전자상거래 상에서의 실시간 데이터 마이닝 활용 모델

김고은[O1], 옥지웅[1], 김응모[1]
[1]성균관대학교 컴퓨터공학과
wowgoeun[C]@hanmail.net, {okjwguy, umkim}@ece.skku.ac.kr

Real-time Data Mining application Model In Electronic Commerce

Koeun Kim[O1], Jeewoong Ok[1], Ungmo Kim[1]
[1]Dept. of Computer Engineering, SungKyunKwan University

요 약

현재 전자상거래는 우리의 생활과 밀접히 연관되어 있다. 최근 인터넷을 기반으로 전자조달, 수출입 브로커 등과 같은 유형의 B2B 전자상거래가 활발히 이루어지고 있으며, 소비자를 대상으로 하는 전자상거래 또한 점차 확산되는 시장을 형성하고 있다. 국제적으로도 전자상거래 시장 규모가 급속도로 증가할 것이라는 전망은 자명한 사실이다.

전자상거래에 대한 의존도가 높아지면서 관리해야 하는 데이터의 양 또한 급속도로 증가하고 있다. 본 논문에서는 실시간으로 유입되는 데이터를 효율적으로 활용하기 위해 실시간 데이터 마이닝 활용 모델을 제안한다. 이 실시간 데이터 마이닝 모델은 지속적으로 유입되는 데이터의 규칙화를 통해 저장 공간의 효율성을 극대화하고 중요도 분석을 통한 총체적인 접근 방법을 시도함으로써 전자상거래 상에서 유용하게 쓰일 수 있는 활용 모델이다. 이 실시간 데이터 마이닝 모델의 바탕은 데이터 마이닝의 기법인 SEMMA를 따르며, 그 특징에 따라 규칙 추출과 의사 결정 나무 기법을 이용하여 전자상거래 상에서 유용하게 사용될 수 있는 모델을 제시하고자 한다.

1. 서 론

현재 전자상거래(Electronic Commerce)는 우리의 생활과 밀접히 연관되어 있다. 최근 인터넷을 기반으로 전자 조달, 수출입 브로커 등과 같은 유형의 B2B 전자상거래가 활발히 이루어지고 있으며, 소비자를 대상으로 하는 전자상거래(B2C) 또한 점차 확산되는 시장을 형성하고 있다. 통계청에서는 우리나라의 올해 1/4분기 전자상거래 총 규모는 115조 9,970억원으로 전년 대비 26조 540억원에 비해 29.0% 증가한 것으로 발표했다. 우리나라 뿐 아니라 국제적으로도 전자상거래 시장 규모가 급속도로 증가할 것이라는 전망은 자명한 사실이다.

이렇게 전자상거래에 대한 의존도(dependence)가 높아지면서 실시간으로 유입되는 데이터를 가공하여 활용하는 것에 대한 관심도도 함께 높아졌다. 데이터 마이닝(Data Mining)이란 거대한 양의 데이터베이스 혹은 자료로부터 의사 결정에 유용한 정보 및 지식을 발견하려는 일련의 자료 분석 및 모형 선정 과정이다[1]. 따라서 판매자와 소비자 간의 만족도를 높이고 상호 간의 거래와 그 이후의 피드백(feed-back) 작용을 위해 데이터 마이닝을 전자상거래에 적용하는 과정이 필요하다.

특히, 전자상거래는 실시간으로 이루어지기 때문에 데이터 마이닝 중에서도 실시간 데이터 마이닝(Real-time Data Mining)이 적합하다. 시간의 흐름에 따라 변하는 연속적이고 잠재적으로 무한히 발생하는 특징을 가지고 있는 데이터를 연속 발생 데이터(Stream Data)라고 하는데 이렇게 자료가 고정되어 있지 않고 계속 증가하는 경우에 마이닝하는 대상 자료의 변화를 다루는 마이닝 기법이 실시간 데이터 마이닝인 것이다.[2]

시간의 흐름에 따라 방대한 양의 데이터가 들어오는 특성 때문에 기존의 데이터 마이닝과는 달리 저장 공간의 문제를 가진다.[3] 또한 이전의 중요하지 않은 정보가 이 후에 들어오는 데이터와 결합하여 중요한 정보가 될 가능성을 내재하므로 마이닝 과정에서의 오차 확률이 일반적인 데이터 마이닝에 비해 상대적으로 높다. 이 전의 연구 방향은 데이터의 주기적인 변화에 초점을 맞춘다는 것 보다 현재의 추세에 초점을 맞추는 경향이 있었다. 하지만 지속적으로 유용한 정보를 추출해내기 위해서는 현재의 정보에 의존하는 것 보다 조금 더 총체적인(ensemble) 접근 방법이 필요하다.

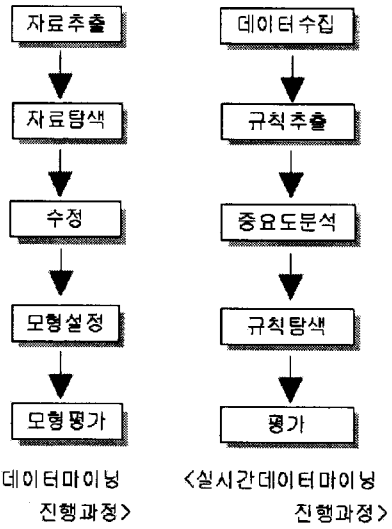
따라서 본 논문에서는 실시간 데이터 마이닝의 단점을 극복하고 지속해서 유입되는 데이터를 효율적으로 활용하기 위해 실시간 데이터 마이닝 모델을 제안한다. 실시간 데이터 마이닝 모델의 바탕은 데이터 마이닝의 기법인 SEMMA를 착안한다[4]. 또한 실시간 데이터 마이닝 고유의 특징에 따라 규칙 추출과 의사 결정 나무 기법(Dision Tree)을 이용하여 전자상거래 상에서 유용하게 사용될 수 있는 모델을 제시하고자 한다.

2. 실시간 데이터 마이닝 모델

2.1 기존 모델과의 비교

실시간 데이터 마이닝(Real-time Data Mining)의 진행 과정을 살펴보기 전에 기본적인 데이터 마이닝(Data Mining)의 진행 과정을 살펴보고자 한다. 데이터 마이닝의 진행 과정은 자료 추출(Sampling), 자료 탐색(Explore), 수정(Modify), 모형 설정(Model), 모형 평가(Assess)의 과정을 따른다. 구체적으로 데이터 마이닝의 진행 과정을 보면 먼저 전체 데이터 중 다루고자 하는 방향과 맞는 자료를 분류한 후, 자료를 분석한다. 그리고 자료의 형태 등을 변환하여 자료들을 가공한 다음, 모형을 선택하고 마지막으로 모형을 평가하는 과정을 거친다.

실시간 데이터 마이닝의 진행 과정은 기본적으로 데이터 마이닝의 진행 과정인 SEMMA 기법을 따르나 실시간으로 자료가 축적되며, 자료의 변화에 따라 그 중요도와 최종적인 정보도 변화해야 한다는 특성에 따라 다소 변형된 진행 과정을 갖는다.



[그림 1] 진행 과정의 비교

[그림 1]은 데이터 마이닝의 진행 과정인 SEMMA와 본 논문에서 제안하는 실시간 데이터 마이닝 진행 과정을 비교해서 보여주고 있다.

2.2 전자상거래 상에서의 실시간 데이터 마이닝 모델

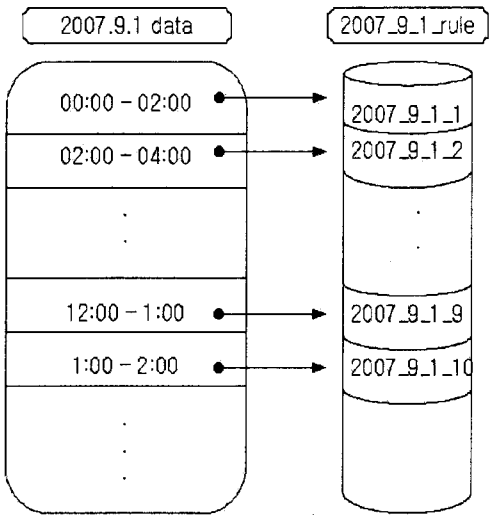
실시간 데이터 마이닝의 진행 과정은 앞서 말한 바와 같이 [그림1]과 같으며, 2.2절에서 전자상거래 상에서 실시간 데이터와 관련하여 유용하게 활용될 수 있는 경우의 예를 들어 단계적으로 설명하고자 한다.

2.2.1 데이터 수집

데이터 마이닝을 위해 선행되어야 할 과정은 데이터 수집이다. 하지만 실시간으로 발생하는 데이터를 모두 수집하는 것은 현실적으로 불가능하다. 유입되는 데이터 그대로를 활용하게 될 때의 가장 큰 문제점은 용량의 문제이다. 또한 그 양이 방대하기 때문에 유입된 대용량의 데이터로 마이닝을 해도 효율성과 정확성이 떨어지는 문제점이 발생한다. 따라서 한정된 데이터를 수집해 데이터 마이닝 작업을 하는 것과는 달리 실시간 데이터 마이닝 작업에서는 데이터를 일정량 수집하게 되면 실시간으로 다음 단계인 규칙 추출을 통해 데이터를 효율적으로 저장하도록 하는 방법을 따른다.

2.2.2 규칙 추출

데이터 수집의 과정 후, 규칙 추출의 단계를 거친다. 규칙 추출 단계는 실시간으로 축적되는 데이터를 일정 기준에 따라 나누고 이것을 규칙이라는 형태로 가공하여 저장하는 단계이다. 이것은 실제 데이터 마이닝의 기법과 유사하며, 한정된 데이터를 통해 규칙을 추출하는 것을 기본 바탕으로 한다. 데이터를 나누는 일정 기준으로 데이터의 양 또는 크기, 시간 등을 고려할 수 있다. 데이터를 나누는 기준 중에서 전자상거래 상에서 유용하게 사용될 수 있는 기준은 시간이다. 그 이유는 시간에 따라 유입되는 데이터의 양이 비슷하다는 점과 시간을 이용함으로써 지속적인 데이터 마이닝이 가능할 수 있는 점을 특징으로 가지기 때문이다.



[그림 2] 전자상거래 상에서 시간 기준의 규칙 추출

특히, 전자상거래 상에서 직접적인 거래나, 고객들의 사이트 동시 접속 수 등은 시간에 의존하게 된다. [그림 2]는 전자상거래 상에서 시간을 기준으로 규칙을 추출하는 것을 설명하고 있다.

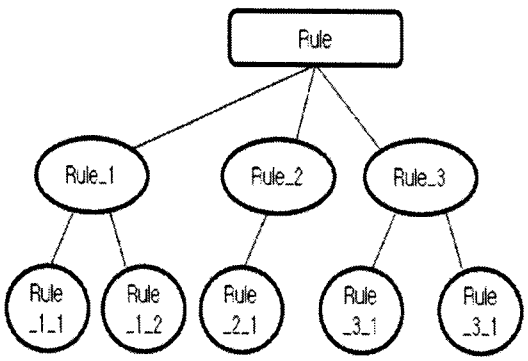
먼저 날짜를 기준으로 2007년 9월 1일의 규칙 집합을 생성한다. 그리고 시간을 기준으로 데이터를 규칙화하여

추출하는데, 24시간을 기준으로 1시간이나 2시간씩 일률적으로 배분하는 것이 아니라 고객의 방문 횟수나 거래의 성립 빈번도에 따라서 새벽 시간대(00:00 - 06:00)에는 2시간씩, 낮 시간대(6:00 - 00:00)에는 1시간씩 데이터를 나누어 규칙을 유동성 있게 추출한다. 이 시간에 따른 분류는 전자상거래가 이루어지는 환경이나 제품, 고객의 성향에 따라 달라질 수 있다. 또한 추출된 규칙은 저장하고, 본래의 데이터는 버림으로써, 한정된 저장 공간을 효율적으로 이용할 수 있고, 용량의 부족 문제를 해결할 수 있다. 또한 추출된 규칙은 상대적으로 압축된 정보를 지니고 있기 때문에 이 후 마이닝의 작업을 거칠 때에 기존의 데이터보다 더 빠르고 정확하게 작업이 진행될 수 있도록 한다.

2.2.3 중요도 분석

다음 단계는 중요도 분석의 단계로써, 앞서 구성한 규칙들 중 상대적으로 중요한 규칙과 그렇지 않은 규칙을 분류하는 것을 수행한다. 중요한 규칙에 가중치를 부여함으로써, 구성된 규칙 집합들을 토대로 데이터 마이닝을 할 때 중요도가 높은 규칙들이 유용한 정보로 분류될 가능성을 높인다.

또한, 덜 중요하게 판단되는 규칙들을 누락시키지 않는 것이 이 과정의 특징이다. 실시간으로 데이터가 유입되면서 규칙들 또한 계속 쌓이게 된다. 이것은 처음에는 중요하지 않게 여겨지는 정보가 후에 쌓여진 축적된 정보로 인해 중요도가 높아지는 경우가 생길 가능성을 시사한다. 따라서 현재 판단되는 중요도에 따라 즉각적인 결정을 실시하지 않고, 규칙들의 순서 매김만을 바꾸어 준다. 즉, 덜 중요하게 판단된 규칙들의 영향력이 일정 시간 후 유용하게 쓰일 경우를 대비한다.



[그림 3] 가중치를 부여한 규칙

중요도를 판단하기 위해 데이터 마이닝의 기법 중 의사 결정 나무 기법(Dision Tree)을 이용한다.[5,6] 의사 결정 나무란 각 규칙들을 사용자가 필요로 하는 정보 중 영향력 있는 규칙을 선정하여 뿌리(Root)에 저장하고 다음으로 영향력을 미치는 규칙을 내부 마디로 지정하는 방법으로 나무를 확장시켜 분류 규칙을 갖는 기법을 의미한다. [그림 3]에서 가장 중요도가 높다고 판단되는 Rule_1, Rule_2, Rule_3를 상위 노드로 하고, 그 연관성을 바탕으로 덜 중요한 노드를 하위 노드로 하여 규칙들을 배치한다.

2.2.4 규칙 탐색과 평가

중요도 분석을 통해 규칙들을 분류하게 되면 규칙 탐색을 통해 유용한 규칙을 추출한다. 새로 유입되는 규칙과 중요도 분석을 마친 가공된 규칙들을 통해 실제 전자상거래 상에서 활용될 수 있는 데이터와 직결되는 정보를 얻는다. 그리고 이 최종적인 정보를 평가하는 단계를 마지막으로 갖는다. 실시간 데이터 마이닝은 지속적으로 정보가 갱신되기 때문에 분석, 평가가 적절한 시점에 이루어져야 하므로 그 중요도가 높다.

2.3 모델의 기여도

이 모델의 기여도는 크게 두 가지로 볼 수 있다. 첫째, 데이터를 규칙화하여 저장함으로써, 사용할 데이터의 총 용량을 줄이는 것이다. 즉, 규칙 추출을 통한 일련의 과정을 거치면서 규칙 추출을 하기 위해 실시간 기준을 통해 데이터의 일정 양을 대폭 감소시키면서 그 규칙을 일정 형태로 저장함으로 저장 공간의 유용하게 활용할 수 있게 된다. 두 번째는 의사 결정 나무 기법을 이용하여 중요도를 분석함으로써, 최종적으로 데이터 마이닝을 하여 유용한 정보를 추출해낼 수 있다는 점이다. 특히, 중요도가 적은 규칙을 하위 노드로 분류하고 그 규칙을 남겨둠으로써, 후에 유입되는 규칙들과 결합하여 더 중요한 가능성이 될 수 있다는 점을 배제하지 않았다.

따라서 이 모델은 데이터를 규칙화함으로써 저장 공간의 효율을 극대화하고, 중요도 분석을 통한 총체적인 접근 방법을 통해 전자상거래 상에서 유용하게 쓰일 수 있는 모델이다.

3. 결 론

본 논문에서는 전자상거래 상에서 발생하는 데이터의 특징에 따라 실시간 데이터 마이닝 기법을 통해 유용하게 사용될 수 있는 데이터를 추출하는 모델을 제안하였다. 특히 데이터를 규칙화하여 최소한의 데이터를 저장함으로써 저장 공간을 활용하였다. 또한 오류 발생률을 줄이기 위해 중요도 분석을 통한 규칙 재배열 과정과 실시간 업데이트 과정을 실시함으로써 정확도를 높일 수 있는 접근 방법을 제시하였다. 따라서 이 모델을 이용하여 전자상거래 상에서 1-1 마케팅을 위한 고객 분석이나, 상품 추천을 통한 고객 관리, 하나의 상품에 대한 고객 성향 분석 등에 사용한다면 판매자와 고객과의 동시 만족이 가능할 것으로 예측한다.

본 논문에서 제안한 실시간 데이터 마이닝의 진행 방향의 초점은 용량의 효율성 극대화와 실시간으로 갱신되는 데이터를 단시간에 유용한 정보로 추출할 수 있는 것이었다. 하지만, 규칙 추출과 중요도 분석을 하는 동안 두 번의 데이터 마이닝이 수행되므로, 각각의 분석과 추출은 시간을 절약할 수 있겠지만, 전체적으로는 과정이 나눠지면서 시간이 지연될 수 있다는 단점을 가진다. 따라서 향후 연구에서는 전체적인 시스템을 일관화하면서

현재 이 모델이 지원하는 장점을 살리는 연구가 필요하다.

4. 참고문헌

[1] 조병엽 최영희, "데이터마이닝을 이용한 데이터 활용에 관한 연구, 조선대 사회과학연구지, 제 22집 1호 통권 27호, p163-179, 2001.
[2] Han, J. and M, Kamber, Data "Mining Concepts and Techniques," Morgan Kaufmann, 2001.
[3] L. O'Callagan, R.Motwani, N. Mishra, Guha S., "Clustering Data Streams", Proceedings of the 41st Annual Symposium on Foundation of Computer Science, 2000
[4] George Fernandez, "Data Mining Using SAS Applications", CHAPMAN & HALL/CRC, 2003.
[5] Domingos, P. and G. Huten, Mining High Speed Data Streams", Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p71-80, 2000.
[6] Hulten G., L. Spencer, and P. Domingos, "Mining Time - Changing Data Streams KDD, 2001
[7] 김진화, 민진영, "연속발생 데이터를 위한 실시간 데이터 마이닝 기법", 한국경영과학회지 제29권 제4호, p41-60, 2004.
[8] 이용준, 서성보, 류근호, 김혜규, "시간간격을 고려한 시간관계 규칙 탐사 기법", 정보과학회논문지, 제 28권 3호, p301-314, 2001.
[9] M. Spiliopoulou, J. F. Roddick, "Temporal data mining : survey and issues", Research Report ACRC-99-007, University of South Australia, 1999.
[10] 김형근, 황환규, "대용량의 데이터셋에서 깊이우선 탐색을 사용한 순차 패턴 마이닝",정보통신논문지, 제9권, 2005.