한글 텍스트 문서 분류시스템을 위한 속성선택
Feature Selection for a Hangul Text Document Classification System

Jae Sik Lee[a] and You Jung Cho[b]

[a]아주대학교 경영학부
수원시 팔달구 원천동 산5, 442-749,
Tel: +82-31-219-2719, E-mail: leejsk@ajou.ac.kr,

[b]아주대학교 일반대학원 경영학과 박사과정, E-mail: mispower@ajou.ac.kr

Abstract

정보 추출(Information Retrieval) 시스템은 거대한 양의 정보들 가운데 필요한 정보의 적절한 탐색을 도와주기 위한 도구이다. 이는 사용자가 요구하는 정보를 보다 정확하고 보다 효과적이면서 보다 효율적으로 전달해주어야만 한다. 그러기 위해서는 문서내의 무수히 많은 속성들 가운데 해당 문서의 특성을 잘 반영하는 속성만을 선별해서 적절히 활용하는 것이 절실히 요구된다. 이에 본 연구는 기존의 한글 문서 분류시스템(CB_TFIDF)[1]의 정확도와 신속성 두 가지 측면의 성능향상에 초점을 두고 있다. 기존의 영문 텍스트 문서 분류시스템에 적용되었던 다양한 속성선택 기법들 가운데 잘 알려진 세가지 즉, Information Gain, Odds Ratio, Document Frequency Thresholding을 통해 선별적인 사례베이스를 구성한 다음에 한글 텍스트 문서 분류시스템에 적용시켜서 성능을 비교 평가한 후, 한글 문서 분류시스템에 가장 적절한 속성선택 기법과 속성선택에 대한 가이드라인을 제시하고자 한다.

Keywords:
Information Retrieval, Feature Selection, Document Classification, Information Gain, Odds Ratio

1. 서론

정보의 홍수 시대를 살아가고 있는 우리는 정보과부하 문제에 직면해서 정보 활용의 비효율성과 관련한 많은 어려움을 겪고 있다.

하지만 정보의 양이 급격하게 증가하는 것에 더불어서 정보 이용자들이 보다 편리하게 필요한 정보를 보다 효율적으로 활용할 수 있게 하기 위한 연구도 최근 들어 더욱 활발하게 지속되고 있다. 대표적인 연구로 자동화된 텍스트 문서 분류시스템의 개발과 관련된 것들을 들 수 있겠다. 이 시스템은 사용자가 원하는 정보를 추출하여 사용자의 정보 활용에 대한 욕구를 충족시켜주는데 일조를 하고 있다. 하지만 단순한 정보추출만으로 사용자의 정보 활용 욕구에 대한 만족을 추구하기는 어렵다. 이는 다시 말하면, 원하는 문서를 얼마나 정확하고 빠르게 추출하느냐 즉, 문서 추출에 대한 효율성과 효과성의 문제가 보다 중요해지고 있다고 할 수 있다.

이와 관련해서 본 연구에서는 기존에 텍스트 문서 분류의 성능향상을 위해 연구된 몇 가지의 속성선택 방법들을 비교 평가해 보고자 한다. 텍스트 문서 분류 시스템이 가지는 가장 고질적인 문제 중에 하나가 고

차원의 속성차원을 가진다는 것이다. 이는 시스템의 성능을 저하시키는 주된 요인이 된다. 따라서 본 연구의 가장 중심적인 초점은 속성차원의 축소, 즉 속성선택에 따른 성능 개선에 있다고 하겠다. 또한 기존의 텍스트 문서를 위한 속성선택은 영문 텍스트문서를 위주로 진행이 되었다. 하지만 본 연구에서는 한글 텍스트 문서를 기반으로 해서 한글 텍스트 문서의 분류성능을 향상시키기 위한 속성선택에 관한 연구를 수행함을 목표로 한다. 기존의 연구들이 영문 텍스트 문서들을 중심으로 진행된 것을 감안하면 한글 텍스트 문서 분류시스템에 적절한 속성선택기법은 영문 텍스트 문서 분류시스템과는 차이를 보일 것으로 기대된다. 이에 본 연구에서는 한글 텍스트 문서 분류시스템에 적절한 속성선택 기법을 발견 및 제시하고, 해당 속성선택 기법을 보다 효율적으로 활용할 수 있는 전략적 지침을 제시하고자 한다.

본 논문은 다음과 같이 구성되어 있다. 제 2절에서는 기존의 관련 연구에 대해 간략한 검토를 하고, 제 3절에서는 텍스트 문서에 적용될 수 있는 속성선택 기법들 가운데 본 연구에 적용될 기법에 대한 검토를 한다. 제 4절에서는 구체적으로 한글 텍스트 문서 분류시스템에 속성선택 기법을 적용시킨 결과에 대한 요약정리 및 비교 분석을 하고, 전략적 활용방안에 대한 제언을 한다. 본 페이퍼에서는 실험이 완료되지 않은 관계로 해당 프로세스에 대한 소개와 기대 결과에 대한 간략한 소개로 대신한다.

2. 본 연구와 관련된 기존연구

텍스트 문서 분류시스템의 구축 및 성능 향상에 대한 연구는 1980년대 이후로 끊이지 않고 지속되고 있다. Dunja *et al.*[4]은 계층적 구조를 가진 온라인 문서들을 대상으로 속성선택을 하는 연구를 하였다. Dunja *et al.* 역시 자동화된 문서 분류시스템을 구축하고자 할 때 가장 큰 문제가 되는 것이 속성들에 대한 차원이 고차원이라는 것을 공감하고 있다. 이에 속성들의 차원을 축소시켜서 자동화된 문서 분류시스템을 보다 실용적으로 활용할 수 있는 방안에 대한 고찰을 하는 것을 연구목적으로 삼고 있다. 총 11개 즉, Information Gain, Cross Entropy, Mutual Information, Weight of Evidence, Odds Ratio, Variants of Odds Ratio(Weighted Odds Ratio, Log-Probability Ratio, Exp-probability Difference, Conditional Odds Ratio), Random Selection 등의 속성선택 지표들을 고안하고 이에 대한 검증을 하였다. 이미 잘 알려진 속성선택 지표들이 있는 반면 "Variants of Odds Ratio"에 제시된 것들은 연구자가 직접 고안한 것들이다. 또한 속성 차원을 축소시키는데 Stop-list를 사용할 뿐만 아니라 기본적으로 빈도가 낮은 단어들은 제거시켰고, 하나의 단어들만을 중심으로 한 것이 아니라 단어의 순서에 따른 결합도 고려를 하였다. 더불어 단어의 순서를 생성하는데 있어서 효율적인 접근법이 무엇인지에 대한 제시도 이루어지고 있다. 해당 연구에 대한 실증적인 검증은 Yahoo! Hierarchy 가운에 일부 Category를 중심으로 이루어졌는데 학습 기법으로는 Naive Bayesian Classifier를 사용하였다. 실제 데이터를 적용해 본 결과, 새롭게 생성한 속성선택 지표에 대한 성능도 만족할 만한 수준이기는 하였으나, 가장 좋은 성과를 제시해 준 것은 Odds Ratio였다.

Yiming *et al.*[5]도 역시 적극적으로 속성차원 축소에 대한 연구를 수행하였다. 통계적 학습 기법에 속성선택 지표들을 적용하여 해당 성능에 대해 비교평가를 하였다. 위의 연구[4]가 구조적인 온라인 문서에 기반을 둔 반면 이 Yiming *et al.*의 연구는 오프라인 문서에 기반하고 있다. Document Frequency Thresholding, Information Gain, Mutual Information, χ^2-test, Term Strength의 지표를 AT&T Lab에서 제공하는 Reuters-22173 데이터집합과 OHSUMED 데이터 집합에 적용시켜서 검증을 하였다. 해당 실험에서 가장 좋은 성능을 보인 지표는 Information Gain과 χ^2-test였다. k-nearest neighbor classifier를 기반으로 Information Gain을 적용하여 실험한 결과, 유일한(unique) 단어를 98%까지 제거하였음에도 분류 성능은 더 향상되는 결과를 가져왔다. Document Frequency도 비슷한 결과를 가져왔는데, 실제로 단어에 대한 Information Gain, Document Frequency, χ^2-test 사이에 강한 상관관계가 있음도 발견되었다. 수행 속도 측면에서는 Document Frequency가 가장 우수한 성능을 보였다. 고로, Information Gain이나 χ^2-test을 사용해서 수행 속도가 많이 느리다면 적극적으로 Document

Frequency 속성선택 지표를 사용할 것을 권하고 있다. 또한 가장 저조한 성능을 보인 지표는 Mutual Information이었다.

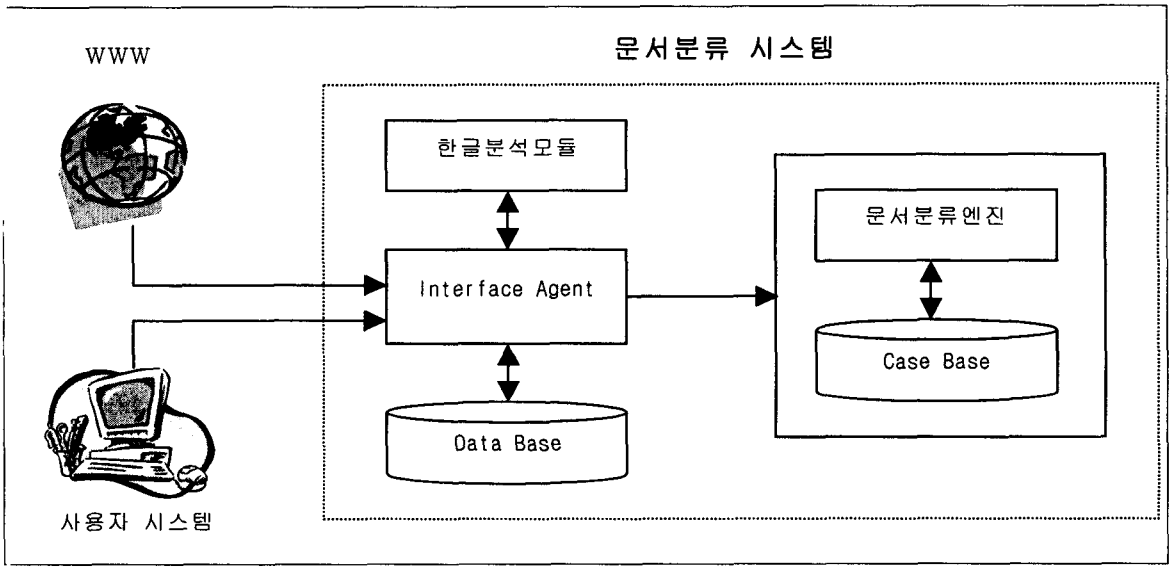
이지식과 이종운[1]의 연구는 본 연구의 모태가 되는 연구이다. 직접적으로 한글 텍스트 문서를 가지고 효율적이면서 효과적인 분류시스템을 Case-based Reasoning으로 구현하였다. 전통적인 문서 분류기법인 Term Frequency and Inverse Document Frequency(TFIDF) 기법을 적용하여 궁극적으로는 CB_TFIDF라는 분류시스템을 구축하였다. 해당 연구에서 구축한 시스템의 개념도는 [그림 1]과 같다. 분류하고자 하는 문서가 입력되면, 인터페이스 에이전트가 해당 문서 속에서 색인어들을 추출하고 이것들에게 가중치를 부여하여 질의문서를 구성한다. 이 질의문서가 문서분류 엔진에 입력되면, 문서분류 엔진은 사례베이스로부터 k개의 문서를 검색한 후, 질의문서와 각 검색문서들 간의 유사도를 계산해서 유사도 점수가 가장 높은 검색문서가 갖는 분류범주(Category)를 입력문서에 할당하게 된다.

전통적인 TFIDF 기법에 따른 분류 성능과 CB_TFIDF 기법에 따른 분류 성능을 비교해 보았을 때, 분류 속도나 정확도 측면 모두 CB_TFIDF가 우수함을 보였다. 하지만 CB_TFIDF는 속성선택을 하지 않았기 때문에, 해당 시스템에 적합한 속성선택 지표가 제시된다면 보다 더 효율적이고 효과적인 시스템으로 완성될 것으로 기대된다.

이에 본 연구에서는 CB_TFIDF 분류시스템에 대한 완성도를 높이기 위한 시도를 해보고자 한다.

3. 텍스트 문서에 적용되는 속성선택 지표

텍스트 문서 분류 시스템에 적용될 수 있는 속성선택 기법에는 많은 것들이 있다. 이미 우리에게 여러 가지 용도로 적용되고 있고, 또한 익숙하게 잘 알려져 있는 속성선택 지표들을 개괄적으로 살펴보고자 한다. 실제로 이 분야에 적용시킬 수 있는 속성선택 지표는 상당수 존재한다. 최근에 발표된 연구결과[4]만 보더라도 총 11의 속성선택 지표들을 적용시킨 바가 있다. Information Gain, Cross Entropy, Mutual Information, Weight of Evidence, Odds Ratio, Variants of Odds Ratio (Weighted Odds Ratio, Log-Probability Ratio, Exp-probability Difference, Conditional Odds Ratio), Random 등이 그것들이다. 또한 속성선택의 비교연구에 관한 연구[5]에선 총 5개의 지표들을 사용했는데, Document Frequency Thresholding, Information Gain, Mutual Information, χ^2-test, Term Strength 등이 그것이다. 두 연구만 보더라도 중복되는 지표들이 있음을 알 수 있다. 본 연구에서는 일반적으로 잘 알려진 속성선택 지표들을 먼저 한글 텍스트 문서 분류시스템에 적용해 보고자 한다. 우선적으로 적용시켜보기 위해 선정한 지표는 Information Gain, Odds Ratio 그리고 Document Frequency



[그림 1] 한글 문서 분류시스템의 개념도

Thresholding이다. 그에 대한 간략한 개요를 보도록 하겠다.

3.1 Information Gain

Information Gain[4, 5]은 속성값(Feature Value) 속에 포함된 분류값(Class Value)에 대한 평균 정보량으로 정의할 수 있다. 즉, 데이터 집합을 하위부분집합으로 분리하는데 특정 속성을 사용한다면 특정 속성을 사용하기 전과 후에 달라지는 엔트로피(Class Entropy)의 차이를 말한다. 해당 엔트로피를 측정하는 식은 다음의 식(1)과 같다.

$$IG(F) = I(C, E) = H(C) - H(C|F) = H(F) - H(F|C) = I(F, C)$$

----- 식(1)

F : 데이터세트를 하위부분집합으로 분리하는데 사용한 속성

$H(C) = \sum_i P(C_i) log_2 P(C_i)$: 분류 속성값에 대한 Shannon의 Entropy

$H(F|C) = \sum_j P(F_j) \sum_i (C_i|F_j) log_2 P(C_i|F_j)$

: 분류 속성값에 대한 조건부 Entropy

$I(F, C) = \sum_j P(F_j) I(F_j, C)$

$= \sum_i P(C_i) I(C_i, F) = I(C, F_j)$

: Expected Cross Entropy

$I(F_j, C) = \sum_i P(C_i|F_j) log_2 (P(C_i|F_j))/(P(C_i))$

: Cross Entropy or Conditional Mutual Information

$P(C_i|F_j)$: j번째 속성값에 주어진 i번째 분류값

$P(C_i)$: i번째 분류값에 대한 확률

$P(F_j)$: j번째 속성값에 대한 확률

3.2 Odds Ratio

Odds Ratio[4]는 분류값(Class Value) 중에 하나의 값을 예측하고자 할 때 사용되는 지표로서 흔히 정보 추출(Information Retrieval)에 자주 사용된다. 이는 문서에 출현한 단어들(words)을 속성으로 사용하여서 양의 분류값(positive(target) class value)에 대한 적절성을 평가하여 순위를 매긴다. 주어진 질의(query)와 추출된 문서간의 적합성에 대해 순위를 매기는 식은 아래의 식(2)와 같다.

$$\text{Ranking}(D, C_1) = \log((P(C_1|D))/(P(C_2|D))) = \log((P(C_1)P(D|C_1))/(P(C_2)P(D|C_2)))$$

----- 식 (2)

여기선 분류값(목표값)이 이진분류값(Binary-valued class)을 가진다는 특성이 존재한다. 위에서 C_1이 의미하는 것은 추출된 문서가 요구한 문서에 대해 적절하다는 것이고, C_2는 적절하지 못하다는 것이다. 질의에 상응하는 단어가 문서에 존재한다면, 각 문서는 이진분류값을 가지는 속성의 벡터로서 표현되어질 수 있다. 하지만, 여기서 각 단어들 사이에 독립성이 존재한다는 가정을 세우면, 식(3)을 얻을 수 있다.

$$\text{Ranking}(D, C_1) = \log\left(\frac{P(C_1)\prod_j P(W_j|C_1)}{P(C_2)\prod_j P(W_j|C_2)}\right) = \sum_j (F_j) Z_j + Const$$

----- 식 (3)

D: 문서

C_1: 양의 분류값(positive class value)

$P(W|C_i) = P(F_j = true|C_i)^{Z_j} * (1 - P(F_j = true|C_i))^{1-Z_j}$

: 주어진 i번째 분류값이 j번째 속성에 대한 값일 조건부 확률

F_j : 속성 F에 대한 단어의 가중치

$Z_j = 1 \quad (F_j = true)$

$= 0 \quad (otherwise)$

위의 식에서 본 바와 같이 단어의 가중치인 F_j가 바로 "Odds Ratio"가 된다. 관련된 식은 아래의 식(4)와 식(5)로 요약할 수 있다.

$$\text{OddsRatio}(F) = \log \frac{odds(W|C_1)}{odds(W|C_2)}$$

----- 식(4)

$$\text{Odds}(X_i) = \frac{\frac{1}{n^2}}{1-\frac{1}{n^2}} ; \quad P(X_i) = 0$$

$$= \frac{1-\frac{1}{n^2}}{\frac{1}{n^2}} ; \quad P(X_i) = 1$$

$$= \frac{P(X_i)}{1-P(X_i)} ; \quad P(X_i) \neq 0 \wedge P(X_i) \neq 1$$

----- 식(5)

여기서, Singularity[4]는 다음과 같이 처리하기로 한다. $P(X_i) = 0$은 $P(X_i) = 1/n^2$으로 대체하고, $P(X_i) = 1$인 경우는 $P(X_i) = 1-(1/n^2)$로 대체한다. n은 X_i가 발생하는 사건에 대한 경우의 수이다.

3.3 Document Frequency Thresholding

DF(Document Frequency)[5]는 보유한 전체 문서 중 해당 단어를 가지고 있는 문서의 비율을 나타내는 것으로 식(6)과 같이 계산된다.

$$DF_i = \frac{n_i}{N}$$

----- 식(6)

n_i : 단어 i를 가진 문서의 수

N : 보유한 전체 문서의 개수

위의 식대로 DF를 계산한 다음, 그 값이 사전에 정의한 임계치(threshold)를 넘지 못하면 해당 속성을 제거하는 방식이다. 이 지표의 기본 가정은 출현 빈도가 낮은 단어들이 분류를 하는데 있어서 적절한 정보를 제공하지 못하거나 전체 분류성능에 큰 영향을 미치지 않는다는 것이다. 이 가정을 바탕으로 Document Frequency Thresholding는 출현빈도가 낮은 속성을 제거함으로서 전체 속성 차원을 축소시킨다.

DF는 이렇게 단어(속성)를 축소하는데 적용할 수 있는 가장 간단한 방법 중에 하나이다. 또한 효율성을 향상시키는 것에만 초점을 둔 임기응변적인(ad hoc) 접근법 중의 하나라고 할 수 있다. 실제로는 이 DF방법이 공격적인 단어 제거를 할 때는 잘 사용되지 않는다. 그 이유는 앞서 한 가정 때문인데, 출현빈도가 낮은 단어가 의외로 해당 문서에 대해 보다 많은 정보를 담고 있을 수 있기에 그러하다. 고로, 이 가정에 대한 보완도 필요하다.

4. 한글 텍스트 문서 분류시스템을 위한 속성선택

4.1 실험 데이터 소개

본 실험에 사용될 데이터는 농업 진흥청 산하 사단법인 농진회에서 제작한 농업기술 CD-ROM에 수록된 문서들로서, 이재식과 이종운[1]의 논문에 사용된 데이터를 그대로 사용하였다. 위의 데이터를 동일하게 사용한 가장 큰 이유는 결과에 대한 비교 평가를 위해서이다. CD-ROM의 내용은 <표 1>과 같다.

<표 1> 농업기술 CD-ROM의 내용

	내 용
CD 1	식량작물, 특용작물, 잠업, 농업환경, 농기계, 농업경영, 농산물이용 등
CD 2	채소, 과수, 화훼, 농업환경, 농기계, 농업경영, 농산물 이용, 농기자개
CD 3	축산, 가축위생, 농기계, 농업경영, 농업환경, 농기자재, 농산물이용

본 문서의 내용은 HTML 형식으로 저장되어 있고 해당 문서마다 8자리의 문서번호를 가지고 있다. 해당 문서번호에 따라 각 분류가 결정되는데 해당 분류범주를 구분하는 기준은 <표2>와 같다.

<표 2> 문서번호 'BA010702'의 의미

분류기호	분류범주	분류단계
BA	식량작물	대분류
BA01	벼	중분류
BA0107	육묘	소분류
BA010702	BA0107의 2번째 문서	

본 연구에 사용된 데이터는 농업기술 CD-ROM 가운데 1번 CD-ROM에 해당하는 문서이다. 1,750개의 문서들을 연구에 사용되었는데 해당 구성비는 다음의 <표 3>과 같다. 그리고 해당 문서의 분류범주의 개수는 <표 4>에 나타나 있다.

<표 3> 연구에 사용된 문서의 구성비

	문서개수	비율(%)	레코드수
사례베이스 문서	1400	80	277,053
훈련용 문서	175	10	29,784
검증용 문서	175	10	30,677

<표 4> 연구에 사용된 문서의 분류범주 개수

	대분류	중분류	소분류
전체 문서	5	28	212
사례베이스 문서	5	28	199
훈련용 문서	5	23	88
검증용 문서	5	26	91

또한 본 연구에서는 훈련용 문서와 검증용 문서를 '질의문서'로 통칭하기로 한다.

4.2 실험의 개요

본 연구에서는 학습을 위한 Classifier로 사례기반추론(Case-Based Reasoning)을 사용한다. 본 실험의 전체적인 텍스트 문서 분류 수행과정이 다음의 [그림 2]에 나타나 있다. 이 그림 가운데 본 연구의 핵심이 되는 부분은 한글 문서 텍스트 분류시스템에 보다 효율적으로 적용될 수 있는 사례베이스를 구성하고 이를 최적화시키는 것이다.

4.3 Information Gain(IG)을 이용한 사례베이스 구성

사례베이스에 포함된 전체 1,400건의 문서에 포함된 속성(단어)들 각각에 대해서 대분류와 중분류 그리고 소분류 각각에 대한 IG 값을 계산한다. 그 값들을 통해 해당 단어가 분류에 미치는 영향력을 가중치 값으로 표현한다.

해당 사례베이스에 대한 크기를 1%, 2%, 3%, 5%, 10%, 20% 등 선별적으로 조정한 다음, 각 크기에 맞는 사례베이스들을 형성한다.

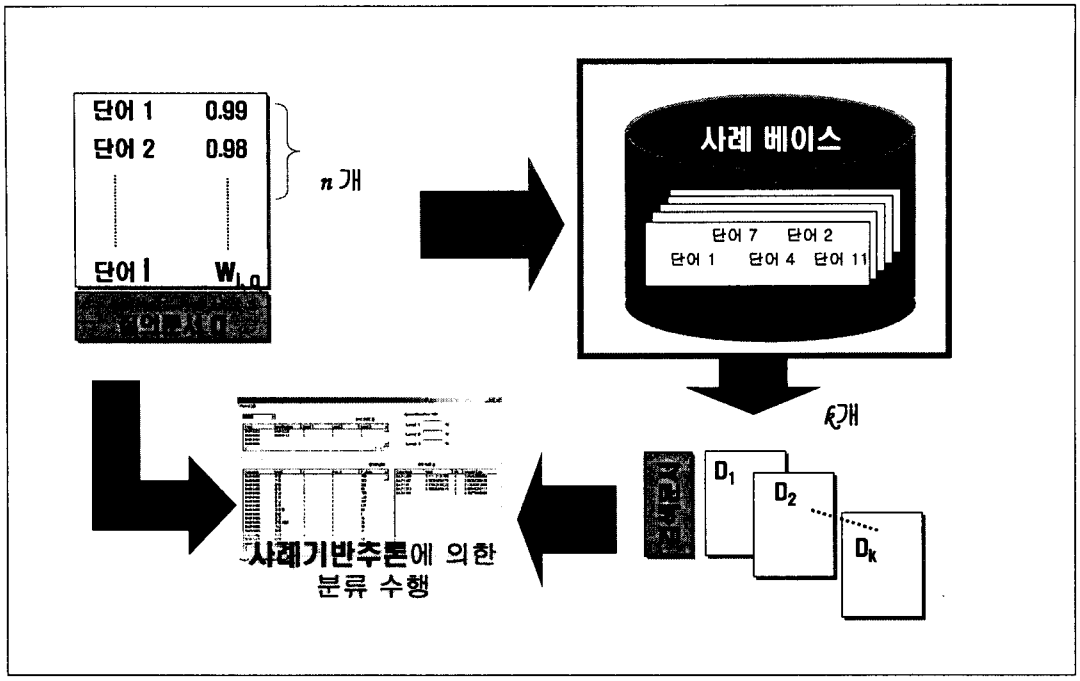
이러한 실험을 통해서는 기존의 한글 텍스트 문서 분류시스템이 가지는 가장 큰 문제점인 고차원의 속성 차원에 대한 부분을 해결할 수 있는 가이드라인을 제시할 수 있을 것으로 기대된다.

실제로 전체 속성에 대한 IG 값을 산출하는데 현재 실험에 사용되고 있는 시스템에서 56시간이 소요된다. 이 또한 비효율적인 부분으로 감안하면, 전체 속성에 대한 IG값을 산출하는 사례베이스를 만들기 전에 구체적인 IG값을 산출할 속성을 선정하는 추가적인 기준을 제시하는 것 또한 의미있는 작업으로 사료된다.

4.4 Odds Ratio를 통한 사례베이스 구성

위의 절에서 설명한 것과 유사하게 Odds ratio를 기준으로 사례베이스용 문서인 1,400개의 문서에 대한 모든 속성의 가중치를 구한다.

Odds Ratio를 이용할 때는 각각의 분류에 대해서 해당 속성이 적절한 분류값을 가지는 것으로 판정할 수 있는 판정기준이 필요하게 된다. 기본적으로 평균



[그림 2] 사례기반추론에 의한 텍스트 문서 분류 수행과정에 대한 개요

의 개념을 도입해 판정기준으로 삼고 이에 대한 평가 작업을 수행하였다. 하지만 적절성 여부를 판정하는 기준을 결정하는 부분에 대한 고찰 또한 꼭 필요한 연구 부분이라 하겠다.

그에 이어서, 한글 텍스트 문서 분류시스템에 적합한 사례베이스의 크기를 정하기 위해 전체 속성에 대한 사례베이스의 크기를 선별적으로 조정한다.

또한 여기서 결정된 사례베이스의 크기와 정확도 그리고 분류 시간 등은 CB_TFIDF와 IG를 통해 구성한 사례베이스를 가진 시스템의 성능과 비교 및 평가가 되어질 것이다.

그리고 각 분류별 특성에 따라 적절한 사례베이스 구성 방법에 대한 제언도 이루어질 것이다.

References

[1] 이재식, 이종운, "사례기반추론을 이용한 텍스트 마이닝", 한국경영정보학회 논문집, 제 12권 (2002)

[2] Jung Yunjae, Haesun Park and Ding-Zhu Du, "An Effective Term-Weighting Scheme for Information Retrieval," Computer Science Technical Report Tr008, University of Minnesota.

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

[4] Dunja Mladenic and Marko Grobelnik, "Feature Selection on Hierarchy of Web Documents," Decision Support Systems 35(2003), 45-87.

[5] Yang Yiming and Jan O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization,"

[6] Cho, W. V., "Knowledge Discovery from Distributed and Textual Data," Ph.D. Dissertation, Dept. of Computer Science, Hong Kong University of Science and Technology, 1999.

[6] Linoff, G. and M. J. A. Berry, *Mastering Data Mining*, Wiley, 2000.

[7] Mena, J., *Data Mining Your Website*, Digital Press, 1999.

[8] Joachims, T. A., "Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization," *Proc. 14th Int'l Conf. Machine Learning (ICML97)*, Morgan Kaufman, San

Francisco, 1997, 143-151.

[9] Lewis, D. D. and M. Ringuette, "Comparison of Two Learning Algorithms for Text Categorization," *Proc. Third Ann. Symp. Document Analysis and Information Retrieval*, Information Science Research Inst., Las Vegas, 1994, 81-93.