사용자 질의패턴 분석을 이용한 효율적인 확장검색어 추천시스템

김 영 안◆, 박 건 우°

An Efficient Extended Query Suggestion System Using the Analysis of Users' Query Patterns

Young-an Kim◆, Gun-Woo Park°

요 약

주요 검색엔진들은 확장 및 연관검색어를 추천하는 서비스를 제공함으로써 질의어 확장에 대한 사용자의 검색 편의성을 제공하고 있다. 하지만 많은 수의 사용자가 많이 찾는 검색어 즉, 대중성을 근거하여 제공되는 확장 및 연관검색어 추천 서비스는 사용자의 만족도를 높이는데 한계가 있다. 왜냐하면 사용자마다 생각하는 차이가 존재하며 선호하는 질의 및 관심 분야도 다르기 때문이다.

본 논문에서는 사용자의 정보요구에 적합한 효율적인 확장검색어를 추천하는 시스템을 설계 및 구현하고 웹 사용자의 정보검색 과정에서 최초 질의어 입력 후 질의어 확장 과정에서 사용자의 편의성을 향상시키고자 하였다. 평가결과 제안시스템은 검색엔진에서 추천하지 못한 구글 41% 및 야후 48%의 확장검색어를 추천할 수 있었으며 사용자의 편의성을 위하여 대중성 기반으로 추천되고 있는 확장 및 연관검색어 추천 서비스의 한계를 보완하여 사용자의 편의성을 향상시킬 수 있었다.

Key Words : Search engines, Related query, Extended query, Users' convenience, Query pattern

ABSTRACT

With the service suggesting additional extended or related query, search engines aim to provide their users more convenience. The extended or related query suggestion service based on popularity, or by how many people have searched on web using the query, has limitations to elevate users' satisfaction, because each user's preference and interests differ.

This paper will demonstrate the design and realization of the system that suggests extended query appropriate for users' demands, and also an improvement in the computing process between entering the first search word and the subsequent extension to the related themes. According to the evaluation the proposed system suggested 41% more extended or related query than when searching on Google, and 48% more than on Yahoo. Also by improving the shortcomings of the extended or related query system based on general popularity rather than each user's preference, the new system enhanced users' convenience further.

Ⅰ. 서 론

일반적으로 검색엔진 사용자들은 그들이 원하는 정보를 대표하는 핵심 키워드를 검색창에 입력하는 방법으로 정보를 검색하고 있으며, 사용자의 웹 검색 질의 유형은 내용검색, 사이트 검색, 서비스 검

◆ 주저자 : 국방대학교 국방과학학과 교수, roundsun@kndu.ac.kr, 정회원
° 교신저자 : 육군 종합보급창, pgw4050@emerald,yonsei.ac.kr, 정회원
논문번호 : KICS2011-11-516, 접수일자 : 2011년 11월 10일, 최종논문접수일자 : 2012년 4월 10일

색 중 하나이다[1].

그러나 일반적으로 사용자가 원활한 검색을 하기 위해서는 검색하는 분야의 정보에 대해 충분한 사전 지식의 파악이 필요하며, 키워드를 식별하는 능력 또한 뛰어나야 한다. 하지만 대부분의 사용자들이 충분한 지식을 갖추고 검색하는 것은 아니며 질의어를 직접 선정하여 입력해야 하는 문제점과 질의어에 적합한 효율적인 키워드를 연상하는데 많은 시간을 소비하는 문제점 등이 나타난다.

이와 같이 질의어는 사용자의 정보요구를 만족시킬 수 있는 최초의 시발점이 되며 정확한 검색 결과를 제공하는 판단의 기준이 되는데 사용자의 질의어는 단일단어 또는 여러 개의 단어와 연산을 복잡하게 결합한 형태일 수도 있다. 선행연구에서 사용자가 입력하는 질의어의 길이는 일반적으로 1개이다. 그러나 단일단어를 사용하여 입력하는 키워드는 사용자마다 서로 다른 의미를 가질 수 있는 의미의 모호성 또는 다중성 문제가 존재한다[2]. 질의어는 사용자의 정보요구를 만족시킬 수 있는 정확한 검색 결과를 제공하기 위한 판단의 기준이 되며 평균적인 질의의 길이는 2.21개 정도이고 사용자들이 입력하는 키워드의 길이가 더욱 짧아지고 있으며 함축적인 의미를 포함하는 경향이 있다[3]. 이와 같이 제한된 환경에서 개인의 필요를 충족시키기 위해 주로 사용되는 방법은 질의에 단어를 추가하여 질의어의 정확을 높이는 방법과 사용자가 클릭한 데이터와 같은 로그데이터를 통해서 적합성을 피드백 하는 방법들이 연구되어 왔다[4].

현재 주요 검색엔진들은 확장 및 연관검색어를 추천하는 서비스를 제공하여 질의어 확장에 대한 사용자의 검색 편의성을 제공하고 있다. 이런 서비스를 통하여 다양한 분야의 정보를 검색 할 수 있도록 함으로써 사용자의 검색 만족도를 높일 수 있다. 하지만 많은 수의 사용자가 많이 찾는 검색어 즉, 대중성에 근거하여 제공되는 추천 서비스는 사용자의 만족도를 높이는데 한계가 발생한다. 왜냐하면 사용자마다 생각하는 차이가 존재하며 선호하는 질의 및 관심 분야도 다르기 때문이다.

본 논문은 정보검색 과정 중 입력되는 질의어를 분석하여 질의패턴을 추출하고, 이를 기반으로 사용자의 정보요구에 적합한 확장검색어를 추천하는 방안을 제시한다. 사용자 질의어 데이터는 AOL에서 공개된 데이터를[5] 연관규칙을 적용하여 대중성 기반의 질의패턴을 분석하고 대중이 선호하는 질의어를 분석한다. 반대로 개별 사용자의 질의패턴을 분석하여 대중성 기반의 제한 사항을 보완하기 위한 확장검색어 추천시스템을 구현하여 사용자의 정보검색 과정 중 최초질의어 입력 후 질의어 확장 과정에서 사용자 편의성을 향상 시키고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구들에 대하여 설명하고, 3장에서는 본 논문에서 제안하는 확장검색어 추천시스템의 구현에 대해 설명한다. 4장에서는 실험환경 구성과 시스템 평가를 실시하고, 마지막 5장에서는 논문의 결론을 맺는다.

Ⅱ. 관련연구

2.1. 확장 및 연관검색어 서비스

'확장검색어'는 사용자의 검색 편의를 위해 검색창에 입력되는 검색어 유형을 분석하여 많은 사용자가 자주 찾는 검색어를 추천해주는 서비스이며 '연관검색어'란 사용자가 특정 단어를 검색한 후 연이어 많이 검색한 검색어를 추출하여 제공하는 서비스로, 해당 검색분야에 확장검색어를 제공하여, 찾으려는 정보에 더욱 쉽게 다가갈 수 있도록 사용자의 검색 편의를 도와주는 서비스이다[6]. 이와 같이 확장 및 연관 검색어 서비스는 사용자의 질의어와 관련 있는 단어를 말하며, 지원하는 목적은 최초 검색어와 관련성이 높은 단어를 제시함으로써 검색의 편의성과 재검색을 통한 결과 값 획득에 용이하게 하기 위해서이다.

2.2. 질의어 확장(Query Expansion)

질의어 확장이란 사용자가 제시한 질의어에 이와 관련된 단어들을 추가해서 문서를 검색함으로서 보다 연관성이 높은 문서들을 검색하고자 하는 것이다. 또 다른 정의로 질의 확장이란 사용자들의 불완전한 최초 질의에 대하여 최초 질의와 관련이 높은 단어들을 선정하여 질의에 추가하는 것을 말한다.

최근에 질의어 확장에 대한 연구는 개인, 그룹의 프로파일을 이용하는 기법과 사용자 피드백에 의한 기법이 있다. 프로파일을 작성하여 사용하는 경우는 사용자가 질의를 입력하면 이미 사용자의 관심 분야에 관련된 키워드들로 작성된 프로파일이나 시소러스를 참조하여 질의어를 확장하게 된다[7]. 사용자 피드백에 의한 질의어 확장 기법은 사용자가 탐색된 결과를 보고 질문을 수정하고 반복적인 탐색을 통해 관련성이 높은 문서를 검색하는 기법으로 시스템의 도움을 받아 보다 좋은 질의어를 입력하거나 질의어를 재 생성한다[8].

2.3. 연관규칙(Association Rule)

데이터마이닝 기법 중 하나인 연관규칙은 데이터들의 빈도수와 동시 발생 확률을 이용하여 한 항목들의 그룹과 다른 항목들의 그룹 사이에 강한 연관성이 있음을 밝혀 주는 기술이다. 기본개념은[9]

(Item set A) = > Item set B)

if A then B: 만약 A가 일어나면 B가 일어난다.

연관 규칙에서 측정의 기본은 빈도수이며 빈도수를 기반으로 연관 정도를 정량화하기 위해서 다음의 두 가지 기준을 고려한다.

(1) 지지도(Support) - 전체 거래 중 항목 X와 항목 Y를 동시에 포함하는 거래가 어느 정도 인가를 나타내주며 전체적 구매도에 대한 경향을 파악할 수 있다[10].

$$\text{지지도} = \frac{\text{품목 } X\text{와 품목 } Y\text{를 포함하는 거래수}}{\text{전체 거래수}(N)}$$

지지도는 한 규칙이 주어진 데이터 집합에 얼마나 자주 적용할 수 있는지를 결정하며 유용하지 않은 규칙들을 제거하는데 종종 사용된다.

(2) 신뢰도(Confidence) - 항목 X를 포함하는 거래 항목 Y가 포함될 확률을 어느 정도 인가를 나타내주며 연관성의 정도를 파악할 수 있다.

$$C = \frac{P(X \cap Y)}{P(X)} = \frac{\text{품목 } X\text{와 품목 } Y\text{를 포함하는 거래수}}{\text{품목 } X\text{를 포함한 거래수}}$$

신뢰도는 규칙에 의해 만들어지는 추론의 확실성을 측정한다. 주어진 규칙 X→Y에서 신뢰도가 높을수록 X를 포함하는 트랜잭션에 Y가 존재할 가능성이 더 높다.

2.4. 대중성 기반 확장 및 연관검색어 문제점

많은 수의 사용자가 자주 찾는 검색어를 추천하는 대중성 기반의 서비스는 사용자가 질의를 확장하는데 많은 도움이 되는 서비스이다. 하지만 개개인의 사용자를 만족시키기에는 한계가 있다. 대중이 선호하는 검색어의 추천이 꼭 자신에게 맞는 검색어라고는 볼 수 없기에 각 검색엔진의 추천 검색어에 대한 한계를 볼 수 있다.

표 1. 사용자 질의의도가 포함된 질의어
Table 1. Query containing users' intention of inquiry

구 분	사용자 질의 의도(중고딕)
사용자 질의어	6자회담 **전망**, 깡통 **주식**, 멜라민 **파동**, 결식아동 **돕기**, 의료보험 **민영화**, 수능 **원서 접수**, 김정일 **건강이상설**

표 1은 정치, 경제, 사회, 문화, 의료, 교육, 국방 분야 등 7개 분야의 확장 및 연관 검색어에 대한 사용자의 질의의도가 포함된 7개의 질의어이다.

사용자의 질의의도를 2단어로 한정하여 사전의 정답 안을 만들고 현재 서비스하고 있는 대표적인 검색엔진(구글, 네이버, 야후, 다음)에서 단일 질의어를 입력하고 질의어를 확장하여 분석한 일부 결과는 표 2와 같다.

표 2. 6자회담에 대한 분석결과
Table 2. Analysis on six-party talk

구 분		'6자회담' 분석 결과
네이버	확장 검색어	X (확장검색어 없음)
	연관 검색어	패션플러스, 개성공단, 북핵, 6차 6자회담, 9. 19 공동성명, 남북경협, 남북정상회담, 남북철도, 대륙철도, 방북, 북핵문제, 핵실험

네이버 분석결과 사용자는 '6자회담 전망'이라는 질의의도를 가지고 최초 질의어로 '6자회담'을 입력한 결과 사용자에게 맞는 질의어 확장을 위한 검색어를 추천하여 주지 못한다.

그림 1은 각 검색엔진 별 확장 및 연관 검색어의 서비스 추천 된 비율을 나타낸다. 분석 된 결과 각 검색엔진 별 다양한 분야에 대한 확장 및 연관검색어 추천 서비스를 제공하고 있지만 사용자가 의도하는 질의에 대한 검색어를 추천하는 서비스는 상대적으로 제한되는 결과 값을 제공하고 있다.

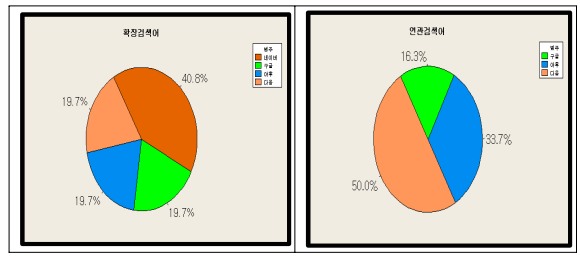

그림 1. 확장 및 연관 검색어 서비스 추천 비율
Fig. 1. Ratio of each search engine's suggestion of expanded or related query

Ⅲ. 제안한 확장검색어 추천시스템

사용자의 확장검색어 추천을 위한 제안시스템은 사용자 질의어를 입력받아 질의어를 DB화 한다. 각 사용자의 사용 빈도수 및 확률적 통계를 바탕으로 순위화 된 키워드를 추출하고 추출된 키워드는 연관규칙 *"if A then B"* 개념으로 연관성이 높은 것으로 판단하여 질의어 확장 과정 중 추천되어 진다.

사용자 질의 패턴을 분석하기 위해 사용자 질의

어 입력을 2단어로 제한하고 일정기간의 사용자가 질의어를 저장한다. 이를 기반으로 사용자 질의패턴을 분석하며 분석과정에서 특정 임계치(Threshold) 이상의 순위화 된 질의어를 기준으로 질의패턴을 추출하여 사용자의 확장검색어를 추천한다.

사용자 PC에 설치 및 운용되면서 사용자 확장검색어를 추천한 시스템을 그림 2에 도식하였다.

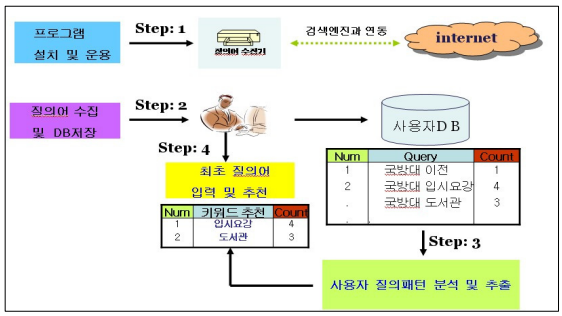

그림 2. 사용자의 확장검색어 추천시스템 구성도
Fig. 2. Schematic diagram of users' expanded query suggestion

3.1. 프로그램 설치 및 운용(step1)

1단계는 사용자가 검색엔진과 연동하여 사용할 수 있는 '질의어 수집기'를 설치 및 운용함으로서 실질적인 사용자의 질의어를 수집한다. 제한사항은 질의어의 수를 2개의 팀으로 고려한다. 이는 질의어의 길이가 대개 1~3로 이루어지고 평균적인 질의의 길이는 2.5단어로 연구되었다[11]. 또한 질의어의 수가 많을수록 정확한 검색결과가 나올 것이라 예측하지만 기존연구 결과 질의어 크기가 최대 3을 초과하면 오히려 검색 효율이 감소한다고 한다.

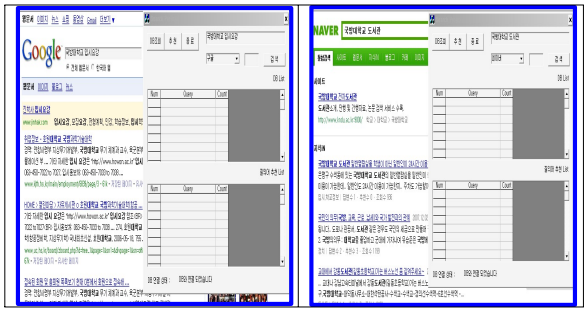

그림 3. 프로그램 실행화면 및 검색엔진 연동 운용
Fig. 3. Running both the program and search engine

그림 3은 프로그램 실행 화면으로 질의어(국방대학교 입시요강)를 입력하고 사용자가 원하는 검색엔진(구글, 네이버, 야후, 엠파스)을 선택하게 되면 검색엔진과 연동되어 사이트 결과를 볼 수 있으며 입력한 질의어는 DB에 저장된다.

3.2. 질의어 수집 및 저장(step2)

2단계는 사용자의 질의어를 수집 및 저장하는 단계로 '질의어 수집기'의 사용자 입력 인터페이스를 통해 입력된 질의어는 사용자의 DB에 저장된다.

그림 4는 '질의어 수집기'에 의해 입력되는 질의어를 DB에 저장하고 시스템에서 조회한 결과이다.

그림 4. 질의어 저장 및 시스템의 DB 조회 결과
Fig. 4. Collection query & viewing collection on the DB

웹 사용자가 검색하기 위한 질의어를 입력하게 되면 '질의어 수집기'는 사용자 PC의 DB를 확인하여 기 존재 여부를 확인한다. 이때 새로 입력한 질의어가 기존에 존재하면 질의어 사용횟수(Count)에 대한 정보를 수정하고 그렇지 않으면 DB에 새로운 질의어를 추가한다.

그림 5는 시스템 입력 인터페이스에 'HTTP 프로토콜'의 새로운 질의어에 대해 DB에 질의어가 추가되는 모습과 '국방대학교 입시요강'의 사용 횟수 수정에 대한 DB 조회 결과이다.

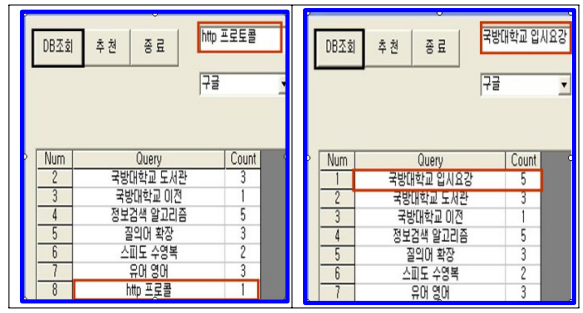

그림 5 사용자 DB 질의처리 결과
Fig. 5. Result of users' DB query process

3.3. 사용자 질의패턴 분석 및 추출(step3)

질의 패턴 분석 과정은 사용자 질의어 사용 빈도 및 확률을 기반으로 분석 및 추출되어 진다. 본 논문에서 제안하는 '질의패턴'이란 연관규칙의 *"if A then B"*의 개념을 의미한다. 예를 들어 "만약 사용자가 국방대학교를 입력하면 입시요강을 입력 할 것이다"로 정의한다. 즉 사용자의 질의패턴 분석이란 사용자가 국방대학교를 입력 시 질의어 확장 과정에서 확률적으로 연관성이 높은 순서로 키워드를

추천하게 된다. 그림 6은 연관규칙의 개념을 적용하여 '국방대학교'를 입력 하였을 경우 연관성이 높은 순서인 '입시요강', '도서관', '이전'의 순으로 질의어를 추천한 예를 보여준다.

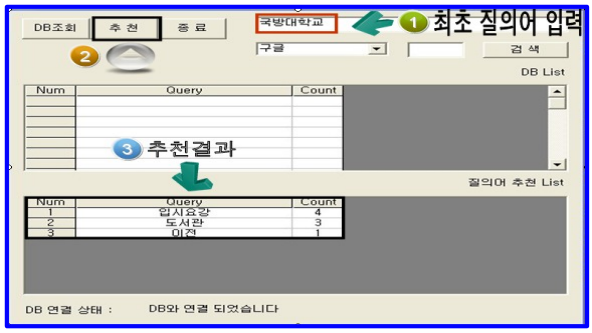

그림 6. 최초 질의어 입력 및 추천 과정
Fig. 6. Input of the initial query & suggestion process

질의 추천단계에서 신뢰성 있는 질의어를 추천하기 위하여 일정한 임계값(threshold)을 두어서 후보 질의어를 걸러내는 작업이 필요하다. 이 값은 일정한 수치가 정해져 있는 것이 아니라, 통계치에 의해서 사용자가 선택 할 수 있다. 그림 7은 사용자가 일정한 임계값을 설정하여 질의어를 추천 받는 과정을 보여준다. 사용자가 최초 질의어 '국방대학교'를 입력하고 하고 추천을 받기위해 임계값 3을 설정한 결과로서 그림 4의 DB 조회 결과와 비교하면 일정한 임계치 이상의 확장검색어가 추천되는 결과를 확인할 수 있다.

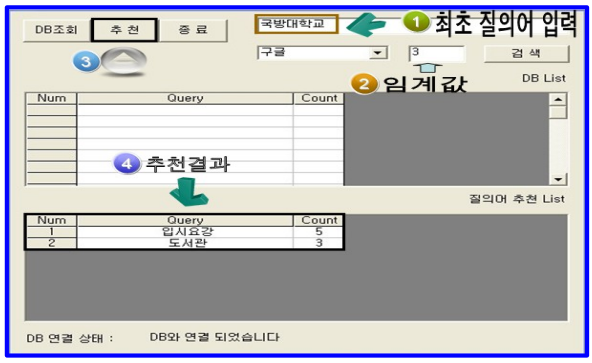

그림 7. 임계값 설정에 의한 확장검색어 추천
Fig. 7. Suggesting extended query by setting up the threshold

Ⅳ. 실험 및 평가

4.1. 실험환경 구성

사용자의 확장검색어의 추천시스템 평가를 위하여 실험은 3가지 과정을 통해서 평가한다. 첫째, 사용자의 질의패턴을 추출하기 위해 데이터마이닝 기법 중 연관규칙을 적용하여 사용자의 대표적인 질의패턴을 추출한다. 사용자 질의어는 AOL 질의어 데이터 중 1001개 이상의 사용자 질의어 데이터를 적용하고 100명의 사용자별 질의패턴을 추출한다. 둘째, 질의패턴을 추출한 사용자 100명의 동일한 질의어를 제안한 시스템에 적용하여 확장검색어 추천 결과와 비교하여 평가한다. 셋째, 제안한 시스템의 확장검색어 추천 결과가 대중성 기반 추천 서비스의 한계를 보완할 수 있는지를 검증하기 위해 검색엔진에서 추천된 확장 및 연관검색어 서비스 결과를 비교 및 평가한다. 추천된 확장 및 연관검색어 서비스 결과를 비교 및 평가를 위해 선택한 검색엔진은 AOL 질의어 데이터의 특성상 검색엔진(구글, 야후)을 선택하였다.

4.2. 단계별 실험 및 평가

4.2.1 사용자 질의패턴 분석

질의패턴을 분석하기 위해 사용한 데이터 마이닝 Tool은 SPSS Clementine 10.1이며 분석을 위하여 전체 AOL Query의 데이터 중 전처리 과정을 통하여 2개 단어를 사용한 질의어를 추출하여 사용자의 질의패턴을 분석하였다. 분석을 위하여 35,864,595개 질의어 중 전처리 과정을 통하여 추출된 질의어는 649,228명이 사용한 질의어 8,010,252개의 질의어를 추출하였다. 전처리 과정은 전체 Query data 중 1개 및 3개 이상의 질의어는 제외하였고 2개의 단어를 사용한 질의어로 제한하였다.

Consequent	Antecedent	Support %	Confidence %
field2 = earth	field1 = google	0.223	19.904
field2 = jeeves	field1 = ask	0.294	85.072
field2 = airways	field1 = us	0.166	14.871
field2 = depot	field1 = home	0.257	52.273
field2 = positions	field1 = sex	0.235	13.062
field2 = swimwear	field1 = sexy	0.214	19.582
field2 = pages	field1 = yellow	0.154	78.189
field2 = quest	field1 = map	0.382	87.178
field2 = channel	field1 = disney	0.172	28.086
field2 = pages	field1 = white	0.227	65.969
field2 = rental	field1 = car	0.174	11.842
field2 = space	field1 = my	0.514	61.549
field2 = layouts	field1 = myspace	0.461	32.036
field2 = rentals	field1 = car	0.174	14.210
field2 = idol	field1 = american	0.613	47.051
field2 = tickets	field1 = cheap	0.188	19.933
field2 = airlines	field1 = southwest	0.197	79.359
field2 = names	field1 = baby	0.192	23.668
field2 = world	field1 = disney	0.172	12.599
field2 = lotto	field1 = florida	0.226	12.855
field2 = mail	field1 = yahoo	0.796	30.046
field2 = search	field1 = google	0.223	16.634
field2 = lottery	field1 = florida	0.226	33.620
field2 = girls	field1 = nude	0.212	17.748
field2 = porn	field1 = free	0.808	15.302
field2 = bank	field1 = us	0.166	14.491
field2 = com	field1 = google	0.223	17.811
field2 = airlines	field1 = american	0.613	13.523

그림 8. 질의패턴 추출 결과
Fig. 8. Result by the extraction of query patters

연관규칙 적용을 통하여 분석된 질의패턴은 일정한 임계값(threshold)을 두어 추출하였다. 임계값은 사용수 범위별로 분류된 3,001개 이상 사용된 질의어(A질의어 308개 단어, B질의어 339개 단어)를 기준으로 일정 지지도 및 신뢰도의 이상의 규칙을 찾

아내어 단어와 단어 사이의 연관관계를 추출하였다. 본 논문에서는 최소 지지도 0.10~0.15%, 신뢰도 10% 임계값 이상의 질의패턴 84개를 추출하였으며 결과의 일부분을 그림 8에 제시하였다. 그림에서의 'Antecedent'는 최초 질의어 이며, 'Consequent'는 이후 입력된 질의어를 의미한다.

4.2.2 사용자 질의패턴 추출

AOL 질의어 데이터 중 1001개 이상의 질의어를 입력한 사용자를 대상으로 전처리 과정을 통하여 2개의 Term을 추출하고 연관규칙을 적용하여 신뢰도 높은 질의패턴을 추출하였다. 그림 9는 사용자의 질의패턴을 분석한 과정이다. 입력한 '사용자 질의어'는 최초 질의어와 이어서 입력한 질의어를 질의어(A)와 질의어(B)로 표기하였으며 최소 지지도 5%, 신뢰도 10%를 적용하여 분석한 결과로서 분석된 결과 중 사용빈도, 지지도 및 신뢰도를 고려하여 대표적인 질의패턴(ligament→supplements)을 하나를 선택하였다.

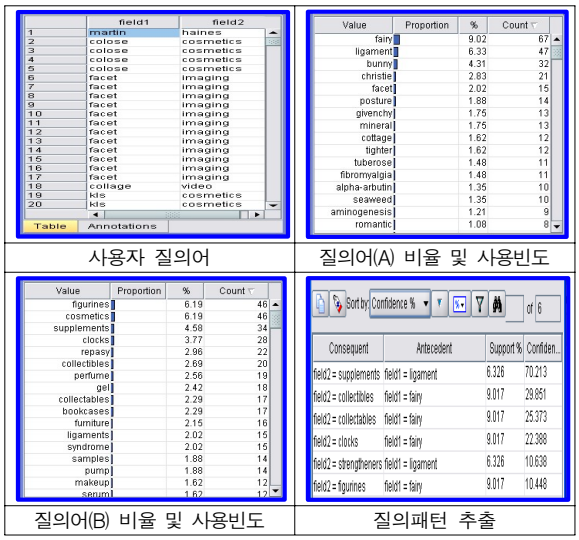

그림 9. 사용자의 질의패턴 추출 과정
Fig. 9. Extraction of users' query patters

4.2.3 질의패턴 VS 사용자 확장검색어 추천 결과 평가

질의패턴을 추출한 사용자 100명의 동일한 질의어를 제안 시스템에 적용하여 확장검색어 추천 결과와 비교하고 평가하였다.

그림 10은 확장검색어 추천시스템에서 추천결과를 확인하는 과정을 나타낸다. 사용자의 확장검색어 추천의 (a)과정은 질의패턴 추출과정으로 그림 9에서 분석에 사용되었던 사용자의 질의어를 확장검색어 추천시스템 DB에 저장하였다. (b)과정은 시스템에서 DB에 저장된 질의어를 조회한 결과이며 (c)과정은 하나의 단일 질의어를 제안한 시스템에 입력하고 (d)과정에서 추천된 결과를 확인 할 수 있다.

그림 9에서 분석된 질의패턴 중 최초 질의어를 'ligament'를 입력하였고 추천된 결과는 supplements, strengtheners, breakthroughs, supplementation, boosters를 순위별로 추천한 결과이다. 위 결과는 제안 시스템의 질의패턴의 *'If A then B'*의 개념을 적용하여, 질의어 확장 과정에서 확률적으로 연관성이 높은 순서로 키워드를 추천하는 과정으로 하나의 단일 질의어를 검색엔진에 입력하고 사용빈도 및 확률로 질의어 확장을 위하여 질의어를 추천하는 과정이다.

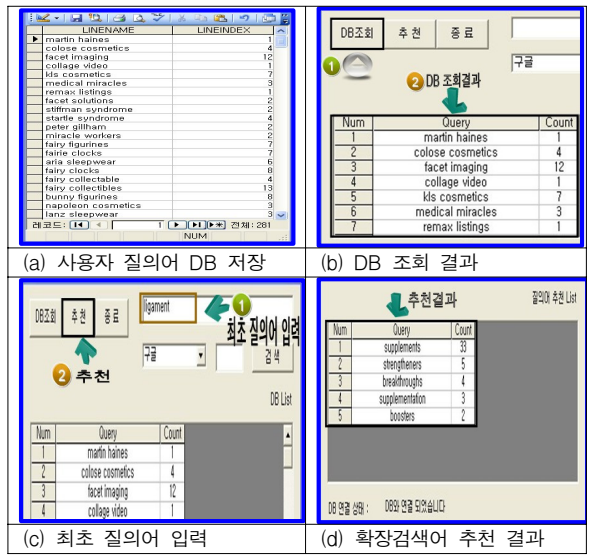

그림 10. 사용자의 확장검색어 추천과정
Fig. 10. Users' extended query suggest process

표 3은 사용자 대표적인 질의패턴 추출 결과와 제안한 시스템의 확장검색어 추천의 결과를 평가한 결과이다. 평가 결과 제안한 사용자의 확장검색어 추천시스템은 100명의 사용자로부터 분석된 질의패턴을 모두 추천 할 수 있었으며 또한 최초 질의어와 함께 이어서 입력한 모든 단어를 사용빈도와 함께 추천함으로써 사용자의 확장검색어 선택의 폭을 다양화 할 수 있는 장점도 가지고 있다.

제안한 사용자의 확장검색어 추천시스템은 2단계 평가 결과 우수하다. 하지만 검색엔진에서 확장 및 연관검색어를 추천하는 서비스가 제공되고 있다면 본 논문의 추천시스템 적용은 어렵다고 할 수 있다.

3단계 실험 및 평가에서는 검색엔진 이용 시 사용자의 편의성을 위하여 대중성 기반으로 추천되고 있는 확장 및 연관 검색어 추천 서비스와 제안한 확장검색어 추천시스템의 추천 결과를 비교하여 사용자의 편의성을 향상시킬 수 있는지 평가하였다.

표 3. 질의패턴 vs 확장검색어 추천 결과 평가
Table 3. Evaluation of query pattern vs. extended query suggestion

※ 중고딕 : 최초 입력 질의어

질의패턴: 확장검색어	결과	질의패턴: 확장검색어	결과
ligament supplements	추천	**emmy** rossum	추천
hydrogen bomb	추천	**fast** cars	추천
hemi challemger	추천	**solderless** breadboard	추천
neck pain	추천	**wonder** woman	추천
yahoo mail	추천	**pat** benter	추천
fernando colunga	추천	**david** hasselhoff	추천
radio veracruz	추천	**steve** burton	추천
john densmore	추천	**jc** penny	추천
arabian stallion	추천	**kid** boys	추천
clive owen	추천	**daily** pictures	추천
jonbenet kamsey	추천	**news** math	추천
matthew fox	추천	**igbo** masks	추천
yahoo sports	추천	**actor** dullea	추천
acura integra	추천	**sasha** cohen	추천
jared padalecki	추천	**copper** fabric	추천
wedding ring	추천	**jonbenet** pictures	추천
baruch college	추천	**american** idol	추천
nyse eslr	추천	**tatto** flames	추천
prostate massage	추천	**jesse** metcafe	추천
kimberly logan	추천	**people** magazine	추천
good charlotte	추천	**left** eye	추천
head lump	추천	**deer** study	추천
picture frames	추천	**telegraph** kit	추천
cottage kitchens	추천	**dinosaur** parties	추천
reggie jackson	추천	**pc** games	추천
donnie mcclurkin	추천	**gold** incest	추천
orlando bloom	추천	**mother** teresa	추천
sheer swimwear	추천	**minesota** bca	추천
spa wrap	추천	**bettie** page	추천
pamela anderson	추천	**ebay** motors	추천
joe afful	추천	**my** space	추천
kelly clarkson	추천	**katrina** looting	추천
belly dancer	추천	**hawaiian** yardage	추천
pasta recipes	추천	**paper** dolls	추천
rento sunglasses	추천	**gerard** butler	추천
:	:	:	:

4.2.4 검색엔진 vs 확장검색어 추천 결과 평가

제안한 시스템의 확장검색어 추천 결과가 대중성 기반 검색엔진의 추천 서비스의 한계를 보완할 수 있는지를 검증하기 위해 구글 및 야후의 검색엔진에서 추천된 확장 및 연관검색어 서비스 결과를 비교 및 평가하였다.

표 4는 검색엔진(구글)에서 제공하고 있는 확장 및 연관검색어 추천 서비스의 일부 결과이다. 표 4의 과정을 통하여 구글 및 야후에서 추천이 가능한 확장 및 연관검색어의 추천 결과를 비교하였으며 중고딕으로 표시한 단어는 검색엔진(구글)에서 추천이 가능한 확장검색어이다. 100명의 대표적 질의패턴으로 추출한 결과를 검색엔진(구글, 야후)에서 추천이 가능한지 평가한 결과는 그림 11과 같다.

표 4. Google.com 확장 및 연관검색어 추천결과 비교
Table 4. Extended and related query suggestion

질의어		확장 및 연관 검색어
ligament	확장	of treitz, flavum, teres, arterisum, damage
	연관	ankle - , foot, - , dlbow -
hydrogen	확장	peroxide, fule cell, **bomb**, sulfide, generator
	연관	fuel, power, uses, element -
kimberly	확장	clark, caldwell, williams, wyatt, page, rogers
	연관	결과 없음
drum	확장	tabs, **sets**, kits, lessons
	연관	- tabacco, ddrum, guitar center
grow	확장	pains, cube, island, **lights**, chart
	연관	- 2, - pot, questions

질의어 확장 과정 중 검색엔진(구글)에서 사용자의 질의패턴을 충족 할 수 있게 추천한 결과는 59%이며 추천하지 못한 결과는 41%이며, 검색엔진(야후)에서 추천 가능한 질의패턴은 52%, 추천하지 못한 결과는 48%로 모든 사용자를 만족하게 추천할 수 없었으며 대중성의 한계를 분석할 수 있다.

대중성 기반의 확장 및 연관 검색어 서비스의 추천 비율이 59% 및 52%로 실험을 위한 사용자 질의패턴의 결과를 추천할 수 있었는데 이는 공동의 관심사에 개인도 역시 공동의 관심을 가지고 있다고 판단 할 수 있을 것이다.

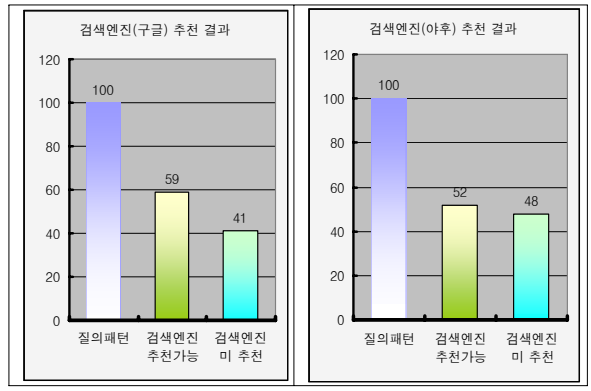

그림 11. 검색엔진 추천 결과
Fig. 11. Result of search engine suggestion

검색엔진에서 제공하고 있는 확장 및 연관검색어 서비스는 많은 수의 사용자가 자주 찾는 검색어를 추천하는 대중성을 기반으로 질의를 확장하는데 많은 도움이 되는 서비스이다. 하지만 대중이 선호하는 검색어가 자신이 선호하는 검색어라고 볼 수 없기 때문에 검색엔진에서 추천한 확장 및 연관 검색어 추천 서비스에 대한 한계를 볼 수 있다. 제한사항은 사용자 질의어가 제한된 범위 내에서 평가가 이루어져 모든 단어를 적용 시 변동이 있을 것으로 예상되고 시간의 흐름에 변화가 있을 것이며, 또한

검색엔진에서는 여러 요소 기술들을 복합 적용하여 서비스를 제공하고 있을 것으로 예상되는데 본 논문은 연관규칙만의 제한된 범위의 적용으로 결과에 차이가 발생할 수 있다.

V. 결 론

본 논문에서는 사용자의 질의패턴을 분석하여 사용자에게 효율적인 확장검색어를 추천하는 시스템을 구현하였다. 제안한 사용자의 확장검색어 추천시스템은 평가 결과 우수하지만 검색엔진에서 확장 및 연관검색어를 추천하는 서비스가 제공되고 있다면 제안시스템의 적용은 어렵다고 할 수 있으므로 검색엔진 이용 시 사용자의 편의성을 위하여 대중성 기반으로 추천되고 있는 확장 및 연관 검색어 추천 서비스와 제안한 확장검색어 추천시스템의 추천 결과를 비교하여 대중성 기반의 추천 서비스의 한계를 보완하고 사용자의 편의성을 향상시킬 수 있는지 평가하였다.

평가결과 논문에서 제안한 질의패턴 분석을 통한 사용자의 확장검색어 추천시스템은 검색엔진(구글, 야후)에서 추천하지 못한 41%, 48%의 확장검색어를 추천할 수 있었으며, 사용자의 편의성을 위하여 대중성 기반으로 추천되고 있는 확장 및 연관 검색어 추천 서비스의 한계를 보완하여 사용자의 편의성을 향상시킬 수 있었다.

향후 연구로 제안시스템에 Collaborative Filtering와 같이 타인의 정보를 활용하여 추천한 결과를 결합한 새로운 방법에 대해 연구할 계획이다.

참 고 문 헌

[1] Broder, A., "A Taxonomy of Web Search", SIGAR Forum Vol. 36, No. 2, 2002.

[2] Hyungil Kim, Juntae Kim "Improving Performance of Web Search using The User Preference in Query Word Senses", KIISE Vol. 31, No. 8, 2004.

[3] Mun HyeonJeong, Lee SuJin, "A Personalized Concept-based Retrieval technique Using Domain Ontology", CALS/EC, Vol. 12, No. 3, 2006.

[4] Zhongming Mai, Gautam Pant, Olivia R. Liu Sheng., "Interest-based personalized search", ACM Transactions on Information systems, Vol.25 Issue 1, 2007.

[5] AOL Query Set, http://www.gregsadesky.com/aol-date

[6] NAVER, http://www.help.naver/service/main.service

[7] P. Wallis. J. A. Tom, "Relevace judgement for accessing recall", Information Processing & Management 32, 1998.

[8] Teevan, J., Dumais, S. T., "Presonalizing search via automated analysis of interests and activities" SiGIR Coference, 2005.

[9] Jihye Kim, Hyun-min Kim "Introduction to Concept in Association Rule Mining", KCC 2002, Vol. 29, No. 1, 2002.

[10] Hwan-Seung Yong, "DATA Mining", Infinitebooks, 2007.

[11] J. R. Wen, J. Y. Nie and H. J. Zhang. "Clustering user queries of a Search Engine". In Proceedings of the Internation World Wide Web conference, 2001.

김 영 안 (Young-an Kim)

1988년 2월 금오공과대학 전산공학과 학사

1996년 3월 Keio University Dept. of Information and Computer Science 석사

2008년 2월 경희대학교 컴퓨터공학과 박사

2009년 2월~현재 국방대학교 국방과학학과 교수

<관심분야> Ad-hoc Network, Routing Protocol, DTN, VANET, WMN, 정보검색, SNA

박 건 우 (Gun-Woo Park)

1997년 2월 충남대학교 컴퓨터과학과 학사

2007년 2월 연세대학교 컴퓨터과학과 석사

2011년 8월 국방대학교 국방과학학과 박사

2011년 9월~현재 종합보급창

<관심분야> 정보검색, 네트워크 보안, SNA