차세대 정보 마이닝 동향

이 종 현* 임 혜 영** 황 준***

◈ 목 차 ◈

1. 서 론

인터넷 일반화에 따른 인터넷 비즈니스에 대한 관심의 대두와 기업 업무현장에서 정보 기술의 발달에 따라, 데이터베이스를 통한 비즈니스 기회 창출을 위하여 웹 마이닝, 데이터 마이닝, CRM, eCRM, PRM, VRM 등 다양한 종류의 마이닝 응용 기술들이 활용되어 왔다. 또한 최근 들어 데이터베이스에 담겨진 데이터 양의 3배 내지 5배를 초과하는 무수한 문서에 담겨 있는 정보(워드 파일, 파워포인트 파일)와 모든 웹 문서의 복잡한 데이터를 분석하고 이해하는 기능을 제공하는 정보 마이닝에 대한 다양한 기술들이 소개되고 있다. 정보 마이닝 기술은 최근 급속도로 해결해야 할 비즈니스의 화두가 되고 있다. 텍스트 정보를 분석하는데 초점을 둔 정보기술 분야는 텍스트 마이닝(수치적 데이터를 의미하는 데이터 마이닝 과는 반대)으로 알려져 있다. 데이터 양의 급속한 증가는 마이닝 알고리즘에 의해 처리되는 순수한 비트 및 바이트의 숫자뿐 만 아니라 중요한 변수들과도 관련이 있다. 변수의 개수가 증가함에 따라 동시에 그것들을 분석하는 우리의 능력도 근본적으로 개선되어야만 한다. 따라서 다루기 힘든 다차원 다변수 공간을 이해하는데 도움을 주기 위해 인간 두뇌의 파워를 이용하는 새로운 접근 방법에 대한 논의가 필요하다. 본 논문에서는 웹 마이닝, CRM, PRM, VRM 등의 마이닝 응용 기술에 대해 간략히 언급하며, 더불어 정보 마이닝의 출현 배경과 기술, 제품 등에 대해 다룬다. 정보 마이닝에서 간단히 다루는 접근 방법들 중에는 의미론적 연산(semantic computing)으로 알려져 왔던 것이 포함된다.

2. 마이닝 응용 기술

일반적으로 마이닝이란 데이터 마이닝을 의미한다. 데이터 마이닝이란 대량의 데이터로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정을 말한다. 여기서 정보는 묵시적이고 잘 알려져 있지 않지만 잠재적으로 활용가치가 있는 정보를 뜻한다. 즉, 데이터 마이닝이란 기업이 보유하고 있는 일일 거래자료, 고객자료, 상품자료, 마케팅 활동의 피드백 자료와 기타 외부자료를 포함하여 사용 가능한 데이터를 기반으로 숨겨진 지식, 기대하지 못했던 패턴, 새로운 법칙과 관계를 발견하고 이를 실제 경영의 의사결정 등을 위한 정보로 활용하고자 하는 것이다. 데이터 마이닝 기법은 다음과 같다.

- 통계 : 엄밀한 의미로 통계 혹은 통계기법은 데이터 마이닝이라 할 수 없다. 하지만 여전히 데이터로부터 패턴을 발견해 예측모델을 만드는데 이용되고

* 중앙대학교 대학원 컴퓨터공학과 석사과정
** 서울여자대학교 대학원 컴퓨터학과 석사과정
*** 서울여자대학교 정보통신공학부 교수

있다. 통계에서의 예측은 회귀분석과 동의어로 사용되기도 한다.

- 군집분석 : 유사한 레코드들을 함께 그룹핑하는데 사용되는 방법이다.
- 최근접 이웃 : 군집분석과 함께 가장 오래된 데이터 마이닝 예측 기법에 속한다. 임의의 레코드에 있는 예측값에 대하여 과거 데이터베이스에서 이와 유사한 예측 변수 값들을 찾아 분류되지 않은 레코드에 가장 가까운 레코드 값을 예측값으로 사용한다.
- 의사결정나무 : 나무의 구조에 기반한 예측모델로써 나무의 가지는 데이터를 분류하기 위한 질문이며, 잎은 분류 결과에 따라 분리된 데이터 세트라고 할 수 있다. 대표적인 알고리즘으로 ID3와 C4.5, CART, CHAID 알고리즘이 사용된다.
- 신경망 : 신경망은 사용하기가 어렵고 모델을 이해하기도 어렵다는 단점을 가지고 있음에도 불구하고 가장 높은 예측력을 제공한다는 장점 때문에 다양한 문제의 영역에서 사용되고 있다.
- 연관규칙 : 연관규칙 유도기법이란 데이터로부터 특정한 연관규칙을 찾아내는 것은 데이터 마이닝의 주요 작업으로 자율학습을 기반으로 하는 지식발견 시스템의 핵심이라 할 수 있다. 가장 '마이닝'이라는 용어와 유사한 개념을 지닌 데이터마이닝 작업이다. 연관규칙을 발견하는 작업은 데이터에 내재하는 모든 유형들을 체계적인 방법으로 추출한 후, 유형별로 정확도와 신뢰도 등을 추가하는 과정을 거친다. 일반적으로 이렇게 발견된 규칙들은 매우 단순하다. 연관규칙 유도기법은 규칙의 정확도와 발생빈도에 따라 규칙들을 정렬한 후 최종 사용자에게 제시한다. 이 기법은 데이터베이스 내에 존재하는 모든 유형들을 찾아준다는 강점을 지닌 반면, 사용자가 손쉽게 이해하고 현실에 적용하기에는 너무나 많은 양의 규칙들을 제공한다는 약점을 동시에 가지고 있다. 따라서 기법이 제시하는 모든 유형들 중에서도 가치 있는 규칙들만을 걸러주는 여과장치가 반드시 필요하다. 이렇게 함으로써 향후 규칙을 활용할 경우 서로 상반되는 결과를 제시하는 규칙들이 난무하는 상황을 미연에 방지할 수 있다. 요즘 상용화된 제품들 대부분은 자동적으로 규칙들을 여과해주는 기능을 갖추고 있으며, 이와 관련한 연구 또한 활발히 진행되고 있다.

CRM(Customer Relationship Management)은 데이터 마이닝을 통해 얻은 결과를 바탕으로 경쟁사와 차별화되는 고객에 대한 광범위하고 심층적인 이해를 바탕으로, 고객의 개별적 요구를 충족할 수 있는 차별적 제품과 서비스를 제공함으로써, 신규고객을 확보하고 기존고객과의 관계를 지속적으로 강화해 나가는 통합적이고 전사적인 마케팅 시스템이다.

웹 마이닝은 웹의 특성상 지속적이고도 비교적 빈번히 데이터가 발생하는 인터넷의 다양한 종류의 고객 로그 정보를 수집하여 실시간으로 고객의 웹 사이트 방문형태 정보와 구매 정보들이 구축되고, 이와 함께 고객등록정보, 구매정보 등의 데이터베이스와 연결하여 능동적으로 고객대응을 전개하는 것을 말한다. 로그 수집 방법에는 TAG 방식, TCP/IP Packet Sniffing 방식, Server-Add In 방식, 웹 서버 로그 수집 방식으로 구분가능하며, 웹 서버 로그는 웹 서버의 종류에 따라 포맷이 약간씩 다르다. 아파치 웹 서버, IIS 웹 서버, Sun One 웹 서버의 로그 포맷은 다음과 같다.

- 아파치 웹 서버 로그 포맷 : Host, RFC931, AuthUser, Time, Request Method, Request Page, Protocol, Status, Volume
- IIS 웹 서버 로그 포맷 : Host, AuthUser, Time, Service, ServerName, ServerIP, DownLoadTime, Receipt, Volume, Status, WindowsNTStatus, Request, Filename, Protocol
- Sun One 웹 서버 로그 포맷 : Host, RFC931, UserName, Date/Time of Request, Request, Protocol, Status Code, Bytes transferred

eCRM(Electronic CRM)은 CRM이 인터넷 기반으로 발전된 형태로, 기존 오프라인 마케팅 채널의 한계를 극복하여 온/오프라인 데이터의 통합을 통한 고객통합관리와 인터넷 비즈니스 시장을 확대하기 위한 다양한 이벤트나 마케팅 캠페인을 전개하여 확보한 고객들을 수익성 있는 고객으로 전환하기 위한 마케팅 활동이다. eCRM은 모든 고객을 똑같이 보지 않고 인터넷을 통해 수집된 개별고객의 회원정보, 구매이력, 캠페인 반응 등의 다양한 데이터를 축적해, 차별화된 고

객서비스 대응을 위한 개인화 서비스 전개, 영업활동 정보, 마케팅캠페인 활동을 수행할 수 있도록 데이터를 분석하며 장기적인 수익고객관리를 통한 충성고객 확보와 수익창출을 꾀하는데 목적이 있다. 또한 기업의 마케팅 비용절감과 고객 가치창출을 위하여 기존에 분산되었던 고객 네트워크, 공급자 네트워크, 파트너 네트워크 등을 효과적으로 관리하여 기업의 마케팅 중심을 기존의 상품중심에서 고객중심으로 전환할 수 있도록 하고 있다.

VRM(Visitor Relation Management)은 인터넷 비즈니스가 성장하면서 오프라인에서 온라인으로 고객접점 채널을 확장하고 있는 기업들에게 기존 오프라인에서의 고객데이터베이스를 활용한 마케팅에서 벗어나 온라인에서 실시간으로 발생하고 있는 방문자들의 다양한 행태정보를 활용한 마케팅 전개는 다양하게 변화하는 고객들의 요구에 빠르게 대처할 수 있을 뿐만 아니라 광고나 마케팅 전개에 따른 고객반응을 실시간 분석해 장기적인 고객관계 관리를 위한 전략데이터로 활용할 수 있다. 방문자 관계관리는 종전 고객이나 회원에 관한 정보를 분석관리 하는 CRM보다 한 단계 발전한 비즈니스 모델로 인터넷 사이트에 들어오는 모든 방문자들의 클릭스트림(Click Streaming) 데이터를 분석하고 고객행동분석 등을 통해 수익성 높은 고객 층을 파악, 이들을 대상으로 집중적으로 개인화된 마케팅을 전개할 수 있다. VRM은 웹 방문자의 이동경로를 추적함으로써 휴면 고객과 이탈고객이 누구이고 어느 사이트가 인기가 높은지, 어떤 광고의 클릭률이 높은지 실시간으로 파악할 수 있을 뿐만 아니라 다차원 분석 및 세부 분석능력을 통해 방문자와의 관계를 강력하게 맺어준다.

PRM(Partner Relationship Management)은 인터넷 비즈니스의 등장은 기존 오프라인 마케팅 채널에서 비용적인 측면이나 효율성 측면에서 어려웠던 온/오프라인의 데이터를 통합하여 고객 뿐만 아니라 분산되었던 공급업체와 유통업체의 파트너 간의 네트워크를 체계적으로 관리할 수 있게 하였다. 그 동안 기업은 협력업체관리에 무관심하였으며 공급업체와 유통업체는 개별적으로 영업, 마케팅서비스, 고객관련 업무를 따로 관리하여 생산관리, 재고관리, 판매관리에서 다양한 문제점들이 발생했다. PRM은 인터넷을 통해 대리점을 비롯한 중간유통업체, 부품공급업체에 이르기까지 모든 파트너의 구매 및 판매이력을 관리하고 이를 기반으로 상세한 전략을 제시하여 협력업체와의 신뢰성 있고 지속적인 거래를 발생하기 위한 관계를 구축한다. CRM이 1차원, eCRM이 2차원이었다면 PRM은 여기서 한 차원 더 나아가 웹, 이메일, 콜 센터 등을 통해 파트너의 고객은 물론 기업 내외부의 조직까지 다차원으로 관리한다는 게 특징이다. 이러한 PRM은 생필품 제조업체처럼 유통업체, 대리점, 전문점 등 다양한 파트너를 갖고 있는 기업들은 물론이고, 보험사나 카드사처럼 그 동안 고객을 기업, 파트너(대리점, 설계사 등)로 각기 따로 관리함으로써 중간 연계과정에 누수현상을 보일 수 밖에 없던 기업에게 매력적인 대상으로 등장했다. PRM을 추진하기 전에 먼저 자사의 비즈니스 형태를 다양한 관점에서 분석하여 추진목적 및 타깃 고객군은 협력업체와 공유되어야 협력업체가 수요사의 고객에 알맞게 프로세스를 정비할 수 있다. 또한 PRM은 다양한 외부 네트워크와 공유하기 때문에 정보공개의 범위와 종류, 중요한 기밀정보를 어떤 방식으로 공유하고 보호할 것인지에 관한 명확한 정책이 수립되어야 한다.

3. 정보 마이닝 출현 배경

모든 비즈니스를 통한 하루하루의 작업들 속에서 많은 양의 데이터가 수집된다(주문 목록, 지불 가능한 계좌, 판매, 고객에 대한 데이터 등). 또 비즈니스를 통해 종종 외부로부터 인구통계학 그리고 우편물 수취인 명부와 같은 데이터도 얻는다. 더 나은 비즈니스 결정을 위해 이러한 데이터를 견고하게 하고 분석하는 능력은 종종 경쟁적인 우위로 이어질 수 있고, 그런 이점을 밝히는 학습이 전략적 비즈니스 인텔리전스의 초점이다.

최근의 매우 경쟁적이고 유동적인 비즈니스 환경에서는 가장 경쟁력 있는 기업만이 지속적인 성공을 이룰 수 있다. 이러한 비즈니스 기회를 이용하기 위해 조직들은 그들의 마켓 장소, 고객, 작업에 대한 데이터를 활용하는 능력면에서 차별화를 꾀할 것이다.

장기간 지속될 수 있는 성공을 위한 전력의 중심부는 활동적인 데이터 저장소, 다시 말해 진보된 데이터 웨어하우스라 할 수 있는데, 이 안에서는 다양한 어플

리케이션이나 비즈니스의 다른 부분으로부터 얻은 데이터들이 합쳐지고 해석된다.

복잡한 데이터로부터 지식을 발견하는 최단 경로는 정보 마이닝이다. 비즈니스 인텔리전스를 위해 요구되는 정보가 취할 수 있는 풍부하고 다양한 형태를 반영하기 위해 단지 데이터 마이닝이라 부르지 않고 정보 마이닝이라 부르는 것에 주목해야 한다. 정보 마이닝은 다음과 같은 일을 수행하는 강력하고 복잡한 도구를 이용하는 것을 의미한다.

- 연관 및 패턴 그리고 경향을 밝힘
- 편차를 검출함
- 정보를 그룹화하고 분류함
- 예측 모델을 개발함

정보 마이닝은 금융이나 건강, 보험, 소매 그리고 통신등과 같은 사업에서 많은 조직에서 경쟁력면에서 상당한 이점을 가져 왔다. 새로운 정보 마이닝 기법은 다른 접근 방법들로는 얻기 힘든 새로운 지식을 밝혀낸다. 그리고 그 새로운 지식은 새로운 경쟁우위를 가진다.

기술적인 면에서 보면 성공적인 정보 마이닝을 위한 진정한 키는 데이터를 비교하고 상호 연관시키는 복잡한 수학적 프로세스에 해당하는 알고리즘에 있다. 정보 마이닝이 기반이 되는 알고리즘에 의해 누가 이 비즈니스의 최상의 고객인가 또는 그들의 무엇을 살 것인가를 결정할 수 있다. 이 알고리즘은 또한 하루 중 몇 시에 그리고 어떤 조합으로 조직이 대상 고객 목록 및 제시가격을 책정하고, 이를 시행하며, 더 구매하도록 하기 위한 판매기술의 최적화를 지원한다.

현실에서 사용되는 많은 양의 정보가 숫자가 아닌 형태(문서, 이미지 그리고 비디오 파일)로 저장되어 있다. 때문에 무수한 문서, 메모, 편지, 계약서, 특허, 연설의 사본, 전자우편 메시지 그리고 비슷한 자원의 형태일 수 있는 텍스트 정보에 대한 마이닝 프로세스를 다룬다.

4. 텍스트 마이닝 및 지식관리

텍스트 마이닝은 정보 마이닝 기술의 한 부분이다. 그리고 정보 마이닝 기술은 지식경영(KM)의 한 부분이다. 이 경우 지식은 집합적인 전문성, 경험, 노하우 그리고 조직의 지혜를 말한다. 지식은 단순한 데이터 및 정보 그 이상이다. 그것은 문맥, 의사결정 프로세스에서 도움을 주는 사실을 포함한다. 비즈니스 세계에서 지식은 전통적인 데이터베이스에서 발견되는 구조화된 데이터 뿐 만 아니라 워드 문서, 메모 그리고 편지, 전자우편, 뉴스, 웹 페이지 등과 같은 다양한 구조화되지 않은 형태로 나타날 수 있다.

Information Week는 지식관리를 "그것이 어디에(데이터베이스, 종이 혹은 사람들의 머리) 있든지 회사의 집합적인 전문성을 찾아내어 가장 큰 이익을 생산할 수 있는 곳에 분해하는 프로세스"라고 설명한다. 기존의 지식관리 도구는 기존 텍스트 객체를 이용하여 작업하며, 개인의 머리에 쓰여지지 않은 텍스트를 공유하기 위한 협력적인 작업을 고무시킬 수 있다.

지식관리 시장에서의 도구와 제품은 탐색 엔진, 문서관리 시스템, 그룹웨어 제품 등을 포함한다.

지식관리를 가능하게 하는 주요 기술로서의 텍스트 마이닝은 정보들 사이의 관계를 밝힌다는 점에서 데이터 마이닝과 유사하다. 그러나 다음과 같은 점에서 차이점을 갖는다.

실제 데이터 마이닝은 통계적 어플리케이션이고, 이전에 확인되지 않았던 연결성이나 상호 연관성을 밝히기 위한 기계적 학습 알고리즘이다. 현재까지 데이터 마이닝은 고객의 행동을 해석하고, 예측 모델을 세우고자 하는 조직에게 귀중한 직관을 제공해 왔다. 그러나 데이터 마이닝은 대부분 구조화된 수치적 데이터를 가지고 작업을 하는데, 이 데이터는 데이터웨어하우스 또는 데이터마트의 중심부에 저장되어 있다.

데이터 마이닝과는 달리 텍스트 마이닝은 텍스트 문서와 같이 구조화되지 않은 데이터를 대상으로 작업한다. 특히 온라인 텍스트 마이닝은 인터넷 상의 구조화되지 않은 데이터를 탐색하고, 그것으로부터 어떤 의미를 유도하는 프로세스를 의미한다. 텍스트 마이닝은 데이터 파일에 통계적인 모델을 적용하는 것 이상의 개념이다. 사실 텍스트 마이닝은 텍스트의 집합에서의 관계를 밝히고 이러한 관계를 찾아내는 지식작업자의 창조성을 이용하여 새로운 지식을 발견하는 것이다. 많은 텍스트 마이닝 알고리즘은 지식작업자의 머리 내에 존재하는 아이디어와 논리를 보완함으로써 새로운 지식발견을 지원한다.

텍스트 마이닝은 조직이나 그 조직의 외부에 저장되어 있는 텍스트 문서에 존재하는 거대한 양의 지식 때문에 더욱 관심의 대상이 되고 있다. 웹과 온라인 출판의 등장은 텍스트 형태로 저장된 정보의 양을 혁신적으로 증대시켰다. 지속적인 연구와 내부 및 외부 소스로부터 얻을 수 있는 정보에 의존하는 조직은 이러한 방대한 양의 텍스트로 작업하는데 있어 상당한 어려움을 갖는다. 조직에게 가용한 텍스트 데이터의 양은 너무 방대하여 쉽게 읽거나 분석할 수 없다. 더욱이 그것은 끊임없이 변하기 때문에(특히 웹 기반의 정보) 지속적으로 검토하거나 분석하는 것이 필요하다. 텍스트 마이닝은 이러한 종류의 유동적인 데이터를 이해하고 분석하기 위해 설계된 도구 및 기법을 정의한다.

5. 텍스트 마이닝의 혜택

텍스트 마이닝 솔루션을 이용함으로써 얻을 수 있는 혜택은 다음과 같다.

- 조직 정보의 증가된 가치 : 텍스트 마이닝을 전개함으로써 조직은 기존의 정보시스템에 회사 투자 가치를 증가시킬 수 있다. 텍스트 마이닝은 회사로 하여금 구조화되지 않은 텍스트의 커다란 집합 내에 숨겨진 회사의 지식을 효율적으로 모으고 사용할 수 있게 한다. 따라서 과거 데이터의 저장소에 비용을 지불하고 재통합한다거나 대체할 필요가 없어진다.
- 다른 텍스트 프로세싱 기법에 비해 적은 통합 비용 : 많은 기존의 텍스트 프로세싱 기법(문서관리 등)들은 엄청난 양의 시스템 통합과 전문적인 자문을 필요로 한다. 많은 전통적인 탐색 제품들은 토픽 트리 또는 수작업에 의한 색인을 만드는데 상당한 비용이 지불된다. 진보된 텍스트 마이닝 제품들은 매우 적은 통합을 요구하며, 기존의 다양한 솔루션과 쉽게 통합된다. 또한 텍스트를 자동으로 처리하기 때문에 비용이 드는 장치나 구성 작업이 필요 없다.
- 지식 작업자들의 증가된 생산성 : 텍스트 마이닝 솔루션은 지식 작업자들이 큰 회사의 데이터베이스 내에서 정보를 찾는 것을 보다 쉽게 해준다. 이 솔루션들은 사용자들이 그들에게 중요한 주요 의미론적 개념을 따름으로 해서 그들의 필요한 정보를 탐색하게 해준다.
- 향상된 경쟁성 : 데이터 마이닝의 주장과 유사하게, 텍스트 마이닝은 빠르고 더 나은 의사결정을 쉽게 할 수 있도록 해준다.

6. 텍스트 마이닝 기술

일반적으로 온라인 텍스트 마이닝을 가능하게 하는 두 가지 주요한 기술이 있다. 하나는 인터넷 탐색 기능이고 다른 하나는 텍스트 분석 방법론인데 두 가지를 겸비한 제품은 거의 없다.

6.1 인터넷 탐색

인터넷 탐색이 시작된 것은 불과 몇 년 밖에 되지 않았다. 과거 몇 년 사이에 World Wide Web 사이트의 폭발로 사용자가 컨텐츠를 찾는데 도움을 주도록 설계된 많은 검색엔진들이 속속 등장했다. 야후, 알타비스타 등은 최초 검색엔진들이었다. 검색엔진은 특정 웹 사이트의 컨텐츠를 검색하고 사용자가 그 인덱스를 탐색함으로써 작동한다. 이러한 도구는 비록 유용하지만 종종 컨텐츠를 부정확하게 색인하기도 한다. 인터넷 탐색에 적용되는 텍스트 마이닝의 진보는 인터넷 탐색 도구의 새로운 세대를 대표하는 온라인 텍스트 마이닝이라는 결과를 낳았다. 이러한 제품으로 사용자들은 더 적은 양의 링크와 페이지 그리고 색인을 프로세싱함으로써 더 관련이 많은 정보를 얻을 수 있게 된다.

6.2 텍스트 분석

텍스트 분석의 역사는 인터넷 탐색보다도 더 오래되었다. 실제 과학자들은 수십 년 동안 컴퓨터가 자연언어를 이해하도록 하기 위해 노력해왔다. 텍스트 분석은 그러한 노력들의 집합인 것이다. 많은 연구가들이 자연언어 문서들로부터 의미를 유도하고 이들이 나타내는 문제를 탐구해 왔지만, 텍스트 마이닝에 대한 많은 기준들과 기술적 접근 방법들이 아직 존재하지 않기 때문에 논쟁의 여지가 없는 실정이다. 텍스트 정보의 자동 분석은 몇 가지 다른 일반적인 목적을 위해 이용될 수 있다.

- 대규모 문서 집합의 컨텐츠에 대한 서론을 제공하기 위해 : 예를 들어 고객 피드백 모음에서 문서의 중요한 군집을 찾는다면 제품 또는 서비스의 어느 부분의 개선이 필요한가를 가려낼 수 있다.
- 객체 그룹들 사이의 숨겨진 구조를 찾기 위해 : 이것은 관련 문서들이 모두 하이퍼링크로 연결될 수 있도록 인트라넷 사이트를 조직하는 것을 도울 수 있다.
- 유사하거나 연관된 정보를 찾는 탐색 프로세스의 효율성 및 효과를 증대시키기 위해 : 예를 들어, 뉴스 서비스로부터 기사를 찾는 것과 지금까지 다른 기사에서는 언급되지 않았던 새로운 경향이나 기술에 대한 힌트를 포함하고 있는 모든 고유한 문서를 발견하는 것이다.
- 저장된 파일 중 중복된 것을 검색하기 위해 요약하면 텍스트 마이닝은 분석되어야 할 많은 양의 텍스트가 있는 곳이라면 어디서나 사용될 수 있다. 자동 프로세싱이 인간의 읽기 분석의 깊이에는 못 미치더라도 요점을 추출하거나 문서를 범주화하고, 요약문을 생성하기 위해 사용될 수 있다.

이제 자동화된 텍스트 분석을 이용하는 텍스트 마이닝이 어디에 적용될 수 있는지 실제 생활의 예를 몇 가지 보겠다.

- 전자우편 관리 : 텍스트 분석이 많이 사용되는 곳이 바로 메시지 라우팅(routing)인데, 컴퓨터는 누가 그것을 취급해야 하는지 결정하기 위해 메시지를 읽는다. 텍스트 마이닝의 다른 응용은 메시지의 성질을 통계적으로 분석하는 것이다.
- 문서 관리 : 텍스트 마이닝은 데이터베이스에 있는 수천 만 개의 문서를 취급하는 것을 돕는다. 이는 문서를 데이터베이스에 넣을 때 문서의 의미와 관련한 다른 문서들을 찾아내는 것으로 언제라도 관련 문서의 위치를 알 수 있는 상세한 인덱스를 만들 수 있다.
- 자동화된 도움 데스크 : 어떤 회사들은 고객의 질의에 대답하기 위해 텍스트 마이닝을 사용한다. 고객의 편지들과 전자 우편들은 텍스트 마이닝 어플리케이션에 의해 처리된다. 어플리케이션은 고객이 무엇을 원하는지 알아낼 수 있다면, 그들에게 적당한 정보를 자동적으로 보낸다.
- 시장 연구 : 시장 연구가들은 World Wide Web 상에 어떤 특정 단어, 구, 개념 또는 주제가 나타나는지에 대한 통계를 얻기 위해 텍스트 마이닝을 사용할 수 있다. 이 정보는 시장 인구 통계 그리고 수요 곡선 등을 추정하는데 유용하게 쓰인다.
- 비즈니스 인텔리전스 수집 : 이것은 텍스트 마이닝의 가장 진보된 이용 형태이다. 현재 많은 회사들은 온라인 텍스트 마이닝의 가장 진보된 형태를 대표하는 자동화된 인텔리전스 웹을 이용하여 그들의 시장, 경쟁자들 그리고 비즈니스 환경에 대한 정보를 모으고 있다. Informations와 같은 회사들은 미리 지정된 주제와 관련된 뉴스를 위해 인터넷을 감시하고 요약된 리포트를 제공하는 제품을 제공하고 있다.

6.3 의미론적 네트워크와 다른 기술들

텍스트 정보 분석, 정리 그리고 탐색을 위한 진보된 시스템을 만드는데 가장 중요한 요소는 조사된 텍스트에 대한 의미론적 네트워크(semantic)이다. 의미론적 네트워크는 분석된 텍스트로부터 유도된 가장 중요한 개념(단어 그리고 단어의 조합)으로 텍스트 상에서의 개념들 사이의 의미론적 관계를 말한다. 의미론적 네트워크는 분석된 텍스트를 간략하고 정확하게 요약하여 제공한다. 인공신경 회로망과 유사하게 의미론적 네트워크의 각 요소는 그것의 비중과 네트워크의 다른 요소(문맥 노드)와의 관계 집합으로 특징 지워지며, 네트워크의 각 요소 사이의 관계에도 비중이 할당된다.

텍스트 분석 작업의 전체적인 스펙트럼은 조사된 텍스트에 대한 의미론적 네트워크가 정확한 집합을 만든 후에 실행된다. 사실 적절히 만들어진(발견된) 의미론적 네트워크는 많은 작업의 가동적인 실행을 지원할 수 있는 어플리케이션을 가능하게 한다. 이러한 작업은 텍스트의 추상화(항해의 용이함, 개인적 지식 베이스의 생성, 문서 전체의 클러스터링, 입력되는 메시지의 분류, 텍스트의 집합 또는 인터넷 상에서 정보의 정확하고 의미론적인 탐색, 전자 서적을 위한 사용자와 친숙한 항해 매커니즘 그리고 많은 다른 것들)을 가능하게 한다.

효과적인 텍스트 분석의 최초 단계는 이 텍스트의 의미론적 네트워크의 개발이다. 물론, 텍스트 분석 시스템의 효율성은 의미론적 네트워크를 만드는데 사용

된 알고리즘의 효율성에 달려 있다. 기존 대부분의 접근방법에 있어서 의미론적 네트워크는 몇 가지 이미 정의되어 있는 규칙이나 개념을 기초로 개발되었다. 그러나 더 강력한 알고리즘은 해당 주제에 대한 사전 배경 지식 없이 조사된 텍스트만을 기초로 하여 자동적으로 완벽하게 의미론적 네트워크를 만들어 낼 수 있다. 이러한 경우에 텍스트 개념의 상대적인 중요성은 단지 네트워크내의 다른 개념에 대한 연관성에 의해 정의된다.

이것이 몇 개의 제품 뒤에 숨겨진 아이디어이다. 그것들 중의 하나인 TextAnalyst system(Megaputer Systems, Russia)은 진보된 동종의 텍스트 프로세싱을 기초로 하여 자동적으로 의미론적 네트워크를 만든다. 사용자는 어떠한 규칙을 지정하거나 네트워크 개발의 씨드(seeds)를 제공할 필요가 없다. 생성된 의미론적 네트워크는 단지 분석된 텍스트의 구조, 어휘 그리고 양에 의존한다. 사실, TextAnalyst는 인간의 두뇌에서 텍스트 분석을 위해 사용되는 알고리즘과 비슷한 알고리즘을 구현하고 있다.

7. 텍스트 마이닝 제품

오늘날 시장에는 많은 텍스트 마이닝 제품들이 있다. 텍스트 마이닝 상품들의 종류는 표 1과 같다.

(표 1) 텍스트 마이닝 제품

회사	제품
Aptex Software, Inc.	SelectResponse
Autonomy	Agentware
Data Junction	Cambio
Excalibur Technologies, Corp.	RetrieveWare
Fulcrum Technologies, Inc.	DCOSFulcrum, Search 서버
IBM Corp.	Intelligent Miner for Text
InsightSoft-M	Cross-Reader
Intercon System, Ltd	DataSet
Megaputer, Inc.	TextAnalyst
Semio Corporation	SemioMap
Soverign Hill Software, Inc.	InQuery
Verity, Inc.	KeyView, Intranet Spider

7.1 Agentware(Automony)

Automony는 온라인 텍스트 마이닝에 관련된 세 종류의 상품을 제공한다. 이들은 "Agentware" 상품군을 형성한다.

- Knowledge 서버 : 완전히 자동화되고 정확한 분류, 참조, 정보의 표현 수단을 사용자에게 제공한다.
- Knowledge Update : 특정 인터넷, 인트라넷 사이트, 새로운 공급 사이트, 내부 자료창고 등을 살펴본다. 그리고 그 컨텐츠에서 개별화된 자료들을 생성한다.
- Knowledge Builder : Agentware의 수용력을 그들 자신의 시스템에 통합시키도록 하는 수간. 자료를 모은 후가 아니라 직접 인터넷 상에서 결부시키며, 어느 정도의 의미분석 능력도 가지고 있다.

Agentware의 상품 구조는 혁신적으로 높은 수행력와 분류, 정보의 참조가 자동화되고, 정보 검색의 효율이 향상되며, 디지털 컨텐츠의 동적 개별화가 가능하도록 하는 문맥상의 분석과 개념 추출을 갖는 패턴 맞춤 알고리즘을 결합시킨다.

Automony의 장점은 높은 수행력과 패턴 맞춤 알고리즘에 있다. 이 알고리즘들은 베이지언 확률 모델, 최근의 신경망의 연구, Shannon의 정보이론 등을 근간으로 만들어졌다. Agentware의 Adaptive Probabilistic Concept Modeling(APCM) 기술은 텍스트를 분석할 수 있고, 단어의 관계와 빈도가 의미와 어떠한 관계에 있는지 이해할 수 있기 때문에 기록 내에 있는 주요 개념을 확인할 수 있다. Agentware는 자료의 디지털 요소를 추출하고 텍스트의 의미를 주는 특징들을 결정하기 위해서 발전된 패턴 맞춤 기술(비선형, 적응 디지털 신호 처리)을 채택한다. Dynamic Reasoning Engine(DRE)은 Autonomy's Agentware 시스템의 핵심이다. DRE는 Autonomy의 모회사인 Neurodynamics에서 개발된 높은 수행능력, 신경망 기술 등을 활용하는 진보된 패턴 맞춤 기술에 기반을 두고 있다. DRE는 APCM을 4개의 주요 기능을 수행하는데 적용시키는 것으로 개념 조화, 에이전트 개발, 에이전트 재교육 그리고 표준 텍스트검색으로 구성되어 있다.

7.2 Cambio(Data Junction, Inc.)

Cambio는 자료를 조사하고 의미 있는 데이터를 데이터베이스 파일에서 추출하는 제품이다. 이는 오프라인 도구이지만 웹 상의 자료 창고에 접속하여 쉽게 작업할 수 있다. Cambio는 위치선정, 패턴인식, 고정과 유동 꼬리표 그리고 텍스트 파일에 데이터 요소를 표시하는 작업들에 주로 사용된다. 그러나 다른 제품들과는 달리 의미분석 기능은 갖고 있지 않다.

7.3 Intelligent Miner for Text(IBM)

Intelligent Miner for Text(IMT)는 소프트웨어 개발도구이다. IMT를 가지고 개발한 어플리케이션은 편지나 웹 페이지 그리고 온라인 뉴스 서비스 같은 텍스트 소스로부터 정보를 얻을 수 있다. IMT는 텍스트로부터 패턴을 뽑아내거나 주어진 화제로부터 문서를 조직화하고, 주어진 주제에 일치하는 서류를 조사할 수 있는 능력을 제공한다. IMT는 문서 분석도구들과 향상된 검색 엔진을 가지며 결과를 나타내는데 있어 기능성과 능력을 향상시켰다. 웹 도구들은 텍스트 도출 능력을 향상시킬 수 있는 모든 요소들을 제공한다. IMT는 또한 사용자의 필요에 의해서 변경 되어질 수 있는 몇몇 어플리케이션을 제공한다.

7.3.1 IMT Language Identification 도구

언어를 분석하기 위해서 연습 문서의 세트에 기초해 언어를 확인하는 서류의 목차에서 단서를 찾는다.

7.3.2 IMT Feature Extractor

두 가지 가능한 모드에서 동작할 수 있다. 그 중 하나는 개별 문서를 분석하는 것이다. 더 향상된 모드에서는 사전에 비슷한 문서들의 수집으로부터 자동적으로 구축된 사전 속에서 발생하는 문서에 단어들을 위치시킨다. 문서의 모음을 사용할 때 feature extractor는 많은 문서들로부터 어휘를 찾기 위한 증거들을 수집할 수 있고, 각각의 아이템들에 대한 통계적인 중요도를 측정할 수 있다.

7.3.3 IMT Term Extraction 모듈

문서 내에서 여러 단어의 기술적인 의미들을 자기 스스로의 간단한 개발법을 사용하여 정의한다. 이것은 영어 단어 사전을 포함하면서 부분적으로 언급된 정보들에 기초해 개발된 방법이다. 이 과정은 다른 접근법들보다 훨씬 빠르다.

IMT의 군집분석은 그룹 내에서 문서의 조합을 나누는데 있어 전적으로 자동화 되어진 과정이다. 각각의 그룹 내에서 문서들은 서로 비슷한 면을 갖는다. 문서의 목차가 군집분석의 기초로 사용되어질 때 다른 그룹들은 그 수집된 것들에서 토론되어진 다른 주제들이나 테마에 일치한다. 그러므로 군집분석은 수집된 것이 포함하고 있는 것을 알아내는 한 방법이다. 그룹의 주제를 정의하는 것을 도와주기 위하여 군집분석 도구는 그룹 내의 문서에서 일반적으로 사용 되어지는 단어나 용어의 리스트를 정의한다. 또한 군집분석 도구는 그룹 내의 문서에서 일반적으로 사용 되어지는 단어나 용어의 리스트를 정의한다. 또한 군집분석은 그들의 길이나 가격, 날짜 같은 문서의 특성 조합의 관점에서 행해질 수 있으며, 이는 또한 데이터 수집에 적절히 사용된다.

도구 모음에서의 IMT 분류자는 분류 되어진 문서로부터 독특한 특징을 확장시키고, 이러한 특징들을 연습구문 내의 예제 문서로부터 확장되어진 각각의 분류에서의 특징과 비교한다. 이 접근은 간결한 인덱스와 빠른 동작을 갖는다.

IMT의 텍스트 검색엔진은 다른 제품에 비해 몇몇 향상된 특징을 제공하는데, 그것은 같은 검색엔진 내에서 많은 다른 검색 범례를 이행하는 것이다. 검색엔진의 핵심은 자유로운 글쓰기와 잡다한 질문들을 제공하는 인덱스 구조이다. 음성적인 검색들이 같은 인덱스 구조에서 가능하며, 특별한 목적의 인덱스는 퍼지 검색이나 일본어, 중국어, 한국어 같이 2바이트를 기반구조로 가지는 언어도 지원한다.

7.3.4 IMT Examples

Finance Wise(Risk Publications and IBM Securities and Capital Markets에 의해서 개발 되어진 검색엔진)는 인터넷상에서 포괄적인 재정 시장 정보를 조사하는 것을 가능하게 한다. Technology Watch라고 불리는 어플

리케이션은 특정 산업에서의 전 세계적인 검색과 투자집중에 대한 조사 그리고 인구 통계학 내에서의 조사와 같은 주요 흐름을 발견하기 위한 특허 어플리케이션들을 분석하기 위해서 IMT 텍스트 알고리즘을 사용한다.

IBM Technology Watch는 파리에 있는 IBM European Centre for Applied Mathematics(ECAM)에 의해서 완벽하게 보완된 상대적인 데이터 분석 테크닉을 사용하는 IMT 어플리케이션이다. 이것은 선택 되어진 특허들이나 텍스트 문서들을 분석하고 자동적으로 그들은 많은 그룹들 안에서 이름에 맞게 분류한다. IBM Technology Watch는 각각의 그룹들에 있는 특허나 문서들이 가능한 한 단순하고 다른 그룹들과 잘 구별될 수 있도록 만들어준다. 간단히 말해, IBM Technology Watch는 다른 그룹들을 보여주는 지도의 형식에서 각각의 그룹을 구분하는 핵심 단어와 각각의 그룹에서 특허 숫자의 형식으로 분석의 비주얼한 요약을 보여준다. 그룹들 사이에서의 뚜렷한 동일성은 그들을 서로 연결하고 있는 색깔 있는 선들에 의해 나타내어진다. 그 색깔들은 관련성의 강함 정도를 나타내며, 노드나 링크에서의 투사를 사용할 수 있다.

7.4 SemioMap(Semio Corporation)

SemiMap는 문서의 큰 틀을 조사하고 검색 가능한 정보의 개념 지도를 만드는 제품이다. 파일을 통해 개념을 조사할 필요 없이 지도를 통해 이들을 탐색할 수 있다. Semio는 모든 문서에서 웹 조사 도구를 사용함으로써 정보를 모으도록 요구한다. Semio가 모든 문서 분석을 마치면 그것은 한 단계씩 작동하고 그 문서의 의미를 해석한다. 그것은 자동적인 문장 분석을 수행하지만 하나의 모드보다는 사용자 상호간에 서로 사용 되어질 수 있도록 디자인 되어진다.

SemioMap는 Semio 개발자인 Claude Vogel에 의해 개발 되어진 특별한 텍스트 채집 기술(SEMIOLEXTM)을 사용한다. 그것은 패턴화된 커뮤니케이션에 의해서 수행 되어진 신호의 형식적인 연구인 수학적 기호학에 적합하다.

SemioMap 소프트웨어는 텍스트 수집으로부터 관련된 모든 구문들을 발췌한다. 그것은 이들 그룹의 가장 정적인 특징들을 강화하고 관련된 구문들로 어휘적인 네트워크를 건설한다. SemioMap에 의해 건설 되어진 패턴은 텍스트 수집의 개념적인 골격을 나타낸다. 그 개념적인 골격은 의미 심장한 관계들과 복잡한 그룹들의 핵심을 나타내며 또한 지식 덩어리인 전 기호학의 네트워크 주제를 나타내는 핵심 개념이다. 강조 되어진 테크놀러지들은 어휘적인 과정과 정보 수집 그리고 그래픽적인 화면을 포함한다.

7.5 TextAnalyst(Megaputer, Inc.)

TextAnalyst는 인텔리전스적인 텍스트 마이닝이고 의미론적인 정보조사 시스템으로 자연어로 쓰여진 텍스트 구조의 프로세싱을 위한 유일한 신경망 기술을 수행한다. TextAnalyst의 기능은 Semio와 비슷하다. TextAnalyst는 이미 얻은 문서의 수집과 더불어 주로 상호 작용을 위해 고안되었다. TextAnalyst는 지식베이스의 구축, 정보의 의미론적인 탐색, 자동 텍스트 추출을 위해서 사용될 수 있다.

이 텍스트 마이닝 소프트웨어는 세 단계로 의미론적 네트워크를 만든다. 텍스트는 한 번에 한 심볼씩 가변길이 윈도우를 통해서 들어간다. 심볼은 글자, 구두점, 여백이 될 수 있다. 윈도우는 2내지 20개 심볼 너비로 구성될 수 있다. 일련의 텍스트가 윈도를 통해 들어갈 때, 텍스트내의 단어와 어근, 어군을 표현하기 위한 스냅샵이 만들어진다.

다음 단계는 하나의 문장과 같은 어떤 텍스트의 의미론적인 조각 안에서 이러한 개념이 얼마나 자주 등장하는가를 식별하는 것이다.

그 다음 단계에서는 예비단계의 의미론적 네트워크를 개발한다. 그 안에서 모든 단어는 가중치를 가지고 모든 발견된 개념은 빈도분석을 기반으로 상응한 가중치를 가진다. 그리고 용어 사이의 관계가 얼마나 자주 발생하느냐에 따라 가중치가 달라진다.

이 과정의 세 번째 단계는 Hopfield 망과 유사한 신경망을 사용한다. Hopfield 망은 모든 마디가 상호 연결된 일차원 신경망이다. 초기 의미론적 네트워크는 신경망의 입력으로 사용되며, 그 결과가 정체된 의미론적 네트워크이다. 관계의 연결 강도 조정을 통해 네트워크가 재규격화 되고, 최종 의미론적 네트워크를 만든다.

의미론적 네트워크의 구축은 가장 중요한 기본적인 요소이다. TextAnalyst 적용에는 지식베이스 만들기, 임의의 텍스트 내용 분석, 텍스트 추출, 특정 주제별로 텍스트 분류, 정보의 의미론적 탐색 수행 등이 포함된다.

8. 인간의 지능 활용

그 동안 모든 산업 분야의 조직들은 많은 양의 데이터를 축적해 왔다. 그들은 데이터를 구성하고, 관리하고, 데이터로의 접근을 제공하기 위해 많은 종류의 데이터 조작 도구와 기술을 개발하고 도입했다. 치열한 경쟁에 대처하고 조직의 효율성 증대를 위해서 기업들은 그들의 데이터베이스에 묻혀 있는 정보를 찾아내기 위해 데이터 마이닝 기술을 사용하기 시작했다.

그러나 저장된 데이터와 그 속에 묻혀져 있는 지식은 인간의 가공품이다. 컴퓨터와 사용자간의 통신이 인간이 인위적으로 만들어 놓은 환경 하에서 진행되어 왔고, 앞으로도 사용될 것이기 때문에 새로운 종류의 도구(semantic computing tool이라 불림)가 필요할지도 모른다는 논쟁이 일어날 수 도 있다. ASOC은 고고학이 광산업과 다른 만큼 여타 데이터 마이닝 도구와 다른 데이터 마이닝 접근법을 제시했다. 이 회사는 의미론적 계산도구를 개발했다. 다음 문장은 ASOC의 출판물에 기술된 내용 중 의미론적 계산의 실례를 언급한 것이다.

"정보화 시대의 등장으로 생존을 위한 기업들의 환경이 전투적으로 바뀐다. 마이닝 도구의 개발은 유용했지만, 그 결과는 아직 미흡하다. 비록 마이닝이란 단어가 폭넓게 사용되는 은유법이지만, 종종 진화 압력과 더 잘 비교된다. 이 진화 압력은 자연환경의 풍부한 정보에 빠르고 유용하게 접근하기 위해서 그리고 그것들에 식물보다 나은 동물의 특권인 빠른 반응을 위해서, 센서 개발, 신경 시스템과 두뇌로 유도한다."

도구들은 훨씬 더 풍부한 데이터 환경에 참여하기 위해 유사한 진보 단계를 겪어야 할 것이다. 이것은 인공 인식 시스템을 고안하기 위해서 더 많은 기술개발을 필요로 하게 될 것이다. 그들의 생존을 위해 중요한, 고차원 데이터 스페이스 안의 사물과 규칙, 기회, 위험 등을 보고 느낄 수 있게 해야 한다.

이 목표를 위한 첫 번째 단계로써 우리의 자연 감각을 확장시켜 인공 데이터 영역을 만들어야 할 것이다. 아직까지 이것은 매우 야심적인 접근법으로 보일지도 모르지만, 이런 원리를 바탕으로 한 데이터 시각화 시스템을 상업용으로 만든 회사가 있다. 그 회사가 ASOC (Associative Computing)이고, 그 상품이 SphinxVision인 것이다. 이 회사는 이 도구로 적절한 알고리즘을 이끌어 내며, 그 알고리즘은 두뇌의 지각모양도의 자가조직 형성의 근거가 되는 많은 과정을 묘사한다고 말한다. 이와 같은 알고리즘은 많은 응용분야에서 이미 매우 성공적으로 적용되었다. ASOC는 이러한 접근법을 개발 확장시켰고, 사용자가 거의 직접적으로 다양한 종류의 데이터 공간에 접촉할 수 있는 고성능 GUI(Graphic User Interface)와 결합시켰다. 비록 많은 다른 접근법들은 또 다른 방법의 다양한 집합체를 가진 도구 박스이지만(신경망과 유전 알고리즘에서부터 다양한 통계 알고리즘까지의 범위를 포함), ASOC의 SphinxVision은 다른 전략을 따른다. 인간 감각의 놀라운 능력 중의 하나는 사물과 기회 그리고 위험을 직접적으로, 사용자 안내서의 도움이나 평가의 어떤 복잡한 절차를 따르지 않고 식별할 수 있다는 것이다. 이 회사의 개발자는 우리의 자연 감각을 인공적인 데이터 공간에 확장시키는 시스템을 위해 비슷한 특성이 필요하다고 말한다

ASOC의 의미론적 계산 개념은 중요한 의미를 갖는다. 사용자에게 의미론적 계산은 사용자가 특정 알고리즘으로부터 독립적으로 일할 수 있고, 사용자들의 도메인을 위한 당장의 관심인 의미론적 개념에 집중할 수 있다는 것을 뜻한다.

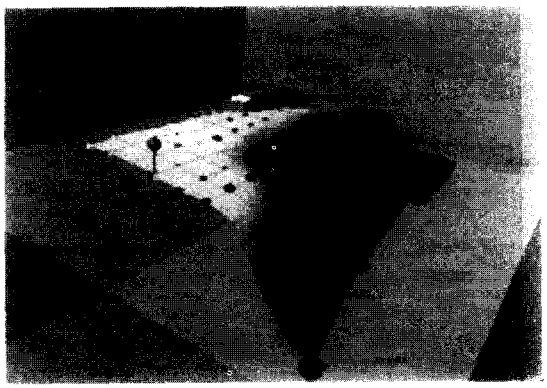
언어로 개인들간 의사소통을 하는 인간의 지식표현에 대해서 생각해 보면, 언어는 일련의 단어들로 구성되어 있고, 각각의 단어들은 어떤 의미와 관련되어 있다. 문장에서 여러 개의 단어들의 조합은 주어진 문법 구조에 순응하게 되면, 문맥 내에 단어들을 포함시킴으로써 더 높은 정도의 의미를 내게 된다. 이러한 문맥의 구성은 단어를 의미있는 문장으로 전환하는데 있어 의미론적 규칙에 의해 지배된다. 지식은 의미론적 문맥에 의해 지배되는 의미의 세계에서 표현되고 소통이 되는 것이다.

만약 사용자가 이미 어떤 의미론적 개념이 적당한지를 알고 있다면, 정보 마이닝 도구는 사용자들에게 개념들 사이의 의존성과 불규칙성을 직접 조사할 수 있도록 해준다. 이것은 한 번 언뜻 보아도 고차원 공간에서의 미세한 관계도 감지될 수 있도록 하는 매우 발전

된 인간의 두뇌 패턴인식 능력을 이용할 수 있는 전체적이고 차원 축소된 조망을 제공함으로써 행해지는 것이다. 이러한 패턴이 확인된 후, 정보 마이닝 도구는 더 세부적인 분석을 돕기 위해 이미 정립된 전통적인 방법을 적용할 수 있는 다양한 방법을 제공한다.

많은 경우에 어떤 분석을 위한 적절한 의미론적 개념은 시작부터 분명하진 않지만, 분석 자체의 일부분으로 간주되어야 한다. 이것은 완전히 자동화될 수는 없다. 항상 사용자가 해당분야의 지식을 레버리징 해야 한다. 의미론적 연산에 있어서 대부분은 작업은 직관적이지만, 사용자의 해당 지식과 데이터베이스 사이에 있어 아주 잘 제어되는 연결쪽으로 방향을 맞추고 있는 것이다. 이러한 작업에 있어 중요한 선결과제는 사용자의 눈앞에 배치된 데이터 객체의 조망을 만들 수 있는 능력인 것이다.

예를 들어, 금융서비스에 있어 전형적인 어플리케이션인 위험고객 예측 프로그램을 생각해보자. 이 어플리케이션의 바람직한 결과는 실제고객 데이터베이스로부터 적절한 레코드(수입, 나이 등과 같은 필드를 포함하는 고객 레코드)들을 추출하여 위험고객에 대한 범주를 확인하는 것이다. 여기서 전체 고객집단은 다른 크기와 색 객체의 모임으로 표현될 수 있다. 이러한 데이터 조망의 그래프적 표현은 명확하고 빠르게 어느 곳에 신용이 좋은 고객들이 위치하고 있는지를 표현한다. 그리고 현재 지금 우리가 파이와 바 차트를 가지고 작업을 하는 것처럼 단지 두 개의 변수 이상을 볼 수 있게 해주는 그림 1과 같은 다차원적이고 유연한 3차원 애니메이션으로 이러한 작업을 하는 것이다.



(그림 1) 정보 마이닝 툴을 이용한 데이터 조망

의미론적 계산에 있어, 사용자는 객체들의 색 조합이 위험고객의 개념을 표현하는 것이라고 말하는 객체들의 한 부분을 확인할 수 있다. 이렇게 조심스럽게 만들어진 의미론적 개념은 다양한 방법으로 평가가 될 수 있다. 예를 들면, 특정한 허용 정책 내에서 고객의 위험을 예측하는데 이용될 수 있다. 또한 부가적인 사용자 정의 개념으로 소개될 수 있는데, 이러한 개념과 이미 존재하고 있는 개념들 간의 상호의존성을 상세히 살펴볼 수 있다. 이러한 평가에 근거해 상호작용으로 현재 데이터 조망의 경계를 변화시킴으로써 정제될 수 있다. 이러한 평가에 근거해 상호작용으로 현재 데이터 조망의 경계를 변화시킴으로써 정제될 수 있다. 물론 각 단계에 있어 관습적인 수치적 표현, 예를 들면, 평균 값, 또는 1·2차원 의존성 도표 등을 추출할 수도 있다. 마지막으로 사용자 정의 개념의 정제가 완료될 때, 추후의 어플리케이션이나 평가를 위해 재사용 가능한 의미론적 모델로 저장될 수 있다. 이러한 것이 사용자가 똑 같은 의미론적 개념의 다양한 변종들을 만들어 내는 것을 가능하게 한다. 예를 들면, 은행은 시간에 따라 몇몇 위험고객 개념을 개발할 수 있는데 각각은 특별하게 서로 다른 금융서비스에 맞춰질지도 모른다. 이러한 방법으로 의미론적 계산은 매우 유연하며 조직 내의 경험적 지식의 중요한 부분의 모듈화를 지원할 수 있고, 또한 이러한 지식을 멀리 떨어진 곳에서도 사용이 가능하게 할 수 있다. 정보 마이닝 툴은 또한 대시 보드를 제공하여 사용자로 하여금 이러한 복잡한 데이터 조망을 살펴볼 수 있도록 해준다.

9. 결 론

결론적으로 2000년대는 Y2K와 같은 문제를 발생시켰을 뿐만 아니라 정보 혁명을 가져왔고, 우리의 삶과 경제적인 면에서 많은 변화를 가져왔다. e-비즈니스 시대는 웹 속도로 빨리 움직이고 종종 많은 놀라움(그것들 모두가 반길 일은 아니지만)을 가져다 주는 반면, 기회를 잃어버린 사람들은 홀로 남겨지게 될 것이다. 그러나 데이터 마이닝, 정보 마이닝과 같은 강력한 기술에 의해 제공된 능력을 십분 발휘하는 사람들은 그 데이터에서 중요한 경향을 발견할 수 있고 고객, 파트너, 생산자의 관계를 강화할 수 있다. 그 결

과 그들은 경쟁력을 갖게 되고, 남들이 부러워하는 위치에 서게 될 것이다. 데이터 마이닝과 정보 마이닝, CRM에 대한 미래는 e-비즈니스의 놀라운 잠재력과 밀접하게 연관되어 있어 그 미래가 아주 밝게 보인다. 특히 인터넷의 광범위한 채택과 함께 필연적으로 발전하는 e-비즈니스에서 지식관리는 창조성, 기억력, 접근하기, 공유하기를 장려하고 기업 이윤을 위해 기업의 무형 자산에 투자하기 위한 통합적 접근법을 촉진하는 훈련이다. 지식 보관소는 잘 정리되고 계층적으로 구조화된 지식 기반 서비스의 복합체이다. 그것은 간단한 업무에서부터 복잡한 예측 모델을 만드는 일까지 두루 포함하고 있고 데이터 유지와 분석 문제에 대한 해결책을 갖고 있다. 예를 들어, 일상 상거래로부터 수집된 현 고객 데이터를 기초로 미래 상품 판매를 예측하거나 통신회사에서 있을 법한 교란 감지를 위한 조기 경보 시스템 등이 있다. 최종 사용자의 견지에서 보면 지식 보관소는 서비스의 집합체로 나타낸다. 이것은 기업이 경쟁에서 우위를 차지하는데 큰 도움을 줄 것이다.

참 고 문 헌

[1] Advances in Knowledge Discovery and Data Mining, Usama M. Fayyad(Editor), et al, 1996

[2] Building Data Mining Applications For CRM, Alex Berson, Stephen Smith, Kert Therling, 1999

[3] Information Graphics, Peter Wildbur and Michael Burke, 1998

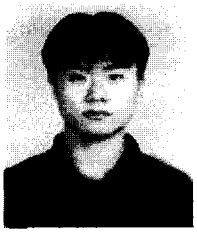
[4] Envisioning Information, Edward Tufte, 1997

[5] The Visual Display of Quantitative Information, Edward Tufte, 1983

[6] Data Warehousing for Dummies, IDG Books, 1997

[7] 효과적인 인터넷 마케팅을 위한 웹 로그 분석, 김형택 · 민옥길, 2001

◐ 저 자 소 개 ◑



이 종 현

2002년 중앙대학교 컴퓨터공학과 졸업(공학사)
2002년~현재 : 중앙대학교 대학원 컴퓨터공학과 석사과정
관심분야 : 분산객체 컴퓨팅, Grid System, 데이터베이스

임 혜 영

2001년 성공회대학교 전산정보학과 졸업(이학사)
2002년~현재 : 서울여자대학교 대학원 컴퓨터학과 석사과정
관심분야 : Grid System, Linux Kernel Programming, High Performance Computing

황 준

1985년 중앙대학교 전자계산학과 졸업(학사)
1987년 중앙대학교 대학원 전자계산학과 졸업(석사)
1991년 중앙대학교 대학원 전자계산학과 졸업(박사)
1992년~현재 : 서울여자대학교 정보통신공학부 교수
2002년~현재 : 서울여자대학교 정보통신대학장
관심분야 : 분산객체 컴퓨팅, Grid System, Linux Kernel Programming, High Performance Computing