정보검색에서 웹마이닝을 이용한 동적인 질의확장에 관한 연구

황 인 수*

A Study on Dynamic Query Expansion Using Web Mining in Information Retrieval

Insoo Hwang*

Abstract

While the WWW offers an incredibly rich base of information, organized as a hypertext, it does not provide a uniform and efficient way to retrieve specific information. When one tries to find information entering several query terms into a search engine, the highly-ranked pages in the result usually contain many irrelevant or useless pages. The problem is that single-term queries do not contain sufficient information to specify exactly which web pages are needed by the user.

The purpose of this paper is to describe the employment of association rules in data mining for developing networks and computing associative coefficient among the terms. And this paper shows how the dynamic query expansion and/or reduction can be performed in information retrieval.

Keywords : information Retrieval(IR), Data Mining, Association Rule, Query Expansion

논문접수일 : 2003년 11월 27일 논문게재확정일 : 2004년 6월 15일

※ 2004년도 전주대학교 해외파견 연구비 지원에 의해 의한 것임.

* 전주대학교 정보기술공학부 부교수

1. 서 론

인터넷을 위한 인프라의 구축이 확대되고 인터넷의 활용이 생활화됨에 따라 웹이나 전자우편을 통해 엄청난 양의 정보가 제공되고 있다. 그러나 정보의 양이 증가할 수록 개인이 필요로 하는 정보를 검색하는 것은 더 많은 시간과 노력을 필요로 한다. 따라서 인터넷에서 정확한 정보를 신속하게 찾아서 제공하는 정보검색에 대한 관심이 점점 더 증대되고 있다. 최근에는 유사문서 검색, 문서파일 검색 등 다양한 기능을 부가한 검색엔진이 운영되고 있다.

일반적으로 정보검색을 지원하는 시스템은 비교적 안정적인 상태를 유지하는 수억 건 이상의 웹문서들을 데이터베이스화한 후, 사용자가 입력한 질의(query)에 대해 유사도가 높은 문서들을 제시한다. 이 때, 정보검색의 속도를 향상시키기 위해 문서에 포함된 주요 색인어들을 추출하여 역파일의 형태로 관리하기 때문에, 하나의 문서는 여러 가지의 질의에 의해 검색될 수 있으며, 하나의 질의에 대해 많은 문서가 검색될 수 있다.

사용자가 정보를 검색하기 위해 입력하는 질의어는 하나 혹은 몇 개의 용어(term)로 이루어져 있기 때문에 검색할 내용을 대표하는 질의어를 찾는 것은 용이한 일이 아니다. 문자적으로 표현된 질의어가 갖는 모호성으로 인해 잘 못된 검색결과를 도출하는 경우도 흔히 발생한다. 이에 따라, 사용자의 질의에 보다 적합한 문서를 검색하도록 질의를 구체화하는 용어를 제안하거나 혹은 자동적으로 부가하여 질의하는 질의확장(query expansion)에 대한 연구가 활발히 진행되고 있다.

질의를 확장하는 방법으로서 용어간의 관계를 나타내는 시소러스를 이용하는 방법, 문서의 주요 색인어를 추출한 후 색인어가 동일한 문서에 출현하는 빈도에 따라 색인어간의 유사도 행렬을 구축하는 방법, 그리고 질의어에 적합한 문서에 포함된 용어를 추가하는 방법 등이 있다.

첫째, 시소러스(thesaurus)에서는 용어간의 관계에 따라 어휘사전을 구축한 후, 동의어나 관련어를 질의어에 추가한다. 그러나 이를 위해서는 먼저 시소러스를 구축해야 하는데, 시소러스는 컴퓨터에 의해 자동적으로 구축되기 어렵다는 문제가 있다. Kristensen[1993] 등은 특정 영역(domain)에 대해 수작업으로 시소러스를 구축하여 활용하는 방안을 연구하였으나, Voorhees & Hou[1993]의 연구에서는 질의의 종류에 따라서 검색성과가 오히려 저하될 수도 있는 것을 나타났다.

둘째, 색인어를 클러스터링하는 방법에서는 색인어들이 동일한 문서에 출현하는 빈도, 즉 동시 출현빈도(co-occurrence frequency)를 이용하여 색인어-색인어의 유사도 행렬을 구성한 후, 클러스터링 알고리즘을 적용하여 유사한 용어를 클러스터링한다. Peat & Willett[1991]의 연구에 따르면 클러스터링 방법은 정보검색 성과의 향상에 크게 기여하지는 못하는 것으로 나타났다.

셋째, 적합도 피드백(relevance feedback) 방법에서는 검색결과에 대한 사용자의 피드백을 이용하여 문서로부터 새로운 질의어를 추출하여 질의에 추가한다[Buckley *et al.*, 1994]. Harman[1992]과 Salton & Buckley[1990]의 연구 등에서 적합도 피드백이 정보검색의 성과를 향상시키는 것으로 나타났다. 그러나 질의에 적합한 문서가 존재하지 않을 경우에는 질의확장이 불가능하며, 부적합한 주제를 선정한 경우에는 질의확장의 성과가 현저히 저하되는 것으로 나타났다.

기타로, Porter[1980]와 Lovins[1968]는 질의

에 포함된 형용사를 명사형으로 전환하여 질의에 사용하는 어근화(stemming) 알고리즘이 정보검색의 성과를 향상시킬 수 있음을 제안하였으나, Hull[1996]과 Keen[1992] 등의 연구에서는 어근화가 성과향상에 영향을 미치지 못하는 것으로 나타났다.

이에 따라 본 연구에서는 비교적 잘 정리된 사전(dictionary), 기사(article), 혹은 웹문서 등 학습을 위한 예제문서로부터 색인어를 추출하여 데이터마이닝의 연관규칙(association rule)에 따라 용어간의 연관계수를 계산하여 색인어 네트워크를 구축한 후, 사용자의 질의에 대한 일반화 및 상세화 질의어를 동적으로 생성하는 방법을 제안한다. 이것은 색인어를 클러스터링하는 방법과 시소러스를 구축하는 방법을 결합한 것으로서, 연관분석을 통해 방향성 그래프의 형태를 갖는 시소러스가 자동적으로 구축되는 장점이 있다. 연관분석의 지지도와 신뢰도를 가중치로 활용할 경우, 확장질의어들의 가중치를 계산할 수 있기 때문에 보다 적합한 확장질의어를 선택함으로써 문서검색의 성과를 크게 향상시킬 수 있을 것으로 기대된다.

본 논문의 구성은 다음과 같다. 제2장에서는 본 연구에서 적용할 데이터마이닝의 연관분석에 대해 기술하며, 제3장에서는 연관분석을 이용한 질의확장 방법을 기술한다. 다음으로, 제4장에서는 질의확장을 위한 검색시스템의 구조 및 실행 예를 기술하고, 제5장에서는 본 연구를 요약하며 향후 연구계획을 제시한다.

2. 연관분석을 이용한 연관계수 계산

2.1 연관계수

두 용어간의 연관성은 각 용어의 개별적인 출현빈도와 동시 출현빈도를 이용하여 계산하는 연관계수로 측정된다. 두 용어의 동시 출현빈도만을 연관계수로 사용하면, 많이 사용되는 용어일수록 동시에 나타나는 빈도가 높아져서 높은 연관계수를 갖게 되므로 연관계수의 값을 상대적으로 비교하기가 어렵다[정영미 외, 1998].

따라서 용어의 개별적인 출현빈도와 전체 용어의 빈도를 함께 이용하여 용어간의 통계적인 연관성을 객관적으로 평가하는 상대 동시출현빈도 방식의 연관계수가 주로 사용된다. 연관계수를 측정하는 방법에는 여러 가지가 있으나, 주로 정보이론에 근거하는 상호정보량과 상대 엔트로피가 사용된다. 상호정보량은 두 독립사건의 확률변수 x와 y 사이의 의존관계를 정량적으로 나타낸 것으로서 다음 계산식에 따라 계산된다.

$$MI(x, y) = \log_2 \frac{p(x, y)}{p(x) \times p(y)}$$

상호정보량은 두 확률이 완전히 독립일 경우 0이 되고, 의존관계가 깊을 수록 큰 값을 가지며, $MI(x, y) = MI(y, x)$의 대칭성을 갖는다. 연관성 분석에서 상호정보량을 이용할 때의 문제점으로는 빈도가 낮은 용어간의 상호정보량이 빈도가 높은 용어간의 상호정보량보다 상대적으로 과대 평가되는 경향이 있다.

상대 엔트로피는 두 확률분포 $p(x)$와 $q(x)$ 사이의 평균적인 차이를 측정하는 것으로서, KL 거리(Kullback-Leiber Distance: D_{KL}) 혹은 교차 엔트로피(cross entropy)라고도 한다[정석경, 1997]. 상대 엔트로피는 다음 계산식에서 보는 바와 같이, 항상 0보다 크거나 같으며 두 확률이 일치할 경우에만 0이 되고 대칭성을 만족하지 않는다.

$$D_{KL}(p(x)||q(x)) = \sum_x p(x)\left[\log \frac{1}{q(x)} - \log \frac{1}{p(x)}\right]$$

2.2 연관분석의 개념

연관분석은 Agrawal et al.[1993]이 제안한 대표적인 데이터마이닝 기법의 하나로서, 장바구니 분석을 통한 상품추천 등에 광범위하게 사용되고 있다. 연관분석은 거래내역을 분석하여 각 거래에 동시에 포함되는 상품들을 연관규칙으로 표현하는 것으로서, 연관규칙에 관한 기존의 연구를 정리하면 다음과 같다.

정의1) k개의 항목으로 구성된 집합을 $I=\{i_1, i_2, \ldots, i_k\}$라고 하고, I로부터 임의로 n개의 항목을 선택하여 구성한 집합을 $B=\{b_1, b_2, \ldots, b_n\}$라고 하면, B의 부분집합은 $b_i \subseteq I$가 되며, 이를 장바구니(basket)라고 한다.

정의2) $i_1 \Rightarrow i_2$의 연관규칙이 존재하기 위해서는 다음의 조건을 만족해야 한다.

① n개의 장바구니 중에서 항목 i_1과 i_2를 모두 포함하는 장바구니가 최소지지도(%) 이상 존재해야 한다.

② 항목 i_1을 포함하는 장바구니의 최소신뢰도(%) 이상이 항목 i_2를 포함하고 있어야 한다.

위의 정의2)에서 항목 ①은 지지도(support)라고 하며, 항목 ②는 신뢰도(confidence)라고 한다. 지지도는 하나의 상품 혹은 일련의 상품이 전체거래에서 차지하는 비율로서 $supp(i_1)$ 혹은 $supp(i_1 \Rightarrow i_2)$ 등으로 표현한다. 신뢰도는 상품 i_1를 포함하는 거래에 상품 i_2가 동시에 포함되는 조건부 확률로서 $supp(i_1 \Rightarrow i_2)/supp(i_1)$를 의미하는 $conf(i_1 \Rightarrow i_2)$로 표현한다.

정의3) 상품 i_1과 i_2간의 향상도(lift) 혹은 관심도(interest)는 다음과 같이 계산된다 [Silverstein et al., 1998].

$$I(i_1, i_2) = \frac{supp(i_1 \Rightarrow i_2)}{supp(i_1)supp(i_2)} = \frac{p(i_1 i_2)}{p(i_1)p(i_2)}$$

향상도는 상품 i_2가 상품 i_1과 함께 구매된 거래와 상품 i_2가 상품 i_1에 관계없이 단독으로 구매된 거래의 비율이다. 따라서, 향상도가 1이라는 것은 상품 i_2가 상품 i_1에 영향을 받아서 구입되는 비율과 상품 i_2가 상품 i_1과 관계없이 단독으로 구입되는 비율이 같음을 의미하기 때문에 이들 두 상품은 독립(independence) 상품이다. 또한, 향상도가 1보다 크면 교차구매의 효과를 있음을 의미하기 때문에 보완재(complementarity)의 성격을 가지며, 향상도가 1보다 작으면 상품 i_1의 구매가 상품 i_2의 구매를 오히려 감소시키기 때문에 대체재(substitutability)의 성격을 갖는다[Brijs *et al.*, 1999].

2.3 연관분석을 이용한 연관계수의 계산

위에서 기술한 연관분석은 하나의 거래에 서로 다른 제품이 동시에 포함되는 확률을 나타낸다. 따라서, 정보검색의 대상인 각 문서를 하나의 거래로 보고, 각 문서에 포함된 색인어를 제품으로 취급할 경우, 연관분석을 정보검색문제에 그대로 적용할 수 있다.

일반적인 연관분석에서는 2개 이상의 제품이 동시에 연관규칙을 형성하지만, 정보검색에서는 두 단어간의 연관관계만을 고려하면 되기 때문에 연관분석의 복잡도가 현저히 감소된다. 뿐만 아니라, 연관분석에서 항목간의 신뢰도가 비대칭적이라는 특성은 질의확장의 방향을 결정하기 위해 사용될 수 있다.

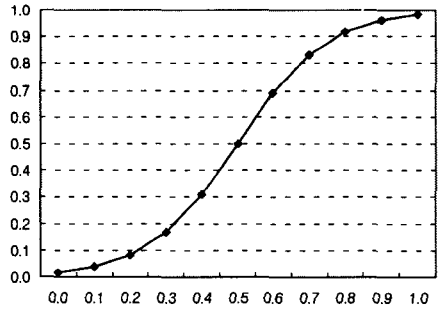
본 연구에서는 예제 문서로부터 색인어를 추

출하여 색인어간의 연관규칙을 생성한 후, 이를 방향성 네트워크로 구성하여 질의확장에 이용하는 방안을 제시한다. 연관규칙을 이용하여 구성한 네트워크에서 특정 노드에 가중치를 부여하면 신뢰도에 따라 연관된 노드로 가중치(연관계수)를 전파하므로, 각 용어들의 연관계수가 동적으로 계산된다.

다음은 임의의 용어 x로부터 관련용어 y로 가중치를 전파하기 위해 이들간의 신뢰도 $conf(x \Rightarrow y)$를 이용하는 예를 보여주고 있다. 이 때, 용어 x와 y간의 신뢰도가 최소신뢰도 $C_{\min}$보다 큰 값들로부터 계산된 값들중에서 최대값을 취하였다.

$$c_y = Max\left(c_x * \frac{1}{1+e^{(\mu - conf(x \Rightarrow y)) * k}}\right) \quad \text{for all } x \text{ where } conf(x \Rightarrow y) > C_{\min}$$

위에서 μ는 시그모이드(sigmoid) 함수의 값을 조정하기 위해 사용하는 오프셋(offset)이며, k는 신뢰도를 시그모이드 함수의 효과를 조정하기 위한 임의의 상수이다. 본 연구에서는 μ를 신뢰도의 평균값인 0.5로 설정하였으며, k는 임의로 8.0으로 설정하였다. (그림 1)은 신뢰도가 0에서부터 1까지 변할 때 시그모이드 함수의 값을 그래프로 나타낸 것이다.



(그림 1) 시그모이드 함수의 계산 결과

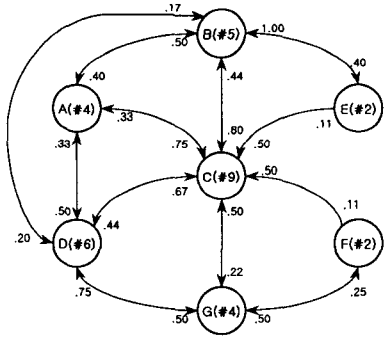
네트워크로 전파되는 값을 정확하게 계산하기 위해서는 자기호출(recursion)을 이용하는 것이 일반적이지만, 네트워크의 규모가 클 경우에는 계산에 많은 시간이 소요되는 단점이 있다.

따라서, 본 연구에서는 특정 노드에 부과한 가중치를 네트워크의 전방향으로 전파(forward propagation)시킨 후, 전파되는 가중치의 값이 일정한 값(e)보다 작을 경우 전파를 종료하도록 함으로써 계산소요시간을 단축하였다.

3. 연관규칙을 이용한 질의확장

3.1 연관규칙 네트워크 구성

데이터마이닝의 연관분석을 통해 문서에 포함된 각 용어간의 지지도와 신뢰도가 계산되면, 이를 이용하여 방향성 네트워크(directed network)를 구성할 수 있다. (그림 2)는 임의로 구성한 문장에서 A~G 등 7개의 용어를 추출하여 각 용어간의 연관분석을 실시한 결과를 그림으로 표현한 것이다.



(그림 2) 용어간의 연관분석 결과

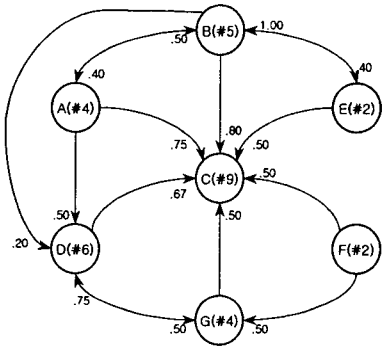
여기서, 괄호 안의 숫자는 용어의 출현횟수이며, 화살표에 부가된 숫자는 용어간 연관규칙의 신뢰도이다. 즉, Term A와 B는 각각 4회와 5회

출현되는데, 문서에 Term A가 있으면서 B가 동시에 존재할 확률이 50%이며, 반대로 Term B가 있으면서 A가 동시에 존재할 확률은 40% 임을 나타낸다.

3.2 질의확장의 방향성

질의확장은 사용자가 입력한 검색어의 동의어 혹은 관련어를 추가하는 것을 의미한다. 그러나 본 연구에서는 연관규칙의 비대칭성을 이용하여 주어진 질의어 보다 일반적인 개념의 용어들로 확장하는 일반화(generalization)와 특정 분야의 용어들로 질의의 영역을 축소하는 상세화(specialization)의 두 가지로 분류하였다.

연관규칙에서 용어간에 상관관계가 있기 위해서는 리프트가 1보다 커야하므로, (그림 2)에서 리프트가 1보다 작은 링크를 제거하면 (그림 3)과 같은 방향성 그래프를 생성할 수 있다.

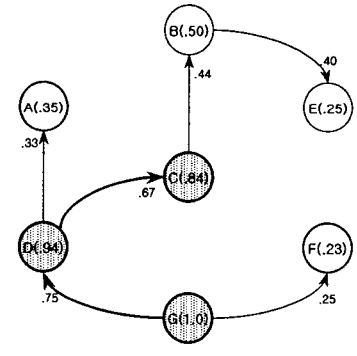


(그림 3) 연관분석을 통한 용어간의 방향성 설정결과

여기서, 대부분의 Term들은 단방향 그래프의 형태로 연결되어 있으나, Term A와 B는 양방향으로 연결되어 있음을 볼 수 있는데, 이는 Term B의 경우 A의 일반화 및 상세화 확장에 모두 사용될 수 있음을 의미한다. 본 연구에서는 Term 간의 리프트를 이용하여 방향성을 도출하였으나, Term간의 신뢰도를 이용하여 신뢰도가 일정한 값보다 크거나 혹은 양방향중에서 신뢰도의 값이 큰 방향을 일반화의 방향으로 설정할 수도 있다.

3.3 일반화 질의확장

앞의 네트워크에서 Term G의 가중치를 1.0으로 설정한 후 네트워크의 각 Term으로 가중치를 전파하면 (그림 4)의 결과를 얻는다. 여기서, Term G에 부과된 가중치는 연관분석에서 도출한 각 용어간의 신뢰도 수준에 따라 네트워크의 F와 D로 전파되며, D로 전파된 가중치는 다시 A와 C로 전파되고, C에서는 B와 E로 순차적으로 가중치를 전파한다.



(그림 4) 신뢰도를 이용한 Term간의 가중치 전파결과

Term D와 C로의 경로는 (그림 3)에서 제시한 Term을 일반화하는 경로에 일치하지만 B와 E, 그리고 A와 F의 경로는 방향성이 일치하지 않는다. 따라서, Term G를 일반화하는 질의확장으로 D와 C를 제안한다.

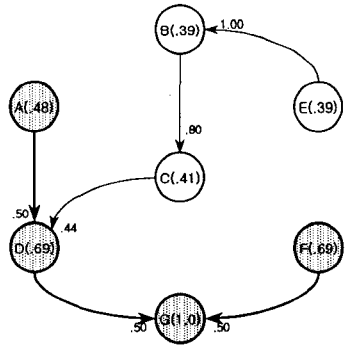
이 때, 문서의 적합도를 계산할 경우에는 추가된 용어의 가중치를 각각 0.94와 0.84로 부과한다. 이것은 검색에 사용한 Term G가 많이 사용되지 않는 국부적인 용어인 경우, 이를 포함

하는 D와 C 등으로 확장할 때 유용하게 사용될 수 있다.

3.4 상세화 질의확장

상세화는 일반화의 반대 개념으로서, 질의영역을 특정분야의 전문화된 용어로 축소하기 위해 도입한 개념이다. 본 연구에서는 연관분석에서 다른 용어로부터 자신이 신뢰되어지는 수준을 의미하는 역신뢰도(inversed confidence)의 개념을 도입하였다.

(그림 5)는 Term G의 가중치를 1.0으로 하고 역신뢰도에 따라 가중치를 각 Term으로 전파한 결과를 보여준다. 예를 들어, Term D 또는 F를 포함하는 문서의 50%는 G를 포함하고 있으므로, 이들은 0.69의 가중치를 갖는다.



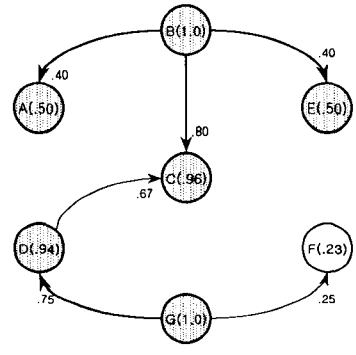
(그림 5) 역신뢰도를 이용한 Term간의 가중치 전파결과

여기서, Term C는 D에 대해 일반화의 방향성을 갖고 있으므로, C를 상세화하는 항목에서는 제외한다. 따라서, G를 상세화하는 질의확장으로 F, D, A를 제안한다.

3.5 다중 질의어에 대한 질의확장

문서를 검색하기 위해 2개 이상의 질의어를 사용할 경우에는 각 질의어에 할당된 가중치를 네트워크의 모든 용어로 전파해야 한다. 이 경우에는 순차적으로 각 용어의 가중치를 1.0으로 설정하여 전파하면서 시그모이드 함수로부터 계산된 최대값(Max)을 취한다.

(그림 6)은 Term B와 G에 대해 질의를 확장한 결과를 그림으로 보여주고 있는데, (그림 3)에서 제시한 방향성에 적합한 용어는 Term B로부터 확장한 C, A, E와 Term G로부터 확장한 D 등 4개이다. 물론, 확장을 하기 위해 최소한으로 요구되는 가중치 혹은 확장할 Term의 개수를 미리 설정할 수도 있다.



(그림 6) Term B와 G에 대한 질의확장의 예

여기서, Term C는 B로부터 확장되었으나, Term G로부터 D를 거쳐 C로 확장하는 경로와도 일치한다. 따라서, 다중검색어로 지정한 Term B와 G는 유의한 다중검색어 조합임을 알 수 있다. 만일, 다중검색어로 지정한 Term들이 방향성 네트워크에서 접속되지 못한다면, 전혀 관련이 없는 단어들을 다중검색어로 지정한 것으로 판단할 수 있다.

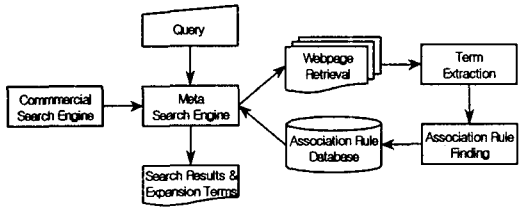
4. 질의확장을 위한 검색시스템 구축

4.1 검색시스템의 구조

인터넷 혹은 검색엔진의 데이터베이스에 존

재하는 모든 웹문서에서 용어를 추출하여 용어간의 연관규칙을 데이터베이스로 구축하는 것은 현실적으로 거의 불가능한 일이다. 따라서, 본 연구에서는 검색의 개인화(personalize)에 초점을 맞추어 검색된 결과로부터 용어간의 연관규칙을 설정함으로써, 검색횟수가 증가함에 따라 관심영역(domain)의 연관규칙 데이터베이스를 자동적으로 구축하는 방안을 제안한다. 향후에는 잘 구축된 영역별 데이터베이스를 공유하거나 혹은 통합하여 운영하는 것도 가능할 것이다.

본 연구에서 제안하는 질의확장 알고리즘을 적용하기 위해서는 검색엔진을 구축하고 인터넷으로부터 웹문서들을 추출하여 데이터베이스를 구축해야 한다. 그러나 이는 매우 방대한 작업으로서, 본 연구에서는 자체의 데이터베이스를 구축하지 않고 상용 검색엔진의 검색결과를 이용하는 메타검색(meta search) 방식을 택하였다. 이에 따라 개발한 질의확장을 위한 메타검색시스템의 구조를 그림으로 나타내면 (그림 7)과 같다. 본 시스템은 2.66GHz CPU를 갖는 윈도우즈 2000서버 운영체제에서 오라클을 데이터베이스로 하고, 프로그래밍은 객체지향언어인 자바(Java)를 이용하였다.



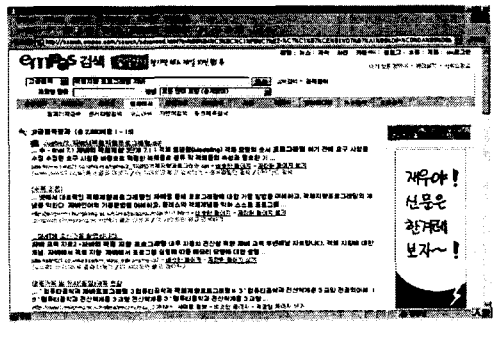
(그림 7) 질의어 확장 및 축소를 위한 검색시스템의 구조

Meta Search Engine 모듈은 사용자의 질의를 상용 검색엔진으로 보내어 검색한 결과와 함께 연관규칙 데이터베이스에서 검색한 확장질의어를 제공한다. Webpage Retrieval 모듈에서는 검색엔진의 검색결과를 전달받아, 각 URL의 컨텐츠를 검색하여 Term Extraction 모듈로 전송한다. Term Extraction 모듈은 웹문서의 텍스트로부터 색인어를 추출하는 기능을 하는데, 일반적으로는 텍스트로부터 색인어를 추출하는 HAM[강승식 외, 1996] 등이 이용된다.

그러나 HAM은 너무 많은 색인어를 추출할 뿐만 아니라, 색인어 추출은 본 연구의 범위를 벗어난다. 따라서, 본 연구에서는 특정 영역의 색인어 목록을 미리 만들어 놓은 후에 웹문서로부터 이들 용어만을 선별적으로 추출하는 방법을 사용하였다. 끝으로, Association Rule Finding 모듈은 각 웹문서로부터 추출한 용어들간의 연관규칙을 생성하여 데이터베이스에 저장한다.

4.2 검색 및 색인어 추출

본 연구에서는 정보검색을 위한 상용 검색엔진으로 엠파스(http://www.empas.com)를 사용하였으며, 검색어로 "객체지향 프로그래밍 자바"를 지정하였다. 이에 따라 (그림 8)에서 보는 바와 같이 총 2,860개의 문서가 추출되었는데, 이중에서 상위 60개의 URL로부터 용어를 추출하여 연구에 사용하였다.



(그림 8) 검색엔진을 이용한 검색결과의 예

자바 객체지향 프로그래밍과 관련 있는 상위 60개의 웹문서로부터 관련 용어를 추출한 결과, 검색어로 제시한 "객체지향", "프로그래밍", 그리고 "자바"는 모든 문서에 포함되어 있었다.

또한 "프로그램", "데이터", 그리고 "컴퓨터" 등의 일반적인 용어들은 출현빈도가 높았으나, "썬", "JVM", "바이트코드" 등 자바의 특성을 나타내는 전문화된 용어들의 출현빈도는 5개 이하로 낮게 나타났는데, 이를 표로 정리하면 <표 1>과 같다.

〈표 1〉 주요 색인어의 출현빈도 현황

빈도	주요 색인어 목록	개수
51-60	객체지향,프로그래밍,자바	3
41-50	프로그램,데이터	2
31-40	컴퓨터,인터넷,소프트웨어	3
21-30	클래스,인터페이스,C++,상속,스레드	5
11-20	함수,변수,메소드,추상,포인터,모바일	6
6-10	속성,오버로딩,플랫폼,메시지, 오버라이딩,정보은닉,다형성	7
1-5	인스턴스,JVM,썬,바이트코드,생성자	5

4.3 연관분석을 이용한 질의확장

본 연구에서는 각 웹문서에 포함되어 있는 용어들의 목록을 장바구니로 가정하고 지지도와 신뢰도에 따라 연관규칙을 도출하였다. <표 2>는 질의어로 "메소드"를 지정할 경우 연관규칙으로 이루어진 용어간의 네트워크에 따라 가중치를 전파하여 제안된 일반화 질의어 및 상세화 질의어를 보여준다.

정보검색 결과의 정확도는 전적으로 사용자의 판단에 따르기 때문에 <표 2>에서 제시한 질의확장의 성과를 객관적으로 검증하는 것은 용이한 일이 아니다.

〈표 2〉 질의어 확장을 위한 용어의 예

우선순위	일반화 질의어		상세화 질의어	
	색인어	가중치	색인어	가중치
1	자바	0.982	오버라이딩	0.982
2	객체지향	0.982	오버로딩	0.961
3	프로그래밍	0.982	추상	0.791
4	클래스	0.971	상속	0.749
5	프로그램	0.953	속성	0.686
6	상속	0.953	포인터	0.671
7	인터페이스	0.924	클래스	0.648
8	데이터	0.790	메시지	0.618

다음의 몇 가지 관점에서 평가할 경우, 본 연구에서 제시한 질의확장이 바람직한 결과를 도출한 것으로 판단된다. 먼저, "자바", "객체지향", 그리고 "프로그래밍"은 정보검색에 사용한 질의어로서 모든 웹문서에 포함되어 있기 때문에 모든 용어들의 일반화 질의어로 사용될 수 있다. 따라서 <표 2>에서 이들이 일반화 질의어의 최우선 순위에 포함된 것은 올바른 결과이다.

다음으로 0.9 이상의 가중치를 갖는 "클래스", "프로그램", "상속", "인터페이스" 등과 "오버라이딩" 및 "오버로딩"은 각각 "메소드"의 상위개념과 하위개념으로 평가될 수 있기 때문에, 본 연구에서 제시하는 질의확장의 결과가 바람직한 결과를 도출하는 것으로 판단된다. 여기서 "상속"과 "클래스"는 질의어의 일반화 및 상세화에 모두 사용될 수 있는 것으로 제시되었으나 보다 큰 가중치를 갖는 일반화 질의확장으로 보는 것이 바람직하겠다.

5. 결론 및 향후 연구계획

본 연구에서는 비교적 잘 정리된 사전, 기사, 혹은 웹문서 등의 학습 예제로부터 색인어를 추출하여 데이터마이닝의 연관규칙에 따라 용어간의 신뢰도를 계산하여 네트워크를 구축한 후, 질의에 적합한 질의확장 용어를 동적으로 제시하는 방안을 제안하였다. 이 방법은 색인어를 클러스터링하는 방법과 시소러스를 구축하는 방법을 결합한 것으로서, 연관분석을 통해 방향성 그래프의 형태를 갖는 시소러스가 자동적으로 생성되는 장점을 갖게 된다.

뿐만 아니라, 연관분석의 지지도 및 신뢰도를 가중치로 활용할 경우, 질의어에 대한 확장질의어의 가중치를 계산할 수 있다. 색인어 네트워크가 잘 구축된 경우에는 질의어에 대한 확장질의 뿐만 아니라, 질의어의 주제를 특정분야로 좁혀 가는 축소질의를 지원하기 때문에 문서검색의 성과를 크게 향상시킬 수 있을 것으로 기대한다.

그러나 연관 네트워크 구성을 위한 좋은 예제가 존재하지 않거나 혹은 네트워크에 포함되지 않은 문제영역에 대한 질의에는 효과적인 질의확장이 이루어지지 못하는 단점이 있다. 따라서, 특정한 값 이상의 신뢰도를 갖는 질의확장이 이루어지지 못한 경우에는 적합도 피드백 등의 방법을 보완적으로 활용하는 방안에 대한 추가적인 연구를 진행하고자 한다.

참 고 문 헌

[1] 강승식, 장병탁, "음절 특성을 이용한 범용 한국어 형태소 분석기 및 맞춤법 검사기", *정보과학회 논문지(B)*, 제23권 5호, 1996, pp. 530-539.

[2] 정석경, *분포정보를 이용한 명사 소프트 클러스터링 연구*, 연세대학교 석사학위 논문, 1997.

[3] 정영미, 이재윤, "한국어 텍스트내 용어연관성 분석을 위한 기초연구," *제5회 한국정보관리학회 학술대회 논문집*, 이화여자대학교, 1998년 9월, pp. 243-246.

[4] Agrawal, T., Imielinski T., and Swami A., "Mining Associations between Sets of Items in Massive Databases," *Proceedings of the ACM SIGMOD International conference on Management of Data*, Washington D.C., May 1993, pp. 207-216.

[5] Agrawal, R., Mannila H., Srikant R., Toivonen H., and Inkeri Verkamo A., "A Fast Discovery of Association Rules," *Advances in Knowledge Discovery and Data Mining*, ed. U. Fayyad *et al.*, AAAI Press: Menlo Park, CA, 1996, pp. 307-328.

[6] Brijs, T., Swinnen G., Vanhoof K., and Wets G., "Using Association Rules for Product Assortment Decisions: A Case Study," *In Proceedings on KDD-99*, ACM, San Diego, CA, USA, 1999, pp. 254-260.

[7] Buckley, C., Salton G., and Allan J., "The Effect of Adding Relevance Information in a Relevance Feedback Environment," *Proceedings of 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Dublin, 1994, pp. 292-300.

[8] Bukley, C., Singhal A., Mitra M., and Salton G., "New Retrieval Approaches Using SMART: TEC 4," *Proceedings of 4th Text REtrieval Conference(TREC-4)*, 1995.

[9] Harman, D. "Relevance Feedback Revisited," *Proceedings of 15th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, 1992, pp. 1-10.

[10] Hull, D. A., "Stemming Algorithms: A Case Study for Detailed Evaluation," *Journal of the American Society for Information Science*, Vol. 47, No. 1, 1996, pp. 70-84.

[11] Keen, E. M., "Presenting Results of Experimental Retrieval Comparisons," *Information Processing and Management*, Vol. 28, No. 4, 1992, pp. 491-502.

[12] Kristensen, J., "Expanding End-Users' Query Statements for Free-text Searching with a Search-aid Thesaurus," *Information Processing and Management*, Vol. 29, No. 6, 1993, pp. 733-744.

[13] Lovins, J. B., "Development of a Stemming Algorithm," *Mechanical Translation and Computational Lingustics*, Vol. 11, 1968, pp. 22-33.

[14] Porter, M., "A Algorithm for Suffix Stripping," Program, Vol. 14, No. 3, 1980, pp. 130-137.

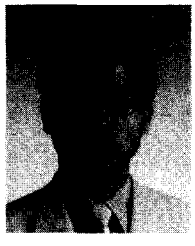
[15] Robertson, S. E. and Sparck-Jones K., "Relevance Weighting of Search Terms," Journal of th *American Society for Informational Science*, Vol. 27, 1976.

[16] Salton, G., *Automatic Text Processing*, Addison-Wesley, 1989.

[17] Salton, G. and Buckley C., "Improving Retrieval Performance by Relevance Feedback," *Journal of the American Society for Information Science*, Vol. 41, 1990, pp. 288-297.

[18] Silverstein, C., Brin S., and Motwani R., "Beyond Market Basket: Generalizing Association Rules to Dependence Rules," *Data Mining and Knowledge Discovery*, Vol. 2, 1998, pp. 38-68.

■ 저자소개



황 인 수

저자는 전주대학교 정보기술공학부 정보시스템 전공의 부교수로 재직중이다. 고려대학교 경영학과를 졸업하고 동 대학원에서 경영정보시스템을 전공하여 석사 및 박사학위를 취득하였으며, 산업연구원(KIET) 물류·유통연구센터의 연구원을 역임하였다. 주요 관심분야는 e-Business, CRM, 데이터마이닝, 웹 에이전트 등이다.

◈ 이 논문은 2003년 11월 27일 접수하여 1차 수정을 거쳐 2004년 6월 15일 게재확정되었습니다.