2001년 한국정보처리학회 춘계 학술발표논문집 제8권 제1호

정보검색에서 사용자 검색 패턴을 이용한 질의 확장

천우관, 김영도, 정인정
고려대학교 전산학과
e-mail:grchun@tiger.korea.ac.kr

Query Expansion Using User Search Pattern in Information Retrieval

Woo-Kwan Chun, Young-Do Kim, In-Jeong Chung
Dept. of Computer Science, Korea University

요약

정보검색에서 가장 많이 사용되는 불리언(Boolean)검색에서는 키워드 일치에 의해서만 검색하는 단점을 가지고 있다. 이를 보완하기 위해 다양한 정보원에서 추출한 관련 용어들을 원질의어에 첨가하여 검색의 효율을 높이기 위한 질의 확장 방법들이 모색되어 왔다. 본 논문에서는 질의 확장을 위하여 사용자가 검색에 사용하였던 질의어들의 연속성을 찾아내어 첨가할 용어를 선택하고 질의 확장을 하는 방법을 제시한다. 사용자가 입력한 질의어의 연속성을 찾아내는 방법으로는 데이터 마이닝 기법중 연관 규칙 탐사 방법을 이용한다. 실험은 현재 구축된 정보통신 기술기준 정보시스템에서 사용자들이 검색한 키워드 정보를 이용하였으며 사용자 검색 패턴(USP) 정보를 이용함으로써 사용자가 검색하고자 하는 질의어와 좀더 연관성 있는 용어로 확장하여 사용자 중심적 결과를 얻을 수 있다.

1. 서론

정보검색은 수집된 정보 또는 정보자료의 내용을 분석한뒤 적절히 가공하여 축적해 놓은 정보파일로부터 사용자의 정보요구에 적합한 정보를 탐색하여 찾아내는 일련의 과정을 의미한다. 사용자가 질의어를 입력하면 시스템은 질의어와 일치하는 색인어가 존재하는가를 탐색하고 일치하는 색인어가 존재하면 그 색인어를 포함하는 문서들을 찾아 사용자에게 제공하게 되는 것이다. 따라서, 사용자가 자신이 필요로 하는 정보요구를 명확히 제시하는 것과 검색 시스템의 성능은 밀접한 관계를 가진다.

이와 같은 정보검색을 위해 불리언(Boolean) 검색모델이 가장 널리 사용되고 있다[7]. 그러나, 불리언 검색은 질의어와 완전히 일치하는 문서만이 검색되므로 부분적으로 일치하는 문서는 검색할 수 없는 단점을 가지고 있다. 이러한 불리언 검색의 단점을 다소 해결한 검색 방법이 퍼지 집합 검색[3]과 벡터 공간 검색[4]이다.

그러나, 퍼지 집합 검색에서도 불리언 검색이 갖는 정보요구 표현상의 부적절성은 여전히 존재하며, 색인어에 가중치를 부여해야만 멤버쉽 함수를 적용할 수 있는 단점이 있다. 또한, 벡터공간 검색은 탐색어들간의 관계성을 표현할 수 없고 각각 독립적으로 표현되고 처리되며, 문서의 수가 많은 시스템에서 전체 문서에 대해 이 기법을 적용하는 것은 많은 시간이 걸리므로 비효율적이다.

이런 이유로 기존의 정보검색 시스템에 인공지능 기법을 도입한 보다 발전된 형태의 검색모델인 지식기반 정보검색[4]에 관한 연구가 진행되었다. 이 시스템은 입력된 질의어와 완전히 일치하는 색인어가 존재하지 않을 경우 질의어를 전향추론 방법에 의하여 동의어나 상위 개념어 혹은 하위 개념어로 확장하여 색인어와 일치여부를 고려하여 문서를 검색한다. 그러나 색인어간의 관계를 고려하지 않고, 질의어를 단순히 동의어나 상위 개념어 혹은 하위 개념어로 확장하므로 전혀 상관이 없는 문서가 검색되어 재현율(Recall)의 저하를 가져올 수 있는 단점을 가지고 있다.

이러한 기존의 검색 시스템을 이용하는 대부분의 사용자는 짧은 질의어를 이용하여 검색을 행하게 된다. 그러나 짧은 질의어는 사용자의 정보 요구를 충분히 표현할 수 없으며, 또한 검색 시스템의 성능을 저하시키는 주요 원인이 된다. 이러한 사용자의 부담을 덜어 주기 위해 사용자의 질의를 가공하거나 대화형 질의 확장(Interactive Query Expansion)을 통해 사용자의 정보요구를 명확히 표현해 주는 기능들이 요구되게 되었다.

본 논문에서는 질의 확장에 첨가할 용어를 선택하기 위해 사용자 검색 패턴을 추출하여 연관 규칙을 생성하고 이 연관 규칙을 이용하여 질의 확장하는 방법을 제시해 본다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 질의 확장에 관한 관련연구에 대하여 알아보고 3장에서는 사용자 검색 정보로부터 검색 패턴을 마이닝하는 알고리즘을 제안한다. 4장에서는 마이닝된 검색 패턴 정보로부터 질의 확장하는 방법을 기술한다. 5장에서는 실험 및 분석를 살펴보고, 마지막으로 6장에서 결론 및 향후과제를 제시한다.

2. 질의 확장 관련 연구

정보검색 시스템의 사용자들은 대부분 찾고자하는 문서와 밀접한 관련이 있는 몇 개의 용어들로 이루어진 질의어를 사용하여 자신의 정보요구를 표현한다. 그러나 대부분의 검색 시스템들은 용어 일치 여부로 검색을 하므로 사용자가 원하지 않는 문서들을 찾아내게 되어 재현율(Recall)이 감소하는 경향이 있다. 이를 위하여 질의어를 구성하는 용어와 다양한 정보원에서 추출한 관련 용어들을 첨가하여 검색의 효율을 높이고자 하였는데, 이를 "질의 확장(Query Expansion)"이 라고 한다. 질의 확장의 문제점은 질의 확장에 첨가할 용어의 선택 문제와 선택된 용어의 가중치 부여 문제로 요약할 수 있다. 이러한 문제점을 해결하기 위하여 여러 가지 방법들이 제시되었다.

지식 베이스(Knowledge-Based)를 이용하는 방법[4]에서는 질의 확장에 시소러스(Thesaurus)를 이용하고 있다. 이를 통해 한글의 동의어 문제를 효율적으로 해결하고 있다. 그러나 시소러스를 구축하기가 용이하지 않으며, 단어의 희귀(Sparseness)문제를 극복하기 어려운 단점이 있다. 또한, 다의어 문제로 인하여 정확률이 떨어지는 경향이 있다.

통계적인 방법[9]은 용어의 동시 출현(Co-occurrence)의 수를 이용하는 것이다. 이 방법은 출현하는 용어들이 같은 주제에 밀접하게 관련되어 있음을 전제로 하고 있다. 그러나 동시 출현 빈도가 높은 단어들은

문서사이의 분별력이 낮은 경향이 있기 때문에 상대적으로 검색효율이 떨어지게 된다. 따라서 질의 확장에 추가되는 용어들은 빈도수가 적은 단어들이 빈도수가 높은 단어들보다 상대적으로 좋은 결과를 나타낸다.

자동적인 질의 확장(Automatic Query Expansion)[3] 방법은 사용자의 관련성 정보(Relevance Feedback)를 이용하여 질의 확장을 실행한다. 이 방법은 사용자에게 문서와 관련되는 용어를 선택하게 하여 검색의 효율을 높이는 방법으로 초기 검색이 무엇보다도 중요하다. 그러나, 이 방법의 단점은 사용자의 개입이 필요하고, 성능이 사용자의 정보 검색 능력과 비례한다는 것이다.

또 다른 연구 방향은 단어와 단어의 잠재적인 의미를 이용하여 유사한 용어를 추가하기보다는 질의어의 개념(Concept)과 가장 유사한 용어를 추가하는 방법[10]이다. 또는 유사 시소러스(Similarity Thesaurus)[11], LSI(Latent Semantic Indexing)의 SVD(Singular Value Decomposition)[3]를 이용하여 질의 확장에 유용한 용어를 선택한다.

본 논문에서는 사용자가 검색하였던 용어들의 순서적 흐름을 파악하여 사용자 검색 패턴 마이닝을 통한 검색 패턴을 추출하고 이를 질의 확장의 추가 용어로서 사용하는 방법을 제시해 본다.

3. 사용자 검색 정보로부터의 데이터 마이닝

본 장에서는 사용자가 검색에 사용한 질의어들로부터 데이터 마이닝[2]기법중의 하나인 연관 규칙 탐사[1,12]를 이용하여 질의어들의 검색 패턴을 추출하는 방법을 제안한다.

3.1 연관 규칙 탐사

연관 규칙은 한 항목들의 그룹과 다른 항목들의 그룹사이에 강한 연관성이 있음을 밝혀주는 규칙이다. 주어진 트랜잭션의 집합을 I 라고 하자. 주어진 집합에서 서로 소의 관계인 두 개의 부분집합 X와 Y사이에 $R : X \rightarrow Y$ 형식이 존재하면 X집합의 트랜잭션이 일어나면 Y집합의 트랜잭션이 일어난다는 뜻을 함축하게 된다. 만일 트랜잭션 T가 X의 모든 항목들을 포함한다면($X \subseteq T$), T가 X를 지지한다(Support)고 한다. $S(X)$는 X를 지지하는 모든 트랜잭션들의 개수를 의미한다. 만일 사용자가 정한 최소 지지도 S_{min}에 대하여 $S(X) \geq S_{min}$이라면 집합 X는 빈발하다(일반적으로 집합 X를 Large Itemset이라고 한다)고 한다. 최소 지지도를 사용하는 이유는 관심 있을 정도로 빈발하게 나타나는 항목만을 고려하기 위해서이다. 항목집합 X의 개수를 $k=|X|$로 나타내고 이를 k-항목집합이라 한다. 만일 한 트랜잭션이 X를 지지한다면, 또한 어떤 확률에 의해 Y도 지지할 것이라는 예측으로 이해될 수 있다면 이런 확률을 신뢰도 $C(R)$이라 한다. R의 신뢰도는 X를 지지하는 T에 대하여 Y 또한 지지할 조건부 확률로 정의된다. 즉, $C(R)=S(X \cup Y)/S(X)$로 정의된다. 규칙의 신뢰도는 얼마나 자주 적용할 수 있는지를 나타내고 지지도는 그 규칙 전부가 얼마나 믿을 만한지를 보여준다. 규칙이 데이터베이스에서 적절해지려면 어떤 주어진 최소 신뢰도 C_{min}과 최소 지지도 S_{min}에 대하여 $C(R) \geq C_{min}$이고 $S(R) \geq S_{min}$을 만족하면 된다. 연관규칙을 탐사하는 문제는 기본적으로 다음의 두 단계로 구성된다[1].

- 단계 1: 트랜잭션 데이터베이스에서 후보 항목집합을 생성한 후 이 집합에서 최소 지지도 S_{min} 이상의 지지도를 가지는 빈발 항목집합들(Large itemset)을 찾아낸다.
- 단계 2: 빈발 항목집합에서 공집합이 아닌 모든 부분집합들을 찾은 후 부분집합의 신뢰도가 최소 신뢰도 보다 큰 모든 부분집합을 찾고 연관 규칙을 출력한다.

기본적인 두 단계 중 단계1이 전체 성능을 좌우하게 된다. 이러한 연관 규칙탐사의 두 과정을 기본으로 본 논문에서는 사용자의 입력 질의어들을 개개인에 따라 순서적으로 기록한다. 첫 번째 단계에서 이 정보를 이용하여 빈발 항목집합을 추출하게 된다. 또한 이렇게 찾아낸 빈발 항목집합들은 두 번째 단계에서 규칙 생성을 위해 이용된다. 일반적인 아이디어는 빈발항목집합 $ABCD$와 AB가 있을 때, 규칙 $AB \rightarrow CD$는 $Conf = Support(ABCD)/Support(AB) \geq minconf$을 계산하여 만족하면 규칙으로 생성하게 된다. 생성된 규칙들은 연속패턴 데이터베이스에 저장이 되어 질의 확장에 사용하게 된다.

3.2 사용자 검색 패턴(User Search Pattern)

기존의 정보 검색에서는 질의 확장을 할 때 지식 베이스를 이용한 시소러스나 용어의 출현 빈도수 등을 계산하여 확장에 이용하였다. 그러나 이러한 방법들은 사용자가 검색하기 원하는 용어와는 전혀 무관한 질의 확장으로 인하여 사용자 입장에서 검색의 효율이 떨어질 수도 있다. 최근에는 이러한 사용자의 입장을 고려해서 검색하는 방법들이 연구되고 있다. 사용자의 개인정보(Profile)를 이용한 방법[4], 억세스패턴(Access Pattern)을 이용한 방법[5], 지식베이스에서 흥미패턴(Interesting Pattern)을 추출하여 이용한 방법[6] 등이 있다.

본 논문에서는 사용자가 입력한 검색 용어들을 시간 순으로 데이터베이스에 저장하여 시간순서에 따른 용어들의 검색 순서패턴을 이용[8]한다. 사용자가 찾기를 원하는 용어A로 검색한 후 그 다음 다른 용어B로 검색을 수행한다는 트랜잭션이 여러 번 존재할 경우 이 검색시스템을 사용하는 사용자들의 일반적인 검색 패턴으로 추출할 수가 있다. 사용자의 검색정보는 다음과 같이 정의한다.

사용자ID	검색시간	검색질의어
1	09:25:30	C
1	09:27:20	I
2	11:25:12	A,B
2	11:28:56	C
2	11:32:24	D,F,G
3	13:40:22	C,D,G
1	17:50:51	C
1	17:55:19	D,G
3	18:03:26	C
3	18:06:11	I

사용자ID	질의어순서
1	<(C)(I)><(C)(D,G)>
2	<(A,B)(C)(D,F,G)>
3	<(C,D,G)><(C)(I)>

(그림1) 트랜잭션 DB T와 질의어순서 DB Q

(그림1) 에서 사용자ID와 그 사용자가 검색한 시간, 그리고 검색한 질의어들로 이루어진 트랜잭션 데이터베이스를 생성한다. 이렇게 만들어진 데이터베이스에서 연관 규칙 탐사에 사용하기 위한 질의어순서 데이터베이스를 생성한다. 질의어순서 데이터베이스는 사용자 각각에 대하여 검색시간의 순서에 따라 질의어순서가 결정되고 이 값이 질의어순서 항목에 저장되어 진다. 질의어 (C)는 단일어를 나타내고 (A,B)는 복합어를 나타낸다. 각각의 복합어들은 단일어로 분리되고 각각의 단어들은 복합어의 구성 순서에 맞추어 입력된다. 같은 사용자의 트랜잭션이 있을 경우 이전의 검색순서와 다른 검색순서로써 질의어순서 데이터베이스에 저장된다. 즉, 사용자1은 첫 번째 검색에서 (C)를 검색한 다음 (I)를 검색하는 패턴을 만들어 냈고 두 번째 검색에서는 (C)를 검색한 다음 (D,G)를 검색하는 패턴을 만들었다. 이 두가지 검색 패턴은 서로 다른 패턴으로 분리되어 사용되어야 한다.

3.3 사용자 질의어 사이의 연관 규칙 탐사

트랜잭션 데이터베이스 T로부터 변환된 질의어순서 데이터베이스 Q는 질의어 사이의 연속 패턴을 추출하는데 사용된다. 질의어순서 데이터베이스로부터 사용자 검색 패턴을 찾기 위해서는 기본적으로 두 단계로 수행이 된다. 첫 번째 단계에서 최소지지도 보다 큰 빈발 질의집합을 발견해 내야한다. 두 번째 단계에서 빈발 질의집합들로부터 최소 신뢰도 보다 큰 모든 연관규칙을 생성해 내야 한다. 먼저 첫 번째 단계로써 빈발 질의집합을 찾아내는 알고리즘은 다음과 같다.

$L_1 \leftarrow$ {Frequent 1-querysets};
/*트랜잭션 데이터베이스 T에 있는 모든 keyword을 count*/

for ($k \leftarrow 2$; $L_{k-1} \neq 0$; k^{++}) do begin

/*빈발질의집합 L_k가 NULL이 아닐때까지 수행*/

for ($i \leftarrow 1$; $i < p.count$; i++) **do**
for ($j \leftarrow i$+1; $j < q.count$; j++) **do**
$C_k \leftarrow (p.item_i \cup q.item_j)$;
/*빈발 (k-1)-queryset을 이용하여 후보 k-queryset을 생성*/

forall candidate $c \in C_k$ **do begin**
count all candidate $c \in Q$;
/*질의어순서 데이터베이스 Q에서 후보 k-querysets의 패턴을 갖는 순차적인 질의어를 계산*/
end;

$L_k \leftarrow \{c \in C_k \mid c.count \geq S_{min}\}$;
/*후보 k-querysets중 S_{min}보다 작은 후보 querysets을 삭제*/
end;
Return $\cup_k L_k$;
/*모든 빈발 질의집합들을 Return*/

빈발 질의집합 추출 알고리즘

빈발 질의집합 추출 알고리즘은 질의어순서 데이터베이스에서 빈발한 질의집합들을 찾아내는 알고리즘이다. 위의 알고리즘에서 후보 질의집합을 생성한 다음 질의어순서 데이터베이스로부터 사용자마다 분류해 놓은 질의어순서를 읽어서 각각의 후보 질의집합의 발생 빈도수를 계산한다. 계산이 끝나면 빈도수가 S_{min}보다 큰 후보 질의집합들을 빈발 질의집합에 추가시키고 S_{min}보다 작은 후보 질의집합들은 제거를 한다.

이 과정을 빈발 질의집합이 더 이상 추출되지 않을 때까지 수행하고 추출된 모든 빈발 질의집합을 저장하게 된다. 이렇게 찾은 빈발 질의집합은 검색 시스템에서 검색 질의어가 들어왔을 때 질의 확장을 위해 추가할 용어를 선택할 때 이용하게 된다.

두 번째 단계에서 빈발 질의집합 추출 알고리즘을 통하여 추출된 패턴 정보들을 이용하여 일반적인 사용자 검색 패턴 규칙을 생성하게 된다. 즉, 추출된 빈발 질의집합 L에서 최소 신뢰도 C_{min}을 만족하는 모든 집합들을 사용자 검색 패턴 규칙으로 생성하여 질의 확장에 사용할 연속 패턴 데이터베이스로 구축을 하게 된다. 다음은 사용자 검색 패턴 규칙을 추출하는 알고리즘이다.

for all k-queryset l_k, $k \geq 2$ **do begin**
/*모든 빈발 질의집합 l_k를 수행*/
A = {k-querysets l_k};

for all $a_m \in A$ **do begin**
/*각각의 a_m에 대한 $conf$ 계산*/
$conf(a_m)$ = support (a_m) / support(a_m - $q.item_m$);
if ($conf \geq C_{min}$) **then begin**
/*최소 신뢰도 C_{min}을 넘는 규칙을 출력*/
output the rule a_m with confidence = $conf(a_m)$
and support = support(a_m);
else
delete a_m;
end;
end;
end;

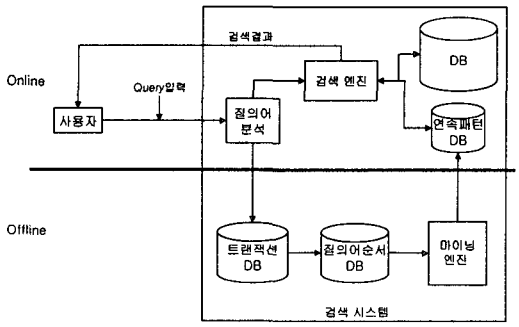
빈발 질의집합에서 사용자 검색 패턴 규칙을 추출하는 알고리즘

사용자 검색 패턴 규칙을 출력하는 두 번째 알고리즘은 각각의 빈발 질의집합에 대하여 질의어의 순서를 바탕으로 최소 신뢰도를 만족하는 사용자 검색 패턴 규칙을 찾아내고 질의 확장에 사용할 수 있도록 연속 패턴 데이터베이스에 규칙을 입력하는 과정을 수행한다.

4. 사용자 검색 패턴 정보를 이용한 질의 확장

본 장에서는 연관 규칙 탐사를 통해 얻은 사용자 검색 패턴 정보를 이용하여 질의 확장하는 방법에 대해 기술한다.

사용자가 입력한 질의어로부터 연관 규칙 탐사를 통하여 얻어진 질의어들의 연속 패턴 정보들은 질의 확장에 추가할 용어로써 사용된다. 질의 확장 방법으로는 불리언 검색에서 가장 기본적으로 사용되고 있는 AND 명령을 사용함으로써 수행이 된다. 즉 사용자가 입력한 질의어와 이 질의어에 의해 찾아질 수 있는 연속 패턴을 연속 패턴 데이터베이스에서 찾아내어 이 단어 다음에 올 수 있는 단어를 선택하여 AND 명령으로 처리하여 검색을 수행한다. 그러면 AND 명령으로 새롭게 재구성된 질의어는 실제로 검색할 문서가 있는 데이터베이스에서 사용자가 입력한 질의어와 연속 패턴으로 찾아진 용어가 모두 포함된 문서를 검색하여 사용자에게 제공하게 된다. 이러한 검색시스템의 전체 구성도는 다음과 같다.



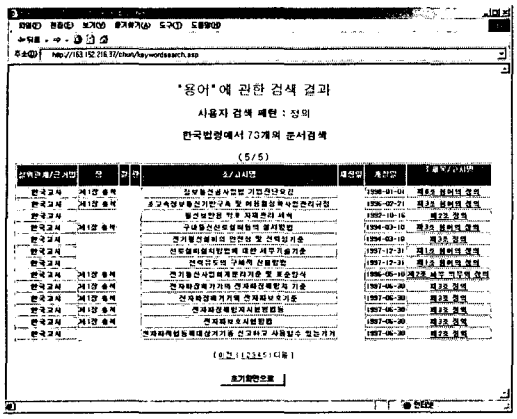
(그림2) USP를 이용한 검색시스템의 전체 구성도

(그림2) 에서와 같이 사용자 검색 패턴(USP) 정보를 이용한 시스템은 온라인과 오프라인으로 구성되어 있다. 온라인 상에서 검색시스템은 사용자로부터 질의어를 입력받으면 트랜잭션 데이터베이스에 저장하고 질의어를 명사단위로 분석을 하게 된다. 질의어가 하나의 명사이면 연속 패턴 데이터베이스를 검색하여 존재하는 가능한 패턴들을 찾아낸다. 만약 복합명사이면 단순명사로 분리한 후 패턴들을 찾아낸다. 질의어 분석 후 연속 패턴 데이터베이스로부터 찾아낸 검색 패턴들 중 가장 신뢰도가 높은 패턴을 선택하고 이 정보를 이용하여 질의 확장을 수행한다. 오프라인 상에서 검색 시스템은 온라인 상에서 발생한 모든 검색 트랜잭션들을 사용자에 따라 질의어순서 데이터베이스로 변환하게 된다. 질의어순서 데이터베이스는 본 논문에서 제시한 빈발 질의집합 추출 알고리즘을 사용하여 질의어들의 연속 패턴을 찾아 낼 때 사용되게 된다. 발견된 질의어 사이의 사용자 검색 패턴(USP) 정보들은 온라인에서 정보 검색시 사용될 수 있도록 연속 패턴 데이터베이스로 구축되고 검색 시스템은 이를 이용하여 질의 확장 검색을 수행하게 된다.

5. 구현 및 분석

5.1 구현

본 논문에서 제안한 사용자 검색 패턴 마이닝 알고리즘을 이용하여 검색 시스템을 구현하고 실험은 한국전자통신연구원에 구축된 정보통신 기술기준 정보시스템에서 사용되고 있는 데이터베이스를 이용하여 구현 및 실험을 한다. 먼저 사용자 검색 패턴 정보를 수집 및 패턴 탐사 할 수 있는 검색시스템을 구축한 후 실험 및 결과 분석한다. 검색 시스템은 Intel 계열의 PentiumIII 700MHz PC에서 Microsoft사의 IIS 4.0을 웹 서버로 사용하였고, DBMS은 MS-SQL Sever 7.0, ASP(Active Server Page)로 구현한다. 이 시스템은 온라인 상에서 수행되는 실제 검색부분과 오프라인 상에서 수행되는 사용자 검색 패턴 탐사 부분으로 구현이 된다.



(그림3) 사용자 검색 패턴을 이용한 정보 검색 결과

(그림3)은 사용자가 "용어"라는 질의어를 입력하였을 경우 질의 확장의 결과로써 사용자 검색 패턴, 검색 문서수, 검색된 문서의 내용등을 출력한 화면이다. 사용자가 "용어"라는 질의어로 검색을 수행하였을 때 검색 시스템은 연속 패턴 데이터베이스에서 사용자 검색 패턴인 "용어->정의"를 찾고, 이때 질의 확장 용어로써 "정의"를 선택하여 확장 검색을 수행하게 된다.

5.2 실험 결과 및 분석

실험은 기존의 키워드 매칭 방법으로 구현된 정보통신 기술기준 정보 시스템과 사용자 검색 패턴 정보를 이용한 검색 시스템을 통해 검색된 문서들의 수를 비교한다. 먼저 사용자 검색 패턴 정보를 이용한 검색 시스템에서는 연속 패턴 데이터베이스가 구축이 되어야하므로 오프라인 상에서 수행되는 사용자 검색 패턴 마이닝을 먼저 수행을 해야 한다. 사용자 검색 패턴 마이닝을 수행할 때 고려해야 할 사항은 패턴을 추출할 때 최소지지도 S_{min}과 최소신뢰도 C_{min}값이다. 이 두 값은 패턴 마이닝에 있어서 상당히 민감한 값으로 값이 너무 적으면 연속 패턴이 너무 많이 추출되어 검색시스템에 반영시 검색 효율이 많이 떨어질 수 있고 값이 너무 높으면 아주 유용한 패턴을 추출하지 못하고 삭제해 버리는 결과를 초래할 수 있다. 본 실험에서는 S_{min}과 C_{min}을 각각 2로 설정을 하였다. (그림4)는 사용자 검색 패턴 정보를 이용하여 검색한 결과와 키워드 매칭 방법을 이용한 결과를 문서수로 비교한 그림이다.

질의어	USP	키워드 매칭 검색		USP를 이용한 검색	문서수의 차이
		1차	2차		
무선	무선->설비	163	207	301	68
전기	전기->통신	80	325	342	63
설비	설비->선로	207	43	221	29
용어	용어->정의	10	72	73	9
방송	방송->무선	92	163	252	3

(그림4) 키워드매칭과 USP를 이용한 검색 방법의 결과 비교

(그림4) 에서 첫 번째 "무선"이라는 질의어에 대해 "무선->설비"라는 USP가 존재한다. 먼저 키워드매칭 방법에서는 1차와 2차를 합쳐 369개의 문서가 검색되었고 USP를 이용한 검색 방법은 301개의 문서를 검색되었다. 두 방법의 문서수 차이를 보면 68개의 문서수 차이를 보이고 있다. 이는 키워드매칭에 비해 한번의 검색만으로 68개의 중복되는 문서들을 줄이면서 1차와 2차 질의어를 모두 만족하는 문서들을 검색한 결과로 볼 수 있다. 위 전반적인 결과에서처럼 검색시스템을 사용한 사용자들의 검색 패턴을 이용하였을 때 입력 키워드에 대해 USP가 존재한다면 질의 확장을 통해 검색 횟수를 줄일 수 있고 횟수를 줄이면서 중복되는 문서 또한 줄일 수 있다는 것을 알 수 있다.

6. 결론 및 향후과제

기존의 정보 검색 시스템에서 질의 확장에 사용한 방법들은 시소러스(Thesaurus)를 이용하는 방법, 동시출현(Co-occurrence)의 수를 이용하는 통계적인 방법, 적합성 피드백(Relevance Feedback)을 이용한 방법, 단어와 단어 사이의 의미 시소러스(Semantic Thesaurus)를 이용한 방법 등이 연구되었고 최근에는 사용자 중심의 정보들을 이용하여 질의 확장하는 방법들이 연구되고 있다. 사용자 중심의 정보들은 사용자의 개인정보(Profile), 억세스패턴(Access Pattern), 지식베이스에서 흥미패턴(Interesting Pattern)등을 말하며 이 정보들을 이용하여 질의 확장에 첨가할 용어를 선택하는 척도로써 사용한다.

본 논문에서는 사용자 중심의 정보로써 사용자가 검색한 질의어들을 시간 순서에 따라 정의한 사용자 검색 패턴(User Search Pattern)을 마이닝 하는 알고리즘을 제안하고 마이닝된 정보를 이용하여 질의 확장 및 검색하는 모델을 제시하였다.

향후 과제로써 사용자 검색 패턴을 탐사하는 알고리즘에서 시공간적 축소를 위한 연구가 이루어 져야 할 것이다. 또한, 마이닝된 사용자 검색 패턴은 한가지가 아닌 다수의 패턴을 만들어 낼 수 있으므로 이를 이용하기 위해 대화형 질의 확장(Interactive Query Expansion)이 가능한 QBE(Query By Example) 인터페이스를 설계해야 할 것이다. 추가적으로, 사용자가 입력한 질의어들이 트랜잭션 데이터베이스에 저장이 되어 있다가 실제 연속 패턴 데이터베이스로 구축하기 위해서는 온라인이 아닌 오프라인에서 수행이 되어야 한다. 오프라인에서 시행되는 연속 패턴 탐사 과정을 실시간으로 수행하여 즉시 사용 가능한 데이터베이스로 변환하는 방법을 고려해 보아야 한다.

참고문헌

[1] 박종수, 유원경, 홍기형, 연관 규칙탐사와 그 응용, 한국정보과학회지 데이터 마이닝 특집 제 16권 제 9호 pp.37-44, 1998

[2] 김정자, 이도헌, 데이터 마이닝 기술 및 연구동향, 한국정보과학회지 데이터 마이닝 특집 제 16권 제9호 pp.6-14, 1998

[3] Jae-Hyun Lim, Hyon-Woo Seung, Jun Hwang, Heung-Nam Kim, Query expansion for intelligent information retrieval on Internet, Parallel and Distributed Systems, pp.656-662, 1997

[4] Montebello. M., Gray. W.A., Hurley. S., Evolvable intelligent user interface for WWW knowledge-based systems, Database Engineering and Applications Symposium, pp.224 -233, 1998

[5] I-Yuan Lin, Xin-Mao Huang and Ming-Syan Chen, Capturing User Access Patterns in the Web for Data Mining, Proceedings of the 11th IEEE International Conference, 345-348, 1999

[6] Bing Liu, Wynne Hsu, Lai-Fun Mun and Hing-Yan Lee, Finding Interesting Patterns Using User Expectations, IEEE Transactions on Knowledge & Data Engineering, V.11 N.6, 817-832, 1999

[7] Valery I. Frants, jacob Shapiro, Isak Taksa, Vladimir G. Voiskunskii, Boolean Search: Current State and Perspectives, Journal of the american society, pp. 86-95, 1999

[8] Rakesh Agrawal, Ramakrishnan Srikant, Mining sequential patterns, Data Engineering, Proceedings of the Eleventh International Conference, pp.3 -14, 1995

[9] Schutze H, Pedersen JO, a cooccurrence-based thesaurus and two applications to information retrieval, Information Processing & Management V.33 N.3, pp.307-318, 1997

[10] Yonggang Qiu, and H. P. Frei, "Concept Based Query Expansion, Proceedings of the Sixteenth Annual International ACM SIGIR Conference, pp.160-169, 1993

[11] Chang CH, Hsu CC, Enabling concept-based relevance feedback for information retrieval on the WWW, IEEE Transactions on Knowledge & Data Engineering V.11 N.4, pp.595-609, 1999

[12] R. Agrawal and R, Srikant, Fast algorithms for mining association rules, In Processings of the 20th VLDB Conference, santiago, Chile, Sept. 1994