자연어 저장소에 기반을 둔 자연어 질의처리를 위한 데이터베이스 활용 방안에 관한 연구

전익진, 이병래
한국방송통신대학교 정보과학과
e-mail : plusstar75@naver.com, brlee@knou.ac.kr

Utilization of A Data Base for Query Processing of natural language on the Repository of natural language

Danny Jeon, Byeong Rae LEE
Dept of InformationScience,KoreaNationalOpenUniversity

요 약

최근 웹을 기반으로 한 계속적인 기술 발전에 따라 의사결정에 필요한 데이터의 요구는 점점 다양해지고 있으며 다양한 요구를 효과적으로 대응하기 위해 데이터 추출 방법에 대한 연구도 지속적으로 이루어지고 있다. 이에 본 논문에서는 자연어를 통해 사용자가 쉽게 원하는 자료를 추출 할 수 있는 방법론을 연구 하였다. 자연어 처리 기술에 대한 연구는 여러 방면에서 이루어지고 있는데 그 중에서도 본 논문에서는 기존의 자연어 처리 연구를 바탕으로 크게 3가지 형태로 연구 진행 하였다. 사용자가 입력한 정보를 바탕으로 유추하여 자연어를 처리하거나 이후 진행될 검색을 선 예측 하는 방법과 사용자 별로 검색되는 자연어를 통해 연관 관계를 설정하여 사용자에게 예측검색을 유도하는 방법 그리고 의사 결정을 위해 구축된 데이터베이스 스키마 정보를 이용하여 사용자가 쉽게 질의 문을 생성할 수 있도록 하는 방법론 연구이다. 본 논문을 통해 연구된 내용은 실제 구축하여 진행 하였고, 연구 결과로 생성된 질의 문이 효과적으로 시스템에서 처리 되는 과정에 대한 연구도 함께 진행하고 검증하였다.

1. 서론

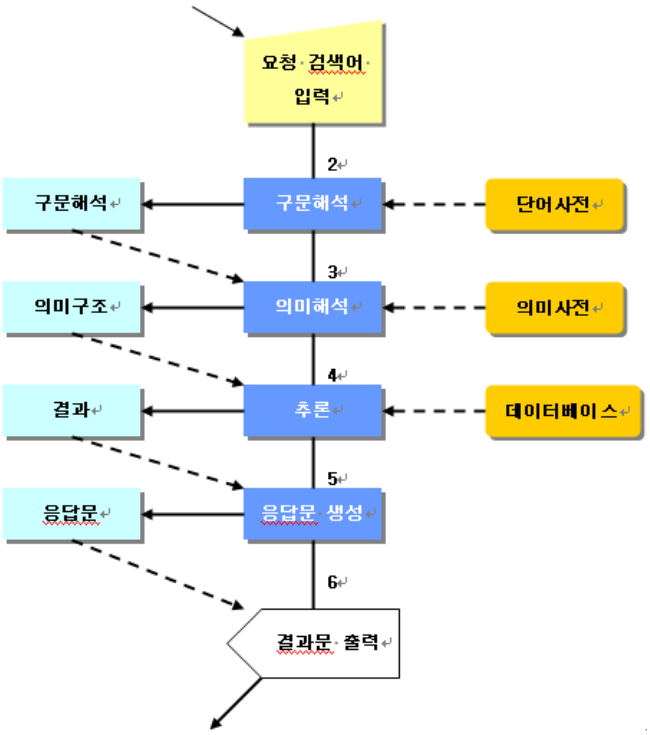
최근 들어 데이터베이스에 저장되는 데이터의 양이 급속도로 증가하고 있다. 이러한 추세는 발생하는 정보의 절대적인 양이 많아진 이유도 있지만 각각의 데이터로는 쓸모없이 보였던 정보가 축적되고 분석된 결과, 가치 있고 기업에서는 이윤이 되는 정보로 변화되었기 때문이라 할 수 있다. 경쟁에서 살아남기 위해 대용량의 정보를 보유하여 기업의 영업활동에 대처할 수 있는 시대를 요구하는 것이다. 이러한 환경의 변화는 더 빠르게 처리 할 수 있는 데이터베이스 시스템을 요구하게 되었으며, 그에 맞는 처리 기능을 요구하게 되었다. 계속 증가되는 대용량의 트랙잭션 처리시스템, 의사 결정 시스템, 멀티미디어 시스템 등 관리 할 자료의 규모가 증가함에 따라 자료 검색 성능 향상을 위한 연구가 지속적으로 요구되고 있는 것이다.
최근 사용자들은 웹 기술을 기반으로 언제 어디서나 의사결정에 필요한 데이터를 요청하며, 더욱 복잡하고 다양한 데이터를 요구하고 있다. 사용자마다 원하는 데이터의 유형과 처리된 결과가 서로 상이하기 때문에 데이터를 추출하고 가공 처리하는 과정을 사용자가 원하는 형태로 맞춰준다는 것은 쉬운 일이 아니다. 그렇기에 방대한 데이터 속에서 사용자가 직접 데이터를 추출 할 수 있다면, 사용자 별 맞춤 프로그램에 대한 압박으로부터 벗어 날 수 있을 것이다. 하지만 사용자가 직접 데이터베이스에 접속하여 데이터를 가져오는 질의 문을 입력한다는 것은 특정 분야이며, 특화된 기능을 요구한다. 따라서 사용자가 질의 문을 생성할 때, 일상에서 사용하는 자연어를 편하게 입력하여 연동처리 될 수 있다면, 이러한 문제는 해결 될 수 있을 것이다. 이에 본 논문에서는 자연어 저장소라는 개념을 도입하여 자동질의응답 처리에 대해 연구하고자 한다. 기존의 자연어 질의처리 연구에서 보편적으로 연구해 오던 패턴 분석 후 조합하는 기능적인 측면보다는 이를 포함하는 활용 방법론에 무게를 두고 연구할 계획 이다. 본 논문에서는 자연어 저장소 활용을 위해 검색 예측(Searching prediction), 연관 검색(Coupling Search), 시각화(Visualization)의 3가지 방법론을 제시하고자 한다.

2. 관련 연구

2.1 질의응답 시스템

질의응답 시스템은 사용자의 질의와 관련된 결과 정보를(일부에서는 문서라고 표현하기도 함.) 검색하는 정보검색 (Information Retrieval) 시스템과는 달리 사용자의 질의에 대한 답변이 될 수 있는 정답을 결과 집합 내에서 사용자에게 제시해주는 시스템이다. 일반적으로 질의응답

시스템은 사용자의 질의에 관련된 결과를 검색하는 후보 검색 단계 (candidate retrieval phase) 와 검색된 문서 내에서 정답을 생성하는 정답추출 단계 (answer extraction phase)로 구성된다. 일반적인 질의응답 시스템의 흐름은 [그림 1]과 같다.



[그림 1] 기본적인 질의응답 시스템의 흐름

2.2 연관 규칙

본 논문에서 다루게 되는 자연어 저장소의 검색 예측과 연관 검색은 연관 규칙(rule)을 기본 골격으로 하여 이루어진다. 연관 규칙은 a => b의 형태를 갖는 패턴으로서, a 와 b는 항목의 집합을 의미한다. 이 a => b 형태의 연관규칙이 갖는 의미는 a항목 집합이 나타날 때는 b항목 집합도 동반하여 나타나는 경향이 있다는 뜻이다. 연관 규칙의 구체적인 예를 확인해 보자. 일반적으로 마트에 들러 상품을 구매하는 가족 고객의 경우 식료품점에 들려 빵을 구매한 고객의 40%는 우유를 함께 구매하게 되고, 전체 구매 행위에 대해 2%가 위 규칙에 적용이 된다고 한다. 여기서 40%는 신뢰도를 의미하고, 2%는 규칙에 대한 지지도라고 한다. 위 규칙에 대한 알고리즘을 표현하면 다음과 같을 것이다.

For i = First Customer To Last Customer Step 1 //전체 고객에 대한 반복문을 실행
If Current Customer Buy() Then //고객의 구매 활동이 이루어지면
Total Buy Count = Total Buy Count + 1 //구매가 이루어진 고객에 대한 전체 구매 횟수를 증가
Purchasing Item = Buy Item // 구매 상품을 저장
Do Until Purchasing Item.eof // 구매 상품의 개수만큼 반복문 실행하여 구매 상품을 검사
If Purchasing Item = Bread And Purchasing Item = Milk Then // 구매한 상품이 빵과 우유일 경우
Association Rule Count = Association Rule Count + 1 // 연관 규칙 횟수를 증가 시키고
Else If Purchasing Item = Bread And Purchasing Item <> Milk Then // 구매 상품 중에 빵을 구매 했지만 우유를 구매하지 않은 경우
Only Bread Buy Count = Only Bread Buy Count + 1 // 빵만 구매한 고객으로 간주
End If
Purchasing Item.move next
Loop
End If
Next
Support Rate = (100 * (Only Bread Buy Count + Association Rule Count)) / Total Buy Count // 빵만 구매한 고객과 연관 규칙 고객을 합산하여 100을 곱하고, 이것을 전체 구매 내역으로 나누면 지지도를 구함
Confidence Rate = (100 * Association Rule Count) / (Only Bread Buy Count + Association Rule Count) //연관 규칙 고객에 100을 곱하고 빵만 구매한 고객과 연관 규칙 고객을 합산한 값으로 나누면 신뢰도를 구함

연관 규칙 탐사는 주어진 데이터베이스에서 사용자가 미리 정의한 최소 신뢰도와 최소 지지도를 초과하는 모든 연관 규칙을 찾아내는 것이다.

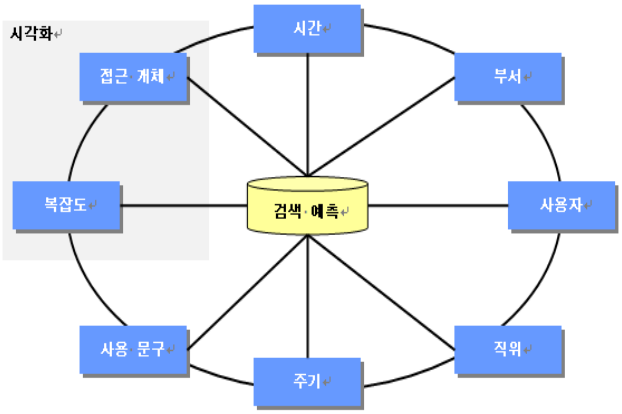
3. 자연어 저장소

사용자가 입력하는 자연어는 예측하기 매우 곤란하다. 사용하는 언어, 입력되는 장소, 시간, 검색하는 사람, 입력 주기 등 많은 것을 고려해야 하며, 모두 충족 되게 예측 되었다 하여도 그 정확도를 장담하기란 매우 힘들기 때문이다. 사용자가 입력한 자연어를 보다 효과적으로 처리하기 위해서는, 사용자가 입력할 자연어에 대한 데이터를 미리 예측 하여 저장해 두는 것이 기존의 구문 분석적인 방법 보다 효과적일 것이다. 본 논문 에서는 이렇게 미리 예측 된 대용량의 자연어를 마이닝하여 구축한 데이터베이스를 자연어 저장소라고 할 것이다. 구축 될 자연어 저장소에는 검색 예측 영역, 연관 검색 영역 그리고 시각화 영역으로 구분되어진다.

3.1 검색 예측

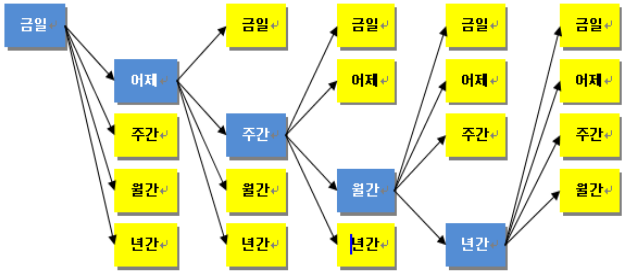
검색 예측이란 이전에 입력 된 자연어를 통해서 새로운 자연어 쿼리를 유추해 내는 것을 의미한다. 자연어를 입력한 사용자나, 반복적으로 입력되는 자연어 주기, 자연어에 포함된 시간, 자연어를 통해 접근되는 개체 정보, 자연어 속에 사용되는 문구, 그리고 얼마나 많은 개체와 상

관관계를 갖는지에 대한 복잡도 등을 통해 유추 가능하게 된다.



[그림 2] 검색 예측 테이블 구성을 위한 항목

입력된 자연어 속에 포함된 내용은 많은 사항을 유추 할 수 있다. 예를 들어 시간과 관련되어 유추 된 단어를 통해 폭넓은 자연어 처리가 가능하게 되는데, 만약 사용자가 금일이라는 단어 대신 '오늘' 혹은 '당일' 이라는 자연어로 대체하여 검색하여도 같은 질의 문을 호출하여야 한다. 금일 = 오늘(당일)이라고 해석이 가능하기 때문이다. 유추어를 뽑게 되면 사용자에게 관련 자연어를 제시할 때 유리해 진다. 반대로 유추 어에 해당 되는 자연어가 입력되어도 연쇄적으로 관련 유추 어를 제시 하게 된다.



[그림 3] 검색 예측에 따른 유추 어 연쇄 효과

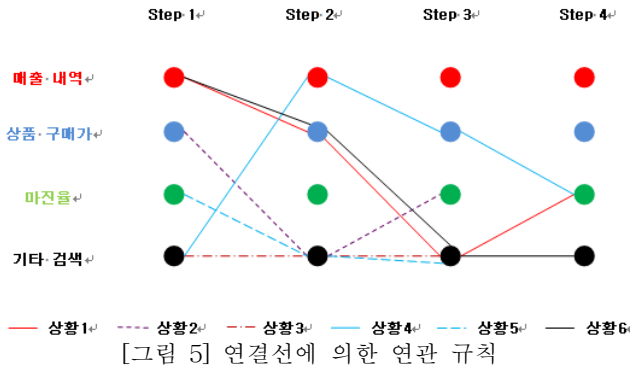
3.2 연관 검색

자연어 저장소는 일정한 규칙을 바탕으로 서로 상호 연결되어 정의 되거나 처리 되는 경우가 많다. 자연어 연관 규칙을 정의하는 대는 시간과 사용자 등의 항목이 효과적으로 사용되게 된다. 주기적으로 입력되는 자연어 검색 내용을 분석해 보면 일정한 연관 규칙을 생성 할 수 있다. 실제로 의사 결정에 필요한 데이터를 검색하는 사용자를 분석해 본 결과 아래와 같은 연관 규칙을 얻을 수 있다.

[그림 4]를 살펴보면 특정 검색어를 입력 할 때 일정하게 등장하는 연관 검색어를 확인 할 수 있다. 즉 매출 내역과 상품 구매가 그리고 마진율이 그것이다. 좀 더 명확하게 결론을 내리면, {매출 내역과 상품 구매가} -> {마진율}이 된다. 상황 별 검색 유형을 [그림 5]와 같이 연결선으로 표현해보면 좀 더 명확한 결과를 얻게 된다.

[그림 4] 검색 상황에 따른 연관 규칙

사용자가 일반적으로 매출 내역을 검색하고 뒤이어 상품 구매 가를 검색 한다면, 결과적으로 마진율을 검색할 확률이 높다는 것을 알 수 있다. 총 4단계의 자연어 검색을 진행 하면, 첫 번째 단계 혹은 두 번째 단계에서 매출 내역을 보고 그 다음으로 곧이어 상품 구매 가를 검색하게 된다는 가정 하에 다음 단계에서 바로 확인되지 않아도 대부분의 사용자는 마진율을 보고자 한다는 것이다.



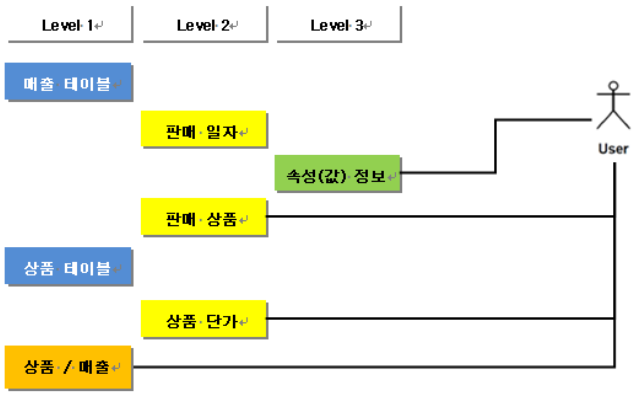
[그림 5] 연결선에 의한 연관 규칙

연관 검색은 트랜젝션 데이터를 분석하고, 그 분석한 데이터가 누적되어 갈수록 더욱 효과적인 데이터를 구축하게 된다. 사용자별 검색 성향, 시간대 별 검색 주기 등등으로 인해 쌓이는 데이터 하나하나가 연관 검색의 데이터로서 활용가치가 높다는 것이다.

3.3 시각화

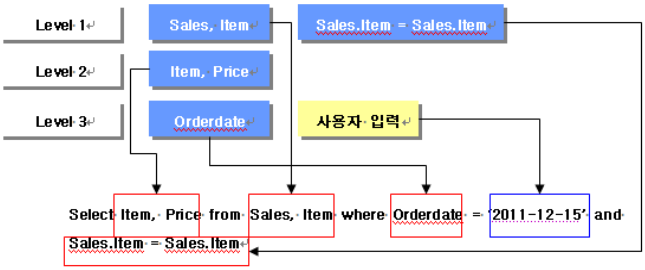
시각화를 간략하게 정의하면 자연어를 통해 해석된 데이터베이스 관련 스키마 정보라 표현할 수 있다. 이러한 시각화는 사용자로 하여금 자연어를 입력하는 수고로움도 덜게 한다. 즉 사용자는 자연어를 입력 할 수 있지만, 제공된 정보를 바탕으로 자연어를 생성 할 수도 있다. 사용자에게 제공된 스키마 정보를 효과적으로 사용하기 위해서 가장 필요한 부분은 각 단계에 대한 설정이 중요하다. 데이터베이스 정보를 가장 상위의 개념으로 보고, 테이블을 중심으로 하여 단계별로 해당 Level 값을 부여하여야 한다. 단계 정보는 사용자가 시각화를 통해 제공되는 자연어를 선택하는 순서를 부여하게 된다. 좀 더 자세한 설명을 위하여, 사용자가 시각화를 통해 제공된 자연어를 다음

그림과 같은 순서에 의해 선택을 한다.



[그림 6] 시각화 따른 사용자 단계별 선택

Level 1에서 선택되는 자연어는 테이블 정보로, Level 2에서의 선택은 조회 항목으로, 그리고 Level 3에서는 조건으로 처리하는 것이다. 시각화에 최대 장점은 사용자가 원하는 데이터를 주어진 정보의 선택만으로 얻을 수 있다는 것이다. 조합된 질의 문 결과는 다음 그림과 같다.



[그림 7] 시각화에서 선택한 자연어 질의 조합

4. 검증

실제 자연어 저장소를 통해 생성된 질의문은 효과적으로 결과물을 호출해야 한다. 이에 따라 자연어 저장소 기능을 통해 생성된 질의 문을 검증 진행 하였다. 자연어 저장소의 검색 예측을 통해서 만들어 진 "금일 매장 별 총 매출 데이터"에 대한 질의 원문은 다음과 같다.

SELECT sBranchCode, bName, sum(convert(bigint, sTotalPrice)) FROM [B_SalesInfo], B_BranchInfo where sBranchCode<>0 and sBranchCode=bCode and convert (varchar(10),sRegDate,120)=convert(varchar(10), DATEADD("yyyy",-1,getdate()),120) group by sBranchCode, bName having sum(convert(bigint,sTotalPrice))>0 order by sum(convert(bigint,sTotalPrice)) desc

실험 내용	Table Scan	Compute Scalar	Index Scan	Hash Match
Node No	8	7	6	5
I/O 비용	4.83	0	0.003	0
CPU 비용	0.32	0.03	0.0003	0.028
연산자	5.16(94%)	0.03(1%)	0.01(0%)	0.29(5%)
하위 트리	5.15	5.18	0.004	5.47
행 수	791.409	791.409	149	670.409
행 크기	21Byte	29Byte	26Byte	32Byte

[표 1] 조인 처리 단계에 대한 비용 분석 결과

위에 제시된 표에서 알 수 있듯이 본 논문에서 제시한 자연어 저장소를 통해 생성된 질의 문을 처리하는데 큰 무리가 없음을 알 수 있다. 전체 CPU에서 6% 미만의 비용으로 처리가 된다는 것을 확인 할 수 있다. 자연어 저장소를 통해 생성된 질의 문은 효과적으로 활용 가능 하다는 것이다.

5. 결론 및 향후 연구과제

논문에 제시된 자연어 저장소의 3가지 제안을 모두 수용하여 하나에 시스템에서 구현하기란 쉽지 않다. 검색의 정확도를 올리기 위해 처리 속도가 낮아질 우려가 있다. 그러나 이중 하나의 제안을 수용하여 시스템에 구현한다면, 자연어 저장소를 통해서 다양한 자연어를 보다 유연하게 처리할 수 있게 되며, 사용자에게 효과적인 자연어를 선 예측하여 제공하게 됨으로써, 자연어조차도 입력해야 하는 번거로움을 최소화 하게 된다. 이와 함께 사용자로 하여금 제안된 자연어로 검색을 유도하여 시스템 성능 또한 향상 시킬 수 있게 될 것이다. 본 논문은 자연어 저장소 구축을 위한 방법 보다 자연어 저장소 활용에 대한 방법론을 제시하고 있다. 자연어 저장소 구축이 선결 처리되어야 하는 문제를 안고 있다. 또한 아무리 많은 자연어를 확보 한다고 하여도 그 한계성은 분명히 존재 할 것이다. 해서 자연어 저장소를 구축하기 위한 방법론이 반드시 연구 되어 져야 한다. 그에 따른 한 방안으로 사용자 정의형 구축방법론을 제시하고자한다. 이는 사용자가 자연어 저장소에 존재하지 않는 새로운 형식의 자연어를 제시하고 이를 사용자간 서로 의미를 부여하며 정의하는 것으로 데이터를 구축해 나가는 것이다.

참고문헌

[1] 이동하 외 3명, "연관 규칙을 이용한 지능적 질의처리 시스템", 포항공대 지능정보시스템 연구실, 1998

[2] 박찬곤 외 4명, "문장패턴을 이용한 자연어 질의 시스템에 대한 연구", 청주대학교 컴퓨터정보공학과, 2003

[3] 김명관, 이영우, "웹 문서 정보추출과 자연어 처리를 통한 온톨로지 자동구축에 관한 연구", 한국인터넷방송통신TV학회지, 2009

[4] 임종현, "자연어 질의로부터 복합 웹 서비스의 자동생성", 연세대학교 컴퓨터과학과 석사졸업논문, 2008

[5] 김민경, "질의문 자동생성 방식의 질의응답시스템의 설계 및 구축", 서울시립대학교 전자전기컴퓨터공학부 석사졸업논문, 2009