웹 로그 분석을 통한 높은 정확도를 가지는 소형 트리 구축

현우석°
한국성서대학교 정보과학부
wshyun°@bible.ac.kr

Constructing A Small Tree with High Accuracy through Web Log Classification

Woo-Seok Hyun°
Dept. of Information Science, Korean Bible University

요 약

웹 마이닝은 e-서비스 시스템에서 고객 활동을 분석하기 위하여 널리 보급된 방법 중 하나로서 궁극적인 목표는 새로운 고객을 얻고 기존 고객을 유지하면서 고객의 생산성을 증가시키는데 도움을 줄 수 있는 유용한 정보를 인식하는 것이다. 그러나 웹 로그 자료와 고객의 구매 패턴 사이에 직접적인 관계가 없고, 실험 데이터 집합이 적고 부정확 할 경우 실험 데이터의 적은 집합만으로 유용한 정보를 인식하는 것은 불충분하기 때문에 유용한 정보를 인식하는 것은 더욱 어렵게 된다. 본 논문에서는 기업들에게 유용한 패턴을 제공할 수 있는 독자적인 분류 방법을 사용하여 기존 고객의 보존력을 높일 수 있는 높은 정확도를 가지는 소형 트리를 구축할 수 있었다.

1. 서론

인터넷은 고객들로 하여금 다양한 기업의 선택을 가능하도록 하였기 때문에 기업들은 고객의 신뢰를 높이기 위해서 노력해야만 한다. 기존 고객은 새로운 고객보다 2배 정도의 수입을 창출해 주기 때문에, 5% 정도 보존력(retention)이 증가된다는 것은 이익을 95% 정도 끌어 올린다는 연구가 있다[1-2]. 그래서 기업들은 고객의 행동, 선호도와 미래의 요구(need)를 잘 이해해야만 한다. 이러한 의무가 많은 기업들로 하여금 자료 수집과 분석을 위한 많은 좋은 e-서비스 시스템을 개발하도록 하였다.

웹 마이닝은 e-서비스 시스템에서 고객 활동을 분석하기 위하여 널리 보급된 방법 중 하나이다[3-4]. 이것은 문서의 내용으로부터 지식을 추출하는 웹 문서 마이닝, 인터넷 연결과 조직으로부터 지식을 추론하는 웹 구조 마이닝, 웹 접근 로그와 다른 웹 사용 정보에서 유용한 패턴을 추출해내는 웹 로그 마이닝을 포함한다.

웹 마이닝의 궁극적인 목표는 새로운 고객을 얻고 기존 고객을 유지하면서 고객의 생산성을 증가시키는데 도움을 줄 수 있는 유용한 정보를 인식하는 것이다[5].

그러나 웹 로그 자료와 고객의 구매 패턴 사이에 직접적인 관계는 없다. 실험 데이터 집합이 적고 부정확 할 경우 실험 데이터의 적은 집합만으로 유용한 정보를 인식하는 것은 불충분하기 때문에 유용한 정보를 인식하는 것은 더욱 어렵게 된다. 본 논문에서는 기업들에게 유용한 패턴을 제공할 수 있는 있는 높은 정확도를 가지는 소형 트리를 구축할 수 있었다.

2. 고객의 특성

실제적인 웹 로그는 요구 시간, IP 주소, 접근 방법, 요구된 파일의 URL, HTTP 프로토콜 버전 수, 서버 반응 상태, 전달하는 바이트 수, 참조하는 페이지의 URL 그리고 사용자 에이전트 등을 포함하는 정보를 가지고 있다. 본 논문의 웹 로그 마이닝 접근은 특정 사이트의 방문자를 분석하여 구매 관심(purchase interest)을 근거로 서로 다른 그룹으로 분류한다. 가장 좋은 잠재적인 고객은 클릭하고 구입할 준비가 된 고객이다. 어떤 고객은 잠재고객으로 다른 브랜드의 정보와 친숙해서 미래에 수익을 낼 수 있는 고객이 될 수도 있다. 또 다른 고객들은 웹상에서 자유롭게 항해하는 것을 즐기면서 어떤 것도 결코 사지 않는 고객이다.

본 연구에서는 구매 관심을 가지면서 패턴에 접근하지 않는 고객을 독자적으로 인식할 수 있도록 하였다. 구매 관심을 가진 고객은 다음과 같이 네 가지 패턴을 나타낸다. 첫째, 이런 고객은 내용을 읽을 시간이 필요하기 때문에 오랜 시간 특정 페이지에 접근한다. 그래서 한 페이지에서 다른 페이지로 넘어가기 위해 드는 시간에 비해서 내용을 읽기 위해서 필요한 시간이 더 많이 소요된다. 둘째, 이런 고객은 특정 주제에 접근할 필요가 있기 때문에 웹상에서 저수준(low-level) 페이지로 아래로 항해한다. 셋째, 이런 고객은 웹 사이트를 등록하는 것에 관심이 있고 자신의 정보로 신청서를 작성하기 때문에 HTTP POST 모드를 자주 사용한다. 넷째, 이런 고객은 이미지와 그래픽 파일에 자주 접근한다.

반면에 구매 관심을 보이지 않는 고객들은 다음과 같은 접근 패턴을 나타낸다. 첫째, 이런 고객은 내용을 보기 위해서 많은 페이지에 빨리 접근한다. 한 페이지에서 다른 페이지로 넘어가기 위해 드는 시간에 비해서 내용을 읽기 위해서 필요한 시간의 비율은 거의 1에 가깝다.

둘째, 이런 고객은 어떤 특별한 주제에도 관심이 없기 때문에 저수준 페이지로 아래로 항해하지 않고 고수준(high-level) 지식 페이지에 여러 번 접근한다. 셋째, 이런 고객은 웹 사이트를 등록하는데 관심이 없기 때문에 POST 모드를 자주 사용하

지 않는다. 넷째, 이런 고객은 이미지와 그래픽 파일에 접근하지 않는다.

이러한 고객의 두 그룹뿐만 아니라 고객의 특별한 그룹을 고려할 필요가 있는 데 그것은 네트워크 로봇이다. 많은 검색 엔진들은 네트워크 로봇을 사용하는데 로봇은 고객 패턴 발견에 중요한 영향을 주는 웹 로그에서 많은 접근 기록을 생성해낸다. 네트워크 로봇을 인식하는 가장 일반적인 방법은 IP 주소와 에이전트에 의한다. 이것은 검색 엔진과 에이전트가 계속해서 증가되기 때문에 모든 가능한 검색 엔진과 에이전트의 선지식을 필요로 한다. 다른 방법은 robots.txt 파일에 접근했는지 여부를 확인하는 것이다. 인터넷 로봇 배제 표준(Internet Robot Exclusion Standard)에 따르면 네트워크 로봇은 사이트 자체를 방문하기 전에 웹상에서 robots.txt 파일을 읽어야만 한다. robots.txt 파일은 일반적으로 웹 사이트의 root 디렉토리에 있는데, 이것은 접근할 수 있고 접근할 수 없는 웹 사이트의 부분을 명시하게 된다. 그러나 표준이 제한적인 것이 아니기 때문에 몇몇 잘못 의도된 로봇은 그 표준을 따르지 않는다. 정확하게 위장된 로봇을 발견하기 위해서 다음과 같은 접근 패턴을 사용해서 데이터를 새롭게 만든다. 첫째, robots.txt 파일을 읽는다. 둘째, 복잡한 기간을 피하기 위해서 네트워크 활동이 가벼운 한밤중에 방문한다. 셋째, 높은 효율성을 위해서 하이퍼링크가 유효한 지를 확인할 때 Get 대신에 Head 접근 모드를 사용한다. 넷째, 전반적인 웹 사이트를 통해서 깊이 있게 보다는 넓게 웹 페이지를 방문한다. 다섯째, 로봇의 목적이 검색된 데이터베이스를 갱신하고 구축하는 것이기 때문에 그래픽 파일보다는 문서 내용에 접근한다.

3. 분류자 구축

본 논문에서는 웹 로그를 결정 트리(decision tree)를 사용하여 분류한다. 첫째, 웹 로그를 분류하기 위해서 속성의 집합을 (A1, A2, A3, ... ,A*n*)으로 인식한다. 구매 관심을 가진 고객, 구매 관심이 없는 고객, 네트워크 로봇의 세 가지 접근 유형의 분석을 근거로 하여 분류자를 만들기 위해서 9가지 속성을 선택했다. 표 1은 세 가지 유형으로 그룹화된 9가지 속성을 보여준다. 여기서 A1-A3는 시간적 속성, A4-A8은 페이지 속성, A9는 커뮤니케이션 속성을 보여준다. 둘째, 표 2에서 보는 바와 같이 속성 값을 분석해낸다. 이 때 고객의 협조로 소형 규모의 실험 데이터 집합을 확인할 수 있었다. 데이터 집합의 레이블은 자신의 행동에 대한 고객의 이해도를 반영하며 최종적으로 분류자를 생성해 낸다.

중요한 쟁점으로는 소형 실험 데이터 집합으로 어떻게 분류자를 구축하느냐는 것이다. 웹 로그가 고객이 어떤 것을 구입할 것인지를 보여주지 않기 때문에 대형 실험 데이터 집합을 획득하는 것은 어려운 일이다. 거의 모든 분류가 긍정적이고 부정적인 예를 지니는 대형 실험 데이터 집합을 필요로 하기 때문에 기존 분류 방법을 적용하는 것은 이런 집합에서는 어렵다. 이러한 문제를 해결하기 위해서 Tim Dates와 David Jensen[7]에 의해 주어진 인자를 적용한다. 이들에 따르면 실험 집합 크기를 증가시키는 것이 트리 크기를 선형적으로 증가시키지 않을 수도 있으며, 심지어 추가적인 복잡도가 분류 정확도에 있어서 중요한 증가를 초래하지 않는다고 한다. 그래서 본 연구에서는 소형 실험 데이터 집합의 부분집합을 사용하였지만 가장 많이 기여하는 클래스인 구매 관심을 보이는 고객에게 특별히 관심을 두었다.

표 1 속성 분류

Type	Attribute	Description
temporal attributes	A1	심야에서 오전 7시 사이에 접근
	A2	전체 세션 시간
	A3	고객이 사이트에 접근한 시간, 고객이 사이트에 머물러 있는 시간, 고객이 다른 페이지에 머물러 있었던 시간 등과 같은 통계
Page attributes	A4	전 세션 동안 접근한 페이지의 총 수
	A5	접근 폭
	A6	접근 깊이
	A7	전체 접근 페이지 수에 대한 요청된 그래픽 파일의 퍼센티지
	A8	robots.txt 파일에 대한 접근 수
Communication attributes	A9	고객이 사이트와 의사소통하기 위해서 사용한 Get, POST, Head 등의 접근 방법

표 2 분석된 속성 값

Attribute	Value0	Value1	Value2	Value3
A1	No	Yes	-	-
A2	≤ 2 min.	2-5 min.	5-16 min.	15-30 min.
A3	≤ 4 sec.	4-35 sec.	≥35 sec.	-
A4	≤3 pages	3-5 pages	≥5 pages	-
A5	≤3 pages	3-5 pages	≥5 pages	-
A6	1 hierarchy	2-3 hierarchies	≥5 hierarchies	-
A7	0%	0-20%	20-50%	50-100%
A8	No	Yes	-	-
A9	Use Get	Use POST	Use Head	-

4. 실험 및 평가

본 논문에서는 수집된 98,567개 레코드를 가지고 실험하였다. 표 1,2는 분류 속성들과 속성값을 보여준다.
사용자와 세션에 의해 로그를 나눈 후에 35,012개 세션을 획득하였다. 이 때 본 연구에서는 112개 세션을 실험 데이터 집합으로 선택하였다. 모드 레코드들은 9개의 속성을 지니며 하나의 레코드에서 i번째 요소의 수는 i번째 속성값(표 1)을 나타낸다. 표 3 은 20개 실험 레코드의 예를 보여준다. 구매 관심을 가지는 고객을 위한 첫 번째 긍정적인 실험 레코드는 9개의 값을 가지는 데, 각각의 값은 표2 에서 보는 바와 같다. 첫 번째 값 0는 속성 A1의 값이 0 값을 가지는 것을 의미하며, 이것은 야간 접근이 없었다는 것을 의미한다.

112개의 세션 데이터 집합을 근거로 하여 실험 부분집합의 크기를 변화시켜 가면서 실험을 하였다. 그림 1, 2, 3 은 그 결과를 보여준다. 그림 1, 2, 3의 그래프 x축은 사용된 실험 데이터의 퍼센티지를 나타낸다. 112개 레코드의 p퍼센트를 분류자로 구축하기 위해서 사용한 경우에는 그것을 실험하기 위해서 (1-p)퍼센트를 사용하였다. 그림 1 은 분류자를 구축하기 위해서 사용한 데이터 레코드의 수를 증가시킬수록 보다 높은 정확도를 나타냄을 보여준다. 그림 2 는 실험 데이터 집합의 p

퍼센트를 사용했을 때 구축한 분류자의 트리 크기를 보여준다. 그림 1, 2 에 의하면 분류자를 구축하기 위해서 사용한 실험 데이터 집합의 크기가 증가됨에 따라 결정 트리의 크기도 증가됨을 보여준다. 하지만 그림 3 은 분류자를 구축하기 위해서 사용한 실험 데이터 집합의 크기가 증가됨에 따라 미분류된 수도 증가됨을 보여준다. 소형 트리가 높은 정확도를 보여준다는 원칙을 고려했을 때 본 연구에서는 표 3 에서 보는 바와 같이 분류자를 구축하기 위해서 20개의 실험 데이터 집합을 선택하여 가장 높은 정확도는 아니지만 합리적으로 높은 정확도를 지니는 소형 트리를 구축할 수 있었다.

표 3 9개 속성을 가지는 20개의 실험 레코드 데이터들

Network robot	With purchase interest	Without purchase interest
0,0,0,0,0,0,0,0,2	0,0,1,2,2,1,3,0,0	0,0,0,0,0,0,0,0,0
0,0,0,2,2,1,0,1,0	0,0,0,2,2,2,3,0,0	0,2,1,0,0,1,3,0,0
0,0,0,0,0,0,0,1,0	0,0,2,1,1,1,0,0,1	0,0,0,0,0,0,3,0,0
0,3,2,2,2,1,0,1,0	0,3,2,2,2,0,1,0,0	0,0,1,1,1,1,0,0,0
1,3,1,2,2,2,0,0,0	1,0,0,0,0,0,0,0,1	0,0,0,1,1,1,0,0,0
1,2,1,2,2,2,2,1,1	0,2,2,1,1,0,2,0,0	0,3,1,0,0,0,0,0,0
	1,2,2,1,1,0,2,0,0	
	1,2,2,2,2,2,2,0,0,	

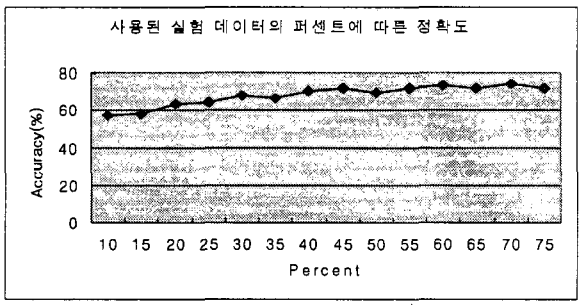

그림 1 사용된 실험 데이터의 퍼센트에 따른 정확도

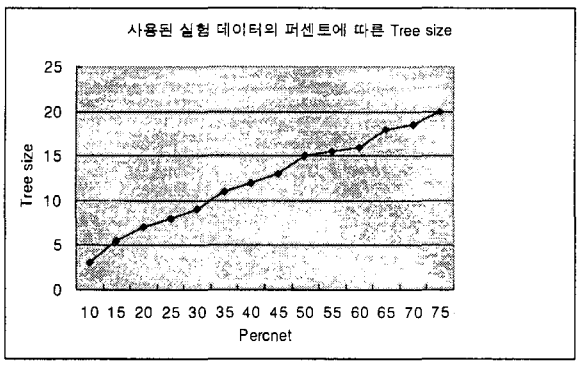

그림 2 사용된 실험 데이터의 퍼센트에 따른 Tree size

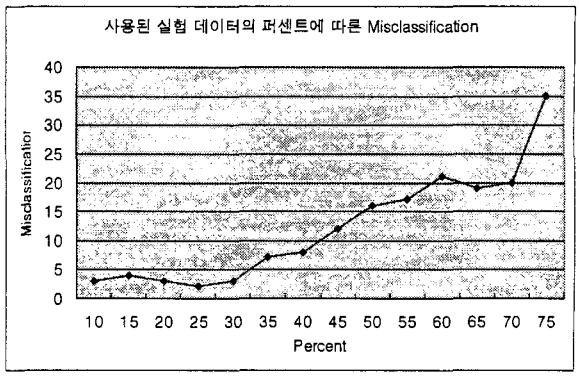

그림 3 사용된 실험 데이터의 퍼센트에 따른 Misclassification

5. 결론 및 향후 과제

본 연구에서는 분류자를 구축하기 위해서 20개의 실험 데이터 집합을 선택하여 가장 높은 정확도는 아니지만 합리적으로 높은 정확도를 가지는 소형 트리를 구축할 수 있었다.

향후 연구과제로는 구축된 소형 트리를 사용하여 고객과 관심이 있는 속성을 선택하는 향후 연구가 남아 있다.

참고 문헌

[1] M. Brohman et al., "data Completeness: A Key to Effectice Net-Based Customer Service Systems," Comm. ACM, vol.46, no.6, pp.47-51, 2003.

[2] C. X. Ling and C. Li, "data Mining for Direct Marketing: Problems and Solutions," Proc. 4th Int'l Conf. Knowledge Discovery and Data Mining(KDD 98), Am. Assoc. Artificial Intelligence, pp.73-79, 1998.

[3] L.Niu et al., "Product Hierarchy-Based Custome Profiles for Electronics Commerce Recommendation," Asian J. Information Technology, vol.4, no.1, pp.18-24, 2003.

[4] Y. Li, C. Zhang, and S. Zhang, "Cooperative Strategy Web-Based Data Cleaning," Applied Artificial Intelligence, vol.17, nos.5-6, pp.443-460, 2003.

[5] P. Resnick and H. Varian, "Recommender Systems," Comm. ACM, vol.40, no.3, pp.56-58, 1997.

[6] X. Y. Jeffrey, O. Yuming, Z. Chengqi and Z. Shichao, "Identifying Interesting Visitors Through Web Log Classification," IEEE Intelligent Systems, pp.55-59, 2005.

[7] T. Oates and D. Jensen, "The Effects of Training Set Size on Decision Tree Complexity," Proc. 14th Int'l Conf. Machine Learning(ICML 97), Morgan Kaufmann, pp.254-262, 1997.