IT 업체정보검색시스템에서 동의어 처리 기법

강옥선(충북대 정보산업공학과)
이현철, 조완섭(충북대 경영정보학과)

요 약

일반적인 정보 검색은 색인어를 통해 이루어지는데 이런 경우 사용자는 정보를 검색하기 위해 데이터베이스에 저장된 정보들이 가지고 있는 색인어를 정확하게 입력해야 한다. 그러나 일반 사용자가 색인어를 정확하기 입력하기는 어렵고, 특히 찾고자 하는 분야가 전문 분야에서 사용되는 용어일 때는 더욱 그러하다. 이럴 때 시소러스와 같은 지식구조를 이용해서 색인어를 탐색하여 검색의 효율을 높일 수 있다.

최근 들어 정보기술 분야의 연구가 활발함에 따라 정보자료의 생산이 급격히 증가하고 이를 관련 주제 분야의 연구정보로 활용하는 경우가 증가하고 있다. 따라서 IT 분야의 정보를 관리할 수 있는 시스템의 개발이 시급하다. 또한 IT 분야와 같은 전문분야일 때 검색 시스템에서 사용할 용어의 관리에 대한 연구의 필요성이 증가하고 있다.

본 논문에서는 IT분야의 정보를 검색할 수 있는 IT 업체정보검색시스템에서 정보 검색시에 생기는 용어간의 불일치 문제를 해결하고, 각 용어들간의 계층 관계를 나타내어 정보 검색시 검색어의 확장을 도울 수 있는 용어 관리 시스템의 구조를 제안하고 그에 대한 검색 알고리즘을 제시한다. 제안된 구조는 사용자의 검색어에 대한 동의어 관계나 상위어, 하위어 등의 계층 관계를 파악하여 검색의 범위에 추가함으로써 검색 효율을 높일 수 있다. 또한 새로운 용어의 생성이나 삭제와 같은 연산이 발생했을 때 시스템을 동적으로 확장할 수 있도록 구현하였다. 제안된 시스템은 단어간의 계층 구조를 효율적으로 검색하기 위하여 객체-관계형 데이터베이스를 사용하였다. 또한 메모리 상주 DBMS를 사용하여 많은 사용자들이 동시에 접근하는 환경에서도 빠른 검색 성능을 유지할 수 있도록 하였다. 제시된 방법은 정보기술 분야뿐만 아니라 다른 전문용어 분야의 연구로도 그 범위를 확장 할 수 있다.

본 문

IT 업체정보검색시스템은 IT 분야의 업체, 제품, 솔루션, 기술, 용역 과제, 전문가 정보 등을 제공하는 시스템이다. 사용자가 검색어를 입력하면 관련 업체 정보나, 제품 정보, 솔루션 정보를 제공한다. 이때 사용자는 데이터베이스에 저장되어 있는 색인어를 정확하게 입력해야 한다. 하지만 일반 사용자가 이러한 색인어를 정확하게 입력하기는 어렵다. 사용자가 'DB'라고 입력한다면 동의어로 쓰이고 있는 '데이터베이스'나 'database'는 검색할 수 없다. 그러나 제안하는 용어 검색 기법은 IT 업체정보검색시스템에서 이러한 용어간의 불일치 문제를 해결한다. 사용자가 입력한 검색어가 색인어와 일치하지 않더라도 동의어나 상위어, 하위어를 검색하여 검색의 범위를 확장하도록 설계하였다.

IT 업체정보검색시스템은 그래픽 사용자 인터페이스(GUI) 환경을 통해 사용자가 검색어를 입력하면 검색어 확장 모듈을 통해 동의어나 상위어, 하위어를 검색하고 새로운 확장 질의를 생성한다. 새로운 질의는 질의 모듈을 통해 IT 데이터베이스를 통해 최종 질의 결과를 출력하게 된다.

본 논문에서는 검색시스템에서 사용자의 검색어를 시소러스와 같은 지식구조를 이용해 동의어나 상위어, 하위어로 확장하는 기법을 제안한다.

기존에 제안된 기법들은 구조가 복잡하거나, 새로운 단어의 삽입이나 삭제와 같은 연산이 발생시 구조의 확장이 어렵다는 단점이 있었다. 그러나 제안하는 구조는 객체-관계형 DBMS를 이용하여 검색이 용이하고, 구조의 확장이 용이하다는 장점이 있다.

여기에서 전문용어라 함은 정보기술(Information Technology : IT) 분야의 용어를 의미하고 단어들은 실제 업체나 연구분야에서 많이 쓰이는 단어들을 중심으로 수집하고 관련 서적들과 전문가의 도움을 받아 수집하였다. 수집된 단어들은 다시 주제별로 나누어서 키워드별로 정리했다. 정리된 단어들은 그 계층 관계를 표현하기 위하여 트리 형태로 재구성하였다. 트리 구조에서 노드(node)는 디스크립터이고 간선(edge)는 단어들 간의 관계를 표현한다.

트리로 표현된 단어들은 데이터베이스에 저장하여 검색 시스템을 구성하였다. 트리에는 표현되지 않은 비디스크립터를 포함하여 모든 단어들을 용어 클래스(Term Class)에 저장하고 이와는 별도로 디스크립터만으로 계층 클래스(Ht Class)를 구성하여 검색의 확장이 용이하도록 하였다. 또한 급속히 성장하는 정보기술 분야에서는 새로운 용어의 생성 또한 급속히 늘어나고, 대부분이 영문 단어이므로 한글화 표현 또한 여러 가지여서 데이터베이스의 지속적인 확장이 가능하도록 설계하였다.

제안한 IT 용어 관리 시스템은 IT분야의 동적인 변화에 신속 정확하게 대응할 수 있도록 다음과 같이 설계하였다.

첫째, 기존에 많이 사용하던 관계형 데이터베이스가 아닌, 객체 지향적 구조로 설계하여 모든 단어와 계층을 하나의 객체로 인식하게 하였다. 둘째, 포인터를 두어 빠른 검색을 지원할 수 있게 하였다. 셋째, 일반 디스크 기반의 데이터베이스가 아니라 메모리 상주 DBMS를 사용하여 성능의 향상을 꾀하였다. 메모리 상주 DBMS는 다른 디스크 기반의 DBMS에 비해 디스크의 I/O를 없애 속도 면에서 빠른 성능을 보인다. 따라서 전자상거래나 본 논문에서 구현한 정보검색 시스템과 같이 많은 사용자들이 빈번하게 이용하는 시스템에서 우수한 성능을 보일 것으로 기대된다. 마지막으로, GUI를 이용한 사용자 환경을 제공하여 사용자가 손쉽게 이용할 수 있도록 하였다.

이를 위해 DBMS는 메모리 상주 객체-관계형 DBMS인 Tachyon을 이용하여 구현하였다.