하이퍼플래인을 이용한 웹 방문 패턴에 대한 사용자 클러스터링

이해각
순천향대학교 정보기술공학부
주영옥
정평모비컴(주) 개발부 R&D팀

A Clustering Method of Web Navigation Pattern Using the Hyperplane

Hae-Kag Lee
Division of Information Technology Engineering, Soonchunhyang University
Young-Ork Joo
R & D Dept., JungPyung Mobile & Computer Technology Co., Ltd

요 약

사용자 웹 방문 패턴 발견으로써의 사용자 클러스터링은 웹 사이트를 이용하는 사용자들의 취향과 행동방식을 얻어내는데 매우 유용하다. 또한 이러한 정보는 웹 개인화나 웹 사이트를 재구성하는데 필수적이다.

본 논문에서 사용자 웹 방문 패스를 클러스터링 하기 위한 시간적으로 효율적이며, 패스 특성을 보다 정확하게 표현하여 클러스터링 할 수 있는 알고리즘이 제안되며, 제안된 알고리즘은 패스 간의 유사도 측정을 통한 클러스터링, 하이퍼플랜을 이용한 K-평균 클러스터링의 2단계 과정으로 이루어져 있다.

1. 서론

초고속 통신망이 널리 보급되면서 인터넷은 급속도로 성장을 하였고, 인터넷에는 컨텐츠 산업의 붐을 타고 많은 기업들이 자사의 사이트를 개설하여 소비자에게 보다 직접적이고 적극적으로 마케팅을 하고 있다. 특히 수익 구조를 인터넷 매매에 갖고 있는 기업 즉, 전자상거래 기반의 기업의 경우 온라인 상의 쇼핑족들에 대해서 실구매까지 이어지도록 많은 투자와 전략을 기획하게 된다. 기업이 인터넷 상에서 사업을 확장하고 많은 투자를 함에 있어서 실질적으로 드러나지 않는 인터넷 사용자들을 어떻게 파악하고 그들의 반응과 관심분야에 대한 정보를 도출해 내는가는 매우 어려운 문제이다.

인터넷 기업의 성패를 좌우하는 중요한 변수가 되는 빠르게 변화하고 개인별로 다양한 사용자의 관심사항을 확인해 볼 수 있는 방법으로 사이트 상에서 이루어지는 그들의 행동양식을 통한 분석 방법은 매우 유용하다. 우리는 사용자 행동양식에 대한 정보를 얻기 위해 웹 방문 패턴(web navigation pattern)에 대한 분석을 요구하게 되며, 이를 위해 웹 로그에 대한 분석을 하게 된다. 웹 로그는 서버에 대한 접근에 관련된 데이터와 참고자료에 대한 데이터 등이 있는데, 그 데이터양이 접속횟수에 대하여 계속적으로 증가하므로 분석에 있어서 대용량의 처리가 필요하고, 처리에 대한 효율성을 제고해 볼 필요가 있다.

이러한 대용량의 데이터를 일반질의(Query), OLAP(*On-Line Analysis Processing*), 다차원 분석(multi-dimensional analysis) 등과 같은 기존의 조회 도구들을 이용하여 분석했을 경우 숨겨진 정보에 대한 한계를 드러낸다. 이때, 데이터 마이닝(Data Mining)은 기존의 조회도구를 보완하며 자동화된 방식으로 숨겨진 정보를 찾아 의사결정이나 정보시스템에 활용할 수 있도록 지원해 준다. 또한 얻어낼 수 있는 정보의 형태도 다양하다. 그러므로 기존 조회도구와 더불어 데이터 마이닝을 이용함으로써 보다 심도 있는 정보가 얻어질 수 있다([1],[2],[9]).

따라서 본 논문에서는 웹에 대한 데이터의 분석으로써 데이터 마이닝 기법을 적용한 웹 사용 마이닝에 대하여 살펴보며, 사용자 웹 방문 패턴 발견으로써 웹

사이트를 이용하는 사용자들의 경향과 행동분석이 이루어지도록 정보를 도출해 내기 위한 사용자 클러스터링 기법을 연구한다. 또한 사용자 클러스터링 기법을 적용함에 있어서 사용자의 웹 패스에 대한 모든 특성을 잘 반영할 수 있으며 보다 효율적인 방법을 모색하는데 중점을 두었다.

2. 본론

2-1. 전처리 과정

웹 사용자 방문에 대한 각 패스의 패턴을 알아내기 위해서 우선 로그 파일의 데이터로부터 원하는 정보만을 취하여 입력데이터로 사용한다. 로그 파일은 하나의 파일에 저장이 가능하며, 이를 일 단위, 주 단위, 월 단위 등의 원하는 기간별로 분리하여 저장할 수도 있다. 웹 서버들로부터 얻은 로그 정보 파일로부터 얻어지는 초기 데이터는 전처리되어 마이닝에 용이하도록 변환되어 재구성된다.

전처리과정은 데이터 정제, 사용자 구분, 세션구분, 세션보정, 패스보정, 형식화 단계로 나눌 수 있다([9],[13]).

2-2. 패스 각을 이용한 유사도 측정에 의한 패스 클러스터링

각각의 패스에 대해 발생하는 서브시퀀스에 대해 벡터를 생성하고 이에 대한 여부를 이용하여 패스를 분류하는 것은 시간적으로나 공간적인 문제에 있어서 매우 어려운 문제이다.

이에 Shahabi[6] 제안한 패스간의 유사도를 측정하는 방법을 사용하였고, 2-3 절에서 이러한 1단계 클러스터링의 결과로 재클러스터링을 함으로써 각 사용자 패스의 특성을 반영한 클러스터링을 하였다. 이로써 local optimal solution에 빠지는 오류를 막을 수 있고 패스 특성을 보다 정확하게 반영한 그룹핑으로써 분석에 효과를 높일 수 있도록 하였다.

두 패스의 각을 계산하기 위하여 feature vector로 사용될 서브시퀀스를 생성한다. 비교할 2개의 패스에 대한 서브패스가 feature가 되며 feature vector로는 m-length까지의 feature를 사용하게 되는데, 이때 m ≥ min(path1-length, path2-length)를 만족해야 한다. 즉, 두 패스의 길이 중 작은 것보다 크거나 같아야 한다.

1단계 클러스터링으로써 2개의 패스간의 유사성 비교는 패스사이의 각을 이용한 내적 계산(inner product)을 통하여 측정할 수 있다.

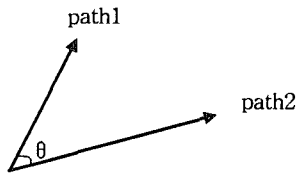

그림1. 2개의 패스간의 각 θ

$$\cos(\theta_{path1,path2}) = \frac{path1 \cdot path2}{||path1||||path2||}$$

내적 계산을 통하여 패스 행렬 결과를 얻으면 cos($\theta_{path1, path2}$)를 이루는 모든 요소는 0~1사이의 값을 갖게 된다. 여기서 0은 두 벡터가 유사성이 없는 관계를 나타내고, 1은 두 벡터가 동일함을 나타낸다.

이 중에서 cos($\theta_{path1, path2}$)에 범위를 지정하여 유사성이 높은 것들끼리 클러스터링 한다. 이를 위해 범위에 속하는 요소를 골라 유사도가 높은 순서로 우선순위를 정하고 패스 수를 1단계씩 늘려가며 가능한 패스집합을 만들고, 우선순위에 의해 정한다.

그룹이 형성되면 그룹별로 해시체이닝 기법을 사용하여 저장한다. 이것은 2단계 클러스터링을 위한 저장을 위해서이다.

자질 벡터의 수가 상당수로 많아지기 때문에 이것을 한번에 관리하기란 용이하지 않다. 또한 2단계 클러스터링 방법은 하이퍼플랜을 이용한 방법으로써 각 그룹 중심값과의 근접도를 보고 가장 가까운 그룹에 그룹핑 하는 작업을 중심값의 차이가 0에 가까울 때까지 반복하기 때문에 레코드들이 많은 수의 그룹간의 이동이 이루어지게 된다. 따라서 그룹이 이동될 때마다 그룹의 중심값을 계산하여주기 위해 그룹별로 저장하고 그룹별로 저장된 데이터를 관리해 주는 방법을 사용한다. 따라서 그룹의 중심값 계산에 대한 반복적인 계산을 피할 수 있다.

각 그룹에 포함되는 자질 벡터의 수는 그룹내의 레코드들은 높은 유사도를 보이기 때문에 임의로 그룹핑했을 경우보다 상대적으로 낮다.

따라서, 그룹별 분할 저장함으로써 저장공간의 효율을 높일 수 있으며 해시 체이닝 기법을 사용하기 때문에 빠른 탐색시간을 확보할 수 있는 장점이 있다.

2-3. Hyperplane을 이용한 클러스터링

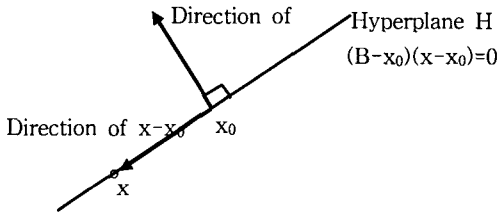

그림2. 하이퍼플랜의 정의

하이퍼플랜은 방정식 $\sum_{j=1}^{n} p_j x_j = k$ 를 만족하는 점들 $x = (x_1, x_2, x_3, \ldots x_n)$로 구성되어 있다. 여기서 p는 E^n안의 0이 아닌 벡터이고 k는 스칼라이다. 또한 p는 하이퍼플랜과 수직을 이룬다. 만약 $x_0 \in H$이면 $px_0 = k$이고 $x \in H$인 어떤 x에 대해서 $px = k$를 취한다. 즉, 하이퍼플랜은 $p(x - x_0) = 0$을 만족시키는 점들의 집합이라고 표현할 수 있으며 여기서 x_0는 하이퍼플랜 상의 고정된 점이다.

하이퍼플랜은 E^n공간을 2개의 영역으로 나누며 이것을 halfspaces라고 부른다. 따라서 x가 어느 영역에 속하게 되는지는 다음과 같이 결정한다. {x : $p(x-x0) \geq 0$}이면 하이퍼플랜을 중심으로 p방향 쪽의 영역에 속하게 되고, {x : $p(x-x0) \leq 0$}이면 p와 반대 방향 쪽의 영역에 속하게 된다.

하이퍼플랜은 하나의 그룹의 중심점 A와 또다른 그룹의 중심점 B의 중간지점인 x_0를 수직으로 지나며 $(B-x_0)(x-x_0) = 0$을 만족한다[그림2]. 즉, 하이퍼플랜은 중심점 A와 중심점 B를 중심으로 공간을 이등분하게 된다. 예를 들어 하이퍼플랜에 의한 클러스터링은 한 개체가 두 개의 중심값에 대하여 $(B-x_0)(x-x_0) \geq 0$를 만족하면 B그룹에, $(B-x_0)(x-x_0) \leq 0$를 만족하면 A그룹에 속하게 되는 방법이다[그림3].

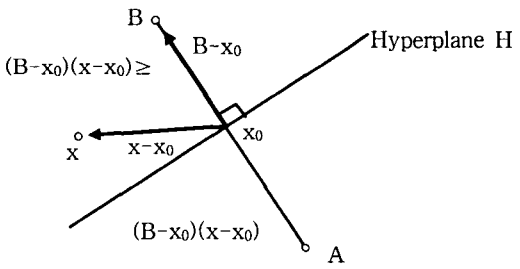

그림3. 두 지점 A와 B사이의 하이퍼플랜에 의한 임의의 값 x의 그룹 결정

하이퍼플랜을 이용하면 2개의 그룹 중심값을 기준으로 $(B-x_0)(x-x_0)$ 값을 계산하여 해당 개체의 2개의 그룹 중심값과의 근접도를 결정할 수가 있다. 이때 모든 가능한 그룹 쌍의 조합에 대한 하이퍼플랜에 의한 그룹 결정이 이루어지는 것이 아니라 비교하여 우위인 것과 비교하지 않은 나머지 그룹의 중심값과의 하이퍼플랜에 의한 그룹 결정을 함으로써 비교횟수를 줄일 수 있다.

유사도 측정에 의한 패스의 클러스터링 후 K-평균 클러스터링을 통한 2차 패스 클러스터링을 하였다. 일반적으로 K-평균 클러스터링 방법은 많은 횟수의 계산과 비교를 하게 된다. 이러한 단점을 줄이는 방법으로 하이퍼플랜을 이용한 방법을 사용하게 되었다.

클러스터링은 데이터를 유사한 특징을 가진 몇 개의 집단으로 그룹화하여 분할하는 것을 말한다. 비유사성의 척도는 연속형 변수의 경우에는 거리가 되며, 범주형 변수의 경우에는 불일치된다.

K-평균 클러스터링은 데이터 마이닝의 클러스터링 작업의 대표적인 기법이다. 여러 클러스터링 방법 중에서 대용량 데이터를 빠르게 처리할 수 있으며, 그 알고리즘도 비교적 간단하기 때문이다. 이 방법은 관찰치들 사이의 거리를 이용해 주어진 기준을 최적화하도록 구현되므로 최적 분리 클러스터링 방법이라고도 한다. 계보적 클러스터링 방법의 단점을 극복할 수 있고, 관찰치의 수가 많을 때 주로 이용하므로, 데이터 마이닝에 유용한 방법이라고 할 수 있다.

하이퍼플랜을 이용한 클러스터링 과정은 다음과 같다.

① 몇 개의 그룹을 생성할 것인지 k를 결정.
② 각 그룹의 중심 값으로 임의의 값을 할당.
③ 2개로 이루어진 그룹 쌍(예를 들어, A, B)의 하이퍼플랜과 수직으로 교차하는 x_0를 구한다.
④ 각각의 데이터 x에 대해 $B-x_0$ 벡터와 $x-x_0$벡터의 내적을 구하여 각 그룹 쌍에서 소속 그룹을 결정, 전체 그룹 쌍에 대하여 반복
$(B-x_0)\ (x-x_0) > 0 \rightarrow B$
$(B-x_0)\ (x-x_0) \leq 0 \rightarrow A$
⑤ 가장 많이 포함되는 그룹으로 소속 그룹 결정
⑥ 각 그룹에 속하는 모든 데이터에 대한 평균값을 구해 새로운 중심 값으로 결정
⑦ 기존의 중심 값과 새로운 중심 값의 차이 계산
⑧ 중심 값의 차이가 0에 근접할 때까지 ③부터 반복

2-4. 실험

하이퍼플랜의 효율성을 검증하기 위하여 K-평균 클러스터링 중 대표적인 거리에 정의로써 유클리드 거리와 맨하탄 거리를 이용한 클러스터링과 하이퍼플

랜을 이용한 클러스터링을 비교 실험하였다.

그림4와 5는 그 실험 결과 중 4000개의 임의의 4차원 데이터를 적용할 경우 클러스터링에 소요되는 loop 회수와 전체 소요 시간을 비교해 본 그래프이며 기존의 방법보다 매우 우수한 성능을 보임을 알 수 있다.

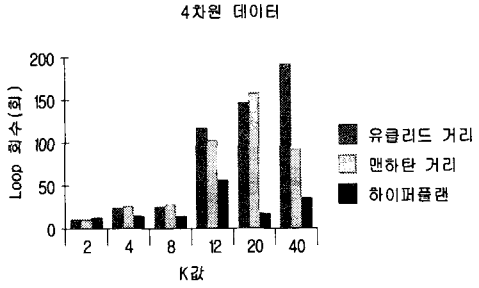

그림4. 4차원 데이터 Loop 회수

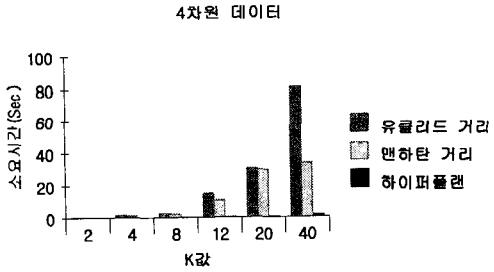

그림5. 4차원 데이터 전체 소요시간

3. 결론

웹 방문 패턴을 인식하기 위하여 각 사용자의 세션별로 방문 패스에 대한 서브시퀀스를 형성하여 이를 feature vector로 사용하였다. 이러한 방법으로써 방문 패스에 내재되어 있는 고유의 순차적인 흐름과 부분적인 특성을 파악할 수 있었다. 또한 첫 번째 단계로써, 두 패스간의 각을 이용하여 유사도 측정을 하고 이를 클러스터링 함으로써 원하는 해에 근접한 초기 해를 얻게 되며 두 번째 단계로써, 보다 정확한 그룹핑을 위하여 하이퍼플랜을 이용한 클러스터링을 하였다.

그 외에도 실험을 통하여 하이퍼플랜을 이용한 클러스터링은 유클리드 거리나 맨하탄 거리를 사용하는 K-평균 클러스터링 기법과 비교했을 때 계산과 비교 회수를 줄여 전체적으로 시간적인 우위를 나타냄을 확인하였다. 또한 K-평균 클러스터링 기법에서는 초기값에 의하여 많은 차이를 보이는 그룹핑이 된다. 때로는 그룹내 큰 편차를 갖는 좋지 못한 클러스터링이 되기도 한다. 따라서 첫 번째 단계에서 각에 의한 유사도에 의한 클러스터링한 결과를 하이퍼플랜을 이용한 클러스터링에 대한 초기값으로 사용하기 때문에 K-평균 클러스터링에 좋지 못한 초기값이 주어졌을 때 그룹 내 큰 편차를 갖는 잘못된 클러스터링 결과가 나타나는 단점을 막을 수 있다.

[참고문헌]

[1] 장남식, 홍성완, 장재호. "(성공적인 지식경영을 위한 핵심정보기술) 데이터 마이닝", 대청

[2] 조재희, 박성진, "데이터 웨어하우징과 OLAP", 대청

[3] Yan Wang(2. 2000), "Web Mining Knowledge Discovery of Usage Patterns"

[4] www.biznet.com

[5] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan "Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data"

[6] Cyrus Shahabi, Amir M. Zarkesh, Jafar Adibi, and Vishal Shah "Knowledge Discovery from Users Web-Page Navigation", RIDE, 1997

[7] Stephen Misener, Ph.D., Roland Somogyi, Ph.D. "Exploration, Inference, Prediction: Making Sense of Large-Scale Gene Expression Data", Molecular Mining Corporation

[8] SANJAY MADRIA, SOURAVS BHOWMICK, W. -K NG, E. P. LIM, "Research Issues in Web Data Mining"

[9] 마이크로 소프트웨어 2001. 5, "새로 쓰는 데이터 마이닝 이야기", 소프트뱅크 미디어

[10] 김종달, "웹 로그에서 웹 방문 패턴을 이용한 사용자 웹 방문 패스 클러스터링"

[11] Erica Kolatch, "Clustering Algorithms for Spatial Databases: A Survey"

[12] Mokhtar S. Bazaraa, John J. Jarvis, "LINEAR PROGRAMMING AND NETWORK FLOWS", JOHN WILEY & SONS

[13] Bamshad Mobasher, Robert Cooley, Jaideep Srivastava, "Automatic Personalization Based on Web Usage Mining"