데이터마이닝기법을 이용한 검색엔진에 관한 연구

A Study on search engines using datamining techniques

이 수 연 동덕여자대학교 문헌정보학과
김 성 희 동덕여자대학교 문헌정보학과 교수

Lee, Su-Yeon
Dept. of Library & Information Science in Dongduk Women's University
Kim, Seong-Hee
Professor Library & Information Science in Dongduk Women's University

본 연구에서는 데이터웨어하우스의 개념에 대해 살펴본 뒤, 데이터마이닝 개념 및 구축방법을 살펴하고, 이를 이용한 검색엔진인 northernlight와 google에 대해 기존의 검색엔진과 비교해서 분석하고자 한다.

1. 서론

정보 기술의 빠른 발전은 업무의 자동화를 촉진시켜 엄청난 양의 데이타를 전자적으로 수집하고 보관하는 것을 가능하게 하고 있다. 그러나, 이러한 데이타의 무제한적인 증가는 우리가 원하는 정보를 찾아내는 일을 더욱 어렵게 만들고 있는 것이 현실이다, 왜냐하면 우리는 대용량의 데이타로부터 의미있는 지식(knowledge)을 찾아내고자 하는 것이 목적인데 반하여 실제적으로는 오히려 데이타만 계속 쌓이고 있는 상황이기 때문이다. 21세기의 가치평가를 자산위주가 아닌 조직내부에 축적되어 있는 지식에 따라 평가되는 지식중심사회로 예견하는 가운데, 지식습득, 공유 및 활용을 위한 지식관리시스템의 효과적인 구축 및 활용을 위하여 최근들어 데이터마이닝 기법이 많이 적용되고 있고 이 새로운 분야에 대한 관심은 지극히 당연한 결과라 생각된다.

본 고에서는 데이타마이닝의 개념과 그 구축방법, 이를 기반으로 구축된 인터넷 검색엔진인 northernlight과 google에 대해 기존의 검색엔진과 비교.분석하고자 한다.

2 데이타마이닝의 이론적 배경

2.1 데이타마이닝 및 데이타웨어하우스 개념

"mine" 이란 의미는 채광하다. 즉, 거대한 더미 속에서 가치 있는 무언가를 캐낸다라는 것이다. 데이타마이닝은 흔히 knowledge discovery in database(정보발견)라고도 불리우며 그외에 knowledge extraction(지식추출), information havesting(정보추수). data archeology(정보고고학),data pattern processing(자료패턴처리)등으로 불리운다. 데이타마이닝(data mining)이라는 것은 방대한 양의 데이타 속에서 쉽게 드러나지않는 유용한 정보를 찾아내는 과

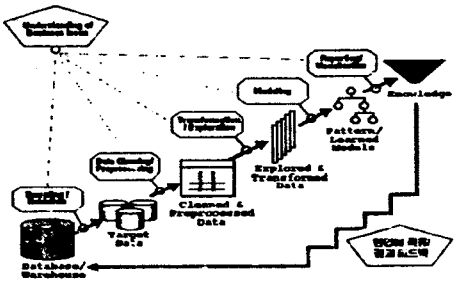
정, 데이타간의 숨겨진 관계, 혹은 겉으로 드러나지 않거나 또는 기존의 통계학적 방법을 통해 뽑아내기에는 너무나 복잡한 관계를 찾아내고, 이 관계를 바탕으로 앞날을 예측하는 기술이며, 대용량의 데이타로부터 이들 데이타 내에 존재하는 관계, 패턴, 규칙등을 탐색하고 찾아내어 모형화함으로서 유용한 지식을 추출하는 일련의 과정들이라고 정의할 수 있다. 이는 하나의 분석기법을 의미하는 것이 아니라 여러기법과 방법들의 적절한 조합으로 이루어진 일련의 과정(process)이라 말할 수 있다(SAS Institute, 1997). 데이타마이닝을 효율적으로 수행하기 위하여 시계열분석등 각종 통계기법과 데이타베이스 기술 뿐만 아니라 산업공학,신경망,인공지능,전문가시스템,퍼지논리,패턴인식,기계적학습(machine-learning),불확실성추론(reasoning with incertainty), 정보검색에 이르기까지 각종 정보기술과 기법들을 사용하게 된다. 또한 경영전략, 마케팅 기법등의 최신 경영기법들의 이용도 필요하다.

데이타웨어하우스(datawarehouse)에 관한 정의는 인먼을 비롯한 여러 학자의 연구보고서에서 살펴 볼 수 있다. 가장 대표적인 정의는 인먼(1992)의 것으로 그는 데이타웨어하우스를 '조직의 의사결정 과정을 지원하기 위한 주제중심적이고 통합적이며 시간성을 가지는 비휘발성 자료의 집합'으로 정의하고 있다. 또한 켈리(Kelly,1994)는 전사적 데이타웨어하우스를 '기업내의 의사결정 지원어플리케이션들을 위한 정보기반을 제공하는 하나의 통합된 데이타 저장공간'으로 정의하고 있다. 이것은 기업내의 상이한 많은 어플리케이션들이 동일한 정보를 공유하게 하는 웨어하우스의 측면을 강조하고 있다. 이상의 정의들에서 알 수 있듯이 데이타웨어하우스는 의사결정에 필요한 정보처리 기능을 효율적으로 지원하기 위한 통합된 데이타를 가진 양질의 데이타베이스라고 정의할 수 있다.

한편, 데이타웨어하우징이란 데이타웨어하우스(정보의 창고)에서 의사결정에 필요한 정보를 적시에 발굴해 나가는 과정이라 할 수 있다. 이러한 데이타웨어하우징에 필요한 구성요소는 크게 데이타 저장을 위한 인프라스트럭처 구축과 데이타웨어하우스에서 데이타의 정보화 및 활용이라는 두가지 범주로 분류할 수 있다.

2.2 데이타마이닝의 구축수행과정

데이타마이닝을 설명할 경우, "마이닝"이라는 것에만 초점을 두어 마치 특정기법(예를 들어, 신경망모형(neural networks))이 데이타마이닝 이라는 오해를 하는 경우가 종종 있다. 그러나 데이터마이닝은 신경망모형(neural networks)이나 의사결정나무(decision tree)와 같은 특정 기법이 아니라 개념적인 정보추출의 방법론이며 일련의과정(process)이다. 실제 데이타마이닝이 적용되는 지식프로세스를 살펴보면 다음과 같은 단계로 나누어 볼 수 있다[그림 1 참조].



<그림1 > 데이타마이닝 구축과정

출처; http://human21.new21.org/datamining/dm_4.htm

1) 표본 선정 및 추출

이 작업은 데이타를 처리하고 분석하기에 앞서 지식발견 프로세스를 통해 얻고자 하는 정보를 정의하고 이의 활용방안을 수립하며, 정보의 원천이 되는 데이타의 종류등을 파악하는 과정이다.

2) 데이타 정제 및 전처리

데이타 베이스에서 일관성이 없고 불완전한 오류가 있는 데이타를 제거하는 데이타 정제과정(data cleansing / preprocessing)을 통해 데이타의 무결성과 질을 보장해주는 데이타충실도가 높은 데이타를 정제하게된다.

3) 탐색 ,보안, 변형

(1) 탐색(exploration).

본격적으로 구축된 데이타베이스에서 데이타를 살펴보는 작업을 수행하게 된다. 데이타의 탐색과정에서

는 이미 알고 있는 사실들을 확인하여 수치화하는 작업을 시작으로하여 보유하고 있는 수많은 변수들의 관계를 살펴보는 단계이다.

(2) 보완(enrichment)

데이타의 정제가 데이타의 정확도를 높이는 단계인데 비해 데이타의 보완은 분석하고자 하는 데이타의 양과 깊이를 늘이는 단계이다.

(3) 변형(transfromation)

데이타의 변환단계에서는 데이타에 포함된 불필요한 레코드와 항목을 삭제하는 작업이 우선적으로 시행된다. 불필요한 레코드와 항목을 찾아내어 보편적인 규칙에 의해 생성된 변수를 이용하여 작업을 수행할 수 있도록 고려하는 단계가 바로 변형 및 조정 단계이다.

4) 모형화(modeling)

데이타마이닝 과정에서 가장 중요한 단계로서, 앞서 선행되었던 단계에서 선정된 주요한 변수를 사용하여 다양한 모형 - neural networks, CHAID, CART, 일반화선형모형등의 전통적인 통계적모형 등을 적용해 보는 단계이다.

5) 보고 및 가시화 (reporting / visualization)

데이타마이닝의 수행결과는 사용자들에게 보기 편하고 이해하기 쉬운 형태로 제공되어져야 한다. 마이닝 결과를 그래프나 각종 차트 형태로 보여주는 것이 가시화이다. 가시화의 장점은 사전지식이 없이 동적인 관찰이 가능하고 인식의 한계에 대한 부담을 경감시킨다는 점이다.

6) 평가(evaluation)

모형의 평가는 지식발견프로세스의 마지막 단계로서 데이타마이닝 기법을 이용하여 구축된 모형이 과연 실제로 현장업무에 적용하기 적절한가를 판단하는 작업이라 할 수 있다. 가장 많이 이용하는 모형평가의 방법으로는 실제 모형을 구축하기전에 별도로 분리하여 확보해 놓은 시험용 데이타를 이용하여 모형이 제공하는 분류규칙이나, 연관규칙등의 정확성을 측정해 보는 작업이 선행되기도 한다.

2.3 데이타마이닝 기법

데이타마이닝에는 특정 문제에 적용하는 기법이 정해져 있지는 않다. 또한 기법이 적용된다고해서 모든 문제가 해결되는 것도 아니다. 얻고자 하는 결과나 데이타의 상태 등에 따라 적용할 수 있는 기법들은 다를 수가 있다. 그러므로 기법들에 대해 어느 정도 이해가 수반되면 문제를 해결하는데 좀 더 최적의 접근으로, 보다 효과적이고 적극적인 데이타마이닝을 수행할 수 있을 것이다. 데이타마이닝의 기법에는 일반적으로 통계학에서 얘기되는 여러 분석 기법들을 포함하여 연관성측정(associations),클러스터링(clustering), 의사결정나무(decision trees), 신경망모형(neural networks)과 같은 기법들이 있다.

먼저 연관성측정(associations)은 어떤 특정 문제에 대해 항목항목별로 기존 데이타로부터 데이터간의 연관성 정도를 측정하여 연관성있는 데이터들을 그룹화하는 clustering의 일종으로서, 향후 예측경향에 대한 문제에 적용될 수 있다.

클러스터링은 어떤 목적변수(target)를 예측하기 보다는 수입, 연령과 같은 속성이 비슷한 데이터들을 묶어서 몇 개의 의미 있는 군집으로 나누는 것을 목적으로 한다.

의사결정나무(decision trees)는 분류 및 예측에 있어서 자주 쓰이는 기법으로, 통계학적인 용어를 쓰지 않고도, DM(direct mail)의 응답여부 등에 영향을 미치는 변수들과 변수들의 상호작용을 쉽게 설명할 수 있다는 것이 장점이다.

신경망 모형은 데이타마이닝에 대한 관심이 모아지면서 가장 일반적으로 언급되어지고 또한 다양한 응용 분야를 가지고 있는 기법이다. 신경망 모형은 인간이 경험으로부터 학습해 가는 두뇌의 신경망 활동을 흉내내어 자신이 가진 데이타로부터의 반복적인 학습과정을 거쳐 패턴을 찾아내고 이를 일반화함으로써 특히 향후를 예측(prediction)하고자 하는 문제에 유용하다. 이상에서 다양한 데이타마이닝 기법에 대해 간략히 설명하였는데 의사결정나무(decisionTrees)나 연관성측정 (associations)은 명쾌하고 쉽게 이해할 수 있는 결과물(일종의 rule)을 제공하는데 반해, 신경망

모형은 인간이 어떠한 현상을 인지하게 되는 것처럼 쉽게 설명되지 않는 내부적인 작업을 수행하고 이를 통해 얻어진 결과물을 제공할 뿐 어떠한 변수가 중요한지, 어떻게 상호작용이 이루어져 그러한 결과물을 주게 되는 지에 대한 설명은 하지 않는다. 따라서 설명력(comprehensibility)보다는 정확한 예측이 중요한 경우에 이용될 수 있다.

3.데이타마이닝기법을 이용한 검색 엔진

3.1 Northern Light Search

(http://www.northernlight.com)

1995년 개발된 검색엔진으로서 1400억이상의 웹페이지와 6,900여개의 원문데이타 베이스를 제공하고 있다. 다른 검색엔진과는 달리 NorthernLight는 전문사서에 의해 미리 주제분류된 folder에 각 정보를 배열하는 데이타마이닝 기술을 사용한 검색엔진이라 하겠다. 이 folder는 주제, 문서형식,소스, 언어별로 분류되며. 대략 20,000개 이상의 광범위한 계층적 관계어와 200,000-300,000에 이르는 첨가어로 수록된다. 이러한 색인은 사람에 의해 수작업으로 생성되지만, 데이타베이스는 컴퓨터에 의해 자동적으로 생성하게 된다. NorthernLight의 Folder는 각각의 검색결과를 미리 지정된 검색결과의 색인을 고려한 알고리즘에 근거하여 구성된다. 이러한 데이타마이닝 기술을 도입한 Folder의 생성과 로봇기술을 이용한 다른 일반검색엔진의 결합은 별도의 통제어를 사용하지 않고 이용자에게 정확한 검색결과를 제공한다.

웹을 대상으로 검색을 실시하는 이 검색엔진의 특징은 Special Collection이라는 6,900여개의 별도 간행물로부터 전문을 제공한다는 것이다. Special Collection에는 Business Magazine, Trade Journal, Newswire, Academic Journal 등이 포함되어 있다. 또한 33개 통신사의 최근 2주간 뉴스를 무료로 이용할 수 있다. 별도의 가입이 필요하며, 가입은 회사, 도서관 등이 포함되어 있는 기관과 개인 등 2가지로 분류된다. 요금은 'pay per document'로 $1~$4까지 지불한다. 구체적인 northernlight search의 특징은 다음과 같다.

1) 일반 로봇검색 방식과 디렉토리 방식통합: 검색어를 입력하여 검색을 하면, 검색결과물을 주제별로 분류하여 출력하기 때문에 막연히 원하는 결과가 나왔을까 찾아야 하는 다른 로봇 방식 검색엔진과는 달리 원하는 결과물을 찾기 쉽다는 것이다. 한글에 대한 지원도 완벽하며 특히 유명 DB, 예를 들면 뉴욕타임즈나 ZD 뉴스, 뉴스바이트 등의 뉴스를 전문적으로 미러링하여 전문기사 검색 등에 강하다.

2) 자연어검색기능: 단어(word)와 구(phrase)검색이 가능하다.불리안 연산자를 이용한 검색이 가능하다 (AND(+), OR, NOT(-),절단기호(*,%), 우선검색("", 괄호를 이용)이 가능하다).

3) 특이한 검색형식제공

(1) Power Search: 웹페이지와 비웹데이 터베이스를 대상으로 검색한다.

(2) Business Search: Business Week, The Economist, Fortune Magazine의 기사내용을 중심으로 제공되는 비즈니스정보원, 전문가에 의해 제공되는 시장동향분석정보, 투자 정보등을 제공한다.

(3) Investext Search: 산업동향, 상품분석, decision-marketing에 필요한 연구보고서를 제공한다.

(4) Stock Quotes Search: NASDAQ , New York Stock Exchange (NYSE.)의 실시장 주식거래정보, 기업정보, SEC Filing 등을 제공한다.

(5) Current News: APOnline, UPI, PR Newswire등 33개의 연합통신사에서 제공되는 최근 2주간의 소식을 제공한다.

(6) Geo Search ; 미국과 캐나다의 지역정보를 제공받음. 주소, 우편번호, 전화번호검색이 가능하다.

4) 마이닝 기법을 사용한 검색결과 배열

(1) Custom Search : Subject, Type, Source, Language로 분류된 Folder검색결과가 분류되어 제공되며 관련내용을 blue folder로 분류하여 제공된다. 이 형식이 타 검색엔진과는 다른 마이닝 기법이 사용된 점이라 할 수 있다.

(2) Special Collection: American Banker, ENR: Engineering News Record, The Lancet, PR

Newswire, and ABC News Transcripts등 6,900여개의 원문저널, 단행본, 잡지, 뉴스와이어, 참고정보원을 대상으로 한다.

5) 다양한 서비스;E-mail 등록을 통해 새로운관심분야 소식을 자동적으로 받아볼 수 있는 Search Alert 서비스와 B2B커머스 공간의 제공과 16개의 카테고리 서비스를 제공한다.

3.2 Google(http://www.google.com)

Google은 1999년 9월 21일 베타 단계를 끝내고 라이브 버전으로 시작한 새내기 검색 엔진이다. 미국 Stanford 대학의 두 연구원, Larry Page와 Sergey Brin이 1998년에 시작한 검색 엔진으로 Yahoo!처럼 특이한 이름을 가지고 있는 이 검색엔진은 Internet, Web, Cars 등 세세한 내용이 아닌 일반 정보를 찾을 때 가장 적합한 검색 엔진으로 평가를 받아왔다. 독특한 "PageRank" 기술을 이용하여 사이트 순위를 매기는 것으로 유명하다.

Google에서는 재미있는 개성을 많이 찾아볼 수 있는데, "I'm feeling lucky" 버튼은 다른 검색 엔진에서 찾아볼 수 없는 것으로 검색 결과를 직접 보여주지 않고 그중 첫번째 웹 페이지로 바로 이동시켜주는 기능이다. 그리고, "Cached" link를 제공하여 이미 사라진 사이트이거나 서버/네트워크 일시장애, 중단으로 접속이 잘 안되는 것을 어느 정도 막아주기도 한다. Google을 이용하면 "404 Not Found" 에러를 그만큼 적게 볼 수 있다.

이 검색엔진의 주요특징은 다음과 같다.

1) 관련성피드백(relevance feedback)검색

정확도를 높이기위한 검색기술로서 여기에서 dataminig기술이 도입되었다고 할 수 있는데, 관련성 피드백을 이용한 랭킹구조를 가지고 정확도를 높이는 검색결과를 출력해내는 것이 바로 그것이다. 구글의 정확도는 하나의 문서에서 발생하는 키워드 정보에 의존하는 것이 아니라 하나의 문서에 대한 정확도를 해당문서를 링크하고 있는 다른 문서수에 의해 결정되는 구조를 가진다. 즉 많은 사람들이 링크를 걸어두고 있는 사이트는 자신한테도 의미있는 사이트가 될 가능성이 높다는 확률적인 이론을 바탕으로 하고 있다.

2) PageRank 라는 사이트 순위 선정 방법

Google의 "PageRank" 기술은 수 많은 웹 페이지가 서로 하이퍼링크로 이어져 있다는 웹의 기본 구조에 기초를 둔다. 좋은 사이트는 링크를 많이 걸어두거나 추천 사이트로 소개해놓기 마련이다. Google은 바로 이것을 이용하여 사이트 순위를 매기는 것이다.

(3) 다국어 검색 서비스

1999년 9월 라이브 버전 이전에 베타 버전으로 미리 서비스를 시작한 Google이 얼마전 다국어 검색 서비스 베타 테스팅을 시작하였다. 현재 영어 이외에 프랑스어, 스페인어, 독일어, 포르투갈어, 이태리어, 네덜란드어, 스웨덴어, 노르웨이어, 핀란드어, 덴마크어 등 유럽어를 중심으로 다국어 검색 서비스를 지원하고 있다.

Google의 새로운 특징인 다국어 검색 서비스는 사용자가 자신이 원하는 언어의 페이지를 골라서 검색할 수 있도록 도와주면서 지역화(localization)를 함께 지원하게된다.

4. 결론

데이타마이닝(datamining)이라는 것은 대용량의 데이타로부터 이들 데이타 내에 존재하는 관계, 패턴, 규칙등을 탐색하고 찾아내어 모형화함으로서 유용한 지식을 추출하는 일련의 과정들이라고 할 수 있다.

데이타마이닝을 이용한 검색엔진의 예로는 northernlight와 google을 들 수 있는데, 이들의 특징을 기존의 검색엔진과 비교해보면 다음과 같다.

1) 데이타마이닝 기법을 이용한 검색엔진들의 공통점은 검색결과의 적합성을 "인간의 판단"에 기초하고 있다는 점이다.

먼저 구글의 접근은 인터넷의 링크 구조에서 그 해답을 찾고 있다. 특정 키워드를 포함하고 있는 웹페이지 A와 웹페이지 B가 있다고 하자. 그런데 웹페이지 A를 많은 사이트들이 링크하고 있고, 웹페이지 B는 거의 링크되어 있지 않다고 한다. 이러한 구조에서 구

글은 특정 키워드에 대해 웹페이지 A가 보다 적합하다고 판단한다는 것이다.

이렇게 하여 검색순위를 정하는 구글의 방식에서는 링크가 많은 웹페이지가 일반적으로 사람들이 우수하다고 평가한다는 가정에 기초한 것이다. 즉, 우수하니까 많은 다른 웹페이지들이 링크를 하고 있을 것이라는 생각이다. 노던라이트의 경우에는 전문사서에 의해 미리 주제 분류된 폴더에 각 정보를 배열하여 검색결과를 보여주는 것이다. 또한 해당주제와 관련된 분야를 정리하여 제공함으로써 기존의 검색엔진에서는 볼 수 없었던 사용자들의 정보나 문서에서 추론하여 결합한 새로운 정보를 생성하는 이른바 지식검색이 가능해졌다는 것이다.

2) 기존의 검색엔진의 단점을 보완할 수 있는 서비스라는 것이다.

최근의 검색엔진들은 대부분 재현율은 좋지만 정확율이 떨어지는 것이 문제가 되고 있다. 검색엔진수가 현재 143개로 증가하고 있는 현재에는 재현율보다는 정확율이 무엇보다 검색엔진에 있어 중요하다고 하겠다. 구글의 관련성 피드백(Relevance Feedback)은 이러한 기존 검색엔진의 단점을 보완해주는 서비스라고 하겠다.

3) 새로운 영역에 대한 검색서비스이다.

최근의 검색엔진은 Yahoo와 Lycos등이 검색서비스로 출발하여 새로운 전자상거래 수행이 가능한 비즈니스모델을 채택하여 서비스하고 있다. 정보검색엔진이라기 보다는 컨텐츠, 커뮤니티, 커머스를 제공해주는 포탈사이트로 지향하고 있는 것이다. 그러나, 구글과 노던라이트의 경우는 이상의 검색엔진들과는 약간 차별화된 접근을 시도하면서 포탈이라는 개념보다는 검색서비스에만 집중하고 있다. 구글의 경우에는 그 흔한 기업광고나 주식시세정보는 볼 수 없고, 단지, 검색서비스를 제공하는 검색창만이 깔끔한 구성으로 제공되고 있다.

<참고문헌>

데이타마이닝. 장남식, 홍성완, 장재호지음, 대청,1999

지식경영과 정보인프라, 정보전문가의 관계. 노정란, 한국비블리어 제9집

지식관리시스템의 단계별 분석 및 구축방안에 관한 연구. 김성희. 정보관리학회지 제16권, 제2호, 1999.

"**Data Mining 방법론과 SAS Enterprise Miner**", 강현철, 박태원, 임난희 한국분류학회 발표논문집,1998.

데이타 웨어하우징과 OLAP , 조재희, 박성진 대청출판사. 1996.

Building the Data WareHouse(2nd Ed.), John Wiley & Sons, Inc., 1996 Inmon,W.H.

Data Mining *Inform*; Silver Spring; Nov/DEC 1999: Gordon Linoff;

Mining online text *Association for Computing Machinery*. Communication of the ACM; New York;Nov 1999; Kevin Knight.

Meets the Web : Data mining Peggy Zorn et al. sep./oct. 1999 Online

http://human21.new21.org/datamining/dm_4.htm

http://human21.new21.org/dataMining/dm_2.htm

http://a-pex.co.kr/choice.htm

http://www.northernlight.com

http://www.google.com