스타일 기반 키워드 추출 및 키워드 마이닝 프로파일 기반 웹 검색 방법

주 길 홍[†] · 이 준 휘[††] · 이 원 석[†††]

요 약

World Wide Web의 대중화로 인해 전자 정보량이 급속하게 증가하였고, 이러한 많은 양의 다양한 정보에 대한 효율적인 검색 시스템의 필요성이 증대되었다. 정확한 검색 결과를 제공하기 위해 사용자 요구 사항의 올바른 분석과 서술이 중요하게 인식되고 있으며, 분산 환경에서의 요구 사항 추출 및 분석의 필요성이 대두되고 있다. 본 논문에서는 웹 검색 방법에 있어서 목표 검색어만을 가지고 검색을 수행하는 기존 검색 방법과 달리 검색어가 나타나는 문맥 정보를 추가하여 검색하는 방법을 제안하고 구현하였다. 또한 본 논문에서는 제안된 새로운 키워드 추출 방법으로 추출된 키워드를 기반으로 키워드 마이닝 프로파일에 기반한 웹검색 시스템을 제안하고 구현하였다. 이는 원하는 정보를 대표하는 목표 검색어만 가지고 검색을 수행하는 기존의 검색방법과 달리 검색어가 포함된 문맥정보를 추가하여 검색하기 때문에 기존의 검색방법보다 정확하고 효율적인 정보를 제공한다. 특정 도메인으로부터 순위가 매겨진 도메인 키워드 리스트를 작성하여 이를 기준으로 기존의 출현빈도기반의 차이를 실험을 통하여 보였으며, 예제 기반 질의를 바탕으로 키워드 마이닝 프로파일을 만들어 검색을 수행하는 검색 방법으로 이의 효용성을 실험을 통해 검증하였다

An Efficient Web Search Method Based on a Style-based Keyword Extraction and a Keyword Mining Profile

Kil Hong Joo[†] · Jun Hwi Lee[††] · Won Suk Lee[†††]

ABSTRACT

With the popularization of a World Wide Web (WWW), the quantity of web information has been increased. Therefore, an efficient searching system is needed to offer the exact result of diverse information to user. Due to this reason, it is important to extract and analysis of user requirements in the distributed information environment. The conventional searching method used the only keyword for the web searching. However, the searching method proposed in this paper adds the context information of keyword for the effective searching. In addition, this searching method extracts keywords by the new keyword extraction method proposed in this paper and it executes the web searching based on a keyword mining profile generated by the extracted keywords. Unlike the conventional searching method which searched for information by a representative word, this searching method proposed in this paper is much more efficient and exact. This is because this searching method proposed in this paper is searched by the example based query included content information as well as a representative word. Moreover, this searching method makes a domain keyword list in order to perform search quickly. The domain keyword is a representative word of a special domain. The performance of the proposed algorithm is analyzed by a series of experiments to identify its various characteristic

키워드 : 키워드 추출(Keyword Extraction), 데이터 마이닝(Data Mining), 웹 정보 검색(Web Information Searching), 웹 문서(Web Document), 패턴 분석(Pattern Analysis), 프로파일 분석(Profile Analysis)

1. 서 론

컴퓨터와 통신 기술이 발전함에 따라 네트워크를 통한 일반 사용자들의 컴퓨터 활용 빈도와 정보의 양이 급격히 증가되었다. 특히, World Wide Web(WWW)의 급속한 발전으로 인터넷을 통해 접하게 되는 전자 정보량의 폭발적으로 증가로 인해 사용자가 필요로 하는 정보를 찾기 위한 시간과 노력이 증가하는 정보과잉(information overload) 상태가 발생하고 있다. 이에 따라 사용자들이 찾고자 하는 유용한 정보를 빠르고 정확하게 찾는 것은 매우 어려운일이 되어가고 있다. 따라서 대용량의 웹 텍스트 데이터에 대한 효율적인 검색방법이 절실히 필요하게 되었다.

기존의 웹 검색 시스템들은 일반적으로 찾고자 하는 내

† 정 회 원 : 연세대학교 대학원 컴퓨터과학과
†† 정 회 원 : 소프트그램 기술연구소
††† 종신회원 : 연세대학교 컴퓨터과학과 교수
논문접수 : 2004년 5월 11일, 심사완료 : 2004년 6월 19일

용을 대표하는 목표 키워드만으로 검색을 수행하였다. 그러나 동일한 키워드가 여러 도메인에서 다양한 의미로 사용될 수 있기 때문에 부정확한 결과를 제공할 가능성이 높다. 따라서 웹 검색 시스템들은 에이전트를 사용하여 검색 대상이 되는 웹 문서로부터 키워드를 추출하여 추출된 키워드를 색인화시키고, 사용자의 검색명령과 색인을 비교하여 질의 결과를 사용자에게 제공하였다[1]. 웹 문서들은 대부분 HTML문서이며, HTML문서는 기존의 plain-text와 다르게 semi-structure 문서로 다양한 스타일을 적용하여 문서를 꾸밀 수 있다. 이렇게 작성된 문서의 스타일에는 문서 작성자의 의도가 반영되어져 있다. 예를들어 강조하고 싶은 단어나 문장은 눈에 잘 띄는 색으로 표현한다거나 글꼴의 크기를 크게 하거나 글씨를 두껍게 표현할 수 있다. 이와같이 다양한 스타일이 적용된 HTML문서를 검색할 때 문서에 사용된 스타일로부터 작성자의 의도를 파악하여 결과를 제공하면 검색 결과의 정확성을 높일 수 있으며, 불필요한 문서의 제공이 줄어들 것이다. 이를 위해 본 논문에서는 검색을 위한 키워드를 추출할 때 기존의 키워드 기반 추출방법과 다르게 스타일 기반 키워드 추출방법을 제안한다.

웹 문서를 검색할 때 문서의 구조화는 검색 시스템의 큰 비중을 차지하기 때문에 문서의 구조화를 고려할 경우 검색의 정확성을 높일 수 있다. 문서 구조화를 통해 문서의 내용을 대표하는 키워드의 추출여부가 효율적인 웹 검색의 중요한 조건이다. 따라서 본 논문에서 제안하는 스타일 기반 키워드 추출방법은 문서를 이루는 단어들 중에서 문서의 전반적인 스타일에서 벗어나는 스타일을 가진 단어들을 키워드로 추출한다. 왜냐하면 다른 스타일을 가진 단어는 중요한 의미를 가지고 있거나 강조하고자 하는 단어일 가능성이 높기 때문이다. 이러한 스타일 차이에 따른 단어의 중요도를 파악하여 키워드를 추출하는 것을 스타일 기반 키워드 추출이라고 정의한다.

스타일 기반 키워드 추출방법을 통하여 추출된 키워드를 바탕으로 프로파일을 생성하고, 생성된 프로파일 기반으로 웹 검색을 수행한다. 이때 본 논문에서는 키워드 마이닝 프로파일 기반 웹 검색을 사용한다. 키워드 마이닝 프로파일 기반 웹 검색이란 먼저 검색자가 찾고자 하는 정보와 유사한 내용을 포함하는 웹 문서들을 예제기반 질의로 제공하고 이로부터 로그를 추출한 후 추출된 로그에 데이터 마이닝 기법을 적용하여 프로파일을 만들고 이를 바탕으로 웹을 검색하여 이와 유사한 문서들을 찾는 방법이다. 이 방법은 기존의 검색 시스템들의 키워드로 질의를 직접 입력하는 방법이 아닌 예제 기반으로 질의를 수행한다. 이는 사용자가 찾고자 하는 내용을 대표하는 실제 웹 페이지들을 질의로 선택하게 함으로써 이루어지며, 질의는 웹 페이지들을 직접 방문하며 질의에 포함될 페이지들을 선택하여 완성한다. 따라서, 본 논문은 스타일 기반 키워드 추출방법을 제안하고 이의 유효성을 실험을 통해 검증하였다. 또한 이 스타일 기반 키워드 추출방법을 기반으로 키워드 마이닝 기반 웹 검색 시스템을 설계하고 구현하였으며 이에 대한 검증을 수행하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대해서 기술하고, 3장에서는 웹 문서의 구조를 분석하여 문서의 구조화를 기반으로 하는 스타일 기반의 키워드 추출방법을 기술한다. 4장에서는 추출된 키워드를 기반으로 프로파일을 생성하고, 생성된 프로파일을 기반으로 웹 검색을 수행하는 키워드 마이닝 프로파일 기반 웹 검색방법을 기술한다. 5장에서는 구현된 프로파일 기반의 웹 검색 시스템의 기능 및 효능에 대해서 설명한다. 6장에서는 3장, 4장 및 5장에서 제안된 내용에 대한 다양한 실험을 수행하고 이에 대한 결과를 분석한다. 마지막으로 7장에서는 최종적인 결론을 맺는다.

2 관련 연구

효율적인 문서 및 키워드 검색을 위해 다양한 연구들[18, 19, 20]이 수행되어졌다. [18]에서는 불린(boolean) 모델의 PRISE 검색 엔진을 이용하여 문서 검색을 수행한 후 검색된 문서들 중에서 제한된 윈도우 크기내에서 해당 불린 질의를 만족하는 단락을 찾아 추출한다. 이렇게 추출된 단락들의 수가 너무 많거나 혹은 너무 적으면 해당 불린 질의를 좀더 강화시키거나 완화시키며 검색을 재수행하여 적절한 수의 단락들이 추출될 때까지 검색 및 추출과정을 반복 수행한다. [19]는 문서검색을 두차례 수행하며 키워드를 추출한다. 첫 번째 단계에서는 모든 키워드를 포함하고 있는 문서를 검색한다. 그러나 모든 키워드를 포함하는 문서는 일반적으로 그 양이 아주 작기때문에 이를 보완하기 위하여 두 번째 검색 단계에서는 키워드 빈도와 가중치를 사용하는 일반적인 문서 검색을 수행하여 문서를 검색한다. 이와같이 검색된 문서들을 문장 단위로 분리를 한 후 각각의 문장들을 대상으로 키워드를 추출한다. [20]은 불린 모델의 검색 엔진을 통하여 문서 검색을 수행한 후 검색된 문서들을 문장 단위로 분리하여 키워드를 포함하고 있는 문장을 추출한다. 추출된 각 문장들을 휴리스틱 방법을 사용하여 순위를 결정하고, 순위대로 키워드를 추출한다. 이와 같이 문서 검색을 수행하는 시스템들의 공통적인 특징은 검색된 문서들에 대하여 자체적인 추출방법을 사용하는 것이다. 즉, 문서 내에서 정답과 관련성이 있는 부분은 문서 전체가 아니라 문서의 일부분이기 때문에 문서에서 필요한 부분들만을 추출하여 사용한다. 그러나 이러한 기존의 키워드 추출방법들은 문서의 스타일을 고려하지 않고 모두 출현빈도

에 동일한 가중치를 부여하였기 때문에 문서 작성자의 의도를 반영하지 못하는 단점이 있다. 따라서 본 논문에서 제안하는 스타일 기반 키워드 추출방법은 단어의 출현빈도를 모두 동일하게 보지 않고 적용된 스타일에 따라 가중치를 부여하여 출현빈도를 계산한다.

문서의 구조화는 문서의 키워드로 구성되어져 있고, 이를 위해 문서의 키워드를 추출하는 방법으로는 단어 빈도수, 곧 *TF*(Term Frequency) 팩터(factor) 기반의 가중치 부여 방식[3, 4]이 널리 사용되고 있다. 이 방법은 문서에서 추출된 단어의 중요도를 반영하기 위한 방법으로 $TF \times IDF$(Term Frequency Inversed Document Frequency)[2-4] 함수를 사용한다. $TF \times IDF$ 함수는 단어의 빈도수와 역 문서 빈도수를 곱하는 것으로 문서 d_i에서 단어 t_j의 가중치 $tfidf_{ij}$는 다음과 같다.

$$tfidf_{ij} = tf_{ij} \times \ln \frac{N}{df_j}$$

이때 N은 전체 문서의 개수를 나타내고, tf_{ij}는 단어 빈도수로서 문서 d_i에서 단어 t_j가 나타난 횟수를 나타내며, df_j는 문서 빈도수로서 N개의 문서들 중에서 단어 t_j가 존재하는 문서 수이다. 이 함수는 하나의 문서에 단어가 많이 포함되어 있다면 그 문서를 대표하는 단어로 사용될 가능성이 높지만 단어가 포함되어 있는 문서가 많을수록 문서를 특정 짓는 능력이 낮아진다는 것을 의미한다. 그러나 $TF \times IDF$ 함수는 처리하고자 하는 문서수가 많을수록, 즉 N 값이 클수록 역문서 빈도수 값이 단어의 가중치를 결정하는데 많은 비중을 차지하는 단점을 가지고 있다.

문서 길이 정규화(document length normalization)에는 최대 빈도수 정규화와 코사인 정규화의 두 가지 방법[5]이 많이 사용되고 있다. 최대 빈도수 정규화(Maximum Frequency Normalization)방법은 문서에서 가장 많이 나타나는 단어의 빈도수로 각 단어의 빈도수를 나눠주는 방법이다. 코사인 정규화(Cosine Normalization)방법은 여러 특성 정보로 구성된 벡터 공간 모델에서 가장 많이 사용되는 방법으로 만약 벡터 W가 $W = \{w_1, w_2, \cdots, w_n\}$와 같을 때, 벡터의 각 원소를 코사인 정규화 원소인 $\sqrt{w_1^2 + w_2^2 + \cdots + w_n^2}$로 나눠주는 것이다. 코사인 정규화는 높은 단어 빈도수에 대해서 정규화할 뿐만 아니라 단어 수가 많은 경우 코사인 정규화 원소가 증가하기 때문에 많은 단어에 대해서도 정규화가 가능하다. 정규화를 통해 분석된 키워드들로 색인을 구성하고 효과적으로 분류하고 구조화하면 검색의 효율을 높일 수 있으며, 이는 유전자 알고리즘이나 신경망과 같은 인공지능 기술에 적용하여 문서인식과 분류에 응용할 수 있다[6, 7]. 또한 예제 기반 질의에 대한 연구는 주로 웹 기반의 문서 클러스터링과 분류화(categorization)와 연관되어 이루어지고 있다[8-10].

데이터 마이닝(Data Mining)은 대량의 실제 데이터로부터 묵시적이고 잠재적으로 유용한 정보를 추출해 내는 작업[11]이라고 정의한다. 이 중 대표적인 것이 장바구니문제(market basket problem)로 소매점의 장바구니 정보, 즉 판매 데이터로부터 숨겨진 연관 관계(association rule)를 찾는 것이다[12]. 연관 관계란 $\{X1, X2, \cdots, Xn\} \rightarrow Y$의 형태로 표현하며, $X1, X2, \cdots, Xn$이 장바구니에 모두 포함되어 있다면 그 안에 Y도 포함될 확률이 높다는 것을 의미한다. 이 규칙을 만족하는 Y를 찾을 확률을 이 규칙의 신뢰도(confidence) 라고 정의하며, 일반적으로 특정한 임계값를 넘는 신뢰도를 가지는 규칙만을 찾는다. 대부분의 경우 이러한 연관관계 중에서 장바구니에 자주 나타나는 항목 집합(Itemset)에 대해서만 관심을 가진다. 예를 들어 아무도 사지 않는 항목들에 대해 연관관계를 찾아낸다고 해도 그로부터 실제적인 이득을 얻어낼 수 없기 때문이다. 따라서 대부분의 데이터 마이닝은 높은 지지도(support)를 가지는 항목 집합, 곧 많은 장바구니에 같이 나타나는 항목들에 대해서만 관심을 가진다. 이는 $\{X1, X2, \cdots, Xn, Y\}$가 전체 장바구니 중 일정 비율 이상에 나타나야만 한다는 것이고, 이때의 비율을 최소 지지도(Minimum support)라고 한다. 이때 최소 지지도를 넘는 항목 집합을 빈발 항목 집합(Frequent Itemset)이라 한다.

연관관계를 찾을때 먼저 최소 지지도를 만족하는 모든 빈발 항목 집합을 찾은 후에 찾은 빈발 항목 집합에서 연관규칙을 찾아내는 것이다[13]. 빈발 항목 집합과 항목들의 지지도를 알고 있다면 연관 규칙은 쉽게 찾아낼 수 있기 때문에 일반적으로 빈발 항목 집합을 찾는 문제에 대해 많은 연구[13, 14]가 이루어지고 있다. 이중 Apriori 알고리즘[13]은 개별 항목의 지지도를 구하여 최소 지지도를 넘는 빈발 항목들을 추출하고, 이를 바탕으로 후보 항목 집합(Candidate Itemset)을 생성한다. 그후 실제 지지도를 실제 데이터로부터 구하여 빈발 항목을 찾는다. 이와 같은 단계별 과정을 반복하여 모든 빈발 항목 집합을 찾는다. 또한 DHP알고리즘[14]은 k-후보 빈발 항목 집합을 생성할 때 점차적으로 트랜잭션 크기를 줄이는 전지(pruning)기법을 사용함으로써 트랜잭션을 효과적으로 줄이며, 다음의 항목집합의 생성을 위해 불필요한 항목집합을 삭제하는 해싱(hashing)기법을 사용함으로써 후보 빈발 항목 집합을 효율적으로 생성한다. 다른 연관규칙 탐색 방법에는 전체 데이터베이스의 스캔이 최대 두번만 발생하는 PARTITION 알고리즘[15], 샘플링 개념을 도입하여 수학적인 정형화에 조정 가능한 정확도를 제시한 DS 알고리즘[16], Apriori 알고리즘보다는 더 작은 패스로 빈발 항목 집합을 찾아내고 샘플링 방법보

다 작은 후보 항목 집합을 사용한 DIC 알고리즘[17] 등이 있다.

3. 스타일 기반 키워드 추출방법

3.1 문서 스타일

웹 검색 시스템에서 가장 중요한 요소인 문서구조화를 수행하기 위하여 가장 필요한 것은 문서의 키워드이다. 문서의 키워드는 문서의 내용을 대표하는 단어로써 정확한 키워드를 추출하는 것은 웹검색 시스템의 효율성을 극대화시킨다. 따라서 기존의 빈도기반 키워드 추출방식[18-20]의 단점을 극복하기 위하여 본 논문에서는 스타일에 기반한 가중치 부여방식을 고려한 새로운 키워드 추출방식을 제안한다. 문서의 단어들 중에서 전체적인 스타일과 다른 스타일을 가진 단어들은 중요한 의미를 가지거나 강조하고자 하는 단어일 확률이 높기 때문에 이러한 스타일의 차이에 따른 단어의 중요도를 파악하여 키워드를 추출하는 것을 스*타일기반 키워드추출 방법*이라고 정의한다. 예를들어 전체 스타일의 글씨 크기가 10포인트일때 임의의 단어의 크기가 15포인트라고 하면 강조된 단어라고 판별할 수 있다. 또한 전체 글씨색이 검정색일때 임의의 단어가 파란색이라면 역시 강조하는 단어라고 판별할 수 있다.

HTML태그들 가운데 Text Formatting 태그들은 적용된 단어를 어떤 식으로 화면에 표시해야 하는 지를 표현한다. 한 단어에 여러 태그가 적용될 수 있고 이들의 종합된 결과로 단어의 스타일이 결정된다. 따라서 스타일기반 키워드 추출방법은 각 Text Formatting 태그별로 가중치를 계산하지않고 최종 결과물로서 브라우저 상에서 실제로 표시되는 스타일로부터 가중치(weight)를 계산한다. 본 논문에서는 스타일을 [12]에서 제안한 7가지 항목으로 나누어 각 항목별로 가중치를 부여한다. 7가지의 스타일 항목은 다음의 <표 1>과 같다.

문서의 키워드는 문서에 단어가 나타나는 빈도로 결정되어진다. 기존의 키워드 추출방법들은 모두 출현빈도에 동일한 가중치를 부여하였다. 그러나 본 논문에서 제안하는 스타일 기반 키워드 추출방법은 단어의 출현빈도를 모두 동일하게 보지 않고 적용된 스타일에 따라 가중치를 부여하여 출현빈도를 계산한다. <표 1>의 7가지 스타일 항목별로 각각 정규화하여 계산된 가중치를 합하여 단어의 가중치를 고려한 출현빈도를 구한다. 본 논문에서는 문서의 단어추출 방법은[13]에서 제안한 방법을 사용하였으며, 문서 내에서 실제로 적용된 스타일을 *스타일 인스턴스*라고 정의한다. 예를들어, 어떤 문서에서 글꼴의 크기가 12, 14, 24의 세 종류가 사용되었다면 글꼴 크기 스타일 인스턴스는 12, 14, 24 세 가지가 된다. 글꼴 스타일(Font Style)은 italic, normal, oblique의 세 가지가 스타일 인스턴스가 될 수 있다.

〈표 1〉 스타일 항목

스타일 종류	스타일 값 표현방법	값 표준화 방법	가중치 부여방법
글꼴 크기(Font Size)	absolute-size, relative-size, length, percenta	지정된 폰트 크기를 pt 단위로 환	식 (1)
글꼴 가중치(Font Weight)	Normal, bold, bolder, Lighter, 100, 200, 300, 400, 500, 600, 700, 800, 9	자연 언어(Natural Language)로 정의된 경우 수치로 환	
글꼴 종류(Font Family)	글꼴 종류		식 (2)
글꼴 스타일(Font Style)	italic, normal, obliq		
글자 정렬(Text Align)	left-aligned, right-aligned, centered, justified		
색(Color)	natural language / CNS, #RGB, #RRGGBB, float range : 0.0 - 1	#RRGGBB 값으로 변	
글자 장식(Text Decoration)	blink, line-through, overline, 또는 underline decoratio		

3.2 정규화(Normalize)된 가중치 수식

각 스타일 종류별로 스타일 인스턴스 i에 대한 가중치를 SW_i라 할때 스타일 가중치 SW_i를 결정하는 방법은 스타일의 종류에 따라 두 가지 방법으로 정의한다. 스타일 인스턴스 i의 값을 SV_i라고 할때 첫 번째 방법은 SV_i가 클수록 중요한 의미를 지난다고 판단할 수 있기 때문에 문서내의 상대적인 중요도를 판단하기 위하여 문서 전체에 사용된 해당 스타일 인스턴스의 평균값을 구하여 이를 기준으로 가중치를 부여한다. SV_i의 값이 평균보다 커질 경우 양의 가중치를 부여하고, 작아질 경우 음의 가중치를 부여하며, 평균값으로부터 멀어질수록 가중치가 커진다. 이때 SV_i의 정규화는 값을 표준편차로 나누어 수행한다. 스타일 인스턴스 i의 평균을 SV_i의 대표값으로 정의하고, 스타일 인스턴스 i가 적용된 단어수를 SC_i라고 정의할 때 SC_i를 도수로 보아 평균과 표준편차를 구한다. 이를 이용한 첫 번째 스타일 가중치 방법은 식 (1)과 같고, 이때 SC_{avg}와 SC_{sd}는 평균과 표준편차를 의미한다.

$$SW_i = \frac{SV_i - SC_{avg}}{SC_{ad}} \quad (1)$$

두 번째 방법은 글꼴 종류(Font Family), 글꼴 스타일(Font Style), 색(Color), 글자정열(Text Align), 글자장식(Text Decoration)과 같이 스타일 인스턴스의 값에 따른 중요도를 판단할수 없는 경우 스타일 인스턴스가 적용된 단어수를 기반으로 중요도를 판단하는 방법이다. 이는 적용된 단어가 적을수록 중요한 의미를 나타낸다고 판단하는 것이다. 이를 위해 스타일 인스턴스의 대표값을 문서의 총 단어수에 대비하여 해당 스타일 인스턴스가 적용된 단어수의 비율로 정하고, 스타일 인스턴스가 적용된 단어수를 도수로 정하여 평균과 표준편차를 구한다. 이때 문서의 총 단어수를 TC라고 할때, SV_i는 SC_i/TC이기 때문에 이를 이용한 두 번째 가중치 방법은 식 (2)와 같다. 식 (2)에서는 식 (1)에서와 달리 SV_i의 값이 중요하지 않기 때문에 음수를 취한다.

$$SW_i = \frac{SV_i - SC_{avg}}{SC_{ad}} = \frac{SC_{avg}}{SC_{sd}} - \frac{SC_i}{TC \cdot SC_{sd}} \quad (2)$$

적용된 태그 인스턴스의 집합을 T라고 할때 인스턴스 j가 부여된 단어 w_i의 출현빈도는 식 (3)과 같이 $fv_{i,j}$로 나타낸다. $fv_{i,j}$는 적용된 스타일 인스턴스의 가중치의 합과 출현에 대한 가중치 1의 합으로 계산된다.

$$fv_{i,j} = 1 + \sum_{k \in T} SW_k \quad (3)$$

따라서, 단어 w_i의 인스턴스의 집합을 I라고 할때 w_i의 인스턴스들의 출현빈도 fv값의 총합이 단어 w_i의 가중치가 부여된 출현회수 fv_i가 되며 식 (4)와 같이 표현된다.

$$fv_i = \sum_{j \in I} fv_{i,j} \quad (4)$$

단어 w_i의 정규화된 가중치가 부여된 빈도 F_i는 fv_i값들이 표준정규분포를 이룬다고 가정하고 정규화를 수행하여 구한다. fv_i값들의 평균과 표준편차를 각각 FV_{avg}와 FV_{sd}라고 정의하면, F_i는 식 (5)와 같이 계산되어진다. 따라서, F_i가 단어 w_i의 스타일 가중치가 된다.

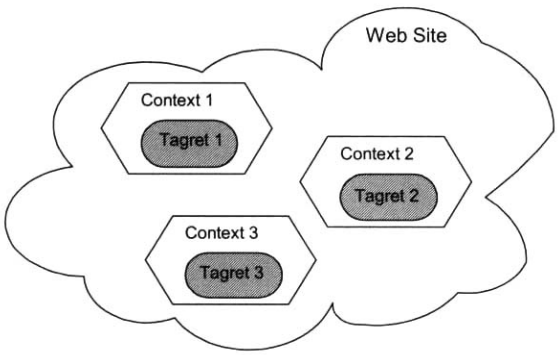
$$F_i = P(Z < a),\ a = \frac{fv_i - FV_{avg}}{FV_{sd}} \quad (5)$$

4. 키워드 마이닝 프로파일 기반 웹 검색

4.1 내용영역과 대상영역

키워드 마이닝 프로파일 기반 웹 검색이란 사용자가 필요로하는 정보와 유사한 내용을 포함하고 있는 웹 문서들을 사용자에게 예제기반 질의로 제공하여 사용자에게 필요한 문서만을 제공하는 방법이다. 따라서, 질의는 직접 키워드를 입력하지 않고 예제 기반으로 이루어진다. 키워드 마이닝 프로파일 기반 웹검색은 다음의 단계에 의하여 수행되어진다. 먼저 검색자가 찾고자 하는 정보와 유사한 내용을 가지고 있는 웹 문서들을 예제기반 질의로 제공하여 로그를 추출한다. 둘째로 추출된 로그에 데이터 마이닝 기법을 적용하여 프로파일을 생성한다. 셋째로 생성된 프로파일을 기반으로 검색을 수행하여 유사한 문서들을 찾는다. 질의방법은 기존의 검색시스템과 같은 직접 키워드를 입력하는 방법이 아니라 예제 기반으로 질의를 하는 것으로써 질의는 찾고자 하는 내용을 대표하는 실제 웹 페이지들을 직접 방문하여 질의에 포함될 페이지들을 선택하여 완성된다.

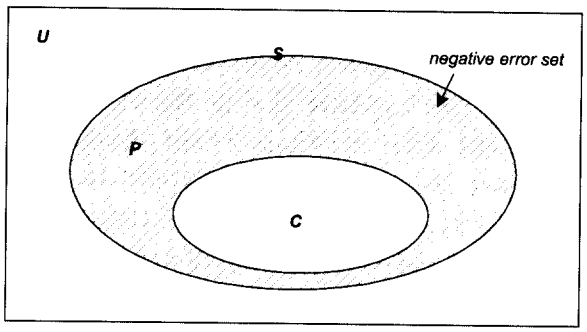
웹사이트는 하나 이상의 주제를 나타내며 하나의 주제에 대하여 세부적인 주제로 분류될 수 있다. 예를들어 신문사 웹 사이트의 문서들은 모두 신문 기사들이지만 정치, 사회, 경제, 스포츠등의 세부 주제로 분류되어져 있다. 따라서 웹 사이트에서 찾고자 하는 문서는 웹 사이트의 주제 또는 세부주제 중에 포함되어 있으며 찾고자 하는 문서들이 속하는 주제가 찾고자 하는 내용의 문맥으로 간주할 수 있다. 따라서 웹 사이트의 구조는 다음의 (그림 1)과 같이 나타낼 수 있다. (그림 1)에서 대상영역(*Target*)은 사용자가 찾고자 하는 문서의 집합이고, 내용영역(*Context*)은 대상영역을 포함하여 문맥정보를 표현하는 문서의 집합이다.



(그림 1) 웹 사이트의 내용영역과 대상영역

기존의 검색방법에서는 찾고자 하는 문서, 대상영역을 대표할 만한 키워드를 사용자가 검색어로 제시하면 검색어가 출현한 페이지들을 검색결과로 제공하였다. 그러나 이러한 방법은 검색어가 어떤 주제에 관하여 사용되는가에 따라 의미가 달라지는 점을 고려하지 못하는 단점이 있다. 예를들어 '환경'이란 단어를 검색어로 입력하였을때 이 단어가 정치, 경제, 학술, 교육등의 다양한 주제 중 어느 주제에 맞게 사용되었는지에 따라서 그 의미가 달라질 수 있다. 따라

서 검색 결과에는 검색어를 포함하고 있으나 사용자에게 불필요한 의미가 다른 문서들이 포함될 수 있다. (그림 2)에서와 같이 전체 웹 문서의 집합을 U라고 하고, 검색어가 출현한 문서들의 집합을 S라고 하면, S의 부분집합이 사용자가 원하는 결과의 집합 $C(C \subset S)$가 된다. 이때 집합 S에서 집합 C를 뺀 차집합 $P(=S-C)$는 검색어가 포함되어 있으나 사용된 의미의 상이함으로 인해 원치 않는 결과 문서의 집합이다. 본 논문에서는 이를 부정(negative)에러집합이라고 정의하며 부정에러집합을 줄이는 것을 목표로 한다.

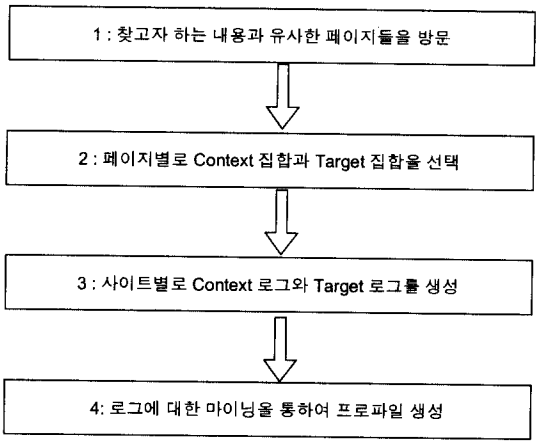


(그림 2) 웹 문서 집합

검색을 수행할 때 사용자가 찾고자 하는 키워드뿐만 아니라 키워드가 포함되어 있는 문맥정보를 포함하여 검색을 수행하면 부정에러집합을 줄일 수 있기 때문에 대상영역 질의뿐만 아니라 대상영역 질의에 내용영역 질의도 포함한 질의도 고려하였다. (그림 1)에서와 같이 대상영역은 기존 검색방법에서 사용되는 목표 검색어와 동일한 의미를 갖는 찾고자 하는 내용을 나타내며, 내용영역은 대상영역을 포함하는 문맥정보를 표현하기 때문에 내용영역을 통하여 대상영역의 정확한 의미를 정의할 수 있으며, 부정에러집합을 줄일 수 있다. 또한 검색의 정확성을 높이기 위하여 내용영역과 대상영역간의 의미관계를 파악하여 검색에 이용한다. 따라서 본 논문에서 제안하는 검색방법은 내용영역 프로파일, 대상영역 프로파일, 내용영역-대상영역 프로파일의 세 가지 프로파일을 생성하여 이를 기반으로 검색을 수행한다.

4.2 웹 사이트의 내용영역과 대상영역

본 논문의 검색방법에서 질의는 선택된 페이지들로부터 구성되며, 페이지는 단어들로 구성된다. 또한 질의내에서는 키워드 추출방법에 의해 추출된 키워드들이 페이지를 대표한다. 따라서 키워드는 질의를 구성하는 최소단위가 되며 질의를 정의할 때 대상영역과 내용영역을 같이 선택해야 하기 때문에 사이트가 사용자가 찾고자 하는 내용을 대표하는 단위가 된다. 키워드를 추출할 때 단순히 여러 사이트에 나타난 키워드들만 추출하는 것이 아니라 여러 사이트에서 나타난 키워드들을 같이 추출한다면 더욱 정확한 정보를 제공할 수 있다. 만약 질의가 하나의 페이지로 구성되어 진다면 페이지를 대표하는 키워드 집합을 프로파일로 선택할 수 있다. 본 논문에서는 검색의 정확성을 높이기 위하여 다수의 예제를 질의로 선택함으로써 다수의 페이지들로부터 프로파일을 생성하여 많은 데이터로부터 유용한 요약을 찾아낸다. 이때 프로파일을 생성하기 위하여 데이터마이닝 기법중의 하나인 Apriori 알고리즘[13]을 사용한다. 예제 기반 질의로부터 프로파일을 생성하는 단계는 (그림 3)과 같이 4단계로 이루어진다.



(그림 3) 프로파일 생성 순서

프로파일 생성 순서에서 단계 1과 단계 2는 질의를 정의하는 단계로 웹페이지들을 방문하면서 찾고자 하는 내용영역 또는 대상영역페이지를 발견한 경우 대표하는 페이지의 전체 또는 필요한 영역에 대하여 키워드를 추출한다. 이때 추출하고자 하는 영역이 연속되지 않은 여러 곳에 있다면 각 부분을 순차적으로 선택하여 추출한다. 모든 페이지에 대하여 동일한 과정을 반복적으로 수행하여 질의를 정의한다. 질의정의가 완료되면 단계 3의 로그생성단계를 수행한다. 단계 3에서는 내용영역과 대상영역으로 선택된 웹 페이지들 사이의 링크 구조와 각 페이지들의 키워드 집합으로부터 내용영역과 대상영역의 프로파일을 구성하기 위한 로그를 생성한다. 이때 생성되는 로그는 내용영역 로그, 대상영역 로그, 내용영역-대상영역 로그 3가지이다. 내용영역 로그는 각 사이트별로 내용영역으로 지정된 페이지들의 키워드 리스트이며, 대상영역 로그는 각 사이트별로 대상영역으로 지정된 페이지들의 키워드 리스트들이다. 마지막으로 내용영역-대상영역 로그는 페이지들의 링크 순서에 따라 페이지의 페어를 구성하여 이를 바탕으로 대상영역 키워드 상위에 나타난 내용영역 키워드의 리스트들이다. 내용영역-대상영역 로그를 생성할 때 페이지의 방문순서는 내용영역 또는 대상영역으로 선택된 페이지들부터 시작하여 깊이우선(depth-first)탐색방법으로 방문한다. 이때 방문하는 링크

의 깊이는 간접링크(indirect link)로 연결된 페이지까지로 제한한다. 간접링크란 직접링크로 연결된 페이지가 아닌 하나의 페이지를 거쳐서 연결된 것을 의미하는 것으로 2레벨 차이가 나는 페이지의 링크이다. 2레벨까지만 고려하는 이유는 3레벨 이상 떨어진 페이지는 내용의 연관성이 낮고, 실제 수행시간이 급격하게 증가하므로 비효율적이기 때문이다. 페이지페어를 생성한 후에 각 대상영역 페이지와 페어를 이루는 상위 내용영역 페이지의 키워드 리스트의 합집합이 대상영역 페이지의 키워드들의 상위 내용영역 키워드 로그가 되고, 내용영역-대상영역 로그는 각 대상영역 키워드 별로 상위의 내용영역 키워드 리스트를 사이트별로 출력한다. 단계 4에서는 생성된 로그들로부터 프로파일을 추출한다. 이때 사이트를 트랜잭션으로 키워드를 항목으로 간주하여 빈번하게 같이 출현하는 모든 빈발항목집합을 찾아서 프로파일로 선택한다. 빈발항목집합은 Apriori 알고리즘[13]을 사용하여 찾는다. 임의의 k 단계에서 Apriori 알고리즘을 사용하여 빈발항목집합 L_k를 찾는 방법은 다음과 같다. 먼저, k-1 단계에서 찾은 빈발항목집합 L_{k-1}로부터 후보항목집합(Candidate Itemset) C_k를 생성한 후 C_k의 지지도를 계산한다. C_k가운데 최소지지도를 넘는 항목집합들로 L_k를 생성한다. 이와같이 단계별로 반복적으로 수행하여 모든 빈발항목집합을 구한다.

4.3 로그와 패턴의 비교 기준 항목

프로파일은 빈발항목집합의 목록으로 이루어진다. 빈발항목집합은 키워드들과 키워드의 지지도로 구성되며, 내용영역-대상영역 프로파일의 경우에는 대상영역 키워드 별로 모든 빈발항목집합을 구한다. 본 논문에서는 이렇게 구해진 빈발항목집합을 패턴(pattern)이라고 정의한다.

패턴을 통해 프로파일을 생성한 후에 생성된 프로파일을 기반으로 사이트들을 비교하여 찾고자 하는 질의 사이트들과의 일치여부를 판정해야 한다. 이때 비교는 사이트별로 로그를 생성한 후 생성된 로그와 프로파일과의 비교를 통해 수행한다. 프로파일의 임의의 패턴에 속하는 모든 단어가 로그에 속할 때 이 패턴은 완전매치(Complete match)한다고 정의하며, 전체 단어의 절반 이상만 만족할 경우 부분매치(Partial match) 한다고 정의한다. 또한 패턴의 전체 단어 수에 대비하여 일치하는 단어의 비율을 로그에 대한 패턴의 매치율(Matching ratio)이라고 정의한다. 사이트의 한 로그가 일정길이 이상의 패턴을 만족시킬때 프로파일을 만족하는 로그로 판정한다. 본 논문에서는 로그와 패턴의 매치도 판정의 정확성을 높이기 위하여 매치패턴비율, 매치단어 수, 패턴 지지도, 매치패턴 포인트 비율, 구간별 매치패턴 지지도 비율, 구간별 단어매치율의 6가지 비교 기준 항목을 제안한다.

첫째로 매치패턴비율(Matched Pattern Ratio : MPR)은 로그와 완전매치 또는 부분매치된 패턴이 프로파일 중에서 차지하는 비율을 의미한다. 둘째로 매치단어 수(Matched Word Count : MWC)는 로그와 완전매치 또는 부분매치된 패턴들의 단어의 평균개수이다. 로그를 만족하는 패턴들이 평균적으로 몇 개의 단어가 일치하는가를 의미한다. 셋째로 패턴 지지도(Pattern Support : PS)는 완전매치 또는 부분매치된 패턴들의 평균지지도를 의미한다. 이는 로그를 만족하는 패턴들이 평균적으로 어느 정도의 지지도를 가지는지를 의미한다. 넷째로 매치패턴 포인트 비율(Matched Pattern Point Ratio : MPPR)을 정의하기 위하여 [정의 1] 및 [정의 2]와 같이 패턴 포인트 및 매치패턴 포인트를 정의한다.

[정의 1] 패턴 포인트(Pattern Point : PP)

여러 단어들과 지지도의 두 가지로 구성되는 패턴의 특성을 하나의 값으로 표현하기위하여 패턴 포인트를 다음과 같이 정의한다. 패턴에 속한 단어의 수를 패턴의 길이라고 정의할 때 패턴 포인트는 패턴의 길이(Pattern Length : PL)와 패턴의 지지도(Pattern Support : PS)의 곱으로 표현되어지며 이는 프로파일이 생성될 때 결정되어진다.

$$PP = PL \times PS \quad \blacksquare$$

[정의 2] 매치패턴 포인트(Matched Pattern Point : MPP)

패턴들은 로그와 부분매치 될 수 있기 때문에 매치율에 따라 차등을 주기 위하여 다음과 같이 매치패턴 포인트를 정의한다. 매치패턴 포인트는 [정의 1]의 패턴 포인트에 매치율을 곱한 값으로 표현된다. 이때 m_i는 패턴 i의 매치율이다.

$$MPR = \sum_{i \in rule} PP_i \times m_i \quad \blacksquare$$

따라서, 매치패턴 포인트 비율은 식 (6)과 같이 매치패턴 포인트를 프로파일의 패턴 포인트의 합으로 나눈 값으로 계산하며 이는 전체 패턴 포인트에 대한 매치된 패턴 포인트의 비율을 의미한다.

$$RMPP = \frac{MPP}{\sum_{i \in rule} PP_i} \tag{6}$$

다섯 번째로 구간별 패턴 매치율(Pattern Matching Ratio in an Interval : PMRI)은 패턴을 지지도 구간별로 나눈 후 구간별로 전체 패턴 대비 매치패턴의 비율을 구한 것이다. 마지막으로 구간별 단어 매치율(Word Matching Ratio in an Interval : WMRI)은 구간별로 패턴에 대해 매치된 단어의 비율을 구한 것이다. 구간별 패턴 매치율과 구간별 단어 매치율은 구간별로 세분화하여 관찰함으로써 상세한 정보

를 얻고 검색의 정확도를 향상시킨다.

이와 같은 6가지 비교 항목들을 기반으로 질의와의 유사 여부를 판정할 기준을 세우기 위해 질의 사이트에 대한 검사를 수행한다. 사이트 단위로 각 항목의 값을 계산하여 전체 질의 사이트에 대한 평균과 표준편차를 구하고, 구해진 평균과 표준편차를 기반으로 질의와의 유사 여부를 판정한다. 이때 각 매치도 항목에 대한 평균을 기준값으로 설정하고, n은 시스템에 주어진 상수, σ는 매치도의 표준편차라고 할때 기준값에서 매치 범위인 $n \times \sigma$이내로 떨어져있다면 매치한다고 판정한다.

위의 단계는 내용영역, 대상영역, 그리고 내용영역-대상영역 프로파일 별로 매치도를 판정하므로 각각의 판정 결과를 총합하는 단계가 필요하다. 대상영역 프로파일을 통한 매치도 판정을 통해 대상영역과의 일치 여부를 판정하고 매치한다고 판정된 경우 내용영역 프로파일과의 매치도를 검사한다. 두 단계에서 모두 매치된다고 판정되었다면 내용영역-대상영역 프로파일과의 매치도를 판정하여 최종 매치여부를 판정한다.

5. 프로파일 기반의 웹 검색 시스템 구현

5.1 질의정의기

질의정의기는 웹브라우저를 내장하여 웹페이지들을 방문하여 웹페이지의 전체 또는 일부분의 내용을 선택하여 내용영역 또는 대상영역으로 설정하여 질의를 구성하는 모듈이다. 크게 질의정의기는 웹 브라우징(Web browsing), 웹 문서 구조 분석, 스타일 기반의 키워드 추출 그리고 내용영역 또는 대상영역 설정의 4가지 기능을 수행한다. 그림 4는 질의정의기 화면을 나타내고 있다. 질의정의기는 그림 4에서 보듯이 웹브라우저를 내장하고 있으며, 네비게이션(navigation)을 위한 버튼들과 질의정의를 위한 버튼들 그리고 주소창으로 구분되어 있다. 질의 정의기 화면을 통하여 웹페이지를 찾아 웹페이지의 구조를 분석하여 (그림 4)의 왼쪽과 같이 트리형태로 표현한다. 트리의 노드(node)는 문서의 엘리먼트 노드 또는 텍스트 노드이며, 트리의 노드를 선택하면 선택된 노드가 브라우저 상에서 선택되어 진다. (그림 4)에서는 <TR>노드가 선택되어져 있으며 브라우저에는 큰 메뉴의 '정치'라고 쓰여진 부분이기 때문에 반전되어 표시됨으로써 선택되어졌다는 것을 보여준다.

(그림 5)와 (그림 6)은 질의정의기에서 키워드를 추출하는 과정을 나타낸다. 웹 페이지의 구조분석을 통하여 페이지의 전체 또는 일부분을 선택한 후 키워드를 추출하는 과정이다. 본 논문에서 제안한 스타일 기반 키워드 추출방법을 사용하여 키워드를 추출하며, 키워드 추출은 먼저 (그림 5)와 같이 내용을 선택한 후에 상단의 'Get'버튼을 누르면 선택된 영역으로부터 키워드를 추출한다. 키워드가 선택된 결과는 (그림 6)과 같다. (그림 6)의 왼쪽 아래에 있는 리스트 콘트롤은 선택된 부분에서 추출된 키워드들을 나타내며, 오른쪽 아래의 리스트 콘트롤은 선택된 부분에 적용된 스타일

(그림 4) 질의정의기

(그림 5) 구조분석을 통한 내용 선택

(그림 6) 키워드 추출

인스턴스의 가중치를 표현한다. 키워드가 추출되어졌다면 브라우저의 텍스트 영역이 아쿠아마린(aquamarine)색으로 반전되어 표시된다. 이와같이 키워드 추출기능으로 페이지의 키워드를 모두 추출한 후 페이지를 내용영역 또는 대상영역 질의로 설정한다. 질의정의기 상단의 버튼을 사용하여 설정할 수 있다.

5.2 로그 생성기 및 프로파일 생성기

로그생성기는 질의정의기에서 정의된 예제기반 질의를 입력받아 로그를 생성한다. 이는 Command-line 프로그램으로 작성되어졌으며, 페이지 페어 생성 알고리즘을 통하여 질의내의 각 페이지를 시작 페이지로 설정한 후 페어생성을 수행한다. 그 후 생성된 페어의 결과로 내용영역, 대상영역, 내용영역-대상영역 키워드 리스트 및 로그를 생성한다. 로그생성기에서 찾아가는 URL은 동일한 페이지가 다른 내용으로 검색되어질 수 있으므로 모두 절대경로로 변환하여 사용한다. 페이지의 링크를 따라 검색을 수행할 때 방문한 페이지가 프레임(frame)으로 이루어진 경우에는 독립적인 URL이지만 내용이 없는 페이지이기 때문에 각 프레임의 페이지들로 현재 페이지를 대체하여 작업을 수행한다. 내용영역 및 대상영역 로그의 형태는 (그림 7)과 같다.

```
Keyword-1 Keyword-2 Keyword-3 · · · · · · Keyword-m
Site-1 (URL)
Keyword-1 Keyword-2 Keyword-3 · · · · · · Keyword-n
Site-2 (URL)
· · · · · ·
```

(그림 7) 내용영역 및 대상영역 로그 포맷

내용영역 및 대상영역 로그와 다르게 내용영역-대상영역 로그는 키워드 별로 상위의 내용영역 키워드 리스트를 사이트 별로 기록한다. 내용영역-대상영역 로그의 형태는 (그림 8)과 같다. 이때 첫줄의 *Site-number*는 전체 사이트의 개수를 의미한다.

```
Site-number = k
[Keyword]
Target_Keyword-1
Keyword-1 Keyword-2 Keyword-3 · · · · · · Keyword-l
Site-1 (URL)
Keyword-1 Keyword-2 Keyword-3 · · · · · · Keyword-m
Site-2 (URL)
· · · · · ·
Keyword-1 Keyword-2 Keyword-3 · · · · · · Keyword-n
Site-k (URL)
[/Keyword]
[Keyword]
Target_Keyword-1
· · · · · ·
```

(그림 8) 내용영역-대상영역 로그 포맷

프로파일 생성기는 로그생성기로부터 생성된 로그 파일의 최소 지지도(minimum support)를 입력받아 Apriori 알고리즘을 사용하여 로그별 프로파일을 생성한다.

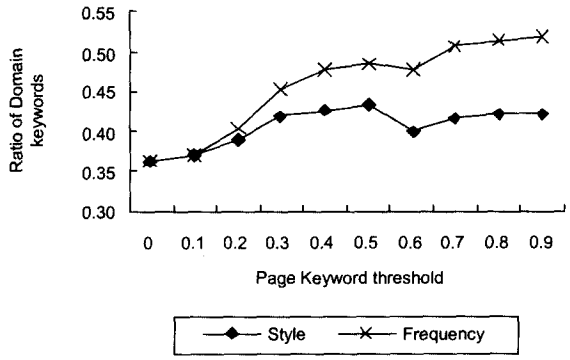
6. 실 험

본 논문에서는 제안된 알고리즘을 다양한 관점에서 평가하기 위하여 여러 가지 실험을 수행하였다. 또한 키워드 추출 및 생성된 프로파일에 대한 다양한 실험을 수행하기 위하여 서로 다른 두 종류의 도메인으로 실험을 수행하였다. 첫 번째 도메인은 미국의 대학 기숙사 페이지로써 *DOMAIN*1이라고 한다. 또 하나의 다른 도메인은 한국 구청의 민원관련 페이지로 이를 *DOMAIN*2라고 하였다. *DOMAIN*1은 총 18개의 대학 기숙사 도메인의 281개 웹 페이지를 대상으로 실험을 수행하였으며, *DOMAIN2*는 총 16개의 구청 도메인의 227개의 웹 페이지를 대상으로 실험을 수행하였다. 이때 키워드의 최소 서버 서포트는 0.6으로 설정하였다. 키워드의 서버 서포트는 전체 사이트에서 키워드가 출현한 사이트의 비율이다.

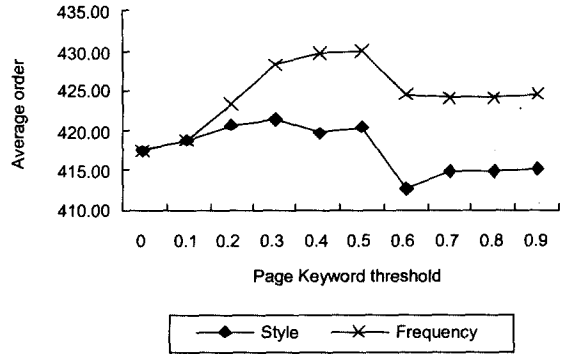
본 논문에서 제안하는 스타일 기반 키워드 추출방법 (Keyword Extraction based on the Style : KES)의 효율성을 검증하기 위하여 기존에 제안된 단어의 출현빈도 기반 키워드(Keyword Extraction based on the Frequency : KEF)추출방법과 비교실험을 수행하였다. 도메인내의 모든 페이지별로 키워드를 추출하여 스타일 기반과 출현빈도 기반으로 비교 실험을 수행하였으며, 결과의 정확성을 검증하기 위하여 양적비교와 질적비교로 나누어 실험을 수행하였다. 양적비교란 추출된 페이지 키워드중에서 도메인 키워드에 포함되는 비율을 측정한 것으로 비율이 높을수록 도메인 키워드가 많이 추출되어진 것이기 때문에 정확한 키워드를 추출했다고 판단할 수 있다. 또한 질적비교는 도메인 키워드에 포함된 키워드의 순위를 결정한 후 순위의 평균을 계산하여 비교를 수행한다. 평균순위가 작을수록 추출된 키워드의 도메인 순위가 높기 때문에 정확한 키워드를 추출했다고 판단할 수 있다. 키워드의 순위를 계산할 때 한 사이트 내의 여러 페이지에 출현한 단어들은 한 페이지를 대표하는 단어가 아니라 일반적인 단어일 가능성이 높기 때문에 낮은 가중치를 부여해야 한다. 따라서 사이트 키워드의 가중치는 전체 페이지 수를 키워드가 출현한 페이지 수로 나누어 계산하고, 사이트 키워드 가중치의 평균을 도메인 키워드 가중치로 하여 순위를 계산한다.

(그림 9)와 (그림 10)은 각각 *DOMAIN*1에 대한 양적비교와 질적비교의 결과를 나타낸다. 실험에서 임계값(threshold)을 0에서 0.9까지 변화시키면서 결과를 비교하였다. (그림 9)의 양적비교에서 페이지 키워드의 임계값이 커질수록 본 논문에서 제안하는 *KES*방법이 *KEF*방법보다 정확

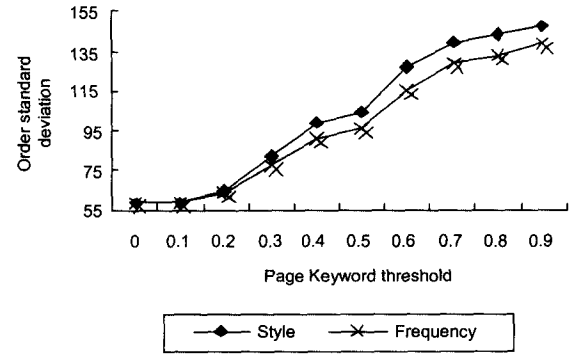
성이 높다는 것을 알수 있다. 따라서 양적비교에서 *KES*방법이 사용자에게 정확한 검색 결과를 제공한다는 것을 알 수 있다. 질적비교에서는 페이지 키워드의 임계값이 증가할수록 *KES* 방법이 *KEF* 방법보다 더 낮은 평균순위를 갖는다. 이는 *KES* 방법이 상위 순위의 키워드를 추출하기 때문에 질적으로도 *KEF* 방법보다 더욱 정확한 정보를 제공한다는 것을 알 수 있다. 그래프로 나타내지는 않았지만 DOMAIN2에서도 DOMAIN1과 비슷한 결과를 나타내었다



(그림 9) 키워드 추출방법의 양적비교



(그림 10) 키워드 추출방법의 질적비교



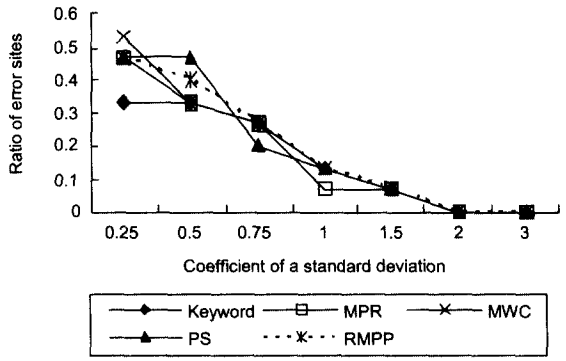
(그림 11) 순위 표준편차 비교

(그림 11)은 *DOMAIN*1의 질적비교에 대한 순위 표준편차에 대한 결과를 나타낸다. (그림 9) 및 (그림 10)과 같이 임계값을 0에서부터 0.9까지 변화시키면서 결과를 비교하였다. (그림 11)에서 보듯이 *KES* 방법의 순위 표준편차가 *KEF* 방법의 순위 표준편차보다 평균 순위차이 만큼 더 큰 값을 갖는다. 이는 본 논문에서 제안하는 *KES* 방법이 넓게 분포하지만 순위는 *KEF* 방법보다 같거나 낮게 분포한다는 것을 의미한다. *DOMAIN*2에서도 *DOMAIN*1과 같은 결과를 보였다.

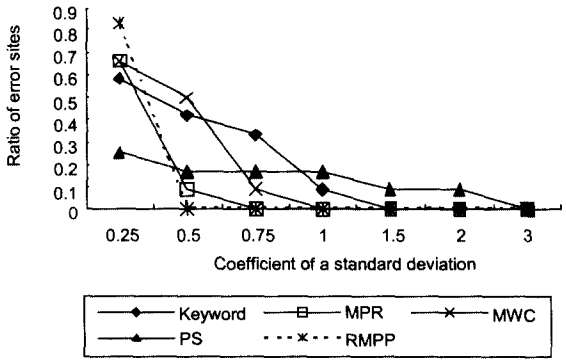
질의 사이트에 대한 프로파일이 사용자가 찾고자 하는 정보에 대한 정확한 정보를 제공하는지를 검증하기 위하여 본 논문에서 제안하는 검색방법에서 기반이 되는 3가지 프로파일인 대상영역 프로파일, 내용영역 프로파일 그리고 내용영역-대상영역 프로파일에 대하여 정확성실험을 수행하였다. 세 가지 프로파일에 대한 매치도 판정을 위하여 4.3에서 제안한 6가지의 비교 기분 항목 중 패턴매치비율(*MPR*), 매치단어 수(*MWC*), 패턴 지지도(*PS*), 매치패턴 포인트 비율(*MPPR*)을 키워드만을 이용하여 매치도를 판정하는 방법(*keyword*)과 비교를 수행하였다. 표준편차계수는 0.25, 0.5, 0.75, 1, 1.5, 2, 3의 7가지에 대하여 실험을 수행하였다. (그림 12), (그림 13) 및 (그림 14)는 각각 순서대로 두 가지 도메인 *DOMAIN*1과 *DOMAIN*2에 대한 내용영역 프로파일, 대상영역 프로파일 및 내용영역-대상영역 프로파일에 대한 결과이다. 세 개의 실험결과에서 Y축은 에러로 판정된 사이트의 비율로 질의 사이트에 대한 판정결과이므로 낮을수록 정확한 결과를 제공한다고 판정할 수 있다. 또한 X축은 판정의 매치 범위를 결정하는 표준편차의 계수로써 X축의 변화에 따라 사이트 매치도의 분포를 파악할 수 있다. (그림 12)의 내용영역 프로파일과 (그림 13)의 대상영역 프로파일의 경우 키워드 기반 판정방법과 비슷하거나 더 좋은 결과를 보였다. 또한 (그림 14)의 내용영역-대상영역 프로파일의 경우 표준편차 계수가 1이하일 경우에는 프로파일 매치도를 통한 판정이 좋은 매치율을 나타내고, 1보다 클 경우에는 비슷한 결과를 나타내는 것을 알 수 있다. 세 가지 프로파일의 결과 그래프에서 키워드 기반 판정방법에 비해 프로파일 기반 판정방법의 기울기 변화가 완만함을 나타내고 있다. 이는 프로파일 기반 판정방법이 기준값에 가까이 접근하며 조밀하게 분포되어 있음을 나타내며, 질의 사이트에 대한 판정결과가 매우 우수하다는 것을 의미한다. 따라서 프로파일이 질의 사이트의 내용을 대표할 수 있음을 알 수 있다.

(그림 15), (그림 16) 및 (그림 17)은 본 논문에서 제안한 검색방법을 통한 부정(negative)에러집합의 포함율을 나타내고 있다. 구글(google)에서 질의 사이트의 대상영역 프로파일의 키워드들로 검색을 수행하여 검색한 결과에서 원치않는 결과 사이트를 추출하여 이를 부정에러집합으로 설정하고, 검색 결과에서 부정에러집합의 포함율을 측정하였다. (그림 15), (그림 16), (그림 17)은 *DOMAIN*1과 *DOMAIN*2에서 차례로 내용영역 프로파일, 대상영역 프로파

일, 내용영역-대상영역 프로파일에 대한 부정에러집합의 실험결과를 나타낸다. 에러 판정비율이 높을수록 많은 부정에러집합을 판정하여 제외시킨 것이기 때문에 좋은 결과이다. (그림 15)에서 내용영역 프로파일의 경우 패턴 지지도(*PS*)가 키워드 기반 판정방법보다 매우 좋은 결과를 나타내고 있다. (그림 16)의 대상영역 프로파일의 경우 패턴 지지도가 키워드 기반 판정방법보다 조금 결과가 떨어지나 전체적인 결과를 나타내는 (그림 17)의 내용영역-대상영역 프로파일의 경우 좋은 결과를 나타낸다. 또한 키워드 기반 판정방법이 프로파일 기반방법보다 더 큰 기울기를 나타낸다. 따라서 프로파일 기반방법은 매치도 결과가 기준값으로부터 멀리 떨어져 있다는 것을 알 수 있으며, 이는 부정에러집합의 매치도가 매우 낮다는 것을 의미한다. 따라서 프로파일 기반의 검색결과는 많은 양의 불필요한 부정에러집합을 판정하여 제외시키기 때문에 효율적임을 알 수 있다.

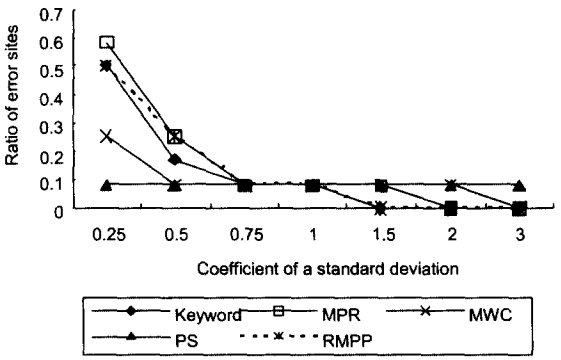


(a) DOMAIN 1

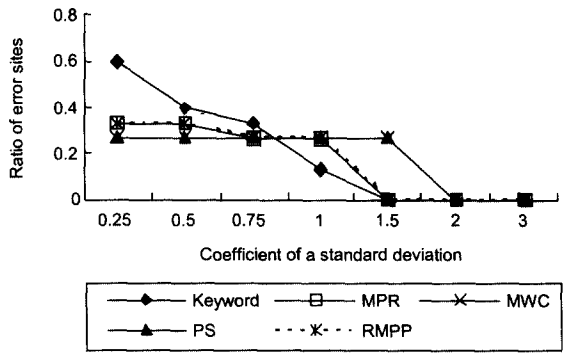


(b) DOMAIN 2

(그림 12) 내용영역 프로파일 결과

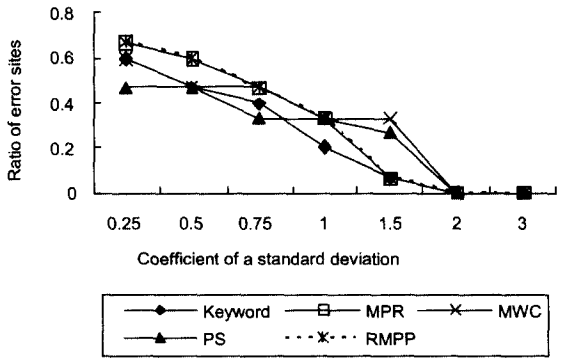


(a) DOMAIN 1

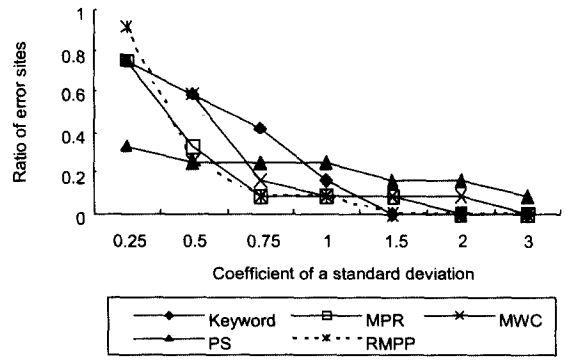


(b) DOMAIN 2

(그림 13) 대상영역 프로파일 결과

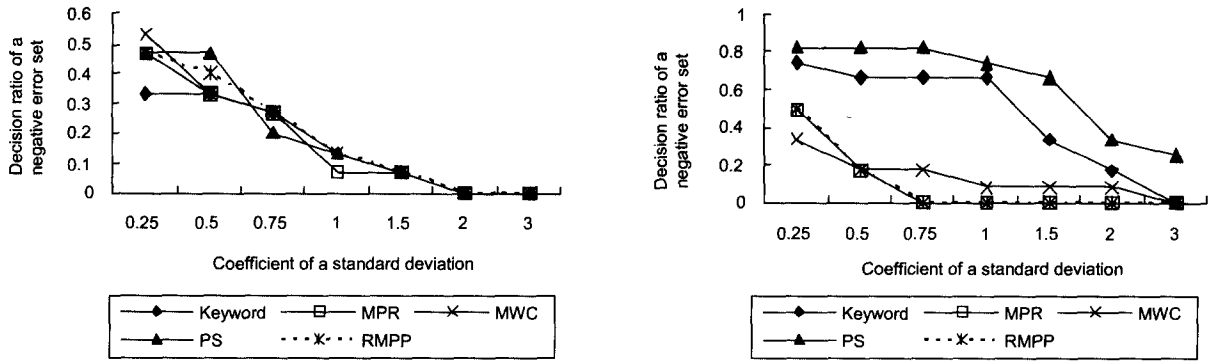


(a) DOMAIN 1



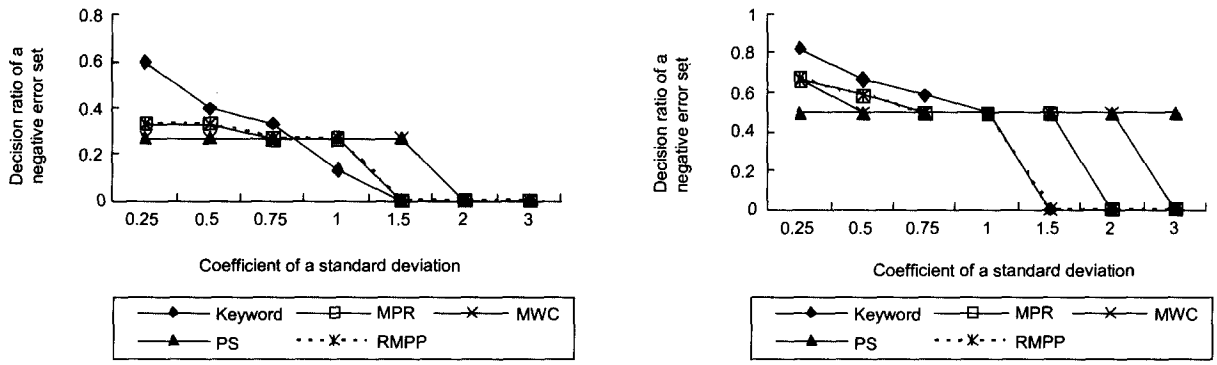
(b) DOMAIN 2

(그림 14) 내용영역-대상영역 프로파일 결과



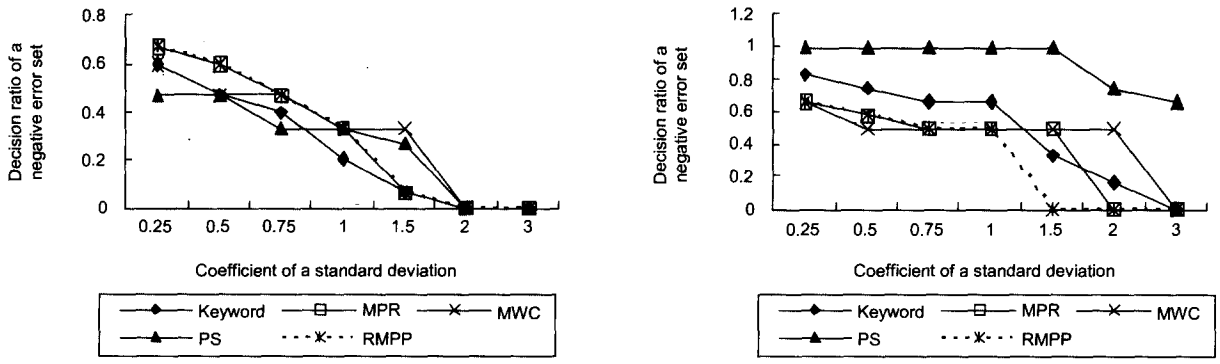
(a) DOMAIN 1 (b) DOMAIN 2

(그림 15) 내용영역 프로파일의 부정에러집합 판정비율



(a) DOMAIN 1 (b) DOMAIN 2

(그림 16) 대상영역 프로파일의 부정에러집합 판정비율



(a) DOMAIN 1 (b) DOMAIN 2

(그림 17) 내용영역-대상영역 프로파일의 부정에러집합 판정비율

7. 결 론

본 논문에서는 키워드를 추출하기 위하여 기존의 빈도 기반 키워드 추출 방법을 바탕으로 하는 스타일 기반 키워드 추출방법을 제안하였다. 스타일 기반 키워드 추출방법은 문서의 스타일을 분석하여 이를 기반으로 중요도를 판정하고 가중치를 부여한다. 스타일 기반 키워드 추출방법과 출현회수에 기반한 키워드 추출 방법을 양적인 면과 질적인 면에서 비교하여 스타일 기반 키워드 추출방법이 더 우수함을 증명하였다. 실험을 통하여 키워드를 추출할 때 양적으로 많은 키워드를 추출하는 것보다 질적으로 우수한 키워드를 추출하는 것이 정확성이 높기 때문에 스타일 기반 키워드 추출 방법은 매우 우수한 방법임을 증명하였다. 또한 스타일 기반의 키워드 추출 방법을 기반으로 내용영역 정보를 고려한 프로파일을 생성하여 키워드 마이닝 프로파일 기반의 웹 검색 시스템을 제안하고 구현하였다. 이때 질의는 단

어기반의 질의가 아닌 예제 기반 질의로 검색을 수행할 수 있도록 하였으며 이를 손쉽게 정의하도록 하였다. 키워드 마이닝 프로파일의 유효성을 검증하기 위하여 질의에 사용된 사이트들을 대상으로 비교 실험을 수행하여 키워드만을 사용하여 판정하는 방법보다 우수하다는 것을 증명하였다.

참 고 문 헌

[1] E. shakshuki and H. Ghenniwa, "A multi-agent system architecture for information gathering," *Database and Expert Systems Applications, Proceedings, 11th International Workshop on*, pp.732-736, 2000.

[2] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, "Modern Information Retrieval," ADDISON WESLEY, pp.29-30, 1999.

[3] I. Aalbersberg, "A Document Retrieval Model Based on Term Frequency Ranks," *17th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.163-172, 1994.

[4] Amit Singhal, Chris Buckley and Mandar Mitra, "Pivoted Document Length Normalization," *Proceedings of 19th ACM International Conference on Research and Development in Information Retrieval*, 1996.

[5] Cazalens S., Desmontils S., Jacquin C. and Lamarre P., "A Web site indexing process for an Internet information retrieval agent system," *Web Information Systems Engineering 2000, Proceedings of the First International Conference on*, Vol.1, pp.254-258, 2000

[6] M. Scmidt and U. Ruckert, "Content-based information retrieval using an embedded neural associative memory," *Parallel and Distributed Processing 2001 Proceedings, Ninth Euromicro Workshop on*, pp.443-450.

[7] Weifeng Li, Baowen Xu, Hongji Yang, Cheng-Chung Chu W. and Chih-Wei Lu at Dept. of Compt. Sci. & Eng. Southeast Univ., Nanjing, China, "Application of genetic algorithm in search engine," *Multimedia Software Engineering, Proceedings, International Symposium on*, pp. 366-371, 2000.

[8] R. Weiss, B. Velez, M. Sheldon, C. Nemprempre, P. Szilagyi and D. K. Gifford, "HyPursuit : A hierachical Network engine that exploits content-link hypertext clustering," In *Proc. Of the 7th ACM Conference on Hypertext and Hypermedia*, Washington, DC, USA, pp.180-193, 1996.

[9] A. Broder, S. Glassman, M. Manasse and G. Zweig, "Syntactic clustering of the web," In *6th Int. WWW Conference*, Snata Clara, CA, USA, pp.391-404, April, 1997.

[10] C-H. Chang and C-C. Hsu, "Customizable mulit-engine search tool with clustering," In *6th Int. WWW Conference*, Santa Clara, Ca, USA, April, 1997.

[11] Jiawei Han, "Data Mining," Encyclopedia of Distributed Computing, Kluwer Academic Publisher.

[12] R. Agrawal and R. Srikant, "Mining association rules betweensets of items in large databases," Proceeding of the ACM SIGMOD Conference on Management of Data, Washington, D.C., pp.207-216, May, 1993.

[13] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," In Proceedings of the 20th VLDB Conference, Santiago, Chile, Sept., 1994.

[14] J. S. Park, M-S. Chen and P. S. Ui, "An effective hash-based algorithm for mining association rules," In Proceedings of ACM SIGMOD Conference on Management of Data, San Jose, California, pp.175-186, May, 1995.

[15] A. Savasere, E. Omiencinsky and S. Navathe, "An efficient algorithm for mining association rules in large databases," In Proceedings of the 21th VLDB Conference, Zurich, Swizerland, pp.432-444, 1995.

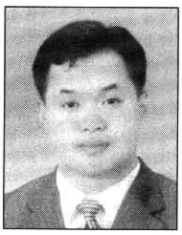
[16] J. S. Park, P. S. Yu and M.-S. Chen, "Mining Association Rules with Adjustable Accuracy," In Proceedings of ACM CIKM '97, Las Vegas, Nevada, pp.151-160, November, 1997.

[17] S. Brin, R. Motwani, J. D. Ullman and S. Tsur, "Dynamic itemset Counting and Implication Rules for Market Basket Data," In Proceedings of ACM SIGMOD Conference on Management of Data, Tucson, Arizona, pp.255-264, May, 1997.

[18] S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus and P. Morarescu, "FALCON : Boosting Knowledge for Answer Engines," In the Proceedings of Text REtrieval Conference (TREC-9), 2000.

[19] S. Alpha, P. Dixon, C. Liao, "Oracle at TREC 10," In the Proceedings of Text REtrieval Conference (TREC 2001), 2001.

[20] E. Hovy, U. Hermjakob, C-Y Lin, "The Use of External Knowledge in Factoid QA," In the Proceedings of Text REtrieval Conference (TREC 2001), 2001.



주 길 홍

e-mail : faholo@amadeus.yonsei.ac.kr

1998년 인천대학교 전자계산학과(공학사)

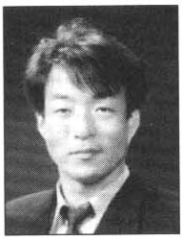
2000년 연세대학교 컴퓨터과학과(공학석사)

2004년 연세대학교 컴퓨터과학과(공학박사)

관심분야 : 분산데이터베이스, 미디에이터시스템, 분산질의처리, 질의최적화, 웹 데이터마이닝

이 준 휘

e-mail : lee@amadeus.yonsei.ac.kr
1999년 연세대학교 컴퓨터과학과(공학사)
2000년 연세대학교 컴퓨터과학과(공학석사)
2002년 소프트그램 기술연구소
관심분야 : 웹 데이터마이닝, 분산질의처리, 분산데이터베이스, 웹 데이터 추출 및 관리



이 원 석

e-mail : leewo@amadeus.yonsei.ac.kr
1985년 미국 보스턴대학교 컴퓨터공학과 (공학사)
1987년 미국 퍼듀대학교 컴퓨터공학과 (공학석사)
1990년 미국 퍼듀대학교 컴퓨터공학과 (공학박사)
1990년~1992년 삼성전자 선임연구원
1993년~1999년 연세대학교 컴퓨터과학과 조교수
1999년~현재 연세대학교 컴퓨터과학과 부교수
관심분야 : 분산데이터베이스, 미디에이터시스템, 데이터마이닝, 침입탐지, 멀티미디어데이터베이스