WMSQL을 이용한 Web Mining System의 설계 및 구현†

o
최성경 박민호 이근호 백인구 한기준
건국대학교 컴퓨터·정보통신공학과
{skchoi, mhpark, khlee, igbaek, kjhan}@db.konkuk.ac.kr

Design and Implementation of a Web Mining System Using WMSQL

Sung-Kyong Choi, Min-Ho Park, Keun-Ho Lee, In-Gu Baek, Ki-Joon Han
Department of Computer Science & Engineering, Kon-Kuk University

요 약

World-Wide Web(WWW)이 발전하면서 웹으로부터 사용자가 원하는 정보를 효과적으로 찾기 위한 정보검색 방법론이 연구가들로부터 중요한 이슈로서 대두되었고 이에 기반하여 여러 상용 정보검색 시스템들이 등장하게 되었다. 그러나, 이러한 정보검색 시스템들은 웹에 존재하는 데이타의 비구조화와 다양성, 사용자의 다양성, 그리고 정보의 질과 양의 문제로 인하여 사용자의 의도와 요구에 맞는 정보를 구하기 어렵다. 또한, 웹 상의 많은 데이타들로부터 단순히 일반적인 정보만을 얻어 이용할 뿐 효과적인 지식의 탐사나 관리 기능을 갖고 있지 않다. 본 논문에서는 이전의 정보검색 시스템들이 갖는 문제점을 분석하고 이를 보완하고자 웹에 대한 지식 발견(Knowledge Discovery)의 새로운 시도인 웹 마이닝(Web Mining)에 대한 관련 연구를 토대로 웹 마이닝 시스템을 설계 및 구현한다. 특히, 사용자의 의도를 정확히 전달하기 위하여 기존의 SQL과 유사한 형태의 질의어인 WMSQL을 사용하여 웹 문서의 내용에 직접적인 웹 마이닝을 수행하는 Web Content Mining을 개발함으로서 웹의 비구조화된 데이타로부터 의미있고 함축적인 지식을 추출할 수 있도록 한다.

1. 서론

WWW 이 발전하면서 웹(Web)상의 데이타의 양은 기하급수적으로 늘고 있으며 그로 인해 사용자들은 원하는 정보를 찾기 힘들 뿐만 아니라 정보를 얻기 위해 지루하고 반복적인 작업을 계속적으로 수행해야 한다. 이러한 이유로 웹으로부터 사용자가 원하는 정보를 효과적으로 찾기 위한 정보검색 방법론이 연구가들로부터 중요한 이슈로서 대두되었고 여러 상용 정보검색 시스템들이 등장하게 되었다[Dan95, Yuw96].

정보검색 시스템의 첫번째 주요한 고려사항은 데이타의 다양성이다. 현재 웹에서 이용되는 데이타로는 문서, 이미지, 사운드, 라이브러리 등 많은 다양한 데이타들이 존재한다. 두번째는 사용자 집단의 다양성이다. 사용자 집단은 빠르게 성장하고 있으며 서로 다른 배경과 흥미와 사용 목적을 가지고 웹에 접근한다. 세번째로는 검색된 결과로서 전달되는 정보의 질과 양에 관한 문제이다. 거대한 양의 비구조화된 데이타내에서 검색으로 찾아지는 전체 혹은 일부의 데이타를 데이타베이스에 저장하고 관리한다는 것은 비현실적이다[Aro98, Etz96a].

로봇(Robot)이라 불리는 에이전트 기반의 정보검색 시스템에서는 데이타 분석, 변형, 일반화 기술의 발전으로 원시적인 다양한 정보를 구조화된 정보 즉, 분류된 정보로서 변형시키고 일반화시키는 것이 가능하지만, 사용자의 의도를 정확히 파악하여 요구에 맞는 정보를 얻기 어려우며 더욱이 체계적인 지식 발견이라는 문제는 해결할 수 없다. 특히, 웹이 빠른 속도로 확장됨에 따라 거대한 문서의 집합을 형성하게 되었고 지식 발견의 새로운 분야로 등장하면서 웹에 있는 데이타로부터 사용자의 요구에 맞는 의미있고 함축적인 지식을 효율적으로 추출하는 것이 점차 필요하게 되었다[Etz96b, Flo98].

웹에서의 데이타 마이닝을 웹 마이닝(Web Mining)이라 하며 데이타 마이닝에서 적용되어지는 여러 기술들이 같은 방식으로 웹에 적용된 분야이다[Lin98, Zai98, Zai99]. 본 논문에서는 기존의 관련 연구를 분석하고, 또한 이를 바탕으로 사용자의 의도와 요구를 정확히 기술하기 위하여 기존의 SQL과 유사한 형태의 질의어인 WMSQL을 사용하여 웹 문서의 내용에 직접적인 마이닝을 수행하는 웹 마이닝 시스템을 설계 및 구현한다. 본 시스템을 사용하면 사용자의 의도와 요구에 맞는 정보의 전달과 체계적인 지식발견이 가능하게 된다.

본 논문의 구성은 다음과 같다. 제 1 장의 서론에 이어 제 2 장 관련 연구에서는 웹 마이닝의 정의와 분류에 대해서 알아보고, 텍스트 마이닝과 WMSQL에 대해서 설명한다. 제 3 장에서는 WMSQL을 이용한 Web Mining System의 설계 및 구성에 대하여 설명한다. 제 4 장에서는 본 시스템의 각 모듈에 대한 구현에 대해 기술하고, 마지막으로 제 5 장에서는 결론과 앞으로의 연구방향에 대해 언급한다.

2. 관련 연구

본 장에서는 관련 연구로서 웹 마이닝에 대한 정의와 분류에 대해서 알아보고, 텍스트 마이닝과 WMSQL에 대하여 설명한다.

2.1 웹 마이닝

웹 상에서의 데이타 마이닝은 일반적으로 웹 마이닝으로서 불려지며 웹의 리소스 안에 포함되지 않은 함축적인 지식을 추출하고 밝히

† 본 연구는 한국학술진흥재단 자유공모과제 (과제번호 : 1997-001-E00322)에서 지원받았음.

는 것을 목적으로 한다. 웹 마이닝은 인공지능, 정보 검색 등과 같은 다른 분야에서 전통적으로 찾을 수 있는 많은 기술을 포함하지만, 아직 명확하게 정의되지 않았고 많은 논제들이 웹 마이닝의 분야로 적용되어져 가고 있다[Lin98, Zai99]. 웹 마이닝은 Web Content Mining, Web Structure Mining, Web Usage Mining의 3가지 방법으로 구분된다. Web Content Mining은 문서들의 내용 또는 문서에 대한 설명으로부터 지식을 추출하는 방법이고, Web Structure Mining은 웹 안의 대상물과 참조 사이의 링크들과 웹 구조로부터 지식을 추론하는 방법이다. 그리고, Web Usage Mining은 사용자와 웹 서버간의 상호작용에 대한 데이타를 기록하고 있는 Web Access Log를 분석함으로써 사용자의 행동방식과 웹의 구조를 추출하는 방법이다[Zai98].

2.2 텍스트 마이닝

텍스트 마이닝은 구조화되지 않은 텍스트 문서상에서 지식을 추출해 내는 방법이다. 이것은 데이타 마이닝의 한 부류로서 데이타 정제에서 지식 시각화까지의 지식 발견 과정의 대부분을 포함하며, 텍스트 마이닝의 기법에는 Link analysis, sequence analysis, anomaly detection, similarity detection, Hypertext analysis 가 있다. 텍스트 마이닝을 통해 생성되는 Rule 에는 Association rule, Classification rule, Clustering rule 등이 있다. 특히, 사용자로부터 질의어가 들어 왔을 경우 텍스트 마이닝을 통해 질의어와 문서의 주제가 일치하거나 또는 가장 근접한 문서를 추출해 낼 수 있다[Wit99, Zai95].

2.3 WMSQL(Web Mining SQL)

WMSQL은 웹 상에서 사용자의 의도를 표현할 수 있도록 SQL과 비슷한 문법 구조를 갖는 구조화된 언어로서 본 논문에서는 웹 마이닝 시스템에서 지식을 발견하기 위한 언어로 사용된다. WMSQL 은 WMSQL 정의문와 WMSQL 조작문을 갖는다. WMSQL 정의문은 크게 CREATE 와 DROP 문으로 구성된다. CREATE 문은 CREATE CATALOG, CREATE RULE, CREATE VIEW 로 세분화되고, DROP 문은 DROP CATALOG, DROP RULE, DROP VIEW 로 세분화된다. WMSQL 조작문은 SELECT, DELETE, INSERT, UPDATE, DESCRIBE 문으로 구성된다. WMSQL 정의문을 통해 생성된 catalog, rule, view 는 데이타베이스내에서 테이블로 존재하고, WMSQL 조작문을 통해 검색, 삭제, 삽입, 갱신된다.

3. WMSQL 을 이용한 Web Mining System 의 설계

본 장에서는 WMSQL을 사용한 웹 마이닝 시스템의 개념적 흐름과 전체 구성을 살펴본다. 그리고, 각각의 구성요소에 관한 모듈별 기능들을 언급한다.

3.1 웹 마이닝 시스템의 개념적 흐름

사용자는 본 논문에서 제시하는 웹 마이닝 시스템을 통해 웹 상의 HTML 문서의 내용으로부터 관심있는 항목에 대한 지식을 얻을 수 있는데 전체적인 시스템의 개념적 흐름은 그림 1 과 같다.

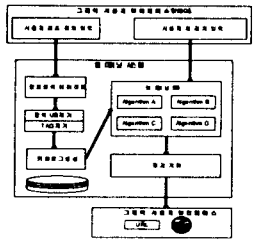

그림 1. 웹 마이닝 시스템의 개념적 흐름

사용자는 그래픽 사용자 인터페이스를 통해 원하는 정보를 포함한 문서의 집합인 catalog를 생성하기 위하여 WMSQL을 사용하여 질의를 수행한다. 입력된 질의는 정보검색 에이전트에 전달되고, 정보검색 에이전트는 웹으로부터 관련된 HTML 문서를 수집한다. 수집된 HTML 문서는 정보검색 에이전트에서 URL 파싱과 문서 파싱 과정을 거쳐 URL 추출과 불필요한 Tag 등을 제거하여 카탈로그를 생성한다. 사용자는 생성된 카탈로그를 기반으로 다시 원하는 항목에 대하여 지식을 추출하기 위한 질의를 그래픽 사용자 인터페이스를 통하여 입력한다. 그러면, 웹 마이닝 시스템의 해당 마이닝 알고리즘 모듈은 입력된 질의에 대한 지식을 생성하고 생성된 지식을 그래픽 사용자 인터페이스를 통해 최종적으로 사용자에게 보여 준다.

3.2 웹 마이닝 시스템의 구성

본 논문에서 개발할 웹 마이닝 시스템의 전체적인 구성은 그림 2 와 같다.

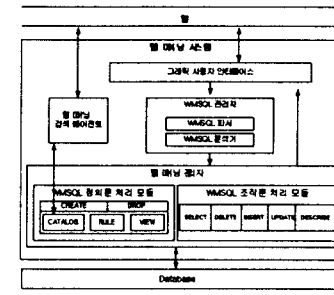

그림 2. 웹 마이닝 시스템의 구성

3.2.1 WMSQL 관리자

WMSQL 관리자는 WMSQL 파서와 WMSQL 분석기로 구성되며, 그래픽 사용자 인터페이스를 통해 입력된 WMSQL의 파싱 및 분석 작업을 수행한다. WMSQL 파서는 WMSQL을 파싱하여 parse tree를 형성함으로서 WMSQL 의 symbol table 을 생성한다. WMSQL 분석기는 WMSQL 파서로부터 파싱된 결과를 구문 분석 후 웹 마이닝 관리자의 해당 모듈을 호출한다.

3.2.2. 웹 마이닝 관리자

웹 마이닝 관리자는 WMSQL 관리자를 통해 파싱되고 분석된 질의에 따라 실제로 웹 마이닝 과정을 수행하는 모듈로서 WMSQL 정의문 처리 모듈과 WMSQL 조작문 처리 모듈로 구성된다. WMSQL 정의문 처리 모듈은 하위 모듈로 catalog, rule, view에 대한 CREATE와 DROP 모듈로 구성되어, WMSQL 정의문에 명시된 catalog, rule, view를 생성 또는 삭제한다. WMSQL 조작문 처리 모듈은 SELECT, DELETE, INSERT, UPDATE, DESCRIBE 모듈로 구성되어, WMSQL 정의문 처리 모듈에 의해 생성된 catalog, rule, view를 검색, 삽입, 삭제, 갱신한다.

4. WMSQL 을 이용한 Web Mining System 의 구현

본 시스템은 Windows NT에서 IIS5.0, JDK1.2, JDBC2.0, JSDK2.1 를 이용하여 Java Servlet 으로 구현되어 웹 상의 브라우저에서 동작하며, 크게 2 개의 WMSQL 관리자와 웹 마이닝 관리자로 나누어 진다. 본 장에서는 이러한 모듈의 구현에 대해서 설명한다.

4.1 WMSQL 관리자의 구현

본 논문에서 제시하는 웹 마이닝 시스템에서는 WMSQL 을 사용하여 웹 마이닝을 수행하기 때문에 기존의 SQL 질의처리기와는 다르게 정확하게 사용자의 요구를 처리할 수 있는 질의처리기가 필요하다. 따

라서, lex&yacc 를 사용하여 WMSQL 의 BNF 표기법에 따라 웹 마이닝 시스템을 위한 질의처리기인 WMSQL 파서와 WMSQL 분석기를 구현하였다. Lex 문법을 통해서 WMSQL의 TOKEN을 구별하고, YACC 구문 구조를 통해서 구문분석 후 질의가 문법에 적합한지를 평가하여 적합한 경우 실행코드를 실행하게 된다.

4.2 웹 마이닝 관리자의 구현

웹 마이닝 관리자는 WMSQL 관리자를 통해 파싱되고 분석된 WMSQL 질의에 따라 실제로 웹 마이닝 과정을 수행하며, WMSQL 정의문 처리 모듈과 WMSQL 조작문 처리 모듈로 구분된다.

4.2.1 WMSQL 정의문 처리 모듈

WMSQL 정의문 처리 모듈은 WMSQL 정의문을 처리하기 위한 모듈로서 CREATE 모듈과 DROP 모듈로서 구성되고 catalog, rule, view 를 생성 또는 삭제한다.

catalog 의 생성은 CREATE CATALOG 모듈을 통해 이루어진다. CREATE CATALOG 모듈은 사용자가 입력한 질의에 대한 HTML 문서를 웹상의 정보검색 시스템으로부터 가져와서 catalog 를 생성하게 되는데, 이러한 역할을 수행하는 것이 웹 마이닝 검색 에이전트이다. 로봇 에이전트가 수집한 URL과 HTML 문서는 URL Parser 와 HTML Parser 를 통해 파싱되어 WMSQL 정의문 처리 모듈의 CREATE CATALOG 모듈에 전달되고, CREATE CATALOG 모듈은 전달받은 URL과 HTML 문서를 관련된 여러 문서 정보와 함께 데이타베이스에 저장하여 catalog 를 생성한다.

rule 의 생성은 CREATE CATALOG 모듈에서 생성된 catalog를 이용하여 CREATE RULE 모듈을 통해 이루어진다. 본 논문에서 CREATE RULE로서 생성되는 지식은 ASSOCIATION RULE 과 CLASSIFICATION RULE 이다. WMSQL 관리자에서 사용자가 WMSQL로 입력한 질의어에 대한 파싱과 분석이 이루어진 후 메소드에 따라 웹 마이닝 관리자의 WMSQL 정의문 처리 모듈 내의 하위 모듈인 CREATE RULE 모듈에서 rule을 생성하게 된다.

먼저 문서 필터링을 통해 catalog 내의 문서가 rule 생성을 위해 입력된 질의어를 포함하였는지를 String Search 알고리즘을 사용하여 선별하고, 선별된 문서는 불용어 사전을 이용하여 문서로부터 불용어를 제거하는 문서 파싱 단계를 거친다. 불용어가 제거된 문서는 ASSOCIATION RULE 을 생성하는 경우에는 Term Analysis, Structure Analysis, Search Engine Ranking Analysis를 이용하여 Rule을 생성한다. Term Analysis 와 Structure Analysis 는 텍스트 마이닝의 Sequence Analysis 를 따른다. Term Analysis 는 문서에 쓰인 단어의 상대적인 빈도수에 따라 rule 생성을 위해 질의어로 입력된 단어에 대한 문서의 가중치를 부여하며, Structure Analysis 는 문서 내에서의 단어의 발생 위치에 따라 rule 생성을 위한 질의어로 입력된 단어에 대한 가중치를 부여하며, Search Engine Ranking Analysis 는 rule 생성을 위한 리소스가 되는 문서의 검색 엔진 순위정보에 대한 가중치를 부여한다.

CLASSIFICATION RULE 의 경우에는 문서 파싱 단계를 거친 후 WMSysDic에 명시된 Classification 정보에 따라 Term Analysis와 Structure Analysis를 이용하여 문서에 대한 Classification Rule을 생성한다. 이때, Term Analysis 와 Structure Analysis 는 ASSOCIATION RULE을 생성하는 경우와 같이 방식으로 적용된다.

4.2.2 WMSQL 조작문 처리 모듈

WMSQL 조작문 처리 모듈은 기존의 SQL 과 같은 방식으로 질의를 수행하는 WMSQL 조작문인 SELECT, DELETE, INSERT, UPDATE, DESCRIBE 문을 처리하기 위한 모듈로 구성되어 WMSQL 정의문 처리 모듈을 통해 생성된 catalog, rule, view에 대해서 질의, 삽입, 삭제, 갱신 등을 수행한다.

5. 결론

인터넷이 빠른 속도로 발전함에 따라서 웹상에 거대한 문서의 집합을 형성하게 되었고, 웹에 있는 데이타로부터 사용자의 의도와 요구에 맞는 의미있고 함축적인 지식을 추출하는 것이 필요하게 되었다. 그러므로, 본 논문에서는 사용자의 의도와 요구에 맞는 정보의 전달과 체계적인 지식발견이라는 문제를 해결하기 위하여 기존의 관련연구를 분석 및 보완하여 WMSQL을 이용한 웹 마이닝 시스템을 설계 및 구현하였다.

본 논문에서 설계 및 구현한 웹 마이닝 시스템은 기존의 데이타베이스 시스템에서 사용하던 SQL과 유사한 WMSQL을 사용함으로써 사용자의 의사를 최대한 반영하여 원하는 정보에 대하여 자유로운 질의가 가능하며, 특히 기존의 SQL 에서 할 수 없었던 지식의 생성이 가능하다. 그러므로, 기존의 정보검색 시스템들이 해결하지 못했던 함축적인 지식 획득과 얻어진 정보에 대한 신뢰를 가질 수 있다. 또한, 본 논문에서 제시한 웹 마이닝의 방법론을 기존의 정보검색 시스템에 적용하여 보완한다면 검색의 정확성 및 사용자의 만족도를 높일 수 있다.

향후 연구 과제로는 웹 마이닝 시스템에서 생성할 수 있는 지식의 종류의 확장과 보다 효율적인 웹 마이닝 알고리즘의 연구가 필요하다.

참고문헌

[Aro98] Arocena, G.O., and Mendelzon, A.O., " WebOQL: Restructuring Documents, Databases, and Webs," Proc. of the 14th Int. Conf. on Data Engineering, Feb. 1998, pp. 24-33.

[Dan95] Daniels, J.J., and Rissland, E.L., " A Case-based Approach to Intelligent Information Retrieval," Proc. of the 18th Int. Conf. ACM SIGIR, Jul. 1995, pp. 238-245.

[Etz96a] Etzioni, O., et al., " Efficient Information Gathering on the Internet," Proc. of the 37th Symp. on Foundations of Computer Science, Oct. 1996, pp. 234-243.

[Etz96b] Etzioni, O., " The World-Wide Web: Quagmire or Gold Mine?," Communications of the ACM, Vol. 39, No. 11, Nov. 1996, pp. 65-68.

[Flo98] Florescu, D., Levy, A., and Mendelzon, A., " Database Techniques for the World-Wide Web: A Servey," SIGMOD Record, Vol. 27, No. 3, Sep. 1998, pp. 59-74.

[Lin98] Lin, S.H., and Shih, C.S.," Extracting Classsification Knowledge of Internet Documents with Mining Term Associations: A Semantic Approach," Proc. of the 21st Int. Conf. ACM SIGIR, Aug. 1998, pp. 241-249.

[Wit99] Witten, L.H., Bary, Z., and Mahoui, M., " Text Mining: A New Frontier for Lossless Compression," Proc. of the DCC, Mar. 1999, pp. 198-207.

[Yuw96] Yuwono, B., Lam, S.L, Ying, J., and Lee, D.L., " A World Wide Web Resource Discovery System," IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No.4, Aug. 1996, pp. 548-554.

[Zai95] Zaiane, O.R., and Han, J., " Resource and Knowledge Discovery in Global Information Systems: A Preliminary Design and Experiment," Proc. of the Int. Conf. KDD, Aug. 1995, pp. 331-336.

[Zai98] Zaiane, O.R., and Han, J., " Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs," Proc. of the IEEE ADL, Apr. 1998, pp. 19-29.

[Zai99] Zaiane, O.R., *Resource and Knowledge Discovery from the Internet and Multimedia Repositories*, Ph.D. thesis, Simon Fraser University, Apr. 1999.