본적인 항목만 전송하여 자동 분류에 많은 어려움이 있다.

NewsML 은 XML 기반의 유연성과 확장성이 있는 구조적인 표준 형식으로 다양한 데이터 표현이 가능하여 자동 문서 범주화에 필요한 중요한 자질 선택이 가능하다.

본 논문에서는 4 가지 기계학습 모델을 이용하여 NewsML 에 적합한 기계학습 모형을 확인하고, NewsML 형식으로 되어 있는 뉴스와 그렇지 않은 뉴스를 구분한 비교 실험을 통해 NewsML 형식이 자동 뉴스 분류에 유용함을 밝힌다.

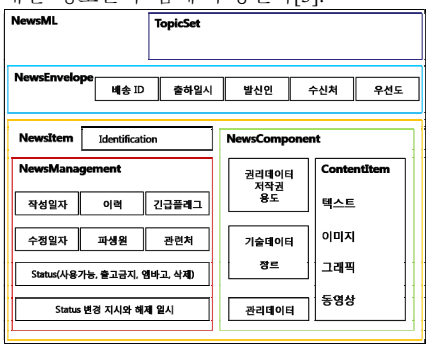
## 2. 관련 연구

### 2.1. NewsML 정의 및 특징

NewsML 은 국제언론통신평의회(IPTC)가 발표한 국제 표준 형식으로 뉴스와 XML 기반을 두고 있다. NewsML 은 원래 뉴스 전송을 위한 표준 형식을 정하는 목적으로 설계되었으나 아카이브 구축, 뉴스의 작성, 편집, 관리, 출판의 전 영역을 지원할 수 있는 것으로 인정받았다. XML 문서는 미리 정의해 둔 문서 구조에 따라 효과적으로 뉴스 콘텐트를 표현할 수 있다[2,3].

NewsML 표준 형식으로 뉴스를 변환하여 전송하는 연구 결과 무결성을 보장하고 데이터베이스를 이용하여 유연성과 확장성이 있고, 전송방식이 단일화 되어 유지보수에 대한 부담이 감소하였다[11].

뉴스 콘텐트의 전송은 콘텐트 그 자체인 내용뿐만 아니라 그림 1 처럼 내용을 설명하는 메타데이터와, 어떻게 그 콘텐트를 처리할 것인가 하는 기사 관리 정보, 그리고 그 콘텐트가 전달되어지는 경로와 과정 등에 대한 정보들과 함께 구성된다[3].




(그림 1) NewsML 구조

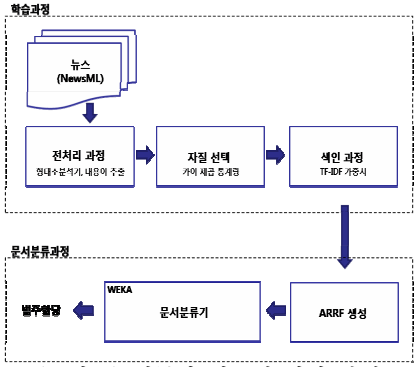
### 2.2. NewsML 문서 검색을 위한 색인 시스템

NewsML 문서를 검색하기 위한 색인 DB 구성[4]은 구조화된 문서를 파악하고 각 요소에 입력되는 뉴스 기사의 특징을 파악할 수 있다. 형태소 분석 가능한 구성요소를 크게 서지정보, 본문, 제목/부제목, 기타정보로 구분한다. 서지정보와 기타 정보는 본문의 주요 내용과 관련이 적은 정보이나 다양한 검색 조건을 맞추기 위해 반드시 필요한 정보이며, 문서범주화에 필요한 제목/부제목/키워드, 검색정보는 색인할 때 가중치를 부여하여 검색 성능을 높였다[4].

그러나 기존 연구는 NewsML 전송에 대한 시스템 구현이나 색인방법에 대한 연구와 표준화 방법론에 대한 연구가 대부분이다. 하지만 방대한 콘텐트가 생성되고 있는 현실을 감안할 때 NewsML 의 특성을 반영한 자동 문서 범주화에 대한 연구가 같이 되어야 한다.

## 3. 자동 문서 범주화의 학습과 분류과정

일반적으로 기계 학습을 이용한 문서 범주화 시스템의 각 처리 단계는 그림 2 과 같이 학습 과정과 문서 분류 과정으로 구분된다[1].




(그림 2) 범주화 시스템 처리 단계

### 3.1. 전처리 과정

뉴스를 정규화 하여 내용어(Content Word)를 추출하는 단계로 HTML, CSS, Javascript, 특수문자 등 불필요한 정보를 제거한 후 NewsML 형식으로 변환하여 별도 저장한다.

내용어를 추출하기 위해 형태소 분석기는 HAM (Hangul Analysis Module)[13]을 사용하여 내용어로 가장 많이 사용하는 명사와 복합명사를 추출한다. 그리고 공통적으로 많이 나타나거나 별다른 정보를 주지 못하는 불용어(Stop Word)는 제거하여 처리 효율을 높인다.

### 3.2. 자질 선택

문서 범주화 성능의 저하 없이 자질의 수를 줄이기 위해 유용한 자질을 선별한다. 여러 자질 선별 방법 중에서 상호 정보(Mutual Information)와 카이 제곱 통계량($X^2$ Statistics)이 비교적 효과적인 것으로 알려져 있다[12]. 본 논문에서는 비교적 구현이 쉽고 빈도 수가 많은 문서에 적합한 카이 제곱 통계량을 사용하여 자질을 선택한다.

### 3.3. 문서 표현 및 색인 과정

선택된 자질을 이용하여 문서를 어떻게 표현할 것인가에 대한 방법으로 일반적으로 가장 많이 사용되는 문서 표현 방법은 벡터 공간 모델이다[1]. 실험에 적용한 가중치 방법은 해당 문서에서 각 자질의 빈도

와 역문헌빈도(IDF)의 곱으로 나타내는 TF-IDF 가중치 방법을 사용한다.

- 색인어 가중치 기법

$$w_{ij} = f_{ij} \times \log \frac{N}{n_i}$$

$f_{ij}$ : Term Frequency

$n_i$ : Number of documents in which the index term $k_i$ appears

$N$ : Total number of documnets

### 3.4. 문서 분류 과정

- 나이브 베이지언 확률 모델(Naïve Bayesian Probability Model)
  대상 문서가 각 범주에 속할 확률을 구해 가장 큰 확률 값을 갖는 범주에 그 문서를 할당하는 기법이다[12]. 나이브 베이지언 확률 모델은 쉽게 구현가능하고 다른 모델에 비해 적은 계산양으로도 효과적인 성능을 기대할 수 있다.
- 지지 벡터 기계(SVM : Support Vector Machines)
  학습 문서를 통해 생성된 양성 자질과 음성 자질을 벡터 공간으로 표현하고 이들의 차이를 극명하게 하는 지지 벡터를 찾는 방법으로 사용한다. SVM은 일반적으로 자질 공간이 아주 큰 경우에 유용하고 문서 분류에 효과가 뛰어난 것으로 알려져 있다[1].
- 의사결정 트리 모델(Decision Tree Model)
  트리 구조로 구성하여 분류를 가능하게 하는 분석방법으로 어떤 변수가 분류를 하는데 영향을 미치는지 트리 구조를 통해 쉽게 파악할 수 있으며, 모델화와 탐색의 두 가지의 특성을 전부 가지고 있다[9].
- k-최근린법(k-Nearest Neighbor)
  입력 문서와 가장 유사한 k개의학습 문서를 구하고 k개 안에 포함된 문서의 범주를 확인하여 입력 문서의 범주를 할당하는 방법이다. 즉, 입력 문서가 주어졌을 때 학습 문서 중에서 실험 문서와의 유사도가 가장 높은 k개의 문서를 추출하고 그들을 사용하여 각 후보 범주의 순위를 매기는 방법이다[12].

## 4. 실험 및 평가

### 4.1. 실험 Data

실험에 사용한 문서는 중앙일보와 연합뉴스를 2007년 1 월부터 12 월까지 기사를 추출하여 문서 집합으로 생성했다. NewsML 의 분류체계는 총 18 개의 대분류로 구성되는데 뉴스정보가 가장 많고 대표성을 띄는 스포츠(Sports), 정치(Politics), 문화(Culture), 과학/테크놀로지(IT), 보건위생(Health)으로 축소하여 실험하였다.

실험에 사용된 자질은 스포츠(392 개), 정치(329 개), 과학&테크놀로지(279 개), 보건위생(354 개)등 총 1,658 개를 사용했으며, 뉴스 기사 건 수는 스포츠 876 건, 정치 2,586 건, 문화 2,388 건, 과학/테크놀로지 2,643 건, 보건위생 2,968 건 등 총 11,461 건을 대상으로 실험을 진행했다. 학습/테스트 문서는 랜덤하게 9:1 로 나누었고, 검증은 10-fold cross-validation 방법을 사용하였다[14].

### 4.2. 평가 척도

본 논문에서 문서 범주화를 평가하기 위한 척도로 정확률(Precision), 재현율(Recall), F-Measure 를 사용하였으며, 척도의 정의는 다음과 같다.

$$\Pr ecision = \frac{\text{Categories assigned by the system and correct}}{\text{Categories assigned by the system}}$$

$$\text{Re } call = \frac{\text{Categories assigned by the system and correct}}{\text{Total Categories correct}}$$

$$F - measure = \frac{2rp}{r + p} \quad \text{r : Recall, p : Precision}$$

### 4.3. 실험 방법

자질 선택 단계에서는 카이 제곱 통계량 기법을 사용하고 문서 색인 단계에서는 TF-IDF 가중치 기법으로 다음과 같이 3 가지로 나누어 실험한다.

- **Test 1** : 제목+부제목+키워드
  NewsML에서 문서범주화에 사용될 수 있는 항목 중 본문을 제외하고 적은 양의 정보를 분석하여 결과 측정한다.
- **Test 2** : 제목+본문
  문서범주화에 일반적으로 사용되는 뉴스 기사 제목과 본문을 분석하여 결과를 측정한다.
- **Test 3** : 제목+부제목+키워드+본문
  NewsML 의 구조화 되어 있는 모든 정보에서 문서범주화에 사용될 수 있는 모든 항목을 분석하여 개선 여부를 확인한다.

### 4.4. 실험 및 결과 분석

Test1(제목+부제목+키워드)의 실험 결과 적은 양의 정보에도 불구하고 지지 벡터 기계 모델이 다른 모델에 비해 정확률, 재현율, F-Measure 가 전반적으로 높았으나 문서의 범위가 작아 각 분류별 또는 기계학습 알고리즘에 따라 편차가 심했다. 특히 스포츠(Sports)의 경우 높은 경우 90.8%의 정확률을 보였으나 문화(Culture)는 76.1%으로 낮았다.

**<표 2> Test1 지지 벡터 기계 Confusion Matrix**

| Class | Sprts | Politics | Culture | IT | Health | Total | Recall |
|---|---|---|---|---|---|---|---|
| Sprts | 655 | 25 | 73 | 24 | 17 | 794 | 82.5% |
| Politics | 13 | 1,971 | 113 | 85 | 43 | 2,225 | 88.6% |
| Culture | 20 | 68 | 1,528 | 131 | 115 | 1,862 | 82.1% |
| IT | 17 | 81 | 147 | 1,918 | 123 | 2,286 | 83.9% |
| Health | 16 | 54 | 148 | 298 | 2,003 | 2,519 | 79.5% |
| Total | 721 | 2,199 | 2,009 | 2,456 | 2,301 | 9,686 | 83.4% |
| Precision | 90.8% | 89.6% | 76.1% | 78.1% | 87.0% | | |

Test2(제목+본문)의 실험 결과 지지 벡터 기계 모델이 다른 모델이 비해 정확률, 재현율, F-Measure 가 높았다. 각 분류 별 편차가 크지 않고 정확률이 전반적으로 높으나, 특이하게 과학/테크놀로지(IT)의 경우 정확률이 70.1%로 낮은 경우도 있었다. 지지 벡

터 기계 모델의 Confusion Matrix 는 표 3 처럼 정확도와 재현율에서 정치(Politics) 분류가 가장 좋은 결과를 보였다

<표 3> Test2 지지 벡터 기계 Confusion Matrix

| Class | Sprts | Politics | Culture | IT | Health | Total | Recall |
|---|---|---|---|---|---|---|---|
| Sprts | 795 | 6 | 12 | 60 | 2 | 875 | 90.9% |
| Politics | 12 | 2,313 | 25 | 207 | 15 | 2,572 | 89.9% |
| Culture | 17 | 38 | 1,999 | 253 | 35 | 2,342 | 85.4% |
| IT | 13 | 44 | 88 | 2,383 | 97 | 2,625 | 90.8% |
| Health | 18 | 10 | 78 | 498 | 2,355 | 2,959 | 79.6% |
| Total | 855 | 2,411 | 2,202 | 3,401 | 2,504 | 11,373 | 86.6% |
| Precision | 93.0% | 95.9% | 90.8% | 70.1% | 94.0% | | |

표 4 는 Test3 에서 각 기계학습 모델에 따른 성능 비교를 나타내며, 지지벡터 모델이 전 범주에 걸쳐 가장 좋은 성능을 보여 주고 있음을 알 수 있다. 지면에 표시하지는 않았지만, Test1 와 Test2 의 실험 결과도, 역시 지지벡터 모델의 결과가 가장 우수한 것으로 드러났다.

<표 4> Test3 Precision, Recall, F-Measure

| Class | 베이지안 확률 모델 | | | 지지 벡터 기계 | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F-Measure | Precision | Recall | F-Measure |
| Sports | 86.5% | 95.4% | 90.8% | 93.2% | 91.8% | 92.4% |
| Politics | 94.8% | 87.5% | 91.0% | 96.2% | 90.8% | 93.5% |
| Culture | 74.1% | 90.0% | 81.3% | 90.0% | 86.5% | 88.2% |
| IT | 79.2% | 77.5% | 78.3% | 72.1% | 91.0% | 80.4% |
| Health | 90.5% | 80.0% | 84.9% | 94.2% | 80.2% | 86.6% |

| Class | k-최근린법 | | | 의사결정 트리 모델 | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F-Measure | Precision | Recall | F-Measure |
| Sports | 92.6% | 84.2% | 88.2% | 85.3% | 91.3% | 88.2% |
| Politics | 83.6% | 87.1% | 85.3% | 88.6% | 90.4% | 89.5% |
| Culture | 78.7% | 74.5% | 76.5% | 82.5% | 81.8% | 82.2% |
| IT | 65.3% | 72.7% | 68.8% | 71.6% | 74.3% | 72.9% |
| Health | 80.2% | 74.8% | 77.4% | 86.4% | 80.6% | 83.4% |

표 5 는 지지 벡터 기계 모델을 사용했을 경우 각 Test 환경간의 성능 차이를 보여준다. NewsML 의 고유한 항목을 활용한 Test3 의 경우가 Test2 에 비해 모든 범주에 걸쳐서 약간의 성능향상을 보여주고 있음을 알 수 있다.

<표 5> 각 Test 환경 별 지지기계 벡터의 성능

| Class | Test1 | | | Test2 | | | Test3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F-Measure | Precision | Recall | F-Measure | Precision | Recall | F-Measure |
| Sports | 90.8% | 82.5% | 86.5% | 93.0% | 90.9% | 91.9% | 93.2% | 91.8% | 92.4% |
| Politics | 89.6% | 88.6% | 89.1% | 95.9% | 89.9% | 92.8% | 96.2% | 90.8% | 93.5% |
| Culture | 76.1% | 82.1% | 78.9% | 90.8% | 85.4% | 88.0% | 90.0% | 86.5% | 88.2% |
| IT | 78.1% | 83.9% | 80.9% | 70.1% | 90.8% | 79.1% | 72.1% | 91.0% | 80.4% |
| Health | 87.0% | 79.5% | 83.1% | 94.0% | 79.6% | 86.2% | 94.2% | 80.2% | 86.6% |

지지 벡터 기계 모델의 Confusion Matrix 는 표 6 처럼 정확도와 재현율에서 정치(Politics), 스포츠(Sports) 분류가 가장 좋은 결과를 보였다

<표 6> Test3 지지 벡터 기계 Confusion Matrix

| Class | Sprts | Politics | Culture | IT | Health | Total | Recall |
|---|---|---|---|---|---|---|---|
| Sprts | 803 | 4 | 21 | 47 | 1 | 876 | 91.7% |
| Politics | 14 | 2,341 | 27 | 181 | 14 | 2,577 | 90.8% |
| Culture | 17 | 36 | 2,056 | 224 | 43 | 2,376 | 86.5% |
| IT | 13 | 41 | 94 | 2,390 | 89 | 2,627 | 91.0% |
| Health | 15 | 11 | 87 | 474 | 2,372 | 2,959 | 80.2% |
| Total | 862 | 2,433 | 2,285 | 3,316 | 2,519 | 11,415 | 87.3% |
| Precision | 93.2% | 96.2% | 90.0% | 72.1% | 94.2% | | |

## 5. 결론

본 논문에서는 NewsML 형식으로 되어 있는 뉴스와 그렇지 않은 뉴스를 구분하여 자동 분류에 대한 비교 실험을 하였다. NewsML 의 구조화된 정보를 활용한 실험(Test3)이 제목과 본문만으로 실험한 결과(Test2)보다 전체 범주에 걸쳐서 좋은 성능을 보여주었으며, 그 중에서 문서 분류에 일반적인 효과가 뛰어난 지지 벡터 기계 모델이 가장 좋은 성능을 보임을 알 수 있었다.

하지만, Test2 와 Test1 의 결과차이를 보면 알 수 있듯이 제목과 본문이 중요한 역할을 수행하고 있기 때문에 그 외의 항목에 비하여 뉴스 분류에 영향을 미치는 비중을 달리 실험해 볼 필요가 있다고 여겨진다. 따라서, 향후 연구로 NewsML 의 정보를 보다 세분화하고 항목별 중요도를 분석하여 가중치 계산을 한다면 좀 더 좋은 성능을 보일 수 있을 것으로 판단되며, 자질 선택 단계에서 자질 수에 대해 재검토하여 다양하게 실험을 해 볼 계획이다.

## 참고문헌


[1] 고영중, 서정연, 문서관리를 위한 자동문서 범주화에 대한 이론 및 기법, 정보관리연구, vol.33, no.2, pp.19-21, 2002

[2] 김명기, 최진순, 뉴스의 혁명 NewsML 뉴스시장의 새로운 패러다임을 연다, 박문각, 2007

[3] 김명기, 뉴스 전송 표준화 모델 개발 연구보고서, 한국언론재단, 2004

[4] 송영록, NewsML 문서검색을 위한 한국어 색인 시스템, 인천대학교 대학원 석사학위 논문, 2003

[5] 김진상, 신양규, 베이지안 학습을 이용한 문서의 자동분류, 한국데이터정보과학회지, Vol.11 No.1, 2000

[6] 조광제, 김준태, 역 카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동 분류, 정보과학회 봄 학술발표논문집, 4 권 2 호, pp.508-510, 1997

[7] 백용규, 한글 인터넷 뉴스 기사 자동 분류시스템에 관한 연구, 고려대학교 대학원 석사학위 논문, 2003

[8] 박수혁, 기계학습 기법을 이용한 문장경계인식, 고려대학교 컴퓨터정보통신대학원, 석사학위 논문, 2008

[9] 송진석, 지능형 개인화 EPG 를 위한 프로그램 정보 장르 분류, 고려대학교 컴퓨터정보통신대학원, 석사학위 논문, 2007

[10] 김상범, 범주간의 상호관계를 고려한 자동 문서 범주화의 개선, 고려대학교 대학원 석사학위 논문, 1999

[11] 안주영, NewsML 표준을 적용한 뉴스 신디케이션 시스템 설계 및 구현, 서강대학교 정보통신대학원, 석사학위 논문, 2001

[12] 박진우, 문장 중요도를 이용한 자동 문서 범주화, 서강대학교 대학원 석사학위 논문, 2001

[13] HAM (Hangul Analysis Module ), http://nlp.kookmin.ac.kr/HAM/kor/

[14] WEKA Machine Learning Software – http://www.cs.waikato.ac.nz/ml

[15] NewsML Forum, http://www.newsml.or.kr