CRM연구
제2권 제2호(2009. 12)

분석 CRM 실무자의 자연어 질의 처리를 위한 기업 데이터베이스 구성요소 인덱싱 방법론

박성혁
경영공학, KAIST 경영대학
(heypark@business.kaist.ac.kr)
황경서
경영공학, KAIST 경영대학
(kh299@business.kaist.ac.kr)
이동원
경영공학, KAIST 경영대학
(tunalee@business.kaist.ac.kr)

분석 CRM 영역에서는 고객 데이터 분석을 통하여 고객 행동과 관련된 통찰력을 얻는 것이 중요하다. 이러한 분석 과정에서, 사용자 스스로 기업 데이터베이스에서 대용량 고객 이력 데이터를 조회하고 추출하기 위해서는 SQL 을 사용하여 자유롭게 질의구문을 작성할 수 있어야 한다. 그런데 일반 사용자들이 이러한 업무를 수행하고자 할 때, 기업 데이터베이스 구성 요소에 대한 전문적인 지식이 부족하기 때문에 정보 탐색에 있어서 어려움을 겪는다. 이를 해결하기 위한 방안으로 본 연구에서는 사용자가 제공하는 자연어 수준의 질의를 분석하고, 데이터베이스를 구성하는 값을 중심으로 올바른 질의 결과를 제공하기 위한 데이터베이스 구성요소 인덱싱 방법론을 제안한다. 구체적으로 기업 데이터베이스를 구성하는 세 가지 요소인 관계, 속성, 값에 대한 정보를 읽어 들여 요약 정보에 대한 인덱스를 구성한 다음 사용자의 자연어 질의에서 분석된 의미 단위 별로 데이터베이스 요약 정보와 연결해주는 TableRank 기법을 소개한다. 실험용 데이터베이스를 대상으로 테스트를 수행한 결과, 사용자의 자연어 질의 결과가 데이터베이스를 구성하는 값 정보와 연결되는 것이 관찰되었다. 논문의 후반부에서는 자연어 질의를 자동적으로 처리하기 위한 선행 모듈 개발의 시사점을 정리하고, 향후 기업 데이터베이스 자동 검색 시스템으로 발전하기 위한 방안에 대해서도 설명한다.

※ 주제어: 자연어 질의 처리, 데이터베이스 인덱싱, 사용자 인터페이스, TableRank, SQL, 데이터 품질

Ⅰ. 서 론

기업의 실무자들은 정보시스템을 통해 자신이 원하는 정보를 얻기 위하여 데이터를 조회하는 등의 업무를 수행한다. 하지만 사용자가 스스로 데이터를 탐색하고 조회하는 과정에서, 찾고자 했던 데이터가 어느 곳에 위치하고 있는지 모르거나 데이터를 조회하기까지 소요되는 검색 시간이 많이 요구되어 불편을 겪는 어려움이 발생한다. 이러한 장애요인은 실무자들이 CRM 활동에 필요한 고객 데이터를 수집하고 분석 하는데 있어서 부정적인 영향을 미치는

데, 기업이 보유하고 있는 IT 인프라의 성능이 부족하기 때문에 발생하는 문제가 아니라, 기술적으로는 부족함이 없는 정보시스템을 사용자에게 친숙한 관점에서 사용하도록 하는 문제인 경우가 많기 때문에 정보시스템 활용 측면에서 그 해결방안을 모색해야 한다. 즉, 하드웨어적인 측면에서 시스템 개선을 고민하기보다는, 주어진 시스템 자원을 활용하는 소프트웨어적인 관점에서 접근해야 한다.

기업 내에서 사용자가 스스로 원하는 정보를 찾기 위한 방법으로는 SQL(structured query language)을 사용한 직접 질의 방식 또는 OLAP(online analytic processing) 솔루션에서 제공하는 UI(user interface) 툴에 의한 방식이 주로 사용된다. 두 경우 모두 사용자가 찾고자 하는 정보를 단순 조회 및 추출할 수 있도록 지원하는 일반적인 방식에 해당한다. 특히 OLAP 솔루션을 활용하는 쪽이 사용자 입장에서 기술적 진입 장벽이 낮기 때문에 더 일반적으로 사용되고 있지만, 사전에 데이터 큐브상에서 정의된 부분에 한해서만 정보가 제공된다는 특징 때문에, 다수의 사용자가 보고 싶어하는 다양한 종류의 정보를 제공하지 못한다는 문제가 있다. 또한, 사용자가 정보를 요청하였을 때 실시간으로 정보를 제공해주는 과정에서, 웹 상에서 데이터 조회 결과를 사용자에게 전달하기까지 시간이 많이 소요될 수 있다는 점도 현실적인 측면에서 실 사용자들이 불만을 제기하는 원인이 된다. 사용자에게 OLAP 방식보다 더 유연하게 데이터에 접근할 수 있도록 해주는 방법으로, SQL 을 이용한 능동적인 데이터 조회방식이 사용된다. SQL을 사용하면 기업의 물리적 데이터 저장소에 존재하는 데이터를 대상으로 사용자가 확인하고자 하는 모든 유형의 정보를 찾아볼 수 있다는 장점이 있지만, 사용자가 찾고자 하는 정보가 테이블 상의 속성 값으로 존재하는 단순한 값 수준에 그치지 않고 여러 종류의 테이블에서 기초 데이터를 가져온 다음에 복잡한 데이터 처리 과정을 거쳐서 만들어질수록, 분석용 질의어 개발을 위해 요구되는 전문적인 지식도 이에 비례하여 증가하게 된다는 단점이 있다. 실제로 이러한 문제 때문에, 다수의 일반 사용자들은 SQL 을 사용하여 원하는 데이터를 가공하고 추출하는데 어려움을 느낀다. 사실 더 중요한 문제는, SQL 을 사용하여 데이터 조회에 필요한 구문을 작성하기 위해서는 데이터베이스의 논리적 설계가 어떻게 되어있는지 알고 있어야 하는데, 일반 사용자가 기업 데이터에 대한 ERD(entity-relationship diagram) 또는 관계형 모형에 대해서 사전에 알고 있거나 학습을 통하여 이해하려고 하는 것이 현실적으로 매우 어려운 일이다.

SQL을 사용하여 정보를 탐색하는 과정에서 발생할 수 있는 문제에 대해서는 접근성과 관련된 데이터 품질(accessibility data quality) 이라는 주제로, 일반 사용자들이 기업의 데이터에 접근하는 과정에서 발생할 수 있는 장애요인에 대해 기업의 실 사용자 관점에서 살펴본 연구를 바탕으로 설명할 수 있다. Wang과 Strong(1996)에 의하면 데이터 품질에서 접근성 이라는 것은, 사용자가 업무에 필요한 데이터 또는 정보를 탐색하는 과정에서 발생할 수 있는 장애 요인에 관한 것으로, 사용자의 데이터 접근 과정에 장애가 되는 요인이 적을수록 접근성과 관련된 데이터 품질이 높다고 볼 수 있다. 일반적으로 데이터 접근성과 관련된 성공요소를 생각할 때, 안정된 네트워크 시설 또는 고성능 서버 장비 수준과 같이 기술적인 요소에 대해서 생각하기 마련이다. 하지만 현실적인 측면에서 우수한 네트워크 시설과 같은 고사양의 IT 인프라가 도입되었다고 하더라도, 실제 사용자를 곤란에 빠뜨리는 것은 데이터 조회와 관련된 하드웨어적인 성능 문제가 아닌, 데

이터 조작과 관련된 소프트웨어적인 문제에 있다. Strong 외 (1997)의 연구에 의하면, 데이터 사용자들이 데이터에 접근함에 있어서 어려움을 겪는 이유가 IT 인프라와 관련된 기술적인 한계점에 있다기 보다, 그들이 원하는 데이터를 확보하는데 장애요인을 제공하는 상황적 요인 때문이라고 한다. 상황적 요인의 예를 들면, 복잡한 코드 값으로 변환된 데이터들의 경우 사용자가 손쉽게 조회해 볼 수 있지만, 코드화된 결과 값을 해석하는 노력이 추가적으로 요구되며, 그 노력이 많이 필요하기 때문에 현실적으로 데이터 접근성 문제가 해결되지 않은 셈이다. 또한, 기업이 보유한 다수의 데이터베이스와 데이터웨어하우스 상에 저장된 데이터에 기술적으로는 모두 접근 가능하더라도, 사용자가 원하는 데이터가 어느 위치에 있으며, 그것을 사용자에게 친숙한 업무 환경으로 추출할 수 있는 지식을 보유하고 있지 않다면, 기술적 가능성과는 독립적으로 그러한 데이터는 접근 불가능한 것이 된다. 정리하면, 기업의 정보시스템에서 보유하고 있는 데이터 중에서 사용자가 원하는 데이터가 어느 곳에 위치하는지 알기 위해서는 실 사용자 관점에서 데이터 관련 지식이 필요한데, 일반적으로 데이터에 대한 수요가 있는 조직의 구성원들이 데이터 조회와 관련된 전문적인 지식을 아는 것은 무리가 있으며, 데이터 접근에 있어서 주된 장애 원인이 된다. 만약, 사용자 입장에서 기업의 데이터를 추출하는데 필요한 전문가 수준의 지식이 없이도 자신이 원하는 데이터가 어디에 있는지 찾을 수 있도록 지원해줄 수 있는 시스템이 존재한다면, 일반사용자들의 데이터에 대한 진입 장벽이 크게 낮아질 것이다. 구체적으로, 사용자가 기업의 정보시스템에서 원하는 데이터 또는 정보를 찾도록 해주는 검색 환경이 주어진다면, 사용자는 접근성과 관련된 데이터 품질 문제와 상관 없이, 자신에게 친숙한 자연어(natural language) 수준에서 데이터를 손쉽게 검색하고 원하는 결과를 확인할 수 있을 것이다. 여기서 손쉬운 검색이라는 것은 마치 웹 상에서 구글이나 네이버와 같은 검색 사이트에 접속하여 사용자가 원하는 정보를 찾기 위하여 핵심 단어 또는 문장의 형식으로 질문하고 그에 대한 답을 제공받는 검색 서비스를 의미한다. 기업 데이터베이스를 대상으로 하는 검색 엔진을 존재한다면, 사람들이 웹에서 자유로운 검색활동을 통해 그들이 원하는 정보를 손쉽게 찾는 것처럼 기업 내 데이터 수요자가 원하는 데이터를 손쉽게 찾을 수 있게 된다. 예를 들어, 기업 마케팅 부서의 담당자가 “서울에서 구매한 이력이 있는 20대 여성”에 대한 리스트를 원하거나, “온라인 매장에서 반품한 경험이 있는 남자 고객”에 대해서 조회하고 싶은 경우, 온라인 검색 엔진에서 관련 결과를 찾아보듯이 기업의 데이터베이스로부터 관련 정보를 제공받을 수 있다면, 누구나 손쉽게 정보를 취득할 수 있으므로 접근성과 관련된 데이터 품질 문제를 데이터 품질을 향상시킬 수 있을 것이다.

본 연구에서는 분석 CRM 강화 측면에서 데이터 접근성 품질 문제를 향상시키기 위한 기업 데이터베이스 구성 요소 인덱싱 기술에 대해서 소개하고, 정보 제공 시스템으로서 실제 기업에서의 활용 가능성에 대해서 다루고자 한다. 단계 별로 살펴보면, 먼저 기업의 실무자들이 데이터를 조회하는 방식에 따라 기업 데이터베이스에 저장된 데이터를 사용자 관점에서 조회할 수 있도록 하기 위한 이론적 가능성에 대해서 설명하고, 대표적인 웹 검색 방식으로 널리 알려진 PageRank 기술을 데이터베이스에 적용해 본 TableRank 검색 방식에 대한 시스템 아키텍처를 소개한다. 그리고 샘플 데이터를 사용한 기초적

인 실험을 통해 자연어 수준에서 요구되는 사용자 질의에 대한 결과를 데이터베이스 자동 검색 엔진이 얼마나 정확하게 제공할 수 있는지를 평가한다.

본 논문은 다음과 같이 전개된다. 제 2장에서는 관련 연구에 대해서 정리하는데, 데이터 조회 방식에 따른 데이터 품질 문제를 정의하는 부분과 자연어 처리를 위한 데이터베이스 검색엔진과 관련된 내용을 언급한다. 제 3장에서는 데이터베이스 구성 요소 인덱싱 기술에 대해서 소개하고, 제안된 TableRank 시스템이 사용자의 자연어 질의를 어떻게 처리할 수 있는지에 대해서 설명한다. 제 4장에서는 제안한 검색 엔진을 대상으로 실험을 수행한 결과를 정리하며, 마지막 장에서는 본 연구의 시사점과 앞으로의 연구 방향에 대해서 정리한다.

II. 관련 연구

2.1 데이터 조회 방식에 따른 데이터 품질 문제

분석 CRM 영역에서 고객 데이터를 분석하는 방식에 있어서 한 가지 특징적인 구분 기준이 될 수 있는 것은 데이터 조회 업무를 직접 수행하는 주체가 본인인지 또는 다른 누군가에게 요청된 것인지 여부이다. 자신이 스스로 데이터를 탐색하는 경우와, 조직 내 누군가에게 요청을 한 다음 관련 정보를 전달받는 두 가지 상황으로 구분되는데, 각각의 특징에 대해서 다음과 같이 설명할 수 있다. 먼저 스스로 데이터를 탐색하는 경우를 보면, 사용자가 원하는 데이터를 찾고 가공하여 필요한 정보를 생성하는데 요구되는 지식의 수준이 높다는 단점이 있지만, 관련 지식과 기술적 소양이 충분한 사용자라면 본인이 찾고자 했던 정보를 언제든지 손쉽게 얻을 수 있다는 장점이 있다. 데이터를 얻는 과정에서 필요한 지식과 기술력을 보유하고 있다면, 데이터 수요자가 직접 탐색하도록 하는 것이 가장 정확하고 효과적인 방법이다. 하지만 대부분의 사용자들은 이와 관련된 전문 지식이나 기술적인 소양이 부족하며, 이들을 대상으로 교육을 통하여 데이터 조회에 대한 전문가로 만드는 일 역시 현실적으로 쉬운 것은 아니다. 또한 모든 사용자가 데이터 조회에 대한 전문적인 역량을 갖추었다고 하더라도, 주관적으로 정보를 가공하여 사용할 수 있다는 문제점이 여전히 존재한다. 즉, 조직 내에서 “작년 신규 고객 수”, “올 해 상반기 VIP 회원 고객 매출 총합”과 같은 질문에 대하여 조사된 데이터 값은 모두 동일해야 할 것인데, 데이터 조회 과정에서의 주관적인 판단으로 인하여 사용자별로 서로 다른 값을 주장 할 수 있다는 것이다.

다음으로, 데이터 수요자 스스로가 아니라 다른 사람에게 데이터 탐색 업무를 요청하는 경우를 살펴보자. 일반적으로 데이터 조회를 전문적으로 수행하는 분석가에게 조직 내에서 업무가 집중될 수 있는데, 최초 질의자가 원하는 정보와 완벽하게 일치하는 값을 얻는데 한계가 있으며, 소수의 분석 전문가들이 조직 내의 수 많은 질의를 담당하는 것이 쉽지 않기 때문에 효율적인 측면에서도 어려움이 발생한다. 또한 질의를 요청하고 결과를 전달받기 까지 소요되는 시간도 이러한 방식의 업무 처리 효율성을 낮추는 원인으로 지적된다. 다만, 중앙 통제 방식으로 다수의 사용자에게 기업 내의 정보를 제공한다는 측면 때문에 정보의 객관성이 확보된다는 장점이 있다.

이상에서 살펴본 두 가지 상황에 대해서, 실제로 많은 기업들이 조직 내 구성원들 스스로가 원하는 것을 얻을 수 있도록 만들어 갈 것인지, 소수의 데이

터 분석가가 데이터 전사적 분석 요구를 담당하도록 할 것인지에 대해서 고민하고 있다. 이에 대한 답은, 사용자가 원하는 정보를 무리 없이 얻을 수 있도록 하는데 더 효과적인 상황으로 설정되어야 한다는 원칙을 바탕으로 고려되어야 한다. 사용자 관점에서 사용 목적에 적합한 데이터가(fitness for use) 고품질의 데이터라는 개념에 의하면, 데이터 품질을 평가함에 있어서 사용 목적에 적합한 데이터인지를 생각해보는 것이 가장 중요하게 요인이 된다(Deming 1986, Juran 1989, Wang & Strong 1996). Wang과 Strong(1996)은 최종 사용자 관점에서 데이터의 품질을 정의하고, 평가하기 위한 개념적 틀로써, 설문을 통해 밝혀낸 다음의 네 가지 데이터 품질 유형: 내재적 품질(intrinsic DQ), 접근성 품질(accessibility DQ), 상황적 품질(contextual DQ), 그리고 표현적 품질(representational DQ)을 설명하고 있다. 구분된다. 〈표 1〉에서는 네 가지 데이터 품질 카테고리와, 각 카테고리 별 세부 데이터 품질 구성 요소가 정리되어 있다.

기업에서 사용자들이 원하는 정보를 찾고자 할 때, 가장 많이 사용하는 두 가지 정보 조회 방식에는 SQL 을 사용한 정보 조회 방식과, OLAP 을 사용한 정보 조회 방식이 있다. OLAP 솔루션은 사전에 데이터 분석가에 의해서 제공된 정보 중에서 개인 사용자 별로 원하는 차원(dimension)의 조합을 선택한 다음, 본인이 확인하고자 하는 값(score)을 볼 수 있도록 지원한다. 사용 방법이 상대적으로 쉽지만, 원하는 정보가 사전에 정의되어 있지 않을 수 있으며, 특정 차원의 조합에 대해서 존재하지 않는 값에 대한 정보를 파악하기 어렵다는 단점이 있다. SQL의 경우는 사용자가 확인하고 싶은 모든 유형의 정보를 추출할 수 있지만, 난이도에 따라서 SQL을 사용해 구문을 작성하는데 어려움이 있을 수 있으며, SQL 언어 구사 능력만큼이나 기업 데이터베이스에 대한 전반적인 이해 수준이 높아야 하기 때문에 일반 사용자들에게는 진입 장벽이 높다. 이상의 두 가지 정보 조회 방식을 정보 가공 프로세스 관점에서 비교해보면, 어떤 방식을 채택하는 것이 데이터 조회에 대한 근본적인 해결책이 될 수 있을지 파악하는데 좋은 시사점을 제공한다.

먼저, 정보 가공 프로세스 개념에 대해서 살펴보도록 하자. 정보 처리 과정에서 정보 제조 시스템에 관여하는 세 가지 역할을 다음과 같이 설명할 수 있다(Ballou 외 1996, Wang 1998). 맨 첫 번째 역할을 담당하는 정보 제공자(information suppliers)는 최종 정보 제품(information products)의 원료가 되는 데이터를 수집하거나 생성하는 역할을 하고, 두 번째 역할자인 정보 가공자(information manufacturers)는 주어진 데이터를 가지고 모형

〈표 1〉 데이터 품질 유형과 세부 요소 (Wang & Strong 1996)

Data Quality Category	Data Quality Dimension
Intrinsic DQ	Accuracy, Objectivity, Believability, Reputation
Accessibility DQ	Accessibility, Access security
Contextual DQ	Relevancy, Value-Added, Timeless, Completeness, Amount of data
Representational DQ	Interpretability, Ease of understanding, Concise representation, Consistent Representation

을 개발하거나 상위 수준의 경영지표들을 파생변수의 형태로 생성하고 관련 데이터를 사용하여 새로운 정보 가치를 창조한다. 이들이 하는 업무는 마치 자동차 조립 라인에서 주요 부품들이 모듈 형식으로 만들어지듯이, 데이터를 가공하여 적정한 규모와 의미를 갖는 상위 수준의 정보로 가공하는 일이다. 마지막 역할인 정보 소비자(information consumers)는 그들의 업무에 필요한 정보를 사용하여 보고서를 작성하거나 의사 결정을 내리는 등의 활동을 수행한다. TDWI(the data warehouse institute) 보고서에 의하면 기업 데이터 품질 문제의 76%가 조직 구성원에 의한 사유로 발생한다고 하는데(Eckerson 2002), 특정 정보를 탐색하는 상황에 대해서 각 역할 담당자들 간에 발생하는 문제를 줄일 수 있다면, 데이터 품질 문제 역시 개선될 수 있을 것이다.

이제, 정보 가공 프로세스 측면에서 세 가지 역할 담당자에 대한 설명을 바탕으로, 정보 탐색에 대한 두 가지 유형을 비교하는 문제를 다시 정리해보자. OLAP 방식은 사전에 정형 장표 형태로 웹 상에 표준화된 정보를 게시하는 역할을 하는 정보 가공자와, 웹에 접속하여 정돈되어 있는 정보 중에서 자신이 원하는 정보를 탐색하는 정보 소비자로 역할이 구분되기 때문에, 실제 사용자 관점에서 고려할 수 있는 모든 유형의 정보를 제공할 수 없다. 따라서, 사용자 지향적인 서비스가 될 수 없다는 구조적인 문제를 가지고 있다. 정보 품질 문제 관점에서 살펴보면, 정보 소비자가 원하는 목적에 완벽하게 부합하도록 정보 가공자가 역할을 수행하기 어렵기 때문에, 정보 가공 단계에서 내재적 품질 중에서도 목적성(objectivity), 신뢰성(believability)에서 문제가 제기될 수 있으며, 그 결과 정보 소비자는 잘못된 정보가 가공되었다고 판단하거나(wrongly value-added), 본인이 원래 생각했던 정보를 받지 못했다고(not relevancy) 생각할 수 있다. 그 결과 정보 가공자 또는 OLAP 자체에 대한 명성(reputation)이 낮아지게 될 것이고, 장기적으로 일반 사용자들이 OLAP 솔루션을 사용하지 않게 될 것이다. 반면에 사용자 스스로 SQL로 질의어를 작성하여 정보를 탐색하는 경우에는, 정보 가공자와 정보 사용자가 일치하게 되므로 앞에서 제시한 목적성, 신뢰성에 대한 문제는 상대적으로 감소하게 될 것이다. 다만, 사용자가 원하는 수준의 정보가 SQL로 구현되기까지 복잡다단한 수준의 노력이 요구된다는 부분이 일반 사용자들이 보편적으로 SQL을 사용하여 데이터를 조회하는데 가장 큰 장애요인이 된다는 점은 분명한 사실이다.

사용자가 SQL을 사용하여 기업 내무 데이터를 조회하고자 할 때에 발생할 수 있는 문제점은 앞서 소개한 네 가지 데이터 품질에 대한 유형 중에서도 접근성(accessibility)과 관련된 데이터 품질 문제와 관련이 있다. 기업의 정보시스템이 제공해주는 데이터 자체는 고품질의 데이터라고 가정할 때, 실제 데이터 사용자가 그 정보를 조회하고 추가 분석을 수행 하는데 어려움을 느낀다면, 정보에 대한 접근성 품질이 낮은 문제로 인하여 데이터에 대한 활용이 어렵기 때문인 것으로 평가할 수 있다. 예를 들어, 분석 CRM 영역에서 우수 고객의 구매 이력 데이터를 조회하고자 할 때, 고객 데이터가 전체 테이블 중에서 어떠한 속성들로 구성되어있는지, 관련 데이터 레코드는 어떻게 확인해야 하는지를 모른다면 아무것도 확인해볼 수 없기 때문이다. 가령, 특정 지역에 거주하면서 A 매장에서의 구매가 활발한 고객을 찾고자 하는 간단한 질문 답을 하기 위해서는 고객과 영업 매장에 대한 지역 정보는 어떻게 확인 할 것이며, 고객 이력은 어떻게 집계해서 보고할 것인지를 데이터 처리 관점에서 알고 있어야 한다.

만약 개인 사용자가 원하는 정보를 얻을 수 있도록 하는 질의어를 SQL로 손쉽게 작성하도록 해주는 기능을 제공할 수 있다면, SQL 을 사용하여 정보를 조회하는 장점을 최대한 강조할 수 있을 것이다. 일반사용자들이 SQL을 손쉽게 작성하기 위해서는 앞에서 다음의 두 가지 문제를 극복할 수 있어야 한다. 구체적으로는, 사용자가 탐색하고자 하는 정보가 기업 데이터베이스 내부에서는 어떻게 표현되는지를 알아야 하는데, 관계형 데이터 테이블을 구성하는 여러 속성 들 중에 어떠한 부분에 해당하는지 식별되어야 한다. 또한, 다수의 데이터 테이블들 간의 관계 정보를 파악하여, 앞 단계에서 식별된 데이터들을 찾아 연결하고, 필요에 따라 2차, 3차적으로 가공해주는 논리가 SQL에 반영되어 작성될 수 있도록 지원받아야 한다. 이 내용에 대해서는 다음 절에서 더 자세히 설명되어있다.

마지막으로, 전략적 측면에서 생각해보면, 더 많은 사람들이 데이터에 대해서 이해하고, 기업 데이터의 활용 가치의 중요성을 인식할수록 데이터 기반의 과학적인 의사결정을 할 수 있게 될 것이다. 이를 위해서는 조직 내의 전략 담당자 또는 분석 CRM 담당자들 중에서 한 명이라도 더 많은 사람들이 의사결정을 위하여 데이터 활용이 중요하다는 것을 인식하고, 고객이 발생시키는 데이터 자체를 조직의 주요 자산으로 인정하도록 해야 한다(Redman 1995). 따라서, 일반 사용자들이 SQL을 사용하여 원하는 정보를 얻을 수 있도록 지원하는 것이 중요하며, 더 많은 사용자가 기업 데이터 분석에 대해서 관심을 가질 수 있게 하기 위하여 데이터베이스에 대한 이해 없이도 정보 검색이 자연어 수준에서 가능하게 해주는 응용 시스템이 설계되어야 한다. 기술수용모형(technology acceptance model) 관점에서 보면, 사용자들이 특정 기술을 받아들이기 위해서는 인지적 유용성(perceived usefulness)과, 사용의 편리성(perceived easy-of-use)이 모두 확보되어야 하는데(Davis 1989, Venkatesh 외. 2003), 데이터베이스 자동검색엔진은 사용자가 SQL 을 사용하여 직접 질의어를 작성하는데 필요한 테이블과 속성에 대한 정보를 제공해줌으로써, 사용의 편리성 인지 수준을 높여주기 위하여 제안되었다. 이를 통해 SQL 을 이용한 질의어 작성의 기술적 난이도가 낮아지면, 기술수용모형에서 밝혀진 인과관계에 따라 인지적 유용성 및 실제 사용 의도가 증가할 것이며, 궁극적으로 정보시스템 상에서 SQL을 이용한 질의 처리 방식이 일반화될 수 있을 것으로 기대한다.

2.2 자연어 질의 처리를 위한 사용자 인터페이스 디자인

데이터베이스를 활용하는 과정에서 자연어 인터페이스의 역할은 사용자가 평소에 사용하는 언어 습관대로 원하는 데이터를 직접 조회를 할 수 있게 해주는 것이다. 이를 통해 SQL을 이용하여 질의어를 작성할 줄 모르는 일반 사용자들에게도 데이터에 대한 접근을 가능하게 한다. 사용자가 입력하는 자연어 질의어의 의도대로 데이터베이스가 문제없이 인식하기 위해서는 자연어 의미 파악 및 데이터베이스 구조와의 원만한 연결을 위한 아키텍처 설계가 잘 진행되어야 한다. 먼저, 자연의 의미 파악 문제에서는, 사용자가 입력한 문장 형식의 질의어에서 각 단어들이 무엇을 의미하고 있는지, 전체 문장에서 특정 단어들의 조합이 새로운 의미 단위로서 역할을 하는지에 대서 확인할 수 있어야 한다. 다음으로, 아키텍처 설계에서는 사용자의 최초 자연어 질의가 데이터베이스의 물리적 구조에서 정보를 조회하는 과정으로 연결될 수 있도록 처리되어야 한다. 하지만, 사용

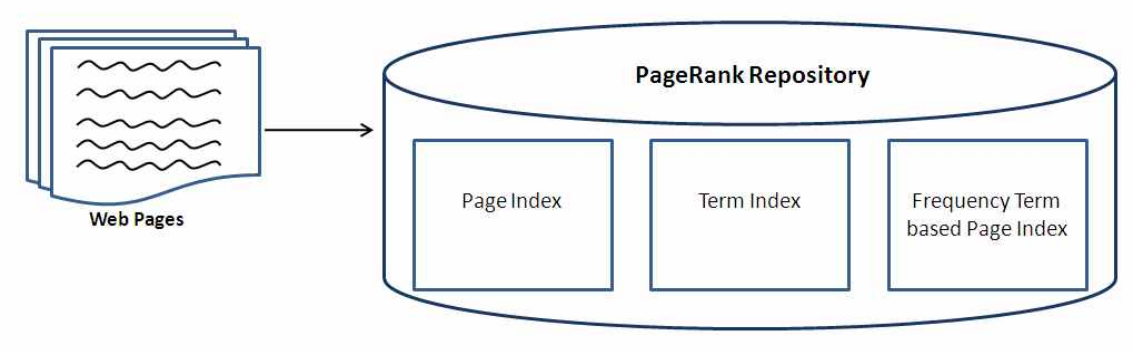
자의 자연어 질의가 어떤 의미로 제안된 것인지를 자동적으로 파악하는 연구 자체가 해결되기 어려운 문제이며, 사용자의 자연어 질의를 완벽하게 파악하고 핵심 단어들을 분리하였다고 하더라도 데이터베이스 상의 정보 값들과 정확하게 연결시키는 것 또한 쉽게 해결하기 어려운 주제이다. 그래서 이러한 두 가지 문제를 한꺼번에 다루기 보다는, 단계 별로 처할 수 있는 문제 상황을 중심으로 연구가 진행되어 온 사례가 있다. 예를 들어, 사용자의 자연어 질의로부터 핵심 개념들이 잘 정리된 상황을 가정한 다음 데이터베이스 스키마에서 이와 관련된 정보들을 자동으로 찾아 제공해주는 연구가 주로 수행되어 왔으며(Li 외. 2005), 사용자의 자연어 질의 의미를 정확하게 파악하기 위한 의미 추론 연구가 진행된 바 있다(Popescu 외. 2003).

사용자가 입력한 자연어 질의로부터 자동으로 결과를 제공하는 과정에 대해서 다음과 같이 설명할 수 있다. 먼저, (1) 사용자로부터 자연어 질의를 입력 받으면, (2) 자연어 분석을 통하여 주요 핵심 단어(key words)들을 분리해내고, (3) 데이터베이스 스키마에서 관련 테이블과 속성들을 파악한 다음, (4) 질의 결과를 제공하도록 하는 SQL을 생성하고, 최종 단계에서 (5) SQL 수행 결과를 사용자에게 제공해준다(Androutsopoulos 외. 1995). 이상의 주요 과정에 대해서 우리의 연구에서는 다음과 같은 상황을 가정한다. 기업 내부의 분석 CRM 담당자들은 최소한 본인의 업무 영역에 대한 전문지식이 있으므로, 자연어 질의를 구성하는 핵심 단어에 대한 패턴 연결 방식으로 질의를 조회할 수 있다. 이 경우, 데이터베이스 스키마와 자연어 질의 분석 결과를 연결해주기 위해서는, 데이터베이스 스키마에 대한 이해가 선행되어야 하며, 데이터베이스를 구성하는 주요 요소들을 사용자 자연어 질의와 연결시켜 주기 위한 방안을 찾는 것이 중요하다.

데이터베이스를 구성하는 세 가지 요소는 관계(relations), 속성(attributes), 그리고 값(values)이다. Popescu 외. (2004)의 연구에서는 사용자의 자연어에서 추출한 언어의 의미 단위(token)와, 데이터베이스를 구성하는 관계, 속성, 값을 적절히 연결해주는 하나의 조합을(T, E, M)으로 표현한다. 구체적으로, T는 의미 단위를 뜻하는 토큰의 약자이고, E는 데이터베이스의 요소(elements)를 의미하며, M은 둘 사이를 연결시켜주기 위한 정보(matching information)를 의미한다. 이렇게 요약된 정보를 바탕으로 질의 결과를 제공하기 위한 SQL이 생성되고, 데이터베이스에서 SQL 수행을 거쳐 최종 결과가 사용자에게 전달된다. 우리의 연구에서는 분석 CRM 담당자가 기업 데이터베이스에 포함된 값에 대한 정보를 이해할 수 있도록 지원하고, 속성이나 관계(또는 테이블)에 대한 지식 없이도 원하는 정보를 찾을 수 있도록 하는 검색 엔진을 제안한다. 즉, 값을 중심으로 사용자의 질의어를 분석하고, 데이터베이스에서도 값을 중심으로 해당 질의와 관련이 높은 결과를 제공해주는 시스템을 디자인하는 것이 목적이다. 실제로 우리가 제안한 데이터베이스 검색 엔진은 구글의 웹 검색 방식으로 많이 알려진 PageRank 방식을 기초로 한 데이터베이스 인덱싱 기술을 사용하고 있다. 여기서는 웹 검색을 위한 PageRank 기술의 작동 원리에 대해서 살펴보도록 하자.

웹 크롤러를 통해 모아진 웹 데이터는 PageRank 알고리즘을 통해 자료가 토큰의 형식으로 정리가 된다. 다음 〈그림 1〉에서는 PageRank 저장소가 어떠한 정보를 담고 있는지를 보여주고 있다. PageRank 방식에 대한 이해를 위해 간략히 도식화된 형태로 웹 페이지 자료가 정리된다. 세부 요소들은 페이지

〈그림 1〉 PageRank 정보 저장소(repository)



인덱스, 용어 인덱스, 그리고 페이지 인덱스 별로 집계된 용어의 빈도 정보의 세 가지 이다. 먼저, 페이지 인덱스는 웹 페이지를 번호로 표시하기 때문에, 데이터 처리 시 빠른 성능이 나올 수 있도록 지원한다. 다음으로, 고유의 용어 인덱스 역시 각 단어 별로 인덱스가 형성되며, 각 페이지 인덱스 별로 용어에 대한 빈도 정보가 집계되어 저장된다(Langville & Meyer 2006). 즉, 마지막 단계에서는 각 용어 별로 페이지에 언급된 횟수를 기입하게 된다. 이 파일은 가장 낮은 수준의 단어 별 검색 과정에서도 이용되는 기초 자료로서 사용된다.

Ⅲ. 데이터베이스 자동 검색 기술

데이터베이스 자동 검색 기술 구현을 위해서 가장 먼저 고려되어야 할 사항은 사용자에게 자연어 질의가 가능할 수 있도록 하는 질의 처리 기술 및 환경 제공이다. 이 장에서는 자연어 처리를 위한 기반 기술에 대한 소개를 한 다음, 본 연구에서 중점적으로 다루고 있는 데이터베이스 자동 검색 기술을 설명한다.

3.1 사용자의 자연어 질의 처리 기술

데이터베이스에 대한 지식을 가지고 있지 않거나 현 데이터베이스 구조를 이해하지 못할 경우 SQL 질의를 통해 정보를 검색하는 것은 불가능한 일이기 때문에, 이러한 문제점을 해결하기 위해 자연어 검색을 통한 데이터의 위치 검색 기술이 제공되어야 한다. 이를 위해 가장 먼저 수행되어야 할 과제는 자연어 검색이 가능하도록 기존의 데이터베이스에 대한 분석 방법이 필요하다. 이러한 자연 검색의 기반 마련을 위해 현재 구글에서 사용되는 PageRank 알고리즘을 기반으로 데이터베이스 자동 검색 기술을 위한 데이터 분석 방법을 적용한다.

3.1.1 PageRank 설명

PageRank는 구글의 웹 크롤러(crawler)를 통해 수집된 다양한 웹 페이지 정보에 대하여 자연어 검색을 위한 용어 인덱싱 및 가중평균 방식의 순위계산법을 위한 기반 기술이다. PageRank는 웹 페이지의 정보를 단어 별로 분할한 후 각 단어의 빈도수

를 페이지 별, 그리고 단어 별로 정보를 정리하게 된다. 영문법의 장점인 문장의 각각의 단어를 기초로 영문 정보를 토큰(Token)별로 정리하게 된다. 예를 들어 "Park works in Accounting in US"라는 검색문장을 사용자가 입력하게 될 경우, "Park" "Work" "Accounting" "US" 등의 단어들을 PageRank에서 정리된 용어 인덱스를 기초로 검색하게 된다. 구체적으로, PageRank로 정리된 웹 페이지 중 "Park" "Work" "Accounting" "US" 등의 단어가 포함된 웹 사이트를 위주로 검색하여 결과를 만든다. 용어에 대한 인덱싱 작업이 완료되면, 각 페이지 별로 참조된 된 횟수를 측정하여 추후 사용자가 검색을 할 경우에 적용될 가중치에 대한 계산 근거를 마련하게 된다. PageRank로 정리된 결과에서 나온 웹 페이지와, 각 페이지 마다 저장되는 참조 가중치 값을 적용하여 최종적인 검색 결과를 만들게 된다. 이 결과를 통해 사용자는 가중치가 높은 조건에 대한 결과 리스트를 받게 되며, 이중에서 사용자가 선정하는 결과가 무엇이냐에 따라 데이터베이스 상에서 특정 값이 위치한 관계와 속성을 보여주어 사용자가 원하는 데이터를 찾는데 도움을 주게 된다.

3.1.2 데이터베이스에 PageRank 방식을 적용 할 때 고려사항

PageRank의 알고리즘은 자동으로 웹 페이지를 수집하여 모은 후, 이에 대한 분석을 하여 추후 사용자 검색 과정에서 사용자 질의에 맞는 웹 페이지를 보여주는 것에 기반을 두고 있다. 또한 영문법을 기반으로 한 자연어 처리에 매우 뛰어난 성능을 자랑하는 만큼 데이터베이스에 저장된 각종 데이터를 대상으로 우수한 성능을 기대해볼 수 있다. 하지만 데이터베이스를 대상으로 PageRank 기술을 적용해 보기 위해서는 먼저 웹 상의 검색과 데이터베이스 상의 검색에서의 차이점을 알아보아야 한다.

먼저 웹 상에서 사용자가 검색을 할 경우, 검색 질의어와 가장 연관성이 높은 URL 주소를 찾아주는 것이 핵심이다. 이를 위해 인터넷 검색 엔진들은 웹 주소와 주소의 이름, 그리고 검색 질의의 단어를 포함하는 정보를 함께 보여주어 검색 결과에 대한 사용자의 이해를 돕는 것이 일반적이다. 또한 검색 결과로 인해 사용자는 각각의 웹 페이지를 방문하여 검색에 대한 결과를 확인하게 된다. 이 과정에서 사용자의 경우 웹 상에서 문서 혹은 페이지에 대한 구조의 이해가 전혀 없이도 결과를 얻을 수 있다는 점이 가장 큰 장점이다.

하지만 데이터베이스의 경우는 검색에 대한 조건과 대상이 조금씩 다르다. 데이터베이스를 검색하고자 하는 사람들은 기업에서 정보를 찾고자 하는 직원들이 일반적이며, 이들은 최소한 자신의 업무 분야에 대한 이해 수준이 높은 사람들이다. 앞에서 설명한 것처럼, 일반 사용자들은 데이터베이스에서 정보를 얻기 위하여 IT 부서의 전문가나 데이터 분석가에게 관련 작업을 요청한다. 타인에 대한 요청 작업이 필요한 이유는 데이터베이스 구조에 대한 완벽한 이해나 SQL 사용 능력 등이 부족한 관계로 개인 사용자가 데이터를 검색 하는 것이 매우 어렵기 때문이다. 우리의 아이디어는, 웹 상에서 대다수의 일반 사용자들이 검색을 통해 원하는 정보를 얻듯이, 기업의 일반 사용자들도 데이터베이스에서 정보 검색을 손쉽게 할 수 있도록 지원하자는 것이다. 즉, 웹 검색 기술에서 URL 경로를 제공해 주었듯이, 데이터베이스 검색에서는 사용자가 원하는 값이 담긴 테이블과 속성의 위치를 제공할 수 있다는 것이다. 두 경우 모두 데이터가 존재하는 "장소"에 대한 검색에는 큰 변화가 없다. 다만 가장 큰 차이라면 데이터들이 한 페이지에 모두 존재하는 웹 페이지에서와는

달리 데이터베이스에서는 여러 개의 관계 혹은 속성들이 구조와 규칙에 따라 분산되어 저장되어 있을 뿐이다. 단일 페이지만 찾아주는 일반 웹 검색 엔진과는 달리 데이터베이스에서는 사용자가 원하는 정보와 연관된 테이블과 속성을 모두 찾을 수 있어야 하고, 각 요소들의 위치를 파악할 수 있도록 결과값을 만들어야 한다. 아래 〈표 2〉에서는 인터넷 검색과 데이터베이스 검색의 차이점에 대해서 검색 대상, 목적, 결과라는 항목 별로 정리해서 보여주고 있으며, 〈표 3〉에서는 수동 검색 방식과 자동 검색 방식에 대하여 구조적 이해, 기술적 이해, 검색 일치성이라는 항목을 구성하여 앞에서 설명해온 내용을 정리하고 있다.

3.2 데이터베이스 자동 스캐닝 기술

3.2.1 데이터베이스에 구성 요소 인덱싱을 위한 PageRank 응용 기술

앞서 설명한 것처럼, 일반적인 PageRank 기술은 웹 페이지 마다 등장하는 단어의 빈도수를 인덱스로 저장해 놓는 것으로 검색에서 활용된다. 하지만 데이터베이스의 경우 여러 개의 서로 연관된 테이블과, 테이블을 구성하는 속성이 값에서 등장하는 단어의 빈도수에 대해서 인덱스 정보를 저장해야 한다는 차이점이 있다. 이러한 차이점을 위해 〈그림 2〉에서 보여주는 것처럼 수정된 PageRank 저장소를 디자인하였다. PageRank 저장소의 페이지 인덱스 개념에서 확장하여, 테이블 인덱스와 속성 인덱스를 각각 저장하게 된다. 또한 단순히 속성 값에 등장하는 단어 별 페이지 빈도수를 기입하는 것이 아니라, 단어 별 속성 빈도수를 저장하도록 설계되었다. 이를 통해 단어 별로 어느 속성에 위치하는지를 추적할 수 있으며, 속성에 대해서도 인덱스에 기록된 위치 정보를 바탕으로 어느 테이블에 위치하고 있는지를 알 수 있도록 지원한다. 이 두 가지 정보로부터, 개별 단어 마다 테이블과 속성에 대한 위치 제공이 가능해진다.

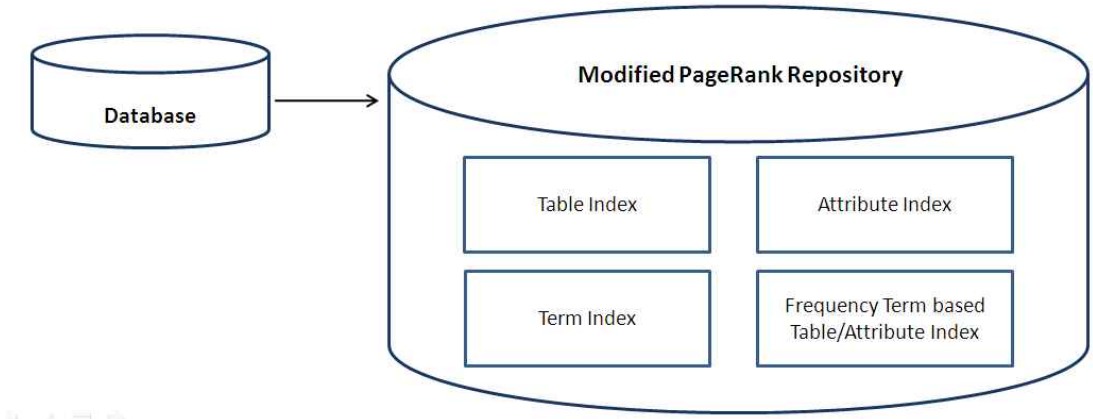
〈표 2〉 인터넷 검색과 데이터베이스 검색의 차이

검색 환경	검색 대상	검색 목적	검색 결과
웹 검색	웹 페이지	페이지 방문	검색당 1개의 웹 페이지가 나옴
데이터베이스 검색	테이블	데이터 포함된 테이블 위치 파악	검색당 N 개의 테이블이 나옴

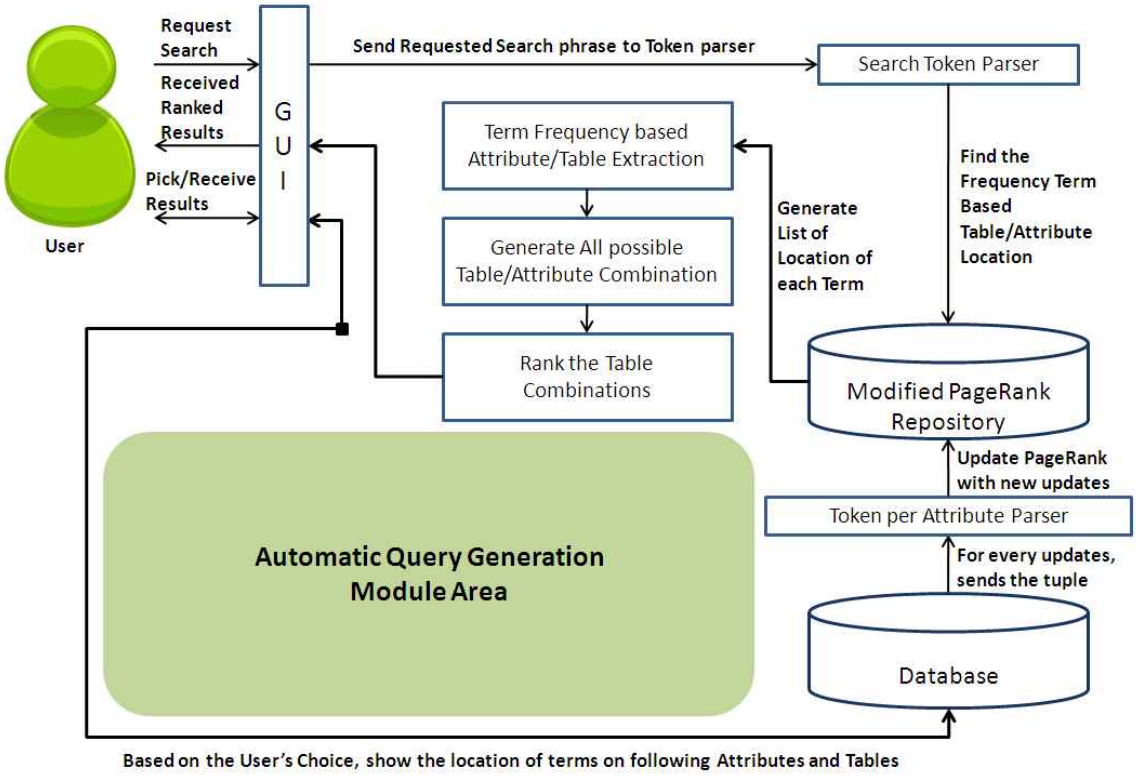
〈표 3〉 수동과 자동 데이터베이스 검색의 차이

환경	구조적 이해	기술적 이해	검색 일치성
Manual 질의	테이블의 구조에 대한 이해가 필요	SQL 작성을 위한 지식이 필요	질의문 작성자에 따라 결과가 다를 수 있음
Automatic 질의	필요 없음	필요 없음	자연어 검색에 대한 결과값이 일정하게 나옴

〈그림 2〉 데이터베이스 구성 요소 인덱싱을 위해 수정된 PageRank 저장소



〈그림 3〉 TableRank 시스템 아키텍처



3.2.2 TableRank 시스템에 대한 아키텍처

〈그림 3〉에서는 TableRank 방식이 적용된 데이터베이스 검색 시스템 아키텍처를 설명하고 있다. 일반적인 검색 엔진과 같이 사용자가 자연어 기반의 검색 질의를 보내게 되면 미리 생성된 수정된 PageRank 저장소에 담긴 위치 정보를 검색하게 된다. 위치 정보 검색 결과를 기반으로 각 단어에 대한 빈도 정보를 기초로 하여 최종 결과물을 생성한 다음, 질의 제공자에게 보내주게 된다. 사용자는 그 결과물을 기반으로 각각의 테이블 및 속성이 정확하게 제공되었는지를 확인하여 자신이 원하는 테이블 혹은 속성의 위치를 찾을 수 있게 된다. 주목할만한 점은, 논문의 서론에서 강조했던 것처럼, 사용자 관점에서는 데이터베이스에 구조에 대한 이해가 없이도 이러한 정보를 제공받을 수 있다는 점이다. TableRank 시스템의 자동화를 위한 주요 정보 처리 단계는 다음과 같다.

3.2.2.1 수정된 PageRank 저장소 생성

TableRank 시스템을 운용하기 위해서는, 신규 생성되는 데이터 레코드에 대해서 수정된 PageRank 저장소에 업데이트해주는 것이 중요하다. 이 과정에 대해서는 다음과 같이 세 단계로 수행된다.

1. 데이터 베이스에 자료가 업데이트 되거나 새롭게 생성 혹은 삭제 될 경우, 각 데이터 레코드에 대하여 수정된 PageRank 저장소 업데이트 리스트에 등록된다.
2. 등록된 업데이트 리스트에 따라 각각의 속성 값에 있는 문자열들을 단어 별로 토큰으로 분리해준다.
3. 각각의 토큰 중 용어 인덱스에 토큰이 존재하지 않는다면 새로운 인덱스와 함께 그 값을 등록시킨다. 빈도 기반의 용어 및 속성 인덱스에 토큰이 존재하는 속성 인덱스를 기입한 후, 해당 빈도수 정보를 함께 기입하는 것으로 수정된 PageRank 저장소에 업데이트 한다.

이상에서 설명한 것처럼, 데이터베이스와의 긴밀한 상호 작용을 통해 사용자는 항상 새롭게 갱신되는 검색 인덱스 정보를 사용할 수 있게 된다. 또한 최초의 모든 데이터베이스를 읽어 들이는 작업과는 달리, 추가적으로 발생하는 정보에 대해서만 업데이트 해줌으로써, 전체 검색 인덱스 정보의 최신성을 유지하는 갱신작업이 매우 빠르고 효과적으로 이루어진다는 장점이 있다.

3.2.2.2 사용자 자연어 질의 검색 처리 프로세스

사용자가 TableRank 방식의 검색 시스템을 활용하여 자연어 질의에 대한 검색 결과를 얻는 과정에서 대해서는 다음과 같이 여섯 가지 단계로 구분하여 설명할 수 있다.

1. 사용자가 검색 질의어를 입력하면, 검색 질의어 중에서 검색에 의미 있는 단어를 분리한 다음 여러 개의 토큰으로 만든다.
2. 분석을 통해 만들어진 각각의 토큰에 대하여 수정된 PageRank 저장소의 인덱스 정보를 통해 빈도 값 별 위치를 파악하게 된다. (본 논문에서는 빈도수가 높을수록 해당 속성 인덱스의 가중치가 높아지도록 설정함)
3. 각 단어 별로 검색된 속성 혹은 테이블 인덱스에 대하여 1차 검색 결과를 정리한다. 이전의 단계에서 측정된 추정치에 대해서, 각 위치 별 빈도 값을 함께 기록한다.
4. 1차 검색 결과를 기반으로 각각의 토큰이 위치하는 테이블 인덱스에 대하여 생성 가능한 모든 테이블 그룹의 조합을 생성한다.

5. 가능한 모든 경우의 테이블 그룹 조합을 기반으로, 각각의 테이블에 대한 토큰 일치율(match rate), 고유 테이블의 수 등 사용자가 검색 시 기대하는 최소한의 테이블 수와 최대한의 단어 인식 수준, 그리고 단어 별 높은 빈도수에 유리하도록 가중치를 계산하여 순위를 측정한다.
6. 최종적으로 측정된 순위를 기반으로 최종 검색 결과 리스트를 생성하여 사용자에게 보여주면, 사용자는 본 검색 결과를 토대로 결과에 대해서 테이블 및 속성의 위치 정보를 얻게 되며, 각 위치에 해당하는 데이터 값은 데이터베이스에서 제공받는다.

이상에서 설명한 사용자 자연어 질의 처리 프로세스를 통해, 사용자는 주요 단어 별 위치 인덱스를 통해 사용자가 원하는 정보가 어떻게 분산되어 있는지를 파악할 수 있다. 또한, 다음 검색을 위하여 어떻게 하면 원하는 정보를 효과적으로 검색할 수 있는지를 설계할 수 있게 된다.

Ⅳ. TableRank 생성과정에서 주요 산출물 및 최종 결과 값

앞 장에서 설명한 데이터베이스 검색엔진의 작동원리 별로 시스템을 테스트 하기 위하여, 기업 샘플 데이터를 대상으로 다음과 같이 실험을 진행하였다. 주요 산출물들은 다음의 〈그림 4〉에 정리되어있다. 먼저 사용자가 입력한 자연어 기반의 검색 질의어에 대해서 단어 별로 검색에 유의한 단어를 선정하게 된다. 이렇게 판별된 각각의 단어를 Token이라 부르며 각 단어에 대하여 위치를 파악한 후 위치 별 단어 빈도수를 검사한다. 이러한 위치 별 빈도수를 기초로 토큰들이 존재하는 테이블들에 대한 조합 가능한 그룹을 만들게 된다. 이 그룹들은 토큰의 수, 고유한 데이블의 수에 따라 다시 한 번 재 정렬되어 사용자에게 전달된다. 예를 들어 최종 결과물의 첫 번째 값은 사용자가 찾고자 하는 세 가지 토큰을 포함하는 고유 테이블들의 개수가 두 개이며, 다른 검색결과들 보다 효율적인 테이블 조합을 보여주기기 때문에 최상위에 등록되어 있다. 즉, 적어도 한 테이블에 두 개 이상의 토큰이 존재하게 되므로, 사용자 입장에서 데이터를 찾거나 검색할 때 드는 비용이 최소한으로 줄어들게 된다. 또한 위치 결과값을 기반으로 자신만의 SQL 구문을 생성하고자 할 때 꼭 필요한 정보를 제공받는 셈이 된다. 반면, 가장 하위에 정렬된 결과값은 "박씨"라는 토큰을 포함하는 테이블이 그룹에 존재 하지 않으며, 나머지 두 개의 토큰 역시 한 테이블이 아닌 분산된 서로 다른 테이블에 존재하게 된다. 이는 사용자가 요구했던 질의어에서 추출된 토큰이 모두 포함되지 않았기 때문에 검색결과가 만족스럽지 않을 가능성이 높다는 것을 의미한다. 하지만 토큰을 모두 포함하는 결과값을 제공받는 것이 항상 검색에서의 최우선이 되는 것은 아님에 주의해야 한다. 만약 사용자의 자연어 질의에서 의미 있는 것으로 판단된 각각의 토큰들 중에서 사용자가 원하는 결과값에 의미가 없는 것이 포함될 위험이 있기 때문에, 다양한 가능성에 대하여 검색결과가 유연하게 대응하기 위한 조건으로 토큰이 포함되지 않을 경우에 대한 조합 또한 최종 결과값 목록에 포함되어있다.

〈그림 4〉 TableRank 생성 과정 및 결과 값 예시

검색어를입력해주세요:
미국에서 회계로 일하고 있는 박씨

의미 있는단어별 속성 결과값

1. 박씨
속성 인덱스23번: 주소
테이블인덱스4번에 1번 나옴

2. 미국
속성 인덱스0번, 나라 코드
테이블인덱스0 번에 1번 나옴

속성 인덱스 27번, 나라 코드
테이블인덱스4 번에 4번 나옴

3. 회계
속성 인덱스4번: 부서 이름
테이블인덱스1 번에 1번 나옴

속성 인덱스19번: 직업 이름
테이블인덱스3 번에 1번 나옴

PageRank 기반 단어별 랭크

박씨	미국	회계
423	427	14
-	00	319

테이블별 가능조합생성

박씨	미국	회계
423	427	14
-	427	14
423	00	14
-	00	14
423	427	319
-	427	319
423	00	319
-	00	319

조합별 순위 결정 및 결과값 생성

박씨	미국	회계		고유 테이블
423	427	14	단어수:3 테이블수:2	(41)
423	427	319	단어수:3 테이블수:2	(43)
423	00	14	단어수:3 테이블수:3	(401)
423	00	319	단어수:3 테이블수:3	(403)
-	427	14	단어수:2 테이블수:2	(41)
-	427	319	단어수:2 테이블수:2	(43)
-	00	14	단어수:2 테이블수:2	(01)
-	00	319	단어수:2 테이블수:2	(03)

Ⅴ. 결 론

분석 CRM 실무자가 기업 데이터베이스에서 고객 데이터와 같은 정보를 조회하고 추출하고자 할 때, 기업 데이터베이스에 대한 지식 없이도 자연어 수준에서 정보를 얻을 수 있도록 지원하는 데이터베이스 구성 요소 자동 인덱싱 기술에 대해서 소개하였다. 이를 위하여, 논문의 앞부분에서는 SQL을 사용하여 정보를 조회하는 방식과, OLAP 와 같이 중앙에서 관리되어 최종 사용자에게 일방적으로 통보되는 정보 조회 방식의 효율을 비교하였으며, SQL 을 이용한 정보 추출 방식을 추구해야 한다는 점을 주장하였다. 그리고 일반 사용자들도 기술적 어려움 없이 SQL 을 사용하는 것처럼 원하는 정보를 탐색할 수 있도록 하기 위하여, 웹 검색에서 사용되어온 PageRank 방식을 개선한 TableRank 방식을 새롭게 제안하였다. 이 기술을 기반으로 데이터베이스 구성 요소인 관계, 속성, 값에 대한 인덱스 정보를 수정된 PageRank 저장소에 정리하는 프로세스를 개발하였으며, 일반 사용자의 자연어 질의 결과를 기업 데이터베이스의 구성 요소와 자동으로 연결시켜준 다음, 그 결과를 사용자에게 제시할 수 있도록 하였다. 기업 예제 데이터를 대상으로 실험을 수행한 결과, 사용자의 자연어 질의에 대한 데이터베이스 구성 요소 인덱싱 결과가 오류 없이 제공된다는 것이 확인되었다. 사용자는 데이터베이스 검색 시스템에서 제공하는 결과 정보를 사용하는 것으로, 데이터베이스 구조에 대한 이해 없이도 데이터 조회 및 추출을 위한 SQL을 작성하는데 중요한 정보를 제공받는다.

향후 연구에서는 사용자의 자연어 질의 결과에 해당하는 데이터베이스 구성 요소 정보를 받아서 SQL을 자동으로 생성해주는 모듈이 개발되어야 할 것이며, 실제 기업데이터를 대상으로 일반 사용자들이 요구하는 다양한 수준의 질의에 대한 결과가 잘 얻어지는지에 대해서 충분한 테스트를 거쳐 시스템의 성능을 평가하여야 한다.

참고문헌

Androutsopoulos, I., Ritchie, G. D., and Thanisch, P. (1995), Natural Language Interfaces to Databases - An Introduction. *In Natural Language Engineering*, 1(1), 29-81.

Ballou, D.P., Wang, R.Y., Pazer, H. and Tayi, G.K. (1998), Modeling information manufacturing systems to determine information product quality. *Management Science*, 44(4), 462-484.

Demng.(1986), *Out of Crisis*. Center for Advanced Engineering Study, Massachusetts Inst. of Technolog3r, Cambridge, Mass.

Eckerson, W. W.(2002), "Achieving Business Success through a Commitment to High Quality Data," *TDWI Report Series*, The Data Warehousing Institute, 5.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319-340.

Juran, J. M., and Godfrey, A. B.(1999), *Juran's Quality Handbook*, Fifth Edition, McGraw-Hill.

Langville, A. and Meyer, C.(2006), *Google's PageRank and Beyond: the Science of Search Engine Rankings*. Princeton University Pres.

Li, Y., Yang, H., and Jagadish, H. V.(2005), NaLIX: an Interactive Natural Language Interface for Querying XML. In *Proceedings of the ACM SIGMOD, Baltimore, Maryland*. p.900-902

Popescu, A., Etzioni, O., and Kautz, H.(2003), Towards a theory of natural language interfaces to databases. *In Proceedings of IUI-2003*

Popescu, A., Armanasu, A., Etzioni, O., Ko, D., and Yates, A.(2004), Modern natural language interfaces to databases: composing statistical parsing with semantic tractability. In *Proceedings of the 20th international Conference on Computational Linguistics,* Geneva, Switzerland, August 23-27.

Strong, D. M., Lee, Y. W., and Wang, R. Y.(1998), Data quality in context. *Communications of ACM*, 40(5), 103-110.

Redman, T. C.(1995), Improve Data Quality for Competitive Advantage, *Sloan Management Review*. 36(2), 99-107,

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003), User acceptance of information technology: Toward a unified view. *MIS Quarterly*. 27(3), 425-478.

Wang, R.Y. and Strong, D. (1996), Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*. 12(4), 5-34.

Wang, R.Y.(1998), A product perspective on total data quality management. *Communications of ACM*. 41(2) 58-65

A PageRank based Data Indexing Method for Designing Natural Language Interface to CRM Databases

Park, Sung-Hyuk*
Hwang, Kyeongseo*
Lee, Dong-Won*

Abstract

Understanding consumer behavior based on the analysis of the customer data is one essential part of analytic CRM. To do this, the analytic skills for data extraction and data processing are required to users. As a user has various kinds of questions for the consumer data analysis, the user should use database language such as SQL. However, for the firm's user, to generate SQL statements is not easy because the accuracy of the query result is hugely influenced by the knowledge of work-site operation and the firm's database. This paper proposes a natural language based database search framework finding relevant database elements. Specifically, we describe how our TableRank method can understand the user's natural query language and provide proper relations and attributes of data records to the user. Through several experiments, it is supported that the TableRank provides accurate database elements related to the user's natural query. We also show that the close distance among relations in the database represents the high data connectivity which guarantees matching with a search query from a user.

※ Key Words: Natural Language Interface, Automatic Query Generation, Analytic CRM, SQL, Data Quality

* Business School, Korea Advanced Institute of Science and Technology