키워드 유사성 검색에 관한 연구

이윤기*, 윤지현*, 정형수*, 염헌영*, 양영규**, 황순욱***
*서울대학교 컴퓨터공학부
**경원대학교 소프트웨어대학
***한국과학기술정보연구원
e-mail: {yklee, jhyoon, jhs, yeom}@dcslab.snu.ac.kr,
ykyang@kyungwon.ac.kr, hwang@kisti.re.kr

A Study on Keyword Proximity Search

Yoon Ki Lee, Ji Hyun Yoon, Hyungsoo Jung, Heon Young Yeom, Young Kyu Yang, Soon Wook Hwang
*School of Computer Science & Engineering, SNU
**School of Software, Kyungwon University
***Korea Institute of Science and Technology Information

요 약

키워드 유사성 검색은 입력받은 키워드에 관련된 의미 있는 데이터를 검색하는 것을 말한다. 데이터들은 매우 다양한 형태로 표현 될 수 있고, 각각의 형태에 대한 키워드 유사성 검색에 대한 많은 연구가 이루어졌다. 이 논문에서는 다양한 키워드 유사성 검색에 대한 연구들의 개관을 살펴보고 그것들을 비교해 볼 것이다. 이 연구들을 비교·분석하는 것은 키워드 유사성 검색을 일반화 하는데 도움이 될 것으로 기대한다.

1. 서론

사용자가 원하는 의미있는 정보를 추출하는 것은 데이터 검색에 있어서 가장 중요하다고 할 수 있다. 이러한 요구를 반영하여 그동안 이 분야의 패러다임에 있어서 상당한 진전이 있었다. 그러나 데이터들이 표현되는 방법이 매우 다양하기 때문에(예를 들어, 관계형 데이터베이스, XML 문서, 웹 사이트 등), 각각에 맞는 특성화된 방향으로 연구가 진행될 수 밖에 없었다. 다양한 분야에서 연구가 진행되었기 때문에, 각각에 맞는 특성화된 알고리즘들이 제시되었다. 그 각각의 연구 성과들을 비교하여 공통점을 분석해 보는 것도 매우 의미 있는 일이라고 할 수 있겠다. 이 논문에서는 관계형 데이터베이스, XML 문서, 웹 사이트 분야에서 이루어진 연구들을 비교해 볼 것이다.

2. 키워드 유사성 검색의 다양한 연구들

(1) 관계형 데이터베이스

관계 데이터에 대한 키워드 유사성 검색에 대한 연구에 대해 살펴보기 전에, 우선 관계 데이터에 대한 키워드 유사성 검색 방법에 대해 살펴본다.

예를 들어 Cities라는 테이블과 Conturies라는 테이블이 있다고 가정해 본다. Cities에는 여러 나라들의 수도에 대한 정보가 들어 있고, Conturies에는 각 나라들에 대한 정보가 들어 있다고 하자. 여기에서 사용자가 “Korea, Seoul” 이라는 키워드를 입력했다면 키워드 유사성 검색에 의해 “Seoul은 Korea의 수도입니다."라는 정보를 얻을 수 있는 것이다. 이렇게 단순히 2개의 테이블에 대한 정보를 표현할 수 있는 것이 아니라, 여러 개의 테이블에 대한 복잡한 정보도 표현할 수 있게 된다.

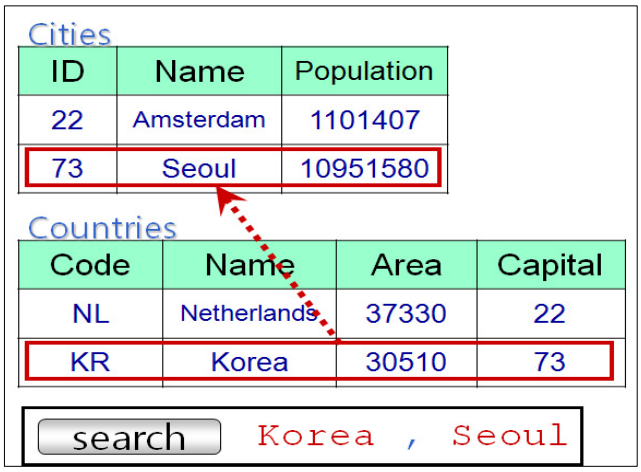

그림 1 관계형 데이터베이스에서의 키워드 유사성 검색

관계데이터에 대한 키워드 유사성 검색에 대한 연구에는 대표적인 것으로 BANKS[1], DISCOVER[2], DBXplorer[3] 등이 있다.

BANKS는 관계형 데이터베이스에서 키워드 기반 검색을 가능하게 해 주는 시스템으로서, 데이터와 스키마 검색을 모두 지원한다. BANKS는 사용자로 하여금 데이터베이스의 스키마를 모르고 복잡한 쿼리를 요구하지 않으면서 정보를 추출할 수 있게 한다. BANKS는 데이터베이

스의 튜플들은 그래프의 노드로 표현하고 각각의 간선들은 테이블간의 관계를 나타낸다. 쿼리에 대한 응답은 루트가 있는 트리로 표현되는데, 이 트리는 각각의 키워드와 맞는 튜플들을 연결하고 있다. 쿼리에 대한 응답은 안쪽연결(inlink)에 기반하여 순위를 매기게 된다.

DISCOVER는 튜플들의 결합 네트워크를 리턴하는데, 이것은 결국 모든 키워드를 포함하는 튜플들의 집합이다. DISCOVER는 2단계로 진행되는데, 첫 번째는 모든 관계 네트워크의 후보들을 생성하고, 그 다음에 그 후보들에 대한 효과적인 평가 계획을 설계한다. 이 연구에서는 최적화된 실행 계획을 설계하는 것이 NP-Complete이기 때문에 근사-최적(near-optimal) 실행 계획을 제시한다.

(2) XML 문서

XML 표현을 통해서 데이터를 계층적으로 표현할 수 있다. 먼저 XML에 대한 키워드 유사성 검색에 대해 살펴본 후에 XML에 대한 키워드 유사성 검색에 대한 연구에 대해 살펴보도록 한다.

XML 데이터가 어떤 논문들에 대한 자료를 표현하다고 하자. 이 데이터는 저자, 제목, 인용 등의 하위 자료 구조를 가질 수 있다. 이때 사용자가 “Yannakakis, Approximation" 이라는 키워드를 입력했다면 키워드 유사성 검색을 통해 ”Yannakakis가 Approximation에 관련된 논문을 작성하였다“ 라는 정보를 얻을 수 있는 것이다. 또한 ”Yannakakis 가 Approximation에 대한 논문에 의해 인용되었다“ 라는 정보도 얻을 수 있을 것이다. 관계 데이터에서 말한 것과 마찬가지로, 복잡한 상·하위 구조에서도 의미있는 정보를 얻을 수 있다.

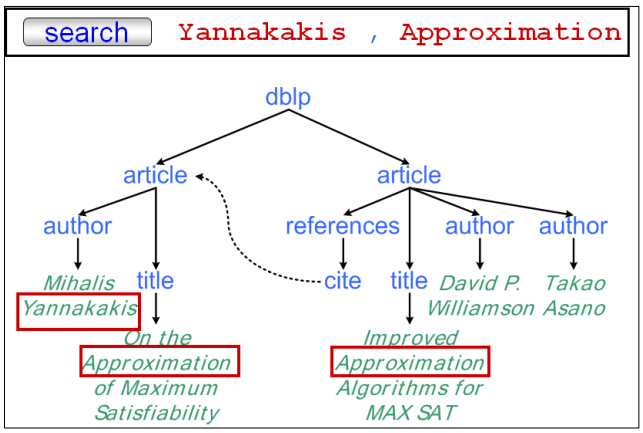

그림 2 XML 문서에서의 키워드 유사성 검색

XML 형태로 표현된 데이터에 대한 키워드 유사성 검색에 대한 연구에는 대표적인 것으로 XKeyword[4]가 있다. XKeyword는 위에서 언급한 DISCOVER에 기반한 시스템이다. XKeyword에서는 그래프를 통해서 쿼리 평가를 최적화한다. XKeyword는 두 단계로 이루어지는데, 전처리 단계에서는 키워드에 대한 인덱스의 집합이 그래프의 특정 패턴들 기술하는 인덱스와 함께 생성된다. 처리 단계에서는 키워드 쿼리 결과를 효과적으로 생성하기 위해 관계의 경로의 근사-최적(near-optimal) 집합을 이용하는 계획을 세운다. 그 결과로서 사용자의 네비게이션에 관련된 도시화된 결과를 제공한다.

(3) 웹 사이트

키워드 유사성 검색은 관계형 데이터베이스, XML 문서에서 뿐만 아니라 웹 사이트로부터 정보를 얻어내는 데도 사용된다. 사용자로부터 키워드가 주어지면, 웹 사이트에 대한 키워드 유사성 검색의 결과로서 웹 페이지들의 집합이 제공된다. 이 페이지들은 키워드와 관련된 페이지들이고 각각의 페이지들은 하이퍼링크로 연결되어 있다.

웹 사이트에 대한 키워드 유사성 검색에 대한 연구에는 대표적인 것으로 Information Unit[5]이 있다. 웹 페이지는 모든 정보가 하나의 물리적인 페이지에 들어 있는 경우와 메인페이지와 관련된 링크 페이지로 나누어진 경우가 있는데 현존하는 웹 검색 엔진들은 오직 물리적 페이지만을 결과로 제공한다. 이 단점을 극복하기 위해 "Information Unit"이라는 개념이 소개된 것인데, 여러개의 물리적 페이지들을 논리적으로 하나의 웹 문서로 생각하겠다는 것이다. 이 연구에서는 점진적인 쿼리 처리를 제안하였는데, 이것은 문서의 의미 유사성과 링크 구조를 모두 고려한 것이다.

3. 결론

데이터가 표현될 수 있는 여러 가지 형태 중에서 크게 관계형 데이터베이스, XML 문서, 웹 사이트의 3가지로 나누어 그것들에서 어떻게 키워드 유사성 검색을 할지에 대한 연구에 대해 살펴보았다. 각각의 연구가 데이터 표현에 특화되었지만, 그것들 모두 그래프로 표현하려는 노력이 있었고, 그 결과 현재 그래프를 이용한 일반화된 키워드 유사성 검색 알고리즘에 대한 연구가 활발히 진행되고 있다.

참고문헌

[1] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan. Keyword searching and browsing in databases using BANKS. In ICDE, pages 431-440, 2002.

[2] V. Hristidis and Y. Papakonstantinou. DISCOVER: Keyword search in relational databases. In VLDB, pages 670-681, 2002.

[3] S. Agrawal, S. Chaudhuri, and G. Das. DBXplorer: enabling keyword search over relational databases. In SIGMOD Conference, page 627, 2002.

[4] V. Hristidis, Y. Papakonstantinou, and A. Balmin. Keyword proximity search on XML graphs. In ICDE, pages 367-378, 2003.

[5] Wen-Syan Li, K. Selçuk Candan, Quoc Vu, and Divyakant Agrawal. Retrieving and organizing web pages by “information unit”. In WWW, pages 230-244, 2001.