'98 International Conference Digital Library & Knowledge
September 22, 1998, Seoul Korea

교사학습 알고리즘을 이용한 텍스트 분류 시스템

(A Text Classification System based on a Supervised Learning Algorithm)

김 진상(金 鎭相), 성 정호(成 定浩)
계명대학교 컴퓨터공학과
김 성주(金 晟柱)
㈜에이전텍
Jin-Sang Kim, Jung-Ho Sung
Department of Computer Engineering, Keimyung University, Taegu, Korea
Soung-Joo Kim
AgenTech, Seoul, Korea

요약

지식경영을 위한 다양한 대상 업무 중에서 텍스트 데이터의 마이닝은 특히 중요하다. 그 이유는 텍스트 데이터가 양적인 면에서 가장 풍부하고, 또 발견할 수 있는 지식을 가장 많이 포함하고 있기 때문이다. 본 논문에서는 텍스트 데이터베이스에서 지식발견을 위한 한 과정으로 텍스트 데이터베이스 내의 텍스트들을 분류하는 기법을 기술한다. 특히 문서 분류 방법은 데이터베이스의 일부 데이터를 훈련 예제로 간주하여 교사 학습 알고리즘을 통해 학습한 후 나머지 데이터를 이용해 분류 정확성을 검증 및 향상시킨다. 시험 데이터로는 인터넷의 뉴스그룹의 기사를 이용하였고, 시험 결과 분류의 정확성은 한글 및 영문 모두 최소 70% 이상으로 나타났다.

1. 서론

지식경영(knowledge management)이란 조직의 모든 처리를 지식 처리로 보는 프레임워크로서, 기업의 경우 해당되는 처리 대상은 조직의 생존을 위한 지식의 생성, 보급, 갱신, 응용 등이 있다[5]. 따라서 지식경영은 지속적이고 불연속적인 환경 변화에 대한 기업의 적응력과 생존경쟁력 향상을

위한 해결책이 될 수 있을 것이며, 정보기술과 사람의 창의력 및 혁신적인 재능이 결합하여 상승 작용을 일으키기를 바라는 기업이나 조직에도 해답이 될 것이다.

현실적으로 지식경영은 기존의 데이터 웨어하우징 기법과 데이터 마이닝 기법을 확장하고 발전시키는 방향에서 그 구체적인 실체를 찾아볼 수 있다. 데이터 웨어하우징 기법을 확장하여 다양한 소스로부터 얻을 수 있는 비정형 텍스트를 관리하고, 또 데이터 마이닝 기법을 발전시켜 데이터와 텍스트 사이의 연관성을 분석한다. 한편 데이터 웨어하우징과 데이터 마이닝 응용은 대량의 트랜잭션을 발생시키는 대기업에서 주로 사용해 왔지만, 지식경영 시스템은 자신의 지식과 정보를 관리하기 위해 유연성이 높은 도구가 필요한 개인, 조직, 그리고 회사 등 거의 모든 전문가 집단이 사용할 것이다.

따라서 비정형 텍스트를 관리하고 분석하는 일은 지식경영에서 본질적이고 필수적인 분야로 대두되었고, 이에 대한 해결책도 다양하게 모색되고 있다. 데이터 마이닝과 데이터베이스에서 지식발견에 관해 무료로 뉴스를 공급하는 Kdnuggets사의 분류에 의하면 텍스트의 관리 및 분석을 목적으로 하는 연구분야를 텍스트 마이닝으로 부르고 있다[1]. 더 구체적으로 텍스트 마이닝은 텍스트에서 핵심 정보를 추출하고, 문서를 카테고리별로 분류하며, 문서 집단을 대표할 수 있는 제목을 찾고, 요청된 관련 문서를 찾아내며, 그리고 이상의 모든 업무의 결과를 효과적으로 보여주는 일 등을 해결하는 분야로 볼 수 있다[2].

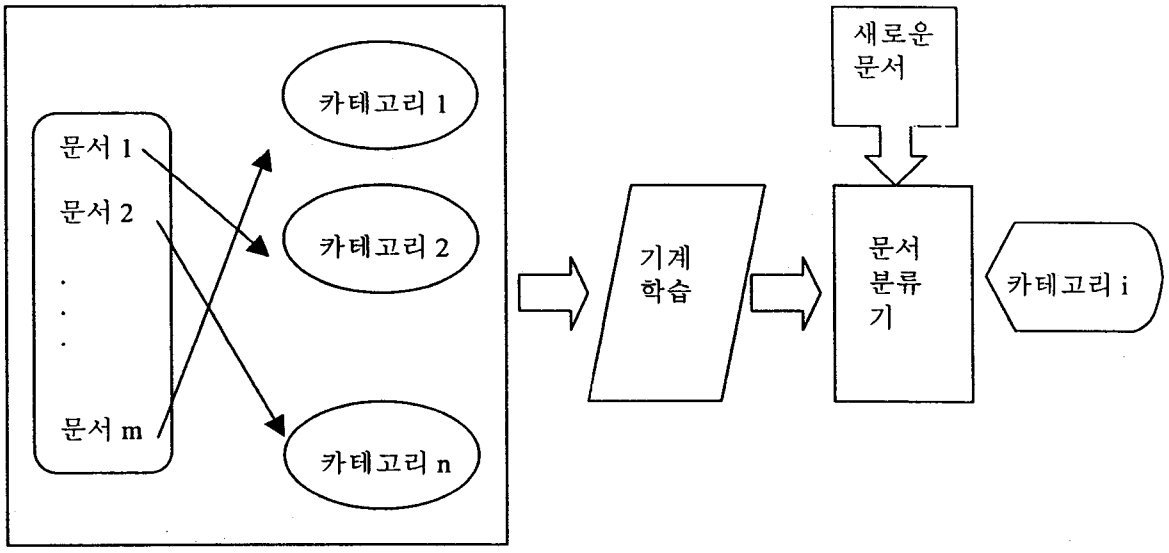
본 논문에서는 앞에서 설명한 텍스트 마이닝 기능 중 텍스트 문서를 카테고리별로 분류하는 기법에 대해 초점을 맞출 것이며, 이는 궁극적으로 문서의 자동분류를 목적으로 연구.개발 중인 시스템에 관한 개략적인 소개이다. 현재까지 알려진 텍스트 문서 분류 기법은 신경망(neural net)이나 통계적 방법, 귀납논리 프로그래밍(inductive logic programming), 결정 트리(decision tree), 규칙 학습(rule learning) 등 매우 다양하며, 또한 이들 방법의 상호 결합이나 새로운 기법의 개발로 더욱 세분화되고 있다. 또한 분류 대상인 문서의 종류도 메모, 편지, 전문 분야의 군서, 뉴스 기사, 웹 문서 등 거의 모든 종류의 문서를 대상으로 하고 있다. 이 논문에서는 인터넷의 뉴스그룹에 등재된 뉴스 기사를 대상으로 시험하였고, 시험 결과 뉴스 기사 분류의 정확성은 영문의 경우 85%이상이

며 한글의 경우 70%를 상회하는 것으로 나타났다. 특히 한글 뉴스들은 형태소 분석과 같은 데이터 클리닝 과정을 거치지 않은 것이어서 본 처리 기법의 견고성(robustness)을 입증하며 전처리 과정을 추가할 경우 영문 뉴스 기사 만큼의 정확성을 기대할 수 있으리라 본다.

본 논문의 구성은 다음과 같다. 제 2 절에서는 텍스트 분류를 위한 기본 개념과 시스템 설계 및 기법을 설명하며, 제 3 절에서는 본 연구에 사용된 시험 데이터의 구조와 특성 및 시험 결과를 기술하고, 마지막으로 결론과 향후 연구과제에 대해 언급할 것이다.

2. 텍스트 분류와 학습 알고리즘

텍스트 분류란 텍스트를 미리 정의된 여러 개의 카테고리 중 어느 하나에 속하도록 분류하는 일을 가리킨다. 텍스트 분류를 위해서는 일반적으로 사전에 분류의 기준이 되는 <문서, 카테고리> 형태의 순서쌍으로 구성된 훈련 예제를 이용해 학습을 시킨 후 그 결과를 이용하여 새로운 텍스트가 어느 카테고리에 속할 것인가를 예측하도록 하는데, 이와 같은 과정은 다음 [그림-1]과 같다.



[그림-1] 텍스트 분류 시스템

[그림-1]에서 보듯이 기계학습을 시키기 위해 입력으로 미리 주어진 문서 및 출력으로 미리 주어

진 카테고리의 쌍을 훈련예제라 부르며, 기계학습을 위해 이와 같이 입력과 출력 훈련예제가 미리 주어지는 경우를 교사학습(supervised learning)이라 한다. 반면에 입력과 출력 훈련예제 중 어느 하나만 주어지는 경우를 비교사학습이라 한다. 교사학습을 위한 하나의 훈련예제를 $<x_i,y_i>$로 나타낸다면 기계학습은 모든 i에 대해 $f(x_i)=y_i$인 가설 함수 f를 구하는 과정이며, 기계학습이 완료되면 임의의 입력 x에 대해 하나의 출력 y를 예측할 수 있게 된다[3].

함수 f가 이산적인 값을 가질 때 출력을 클래스 혹은 카테고리라 부르며, 이 때의 학습을 분류라고 한다. 따라서 텍스트 분류의 경우 입력은 텍스트 문서가 되며 출력은 입력 문서가 속할 문서 카테고리가 된다. 학습시킬 훈련예제의 출력, 즉 카테고리가 미리 주어지지 않는 경우는 입력문서의 패턴을 분석하여 공통적인 속성을 가진 문서들 끼리 군집화시키는 비교사학습 알고리즘을 사용하게 된다[4].

문서를 분류할 때 고려해야 할 중요한 사항 중 하나는 텍스트 문서를 특징 벡터로 나타내는 일인데, 이 경우 가장 단순한 방법은 하나의 단어를 하나의 특징으로 간주하는 것이다. 가령 n개의 단어로 구성된 텍스트 문서가 있다면 같은 수 만큼의 원소로 구성된 특징 벡터를 구성하는 것이다. 그러나 특징 벡터의 원소들 중에서 영어의 관사와 전치사 혹은 한글의 조사와 접미사 등은 텍스트 문서의 내용을 결정하는데 별 다른 영향을 미치지 않기 때문에 제거하여도 정확성에 영향을 미치지 않으며 오히려 효율성을 높일 수 있다. 즉, 텍스트 문서에서 불필요한 단어를 제거하여 그 문서가 가진 정보를 나타낼 수 있는 단어들로만 구성된 특징 벡터를 구성하는 것이 텍스트 분류를 위한 데이터 클리닝 과정에 해당된다.

문서 분류를 위한 다음 과정은 특징 벡터를 이용한 압축 과정인데, 여기에는 다양한 기법이 사용될 수 있다. 특징 벡터의 단어에 무게를 계산하여 처리하는 TFIDF의 경우 문서 집합 D에 속한 하나의 문서 d내의 i-번째 단어의 무게 $w(i,d)$는 다음과 같이 정의한다[7].

$$w(i, d) = tf(i,d)*idf(i)$$

여기서 $tf(i,d)$는 문서 d내에 나타나는 단어 i의 출현 빈도수이며, $idf(i)$는 $\log(|D|/df(i))$이며, 또한 $df(i)$는 단어 i를 포함하는 D내의 문서 수를 나타낸다.

한편 정보 이론에서는 특징 벡터의 압축을 위해 $-\log_2 p$ 비트가 최적이라고 알려져 있다[6]. 예를 들어, 8 개의 문서가 있고 각각의 문서가 나타날 확률이 모두 동일하다면 그 중 어느 한 문서가 나타내는 정보는 3 비트 만으로 표현할 수 있다.

본 연구에서는 일종의 교사학습법을 이용하여 훈련예제를 먼저 학습시킨 후 학습결과를 이용하여 주어진 문서를 가장 잘 압축시키는 카테고리로 그 문서를 분류하였다. 다음은 훈련예제 학습과 분류과정을 보여주는 알고리즘이다:

[훈련예제 학습]

각 카테고리 i에 속하는 문서들에 대해

- ✓ 단어 j 하나 하나에 대해
  - ➢ 확률 $Pr(j \mid i)$를 계산

[분류]

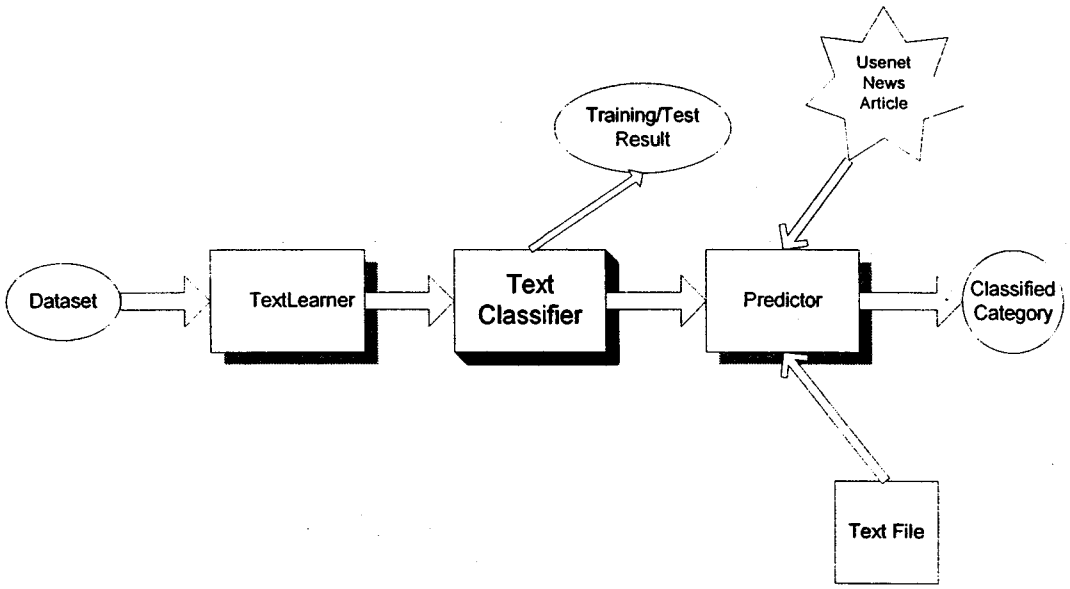
문서 d가 주어지면

d에 속하는 모든 단어 j에 대해 $-\log_2 Pr(j|i)$의 합을 구해 이 값이 최소가 되는 카테고리 i로 문서 d를 분류

이상의 개략적인 텍스트 분류 과정과 방법을 이용한 뉴스그룹의 뉴스 기사 분류 실험 환경과 결과는 다음과 같다.

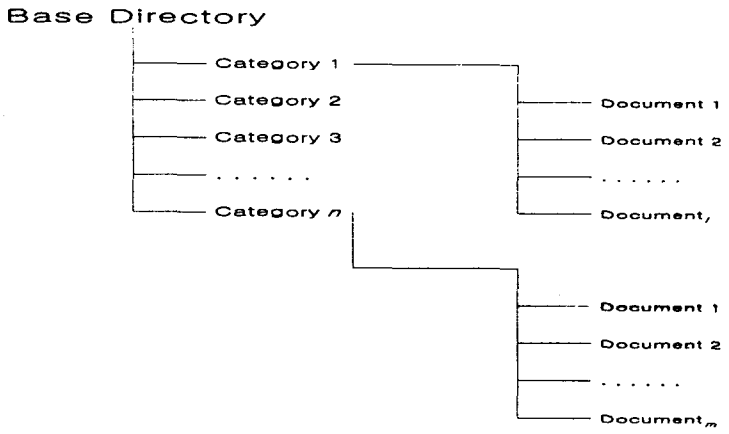
4. 실험 및 결과

본 논문에서 구현된 텍스트 분류 시스템은 [그림-2]에서와 같이 크게 학습을 위한 Learner 와 학습을 통해 생성된 Text Classifier 를 읽어 들여 뉴스그룹 기사나 아스키 텍스트 문서의 분류를 위한 Predictor 두 부분으로 구성되어 있다. Learner 와 Predictor 모듈은 모두 Windows NT 4.0 환경에서 Visual C++을 이용하여 작성되었으며 특히 Predictor 는 뉴스그룹 기사의 분류를 테스트하기 위해 간이 뉴스그룹 Reader 의 기능을 가지고 있다.



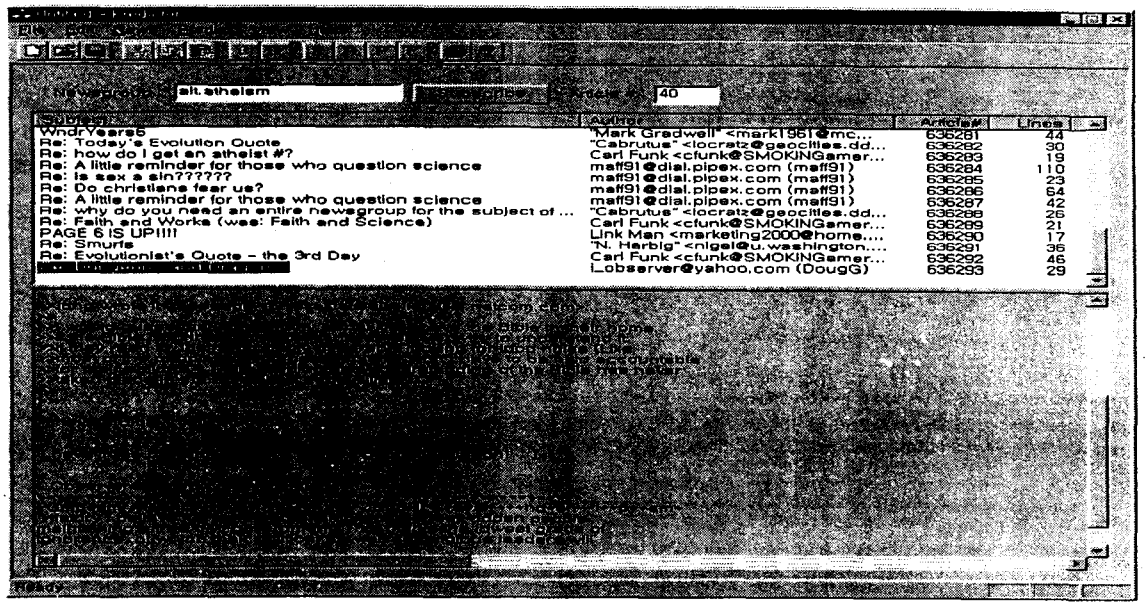
[그림-2] 텍스트 분류 시스템 구조

테스트를 위해 사용된 Dataset 은 뉴스그룹의 기사를 이용했는데 기사가 포스팅된 그룹이 카테고리가 되며 그 그룹의 기사들이 그 카테고리에 속하는 텍스트 문서에 해당되게 된다. TextLearner 의 입력으로 이용된 Dataset 의 구조는 [그림-3]과 같이 파일시스템의 디렉토리와 동일한 구조로 되어 있는데 그림 각각의 카테고리에 해당하는 이름으로 서브디렉토리를 구성하고 그 아래에 텍스트 문서들을 해당하는 카테고리의 이름으로 된 디렉토리 아래에 저장하는 구조이다.

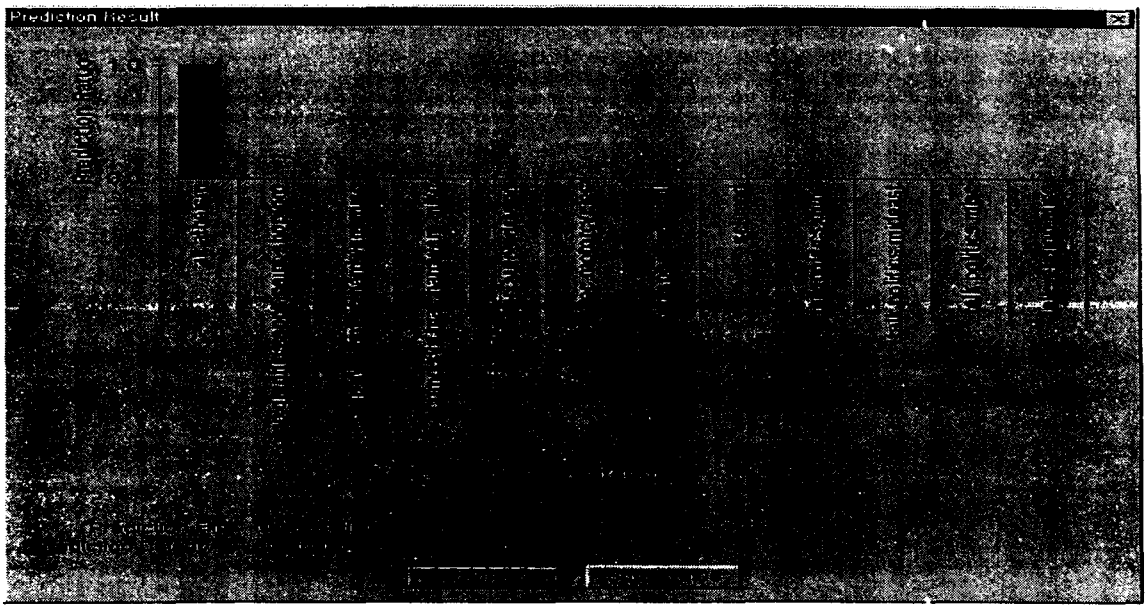


[그림-3] Dataset 구조

[그림-4]는 Predictor 가 실행된 상태에서 alt.atheism 이란 뉴스그룹의 기사를 구독해 가장 최근의 기사를 읽어 들인 것으로, 그 기사를 미리 학습된 카테고리에서 어떤 그룹에 속하는 기사인가를 분류해 그 결과를 수직막대그래프 형태로 표시한 것이 [그림-5]이다.



[그림-4] Predictor 의 실행화면



[그림-5] 특정 기사를 분류한 결과

[그림-5]의 그래프에서 알 수 있듯이 [그림-4]에서 읽은 기사의 카테고리가 alt.atheism 임을 제대로 분류했음을 확인할 수 있다.

본 논문에서는 3 가지 종류의 Dataset 을 이용해 분류의 정확성을 테스트 했으며 각 Dataset 의 특성은 아래 [표-1]과 같다. 그리고 학습과 테스트의 정확성을 위해 각각의 뉴스그룹의 기사에서 'Subject'를 제외한 모든 헤더 정보는 포함시키지 않았음을 밝혀둔다. 헤더 정보를 제거한 이유는 학습알고리즘에서 뉴스 기사라는 문서의 특징적 정보를 배제하기 위함이고 따라서 본 시스템의 일반성을 높일 수 있기 때문이다. 또한 [표-1]의 결과 중 9han_newsgroups 의 경우는 형태소 분석을 통한 데이터 클리닝 과정을 적용하지 않은 것이기 때문에 상대적으로 본 분류 시스템의 견고성을 입증한다고 볼 수 있다. 다시 말해 한글 뉴스의 경우 전처리를 하게되면 정확성은 제시된 것 이상으로 향상될 수 있다고 기대한다.

	카테고리 수	전체 기사 개수	평균 정확성	테스트 비율	언어
20_newsgroups	20	19,997	77.79%	33%	영문
12_newsgroups	12	10,800	85.56%	20%	영문
9han_newsgroups	9	7,850	71.36%	20%	한글

[표-1] Dataset 의 특성

5. 결론 및 전망

이 논문에서는 인터넷의 뉴스그룹에 등재된 뉴스 기사를 대상으로 텍스트 분류를 시험하였고, 시험 결과 뉴스 기사 분류의 정확성은 영문의 경우 85%이상이며 한글의 경우 70%이하로 떨어지지 않음을 확인하였다. 본 논문에서 소개된 텍스트 분류 시스템은 뉴스그룹의 텍스트 형태의 기사들을 대상으로 하고 있는데, 이것은 향후 인터넷 상의 다양한 웹 문서들에 대한 분류에 쉽게 접근하기 위함이며, 일반 문서가 아닌 구조화된 웹 문서들에 대한 추가적인 처리 방법들이 앞으로 보완되어야 할 것이다. 또한 대부분의 분류 목적 학습 알고리즘들은 배치 처리 방식에 기반을 둔 교사학습 방법을 이용하고 있지만, 분류의 일반성을 높이기 위해서 비교사학습 방법과의 결합이 필요할 것이다.

이러한 텍스트 분류 시스템은 텍스트 요약(Summarization)이나 텍스트 타일링(Tiling)과 같은 다른 텍스트 처리 기법과의 연동으로 지능적인 문서관리 시스템을 구성할 수 있을 것이다. 즉, 어떤 조직의 문서 Pool 에서 주 개념의 추출을 자동화 할 수 있으며, 이것은 최근 많은 기업들이 관심을 기울이고 있는 지식경영시스템의 핵심 기술로 이용될 수 있다.

참고문헌

[1] —, www.kdnuggets.com, 1998.

[2] —, www.software.ibm.com/data/iminer, 1988.

[3] T. Dean, J. Allen, Y. Aloimonos. *Artificial Intelligence: Theory and Practice*, Benjamin/Cummings Pub., 1995.

[4] D. Freitag. Multistrategy Learning for Information Extraction, 1998 *Int'l Conference on Machine*

Learning(ICML-98), 1998.

[5] Y. Malhotra. Knowledge Management, Knowledge Organizations & Knowledge Workers: A View from the Front Lines, *Maeil Business Newspaper*, Feb 19, 1998.

[6] R. Quinlan. C4.5: Programming for Machine Learning, Morgan Kaufmann, 1993.

[7] G. Salton. Developments in Automatic Text Retrieval, Science 253, pp. 974-979, 1991.