Journal of Korea Multimedia Society Vol. 14, No. 6, June 2011(pp. 730-741)

연관규칙을 이용한 뉴스기사의 계층적 자동분류기법

주길홍[†], 신은영[††], 이주일[†††], 이원석[††††]

요 약

인터넷과 컴퓨터 기술이 발전함에 따라 정보의 양이 폭발적으로 증가하였으며 사용자의 다양한 요구가 생겨나게 되었다. 이로 인해 대용량의 문서를 효과적으로 분류하기 위한 다양한 방법의 연구가 필요하게 되었다. 기존의 문서 범주화는 문서의 분류를 위해 연관된 문서의 키워드를 중심으로 하는 방법을 사용하였다. 그러나 본 논문에서는 연관규칙을 이용하여 범주 내의 문서들 간에 연관성 있는 키워드들의 집합을 추출하고 각 범주 별로 의미적으로 대표성을 가진 키워드들로 분류 규칙을 생성한다. 또한 효율적인 키워드 생성을 위한 데이터 전처리 방안을 제시하고, 새로운 문서 범주를 예측한다. 프로파일의 분류성능을 높이기 위한 분류함수를 설계하고 실험을 통하여 성능을 측정한다. 마지막으로 평면적인 범주 구조에서 확장하여 계층적인 분류체계 구조에서도 적용할 수 있는 자동분류 방안을 제시한다.

Hierarchical Automatic Classification of News Articles based on Association Rules

Kil-Hong Joo[†], Eun-Young Shin[††], Joo-Il Lee[†††], Won-Suk Lee[††††]

ABSTRACT

With the development of the internet and computer technology, the amount of information through the internet is increasing rapidly and it is managed in document form. For this reason, the research into the method to manage for a large amount of document in an effective way is necessary. The conventional document categorization method used only the keywords of related documents for document classification.However, this paper proposed keyword extraction method of based on association rule. This method extracts a set of related keywords which are involved in document's category and classifies representative keyword by using the classification rule proposed in this paper. In addition, this paper proposed the preprocessing method for efficient keywords creation and predicted the new document's category. We can design the classifier and measure the performance throughout the experiment to increase the profile's classification performance. When predicting the category, substituting all the classification rules one by one is the major reason to decrease the process performance in a profile. Finally, this paper suggested automatically categorizing plan which can be applied to hierarchical category architecture, extended from simple category architecture.

Key words: Keyword Extraction (키워드 추출), Association Rule(연관규칙), Web Information Searching (웹 정보 검색), Clustering(클러스터링)

※ 교신저자(Corresponding Author) : 주길홍, 주소 : 경기도 안양시 만안구 경인교대길 353(430-804), 전화 : 031470-6294, FAX : 031)470-6299, E-mail : khjoo@ginue.ac.kr
접수일 : 2011년 3월 7일, 수정일 : 2011년 4월 15일
완료일 : 2011년 5월 31일
[†] 정회원, 경인교육대학교 컴퓨터교육과
[††] 정회원, YTN 정보시스템팀
(E-mail: nina97@ytn.co.kr)
[†††] 정회원, 연세대학교 컴퓨터과학과
(E-mail: tad@database.yonsei.ac.kr)
[††††] 정회원, 연세대학교 컴퓨터과학과
(E-mail: leewo@database.yonsei.ac.kr)
※ 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2011-0016648)

1. 서 론

방송산업에서 정보기술과 통신기술이 발전함에 따라 뉴스 기사 및 멀티미디어 데이터 같은 뉴스보도에 관련된 데이터를 체계적으로 저장 및 관리가 가능하게 되었고 사용자들은 인터넷, 핸드폰, 휴대용단말기 등을 이용하여 언제 어디서나 신속한 뉴스기사 서비스를 받을 수 있게 되었다. 이에 따라 국내외 언론사들은 인터넷 기사 서비스를 위한 별도의 체계를 구성·운영하여 정규 뉴스보도 외에도 사용자들이 시간과 장소에 구애 받지 않고 신속하게 뉴스 서비스를 이용할 수 있도록 지원하고 있다. 대부분의 언론사에서는 기사를 인터넷에 게시하기 전에 분류 전문가를 통해 기사를 분류하고 검증하는 단계를 거친다. 그러나 이러한 방법들은 정보시스템의 급속한 발달로 인해 처리해야 할 정보와 문서의 양이 점점 방대해지고 복잡해지는 현대 시대에 빠르게 전달해야 하는 뉴스의 속도를 저하시킬 뿐만 아니라 많은 비용을 소비하고 있다. 따라서 문서 분류의 자동화에 대한 필요성은 더욱 증대되고 있다. 본 논문에서는 단순히 문서에 나타나는 단어의 빈도를 이용하여 분류 범주를 지정하는 통계적인 분류방법과는 달리 연관규칙분석 알고리즘인 Apriori 알고리즘을 이용하여 비구조적 형태인 기사내용으로부터 각 범주에서 대표성을 지닌 키워드들을 추출하여 프로파일을 생성하고 이를 바탕으로 기사를 적합한 범주로 자동 분류하는 방안을 제시하고 유효성을 검증하는 것이 목적이다. 또한 평면적인 범주 구조에서뿐만 아니라 계층적인 분류체계 구조에서도 적용 가능한 자동분류 방안을 제시하여 보다 세부적인 정보이용을 가능하게 하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문과 관련된 문서분류, 연관규칙에 대하여 소개하고 3장에서는 빈발패턴 마이닝을 이용한 기사 자동분류의 설계에 관한 내용을 기술한다. 4장에서는 3장의 설계내용을 바탕으로 실험을 수행하고 결과를 분석하고 마지막으로 5장에서는 본 연구의 결론 및 향후 연구방향을 제시한다.

2. 관련연구

문서 자동분류는 컴퓨터를 이용하여 문서를 대표하는 특징으로 구성된 색인어 집합과 유사한 문서들을 같은 그룹으로 분류하는 기법으로 대량의 문서를 효율적으로 관리하고 검색할 수 있게 하는 동시에 방대한 양의 수작업을 감소시키는데 그 목적이 있다. 컴퓨터를 이용하여 문서를 자동으로 분류하기 위한 시도는 1970년대 말 Salton에 의해 체계화되기 시작하였다[1]. 문서의 자동분류 방법은 사전 분류기준의 유무에 따라 크게 문서 클러스터링(Document Clustering) 기법과 문서 분류(Document Classification) 기법으로 구분할 수 있다[2]. 문서 클러스터링은 사전에 정의된 분류기준 없이 문서 간 유사도에 기반하여 연관된 문서들의 집합을 생성하여, 생성된 집합을 반복적으로 합쳐나가는 방법이고 문서분류는 미리 정의된 사전 분류에 기초하여 분류대상이 되는 문서를 가장 적합한 범주에 할당하는 방법이다. 문서 클러스터링에서 문서 그룹은 문자와 단어의 관계를 정의하는 알고리즘과 둘 사이의 순환 패턴에 의해 정의되고, 단어 자체의 개념에는 크게 영향을 받지 않는다. 그러나 어떤 범주로 분류되어야 할 문서에 반드시 그 관련 단어가 많이 포함되어 있는 것은 아니기 때문에 단어나 문자에 의한 클러스터링은 실제 문서의 분야를 나타내는 데에는 한계가 있다. 반면 문서분류를 이용한 범주화에서는 각각의 범주 내의 문서간의 관계를 정의하여 새로운 문서의 분류에 적용하므로 문서의 양이 많아질수록 정확한 분류가 가능해진다.

문서 자동분류 방법은 앞서 설명한 사전분류기준의 유무에 따른 구분 법 외에 문서들이 가지고 있는 문장의 뜻을 파악하여 분류에 이용하는지 여부에 따라 다음의 두 가지로도 구분이 가능하다. 이미 분류되어 있는 문서들로부터 각 분류 범주에 나타나는 단어들의 출현 빈도에 대한 정보를 추출하여 분류에 이용하는 통계적인 방법(statistical categorization)[3-8]과, 전문가가 행하는 것처럼 문서의 내용을 기반으로 하는 분류 규칙에 따라 분류를 수행하는 지식기반 분류 방법(knowledge-based categoryzation)이 있다[3,4,9,10]. 통계 적인 분류 방법은 사람에 의해 이미 분류되어 있는 문서들로부터 각 분류 범주에 나타나는 단어들의 출현 빈도에 대한 정보를 추출하고, 분류하고자 하는 문서로부터 주요 단어들과 단어들의 출현 빈도를 추출한 뒤 이러한 정보를 이용하여 가장 적합한 범주를 찾거나 각 범주에 대하여 포함

여부를 판단하는 것으로, 많이 사용되는 통계적인 분류 방법으로는 베이즈확률(Bayesian Probability)를 이용하여 문서가 각 범주에 속할 확률을 계산하는 방법[6,8]과, 분류하려는 문서와 각 범주에 포함된 문서들간의 유사도를 계산하는 방법[5,7]등이 있다.

지식 기반 방법은 분류 대상 문서의 샘플들을 분석하여 분류 규칙들을 만들고 이러한 규칙을 이용하여 문서 분류를 수행하는 것으로, 문서의 내용에 따른 분류 규칙을 만드는 방법[3,4,9,10]과, 문서 내용 외의 정보들을 이용하는 방법[11-13]이 있다. 문서의 내용에 따른 분류 방법으로는 특정 범주로의 분류에 결정적인 단서가 되는 핵심 단어들을 추출하여 이러한 단어들의 출현 여부에 따라 분류를 수행하도록 하는 방법, 그리고 특정 범주로 분류되는 문서들에 자주 나타나는 구나 문장 형태를 패턴으로 표현하여 패턴 매칭에 의해 문서를 분류하는 방법, 문서의 내용을 파악하여 문서를 분류하는 방법이 있다. 문서 내용 외의 정보들을 이용하는 방법으로는 문서의 작성 부서와 같은 정보들을 이용하는 규칙을 만들어 문서를 분류하는 방법으로 전문가 시스템 형태로 구현될 수 있다.

일반적으로 통계적인 방법은 단어들의 출현 빈도를 기반으로 각 범주로 분류될 확률이나 각 범주와의 유사도를 계산하므로 가장 높은 값을 갖는 단일 범주로 문서를 분류하는 경우, 모든 문서를 분류할 수 있으나, 문서의 내용을 분석하는 것은 아니므로 분류의 정확도에는 한계가 있다. 지식 기반 방법은 사람에 의해 분류 대상 문서들에 대한 분석이 이루어 진 후 분류 규칙을 만들어 사용하므로 규칙에 따라 분류된 문서들의 경우 높은 정확도를 나타내지만 충분한 규칙을 제공하지 못하면 분류되지 못하는 문서들의 비율이 높을 수 있다는 단점이 있다.

연관규칙이란 동시에 발생하는 사건들을 규칙의 형태로 표현한 것으로 특정사건이 발생하면 동시에 혹은 일정한 시간 간격으로 다른 사건이 일어나는 관련성을 의미한다. 데이터베이스가 총 n개의 트랜잭션 데이터로 구성되고 전체 m개의 항목으로 구성된 상품들의 집합을 I라고 할 때 연관규칙 R은 "R : $X \Rightarrow Y$" 와 같이 표현할 수 있다. 연관규칙 R은 조건부와 결과부로 구성되며 항목집합 X와 Y에 대하여 X가 일어나면 Y도 일어난다는 연관성을 나타낸다. 여기서 $X,Y \in I$이고 $X \cap Y = \Phi$ 이어야 한다. 따라서 연관규칙을 탐사하는 것은 적절한 항목집합 X와 Y를 선택하는 문제로 볼 수 있으며 이를 위해 두 가지 척도인 지지도(support)와 신뢰도 (confidence)를 사용한다. 지지도는 데이터베이스에서 관심 있을 정도로 빈발하게 나타나는 항목을 고려하기 위한 중요한 척도로서 X와 Y를 동시에 포함하는 트랜잭션 수의 비율을 말하며 다음 식과 같이 표현된다.

$$\text{Supp}(R) = P(X \cap Y) = \frac{n(X \cap Y)}{N}$$

신뢰도는 규칙의 강도를 나타내는 척도로서 조건부 확률의 개념으로 X(조건)가 발생한다고 할 때 Y(결과)도 동시에 발생할 확률을 의미한다. 즉, 트랜잭션에 X의 항목들을 포함하는 경우 Y의 항목들도 동시에 포함할 확률을 나타내며 신뢰도가 높은 규칙일수록 의미가 크다고 할 수 있다.

$$\text{Conf}(R) = P(Y \mid X) = \frac{P(X \cap Y)}{P(X)}$$

대부분의 경우에 이러한 연관관계 중에 장바구니에 자주 나타나는 항목집합에 대해서만 관심을 가진다. 이는 항목들이 전체 트랜잭션 중 일정 비율 이상에 나타나는 항목들을 대상으로 해야 한다는 것이고 이 비율을 최소지지도(minimum support)라 하고 이 최소 지지도를 넘는 항목 집합을 빈발항목집합(Frequent Itemset)이라고 한다.

연관 관계를 찾아내는 문제는 두 가지로 나누어진다. 하나는 최소 지지도를 넘는 모든 빈발항목집합을 찾아내는 문제이고 다른 하나는 이 빈발항목에서 연관규칙을 찾아내는 문제이다. 빈발항목집합과 항목들의 지지도를 알고 있다면 연관규칙은 쉽게 찾아낼 수 있기 때문에 일반적으로 첫 번째 문제인 빈발항목집합을 찾는 문제에 대해 많은 연구가 이루어지고 있다. 빈발항목집합을 찾는 방법으로는 대표적인 알고리즘인 Apriori알고리즘은 우선 항목들의 지지도를 구해 최소 지지도를 넘는 빈발항목을 뽑아내고, 이를 바탕으로 후보항목집합(Candidate Itemset)을 생성한 후 최소 지지도를 만족하는 항목들로 다시 빈발항목을 뽑아내는 과정을 반복하여 모든 빈발 항목집합을 구한다.

3. 연관 규칙을 통한 뉴스 기사의 자동 분류

문서 자동분류를 위해서는 먼저 훈련용 문서를 수

집하여 범주별로 출현하는 키워드들을 추출한 후 이를 바탕으로 프로파일을 생성한다. 새로운 문서가 들어오면 각 범주별로 생성된 프로파일과 일치되는 정도를 계산하여 가장 높은 일치도를 가지는 범주로 문서를 분류한다.

3.1 문서 추출을 통한 키워드 집합 생성

문서추출은 뉴스기사에 대한 분류규칙 생성을 위한 문서들을 수집하는 단계이다. 수집한 문서 들을 대상으로 정제과정을 통하여 연구에서 사용할 <범주>정보와 <내용>정보 외의 속성은 모두 제거하여 순수 기사 내용만을 유지하도록 한다.그 후 빈발항목을 추출하기 위해서 각 범주 내의 문서들에서 의미있는 키워드들을 추출하여 범주별 키워드 집합을 생성한다. 기사 내용은 자연어로 이루어진 비정형 데이터로 많은 불용어를 내포하고 있어 빈발패턴을 찾기에는 매우 부적당하다. 여기서 의미하는 불용어란 문서 또는 문장에서 의미 없이 쓰이는 관사, 조사, 접속사 등을 말하는 것으로 문서분류나 정보검색에서 처리시간을 지연시키며, 정확성을 떨어드리는 요인이 된다. 이러한 불용어를 제거하여 문서별 키워드를 추출하는 과정은 전처리, 키워드추출, 후처리의 세 단계로 이루어진다.

전처리 단계는 키워드 생성 전에 문서 내용에 공통적으로 속하는 단어나 문장을 제거하는 과정으로 모든 문서에 공통적으로 포함되는 단어는 변별력이 없어 문서분류에 영향을 주지 못하고 키워드 생성시에 성능저하를 유발하게 되므로 불용어로 간주하여 처리해야 한다. 뉴스기사를 예로 들면 모든 기사 내용에 공통적으로 포함되는 <시작태그>, <종료태그>, <기자이름>, <기자메일>, <저작권표시> 등이 그것이다. 전처리를 통하여 1차 불용어를 제거한 문서내용은 형태소분석 단계를 통하여 명사만으로 구성된다. 마지막으로 후처리 단계에서는 추출된 명사들 중 의미적으로 문서의 내용을 대표할 수 없는 단어들 즉, <숫자>, <특수문자>, <1음절단어>를 2차 불용어로 분류하여 제거하고, 하나의 문서를 트랜잭션의 단위로 구성할 것이므로 문서 내에서는 중복되는 단어가 발생하지 않도록 정제한다.

범주 내에 문서의 수가 많아질수록 추출되는 키워드의 수도 많아지게 된다. 키워드 수의 증가는 프로파일 생성시 성능저하를 초래하는 원인이 되므로 빈도수가 낮고 범주간에 서로 중복되는 키워드는 프로파일 생성 과정에서 배제시켜야 한다. 본 논문에서는 범주간의 서로 중복되는 키워드들의 전처리 정도가 프로파일의 정확도에 미치는 영향을 측정하였다.

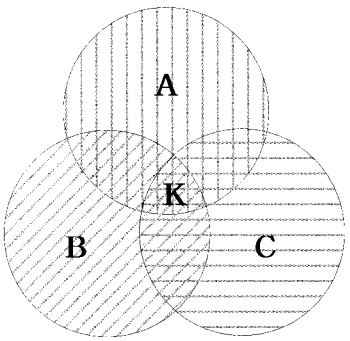


그림 1. 범주 간 키워드의 집합관계

문서별로 생성한 키워드들 중에서 범주별로 출현횟수가 일정횟수 이상을 만족하는 키워드를 범주 키워드로 선택한다. 그림 1과 같이 일정빈도 이상의 키워드들의 집합으로 이루어진 동일한 단계의 범주 A, B, C 가 있을 때, 모든 범주 안에서 동시에 나타나는 키워드 집합 K 대한 처리방법을 달리하여 다음과 같이 3개의 실험집단을 구성한다.

[그룹A] 중복 키워드 K를 모두 허용한 집단
[그룹B] 중복 키워드 K 중, 30% 미만의 낮은 비중을 차지하는 키워드를 제거한 집단
[그룹C] 중복키워드 K를 모두 제거한 집단

K중에서 문서지지도가 높은 단어들은 출현빈도가 매우 높기때문에 범주를 대표하는 키워드라고 볼 수 있다. [그룹B]는 K 중에서 문서지지도가 0.3 이상인 키워드들로 구성한 집단이다. 본 논문에서 문서지지도는 다음과 같이 정의한다.

[정의 1] 문서지지도

문서지지도 $Sc_i(w)$는 범주 c_i에서 키워드 w가 여러 문서에서 공통으로 사용되는 단어인가를 측정하기 위한 변수로 [그룹B]를 생성할 때 K에서 불용어로 처리할 단어들의 측정기준을 정할 때 사용한다. 범주 c_i에서 키워드 w를 포함하는 문서수를 DC_w, 범주 c_i의 전체 문서수를 DC_{all}이라고 할 때, 문서지지도 $Sc_i(w)$는 다음 식(1)과 같이 나타낸다.

$$Sci(w) = \frac{DC_w}{DC_{all}} \quad (1)$$

3.2 범주별 분류규칙 생성

본 논문에서 제시하는 분류규칙(프로파일) 생성은 빈발항목추출과 규칙의 특성부여의 2단계로 이루어진다. 먼저 빈발항목추출단계에서의 프로파일 생성은 문서를 장바구니, 즉 트랜잭션으로 보고 키워드를 항목으로 보아 빈발하게 같이 출현하는 모든 빈발 항목집합을 찾은 후, 이를 프로파일의 규칙으로 매핑한다. 빈발항목집합을 찾는데는 연관성 분석 알고리즘인 Apriori 알고리즘을 사용한다. 다음의 그림 2는 본 논문에서 사용한 Apriori 알고리즘이다. 첫 번째 단계에서 빈발 1 항목 집합을 결정하기 위해 단순히 항목들의 출현횟수를 센다. 다음 단계에서는 먼저 *k-1* 단계에서 찾아진 빈발항목집합 L_{k-1}로부터 후보 항목집합 (Candidate Itemset) C_k를 생성한다. 다음으로 데이터를 읽어서 C_k의 지지도를 계산한다. C_k의 항목 중 최소 지지도를 넘는 항목집합만이 L_k를 이루게 된다. 위의 단계를 더 이상 후보 항목집합을 만들 수 없을 때까지 반복하면 모든 빈발 항목 집합을 구할 수 있게 된다. 이렇게 생성된 범주별 빈발항목 집합을 원시 프로파일이라고 하며 원시 프로파일은 빈발 항목 집합의 목록, 즉 규칙의 목록으로 구성되고 하나의 규칙은 키워드들과 지지도로 구성된다. 그림 3은 원시 프로파일의 내용을 보여준다.

```
Iuput :
    D : a database of transactions.
    min_sup: the minimum support count threshold.
Output :
    L : frequent itemsets in D.
Method :
    L1 l= find_frequent-1-itemset(D);
    for (k = 2; Lk-1 ≠ ∅∅ k++){
    Ck = apriori_gen(Lk-1);
            for each transaction t ∈ D { Scan D for counts Ct = subset(Ck, t );
                                        // get the subsets of t that are candidates
                    for each candidate c ∈ Ct
                            c.count++;
            }
            Lk = {c ∈ Ck | c.count >= min_sup}
    }
    return L = ∪k Lk ;

procedure apriori_gen(Lk-1 : frequent (k-1)-itemsets)
    for each itemset I1 ∈ Lk-1
    for each itemset I2 ∈ Lk-1
    if (I1[1]=I2[1]∧(I1[2]∧...∧(I1[k-2]=I2[k-2])∧(I1[k-2]∧I2[k-1]) then {
            c = I1 ⋈ I2 // prune step: remove unfruitful candidate
            if has_infrequent_subset(c, Lk-1 ) then
                    delete c ; // prune step : remove unfruitful candidate
            else add c to Ck
    }
    return Ck ;

procedure has_infrequent_subset(c: candidate k-itemset ; Lk-1 : frequent (k-1)-itemsets )
    for each (k-1)-subset s of c ;
            if s ∉ Lk-1 then
                    return TRUE;
    return FALSE;
```

그림 2. Apriori 알고리즘

keyword1 keyword2 keyword3 ... keyword_1 (support)
keyword1 keyword2 keyword3 ... keyword_m (support)
...
keyword1 keyword2 keyword3 ... keyword_n (support)

그림 3. 원시 프로파일 내용

생성된 원시 프로파일은 규칙의 집합, 규칙 길이(규칙이 포함하는 단어 수), 지지도로 이루어져있으며 임의의 규칙 $rule_i (1 \le i \le n)$는 다음의 속성값으로 표현된다.

$$rule_i = (K_{rule_i}, length_{rule_i}, sup_{rule_i})$$

$length_{rule_i}$: 규칙 길이

sup_{rule_i} : 지지도

기본속성인 규칙길이와 지지도만으로는 범주별 규칙의 특성을 나타내는데 충분하지 않기 때문에 기본속성 외에 규칙 스코어와 범주별 규칙 가중치를 추가하여 분류별 프로파일의 일치도를 향상시키는데 이용하였다. 아래는 속성이 추가된 규칙의 최종적인 형태이다.

$$rule_i = (K_{rule_i}, length_{rule_i}, sup_{rule_i}, score_{rule_i}, weight_{rule_i})$$

$length_{rule_i}$: 규칙 길이, sup_{rule_i} : 지지도.

$score_{rule_i}$: 규칙스코어, $weight_{rule_i}$: 규칙가중치

[정의 2] 규칙 스코어(Rule Score)

규칙 스코어(Rule Score)는 규칙의 길이와 지지도의 차이에 따라 일치율에 차등을 주기 위한 개념으로 규칙 길이가 길거나 지지도가 높은 규칙은 그렇지 않은 규칙들에 비해 더 높은 점수를 받게 된다. 규칙 스코어는 프로파일 생성시, 규칙의 속성으로써 함께 결정되며 다음 식(2)와 같이 표현된다.

$$Rule\ Score = Rule\ Length \times Rule\ Support \quad (2)$$

[정의 3] 범주별 규칙 가중치(Category Rule Weight)

특정 범주에서만 출현하는 것이 아니라 여러 범주에 걸쳐 출현하는 규칙들은 해당 범주를 대표하는 규칙이라기 보다는 범용적인 규칙일 가능성이 높다. 따라서 여러 범주에 나타난 규칙일수록 낮은 가중치를 부여하는 것이 타당하다. 범주별 규칙 가중치는 아래 식에 따라서 부여한다. $TotalCategoryCount$는 전체 범주의 수이고, RC는 규칙이 출현한 범주의 수이다. 아래 식(3)에 의해 상대적으로 적은 범주에 출현한 규칙일수록 높은 가중치를 가지게 된다.

$$Category\ Rule\ Weight = \frac{Total\ Category\ Count}{RC} \quad (3)$$

3.3 프로파일의 생성 및 적용

문서분류를 위해서는 분류대상 문서들에 범주별 프로파일을 적용하여 일치되는 규칙이 존재하는지 여부를 판정하고, 범주별로 일치점수를 계산하여 가장 높은 점수를 얻은 범주로 문서를 분류한다. 프로파일의 하나의 규칙에 속하는 모든 키워드들이 문서내에 모두 속할 때 이 규칙은 문서를 완전일치(Complete match)한다고 정의하고, 문서의 키워드가 2개 이상의 빈발항목으로 구성된 규칙과 일치되면 프로파일을 만족하는 것으로 판단한다. 하나의 문서가 여러 범주의 프로파일에 일치될 때에는 일치된 프로파일의 중요도를 고려하여 더 높은 중요도를 갖는 프로파일의 범주로 문서를 분류할 수 있다. 이 중요도를 계산할 때 단순히 완전 일치된 규칙의 수가 많은 프로파일의 중요도가 높다고 판단할 수는 없기 때문에 규칙의 특성을 고려하여 일치점수에 가중치를 가감해야 한다. 본 연구에서는 기사의 일치도를 평가하기 위해 일치율, 규칙 스코어 비율, 규칙 가중치 비율의 3가지를 기준항목을 정의한다.

[정의 4] 규칙 일치율(Matched Rule Ratio)

규칙 일치율은 기사와 완전일치되는 규칙이 프로파일의 전체 규칙 중에 차지하는 비율을 나타낸다. 프로파일 안에서 기사와 완전일치되는 규칙의 개수를 RC_{match}, 전체 규칙 개수를 RC_{all} 이라고 하면 규칙 일치율은 다음 식(4)와 같다.

$$Matched\ Rule\ Ratio = \frac{RC_{match}}{RC_{all}} \quad (4)$$

[정의 5] 규칙 스코어 비율(Matched Rule Score Ratio)

규칙 스코어 비율은 기사와 완전일치되는 규칙들의 규칙 스코어의 합을 프로파일 내의 전체 규칙 스코어의 합으로 나눈 값으로 규칙 길이와 지지도에 따라 일치율에 차등을 주기위한 비율이다. $rule_{all}$을 범주 c_i의 프로파일의 전체 규칙의 집합, $rule_m$을 범주 c_i의 프로파일에서 완전일치되는 규칙의 집합,

$scr(c_i,k)$을 범주 i 내의 규칙 k의 규칙 스코어라고 할 때 규칙 스코어비율은 다음 식(5)와 같다.

$$MatchedRuleScoreRatio = \frac{\sum_{k \in rule_m} scr(c_i,k)}{\sum_{k \in rule_{all}} scr(c_i,k)} \quad (5)$$

[정의 6] 가중치를 부여한 규칙 스코어 비율 (Weighted Rule Score Ratio)

가중치를 부여한 규칙 스코어 비율은 식(5)의 규칙 스코어 비율에 범주별 규칙 가중치(Category Rule Weight)를 적용하여 범주별로 대표성을 띄는 규칙들이 일치 시에 더 높은 점수를 부여받을 수 있도록 하기 위한 기준이다. 완전 일치되는 규칙의 규칙 스코어를 $scr(c_i,k)$, 분류별 규칙 가중치를 $wgt(c_i,k)$라고 할 때 이 둘을 곱한 값을 전체 규칙 스코어의 합으로 나눈 값이다.

$$WeightedRuleScoreRatio = \frac{\sum_{k \in rule_m} (scr(c_i,k) \times wgt(c_i,k))}{\sum_{k \in rule_{all}} scr(c_i,k)} \quad (6)$$

범주별로 생성된 프로파일은 각 범주의 특성에 따라 규칙의 최대 길이 및 총 개수가 서로 다르게 되므로 문서를 분류 할 때 어떤 기준으로 프로파일을 적용하여야 할지 모호하게 된다. 일치도 계산시, 프로파일 안에 존재하는 1부터 n까지의 모든 규칙들을 일일이 적용시켜보는 방법은 매우 비효율적이다. 분류규칙 내의 규칙의 숫자가 많아질수록 매칭되는 단어수가 많아지므로 정확도는 높아지지만 비교해야 할 규칙의 개수가 증가하므로 처리성능은 떨어지게 된다. 그러므로 정확성이 높으면서 처리성능을 높일 수 있는 일정 범위 내에 존재하는 규칙들을 찾아낼 수 있는 비교기준이 필요하다. 본 연구에서는 휴리스틱한 접근법으로 규칙 길이별로 규칙들을 그룹화하여 길이별 규칙그룹을 만든 후, 길이별 규칙그룹에 해당되는 규칙의 수를 프로파일 내의 전체 규칙 개수로 나누어 길이별 규칙그룹이 전체 중 차지하는 비율을 정의하여 그 값을 실험에 활용하기로 한다. 아래 식(7)은 특정 규칙 길이 n인 그룹이 전체 프로파일 중에 차지하는 비중을 나타낸다.

$$\text{룰길이}n\text{인그룹의비율} = \frac{\text{길이}n\text{인그룹에속하는룰의개수}}{\text{프로파일내의총룰개수}} \quad (7)$$

본 논문에서는 위에서 제시한 프로파일들을 계층적 범주 형태에서도 사용할 수 있다. 계층형 범주는 분류체계가 계층적인 트리형태의 구조로 구성된 것으로 다음의 그림 4와 같은 구조를 갖는다. 계층형 범주는 평면적인 분류체계에 비해 하위 범주로 갈수록 세부적인 정보를 포함하기 때문에 보다 정확한 정보 검색이 가능하도록 해준다. 본 논문에서는 분류체계가 평면적이지 않고 계층적인 경우에 대하여도 문서의 분류 방법을 정의한다.

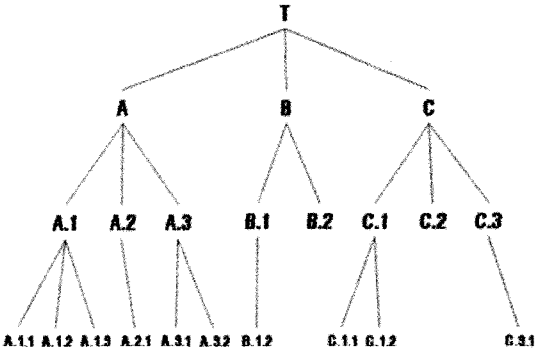


그림 4. 계층형 범주 구조

계층형 범주 구조 상에서의 문서분류방법은 평면적인 분류체계에서의 방법과 거의 동일하나 프로파일 생성과 적용 시의 처리부분에서 차이가 있다. 상위 범주에서 불용어로 간주된 단어들은 하위 범주에서도 분류의 단서가 될 수 없으므로 핵심단어에서 제외되어야 한다. 그러므로 프로파일 생성 전에 상위 범주의 불용어를 이용하여 전처리를 해주어야 한다. 위의 그림 4에서 단계 2(A.3)와 단계 3(A.3.2)의 프로파일을 생성한다고 하면, A에서 불용어로 처리된 단어를 A.3의 프로파일을 생성하기 전에 불용어로 간주하여 제거하고, 마찬가지로 A.3.2의 프로파일을 생성하기 전에는 A.3에서 불용어로 분류된 단어를 제거한 뒤 프로파일을 생성한다.

프로파일 적용 시에는 상위 범주부터 시작하여 각 단계별로 일치도를 계산하여 가장 일치도가 높은 경로를 선택해 나가는 과정을 최하위 범주까지 반복한다. 그림 4에서 예를 들면 분류하고자 하는 문서를 처음 단계1의 A, B, C 의 프로파일에 적용할 때 일치도를 비교해 본 결과 문서가 A로 분류되었다면 다음 단계인 A.1, A.2, A.3의 프로파일을 이용하여 일치도를 계산한다. 단계 2의 분류결과가 A.3로 결정되었다

면 마지막으로 단계3인 A.3.1, A.3.2의 프로파일을 적용시켜 일치되는 범주를 선택한다. 단계3에서 A.3.2로 분류되었다면 이 문서의 분류결과는 A - A.3 - A.3.2가 된다.

4. 실험 결과 및 분석

실험은 2009년 한 해 동안 연합통신사에서 작성된 기사 275,996건을 대상으로 실시하였다. XML형식으로 구성된 연합기사를 별도로 개발한 프로그램으로 파싱하여 본 연구에서 필요한 정보인 <기사ID>, <범주>, <기사내용>에 해당되는 정보를 추출하여 실험 데이터베이스를 구축하였다. 연합기사의 범주는 다음의 표 1에서 보여지는 것과 같이 대·중·소분류 3개의 계층형 구조를 제공하고 있으며 대분류 12개, 중분류 88개, 소분류 412개 구성되어 있다. 이 중 실험을 위해 대분류는 5개, 중분류는 대분류당 각 5개로 지정하고 소분류는 훈련 샘플문서의 수가 규칙 형성에 미치는 영향을 비교하기 위하여 2개로 구성된 집단과 4개로 구성된 집단의 두 그룹으로 지정하였다. 분류 실험의 정확성을 위해서 실험데이터의 선별에 연구자의 주관적인 의견이 개입되지 않고 특정 기간이나 사건에 영향을 받지 않아야 하며 범주간의 실험문서 수도 동일하도록 하였다. 이를 위해서 본 연구에서는 전체 문서집단에서 각 범주별로 무작위로 일정 개수의 문서를 선별하되 추출된 문서간에 중복되는 문서가 없도록 실험문서를 추출하였다.

표 1. 기사 범주별 건수

대 분 류	중 분 류	소 분 류
정치	9	43
경제	5	28
사회	3	19
금융,증권	1	54
산업	1	44
사건사고	7	31
문화	1	48
생활건강	5	23
IT, 과학	3	13
북한	8	24
국제	2	14
스포츠	1	71

계층형 범주 구조를 가진 분류체계에서는 대분류를 기준으로 실험문서를 추출하게 되면 서브범주인 중·소분류별로 문서의 건수가 일정하지 않게 되어 정확한 실험을 수행할 수 없다. 그러므로 미리 실험에 이용할 범주의 수를 지정해놓고 거꾸로 가장 하위 단계의 범주에서부터 문서를 추출하여 상위 범주 간에 실험문서의 건수를 동일하게 설정되도록 하는 방식을 사용하여 최상위 범주에서의 총합을 같도록 만들었다. 대분류 범주의 문서가 각 1000건, 중분류가 각 200건이 되도록 소분류에서부터 샘플문서를 무작위로 추출하여 각 분류별로 샘플 문서 개수를 동일하게 설정되도록 하는 방식을 사용하여 최종적으로 각 대분류에서의 총합을 같도록 하였다.

범주별로 생성된 문서집합은 프로파일을 생성하기 위한 실험용 문서집단과 프로파일 검증을 위한 검증용 문서집단으로 구분해야 한다. 본 실험에서는 이를 위해 교차검증법의 하나인 k-묶음 교차 검증법(k-fold Cross Validation)을 사용하였다. 추출한 5000건의 샘플문서를 5개의 집단으로 나눈 후, 4개는 실험용으로 1개는 검증용으로 사용하였다. 기사 및 생성된 프로파일을 저장하기 위한 데이터베이스는 Oracle 10g Standard Edition을 사용하였고 한글에 대한 전처리 작업은 국민대학교 자연어처리 연구실의 형태소분석기 KLT 2008-Test version 를 이용하였다. 본 실험에서는 문서 분류의 결과에 대한 성능 평가 척도로 분류율을 사용한다. 분류율은 검증에 사용한 전체 문서 중에서 정확하게 분류된 문서가 차지하는 비율을 말하며 다음 식으로 나타낸다.

$$\text{분류율} = \frac{\text{정확하게분류된문서수}}{\text{검증에사용한전체문서수}} \times 100$$

일치함수의 성능을 비교하기 위하여 대분류 범주에서 최소지지도 0.5를 기준으로 규칙 일치율, 규칙 스코어 비율, 가중치를 부여한 규칙 스코어 비율의 세 가지 함수의 분류율을 측정하였다. 그 결과, 그림 5에서와 같이 규칙 일치율은 86.9%, 규칙 스코어 비율은 87.7%, 가중치를 부여한 규칙 스코어 비율은 91.8% 로 가중치를 부여한 규칙 스코어 비율이 좋은 분류 성능을 나타냈다.

프로파일 생성 전에 모든 범주에 동시에 출현하는 중복 키워드 K를 처리하는 방식이 연관단어 생성에 미치는 영향을 분석해보기 위하여 실험집단을 A, B, C의 세 그룹으로 나누어 분류성능을 측정하였다. 최

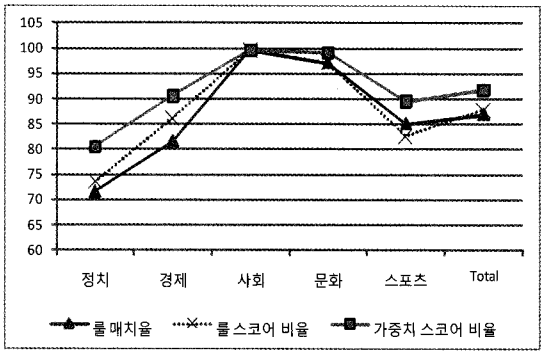


그림 5. 일치함수 성능비교

소지지도는 0.5, 일치함수는 가중치를 부여한 규칙 스코어 비율을 사용하였다.

[그룹A] 중복 키워드 K를 모두 허용한 그룹
[그룹B] 중복 키워드 K 중, 30%미만의 낮은 비중을 차지하는 키워드를 제거한 그룹
[그룹C] 중복키워드 K를 모두 제거한 그룹

다음의 그림 6은 중복키워드 처리에 따른 일치율을 비교한 실험으로 중복키워드를 모두 허용한 집단인 [그룹A]의 분류율이 65.2%로 가장 낮았고 중복을 모두 제거한 [그룹C]는 66.3%로 [그룹A]와 큰 차이가 없었으나 중복키워드 중 대표성이 없는 키워드를 정제한 [그룹B]의 경우는 91.8%로 [그룹A]보다 26.6% 향상되었다.

프로파일 내에서 분류규칙을 선택하는 기준을 정하기 위하여 규칙 길이별로 규칙들을 그룹화하여 길이별 규칙그룹을 만든 후, 길이별 규칙그룹에 해당되는 규칙 그룹이 전체 중 차지하는 비율을 정의하여 각 비율별로 분류실험을 수행하였다. 다음 표 2는 규칙 길이별 비중에 따른 규칙 개수를 나타낸 것으로 규칙 길이별 비중이 증가함에 따라 선택되는 규칙의 개수는 감소하였다.

표 2. 길이별 비중에 따른 규칙 개수 (단위: 천개)

	전체	5%	10%	15%	20%
규칙 개수	1987	1883	1748	1745	1037

그림 7은 실험결과를 나타낸 그래프로 각 범주별 특성에 따라 높은 분류성능을 보이는 규칙 길이별 비중에 차이를 보였다. 그러나 전체적으로 보면 모든 규칙을 사용했을 때, 5%, 15%, 10%, 20% 순으로 선택한 규칙의 수가 줄어드는 것과 비례하여 분류성능이 감소하는 경향을 보였다. 그러나 15%이상의 비중을 차지하는 규칙들을 선택했을 때(84.1%)는 10% 일 때(80.6%)보다 3.5%만큼 오히려 높아졌다.

그림 8은 최소지지도 설정에 따른 일치율을 비교한 실험으로 최소지지도 설정을 0.5에서부터 5까지 0.5 간격으로 변경하면서 범주별 일치율을 측정하였다. 그 결과 규칙의 수가 가장 많은 0.5에서 가장 좋은

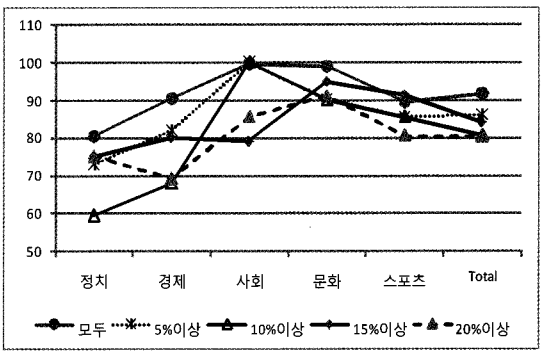


그림 7. 분류규칙 선택에 따른 일치율 비교

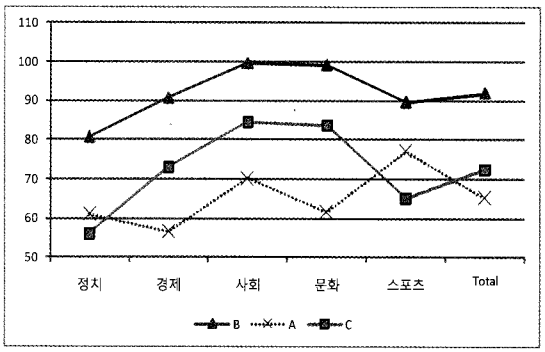


그림 6. 중복 키워드 처리에 따른 일치율 비교

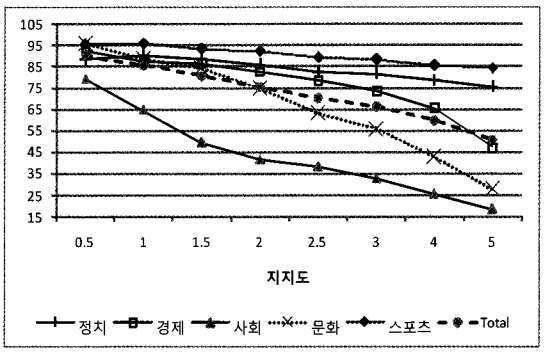


그림 8. 지지도 설정에 따른 일치율 비교

일치율을 보였고 최소지지도가 증가함에 따라 일치율은 낮아졌다. 이는 지지도가 증가함에 따라 프로파일 내의 규칙의 개수가 줄어들게 되어 새로운 기사와 일치되는 단어를 찾을 수 없는 경우가 많아지기 때문으로 타 범주에 비해 '문화'와 '스포츠'에서 편차가 더욱 심하게 나타났다.

아래의 그림 9부터 그림 11까지는 계층형 범주에서의 일치율을 비교한 실험으로 3개의 계층구조로 이루어진 범주를 대상으로 일치율을 측정하였다. '정치' 범주의 경우, 단계2에서 규칙 길이별 비중이 30% 이상인 분류규칙을 대상으로 분류실험을 진행한 결과, 단계1에서 정확하게 분류된 기사 80.5% 중 단계2 범주로 정확하게 분류된 기사는 80.74% 였다.

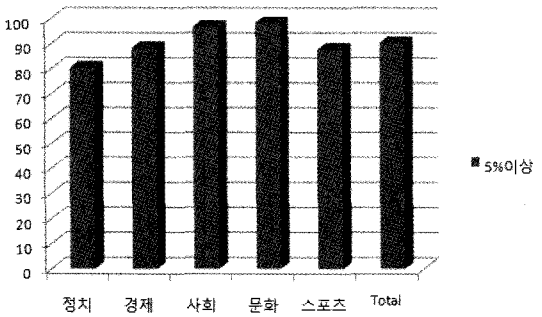
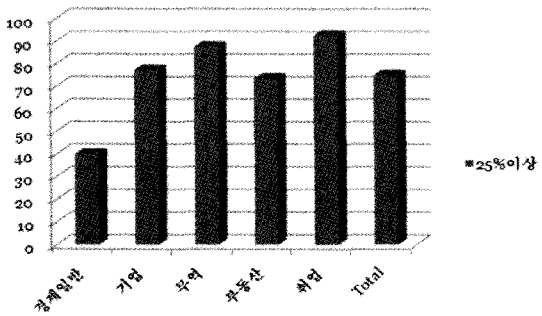
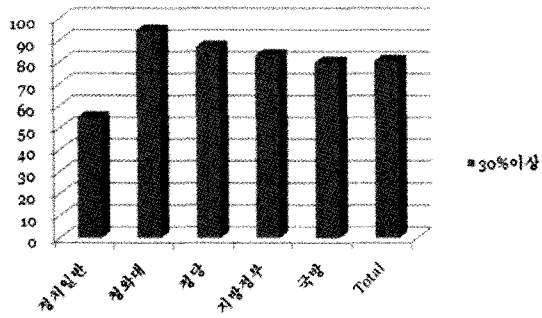


그림 9. 단계 1에서의 일치율 비교

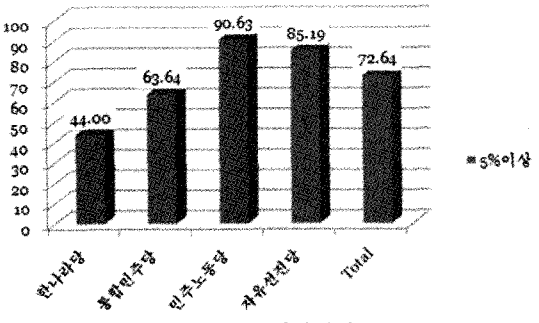


(a) 정치

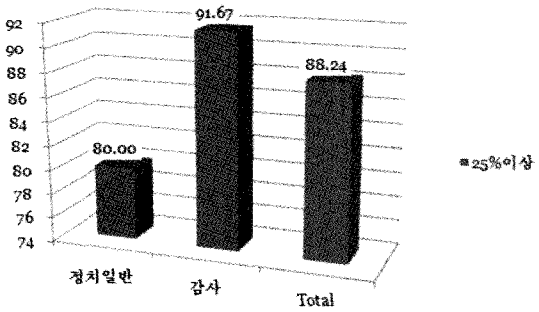


(b) 경제

그림 10. 단계 2에서의 일치율 비교



(a) 정치-정치일반



(b) 정치-정당

그림 11. 단계 3에서의 일치율 비교

단계3인 '정치-정치일반'과 '정치-정당'은 규칙 길이별 비중을 25%와 5%로 기준으로 하여 실험하였으며, 각각 88.24%, 72.64% 의 일치결과를 얻었다. '정치-정치일반'은 2개의 하위 범주로 '정치-정당'은 4개의 하위 범주로 구성되어있기 때문에 단계3에서 실험데이터 추출 시 '정치-정치일반'의 경우 각 100개의 학습용 기사를 이용하였고 '정치-정당'의 경우 각 50개의 학습용 기사를 이용하였다. 실험결과 학습문서를 적게 추출한 범주에서의 일치율이 낮게 나타남을 알 수 있다.

5. 결 론

본 논문은 통계적인 방법이 아닌 데이터마이닝 기법 중에 하나인 연관규칙분석을 이용하여 분류규칙을 생성하였다. 비구조적 형태인 문서 내용으로부터 의미적으로 빈번하게 사용되고 동시에 출현하는 단어들을 범주의 핵심 키워드로 선정하여 이를 새로운 문서의 분류예측에 사용함으로써 분류의 정확도를 높일 수 있었다. 이때 중복 키워드를 모두 제거하거

나 모두 허용한 경우에 비해서 범주 내에서 발생빈도가 낮은 중복 키워드를 제거한 경우에 프로파일의 정확도가 향상되었다. 원시 프로파일의 기본속성인 규칙 길이, 지지도에 스코어와 가중치의 두 가지 속성을 추가하여 일치점수 계산시 차등을 주는데 이용하였으며, 각 일치함수의 성능을 측정한 실험에서 기본 속성만을 이용한 함수인 규칙 일치율에 비해 스코어와 가중치 속성을 적용한 함수인 규칙 스코어 비율과 가중치를 부여한 규칙 스코어 비율에서 좋은 분류성능을 보였다. 또한 최소지지도의 설정을 변경하면서 범주별 일치율을 측정하여 최소지지도가 증가함에 따라 일치율은 낮아지는 현상을 보였는데 이는 지지도가 증가함에 따라 프로파일 내의 규칙의 개수가 줄어들게 되어 새로운 기사와 일치되는 단어를 찾을 수 없기 때문이다.

계층형 범주에서 자동분류를 적용하기 위하여 하위 범주에서 프로파일을 생성하기 전에 부모 범주에서 불용어로 처리된 단어들은 하위 범주에서도 불용어로 분류하여 제거하였다. 새로운 기사가 들어왔을 때 상위 범주에서부터 각 범주 단계별로 일치도를 계산하여 가장 일치도가 높은 경로를 선택해 나가는 과정을 최하위 범주에 이를 때까지 반복 수행하여 계층형 구조에서도 각 범주를 대표하는 키워드를 생성할 수 있었으며, 부모 범주의 불용어를 자식 범주의 전처리에 사용함으로써 자식 범주의 프로파일의 크기를 줄여 분류성능을 높이고 프로파일 생성시간을 단축시킬 수 있었다.

참 고 문 헌

[1] 윤종찬, 윤성대, "스퀀스 연관규칙을 이용한 개인화 웹 마이닝 설계," 한국멀티미디어학회논문지, 제11권, 제11호, pp.1566-1574, 2008.

[2] 이형우, 김태수, "온톨로지 기반에서 연관 마이닝 방법을 이용한 지식 추론 알고리즘 연구," 한국멀티미디어학회논문지, 제11권, 제11호, pp.1601-1614, 2008.

[3] P. Hayes, P. Anderson, I. Nirenburg, and L. Schmandt. "TCS: A Shell for Content-based Text Categorization," Proceedings of the 6th IEEE Conference on Artificial Intelligence.

[4] J. R. Hobbs., D. Appelt, M. Tyson, J. Bear and D. Israel, "FASTUS: System summary," Proceedings of Fourth Message Understanding Conference, 1992.

[5] L. Larkey. and W. Croft, "Combining Classifiers in Text Categorization," SIGIR'96, 1996.

[6] D. Lewis. "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task," SIGIR'92.

[7] B. Masand., "Classifying News Stories using Memory Based Reasoning," SIGIR'92.

[8] M. Maron, "Automatic Indexing: An Experimental Inquiry," Journal of the ACM, 1961.

[9] R. Hoch., "Using IR Techniques for Text Classification in Document Analysis," SIGIR'94, 1994.

[10] P. Jacobs., Using statistical methods to improve knowledge-based news.

[11] M. Blosseville. G. Hebrail, M. Monteil, and N. Penot., "Automatic Document Classification: Natural Language Processing, Statistical Analysis, and Expert System Techniques used Together," SIGIR'92, 1992.

[12] 김국희. "웹 기반 문서 자동분류시스템 설계 및 성능실험," 국방대 국방관리 대학원, 2005.

[13] 명진. "인공지능을 이용한 웹 문서의 자동분류," 서강대학교 경영대학원 석사학위 논문, 2004

[14] 황성하. "인터넷 문서의 자동분류 서비스 시스템에 관한 구현," 한국 컨텐츠학회 추계종합학술대회 논문집 제3권, 2005.

[15] 한정기. "구문 패턴과 키워드 집합을 이용한 통계적 자동 문서 분류의 성능 향상," 한국정보처리학회 학술대회 논문집, 2000.

[16] 박흠, "문서 자동분류에서 자질의 대표성 향상을 위한 자질 축소와 자질 필터링 방법," 부산대학교 정보통신대학원 박사학위 논문, 2008.

[17] 하원식, "협력적 필터링을 위해 연관단어 빈도를 이용한 웹 문서 분류," 한국정보과학회 학술대회 논문집 Vol.31, No.2, 2004.

[18] 김홍남, "가중치가 부여된 단어 연관 규칙 기반의 문서 분류," 인하대 대학원 석사학위 논문, 2004.

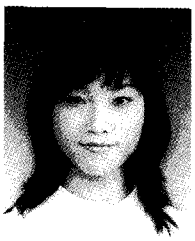
[19] 백용규, "인터넷 뉴스기사에 대한 자동 분류 정

보 시스템에 관한 연구," 한국경영정보학회 학술대회 논문집, 2003.

주 길 홍

2000년 2월 연세대학교 컴퓨터과학과(공학석사)
2004년 8월 연세대학교 컴퓨터과학과(공학박사)
2005년 3월~현재 경인교육대학교 컴퓨터교육과 부교수

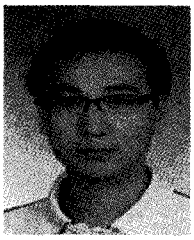
관심분야 : 분산 데이터베이스 시스템, 데이터마이닝, 스마트 러닝, ICT 활용교육



신 은 영

2000년 2월 신구대학 방사선과 졸업
2005년 2월 방송통신대학교 컴퓨터과학과 졸업(학사)
2010년 8월 연세대학교 공학대학원 컴퓨터공학과 졸업(석사)

2001년~2004년 씨아이테크놀로지 근무
2004년~현재 YTN 정보시스템팀 근무
관심분야 : 정보검색, HCI, 데이터마이닝



이 주 일

2006년 2월 홍익대학교 컴퓨터공학과(공학석사)
2009년 3월~현재 연세대학교 컴퓨터과학과 박사과정
관심분야 : 데이터베이스 시스템, 데이터스트림 마이닝, Context Awareness

이 원 석

1985년 7월 Boston University 컴퓨터과학과(공학사)
1987년 7월 purdue University 컴퓨터과학과(공학석사)
1990년 7월 purdue University 컴퓨터과학과(공학박사)

2004년 3월~현재 연세대학교 컴퓨터과학과 교수
관심분야 : Sensor Data Stream Processing, Data Stream Mining, OLAP, Data Warehouse