# 클러스터의 점유시간을 이용한 웹 페이지 추천 기법

신형섭[0] 이충세
충북대학교 전자계산학과
shinbeauty[0]@hanmail.net csrhee@cbucc.chungbuk.ac.kr

# Web Page Recommendation Using Percentage Of The Time In The Cluster

Hyung-Sup Shin[0] Chung-Sei Rhee
Dept. of Computer Science, Chung-Buk National Univ.

**요 약**

기존의 여러 동적 추천 시스템에서는 웹 페이지들 사이의 유사도와 로그 파일 안에 들어 있는 사용자들의 패턴을 이용하였기 때문에 연관된 페이지 뿐 아니라 단순히 순차적으로 연결되는 문서를 추천 페이지로 제공할 수 있었다.

본 논문에서는 기존의 방식에 각 페이지가 점유하는 시간의 분석을 더하려 한다. Data를 여러 분야로 나눌 수 있는 전자상거래의 특성을 이용하여 개개의 클러스터로 분류된 사이트들의 로그파일을 분석하여 점유시간의 크기와 무의미하게 보내어 지는 시간을 가중치를 주어 구별해내는 결과를 바탕으로 사용자가 주로 방문하는 연관성이 높다고 판단되는 웹 페이지를 추천하는 방법을 제안한다.

## 1. 서 론

컴퓨터 정보 통신 기술이 발달하면 할수록 정보의 양이 급증을 한다. 인터넷 상에서의 정제되지 않은 데이터에는 사용자들이 필요로 하는 정보도 있지만 필요로 하지 않은 정보가 무수히 많다. 필요한 정보가 많아지면 많아질수록 필요하지 않은 정보 또한 많아진다. 추천 시스템은 이러한 정보 속에서 사용자가 필요로 하는 정보를 데이터 마이닝 기법이나 다른 여러 가지 기법들을 사용하여 원하는 정보만을 찾아주는 것이다.

실시간으로 많은 양의 문서를 처리해야 하는 전자상거래나 쇼핑몰 시스템의 경우 에이전트 기법을 통해 사용자들에 좀 더 쉽게 원하는 정보에 접근할 수 있게 서비스를 제공한다. 어떤 사이트에서 사용자들이 적은 노력으로 자신이 원하는 정보를 손쉽게 얻을 수 있다면 자주 찾아 올 수 있을 것이다.

기존의 추천 시스템은 페이지를 단위로 사용자들의 경로를 연관규칙을 통해 분석하는 방식을 많이 사용해왔다. 대부분의 기업과 웹사이트에 있어서 기존의 데이터베이스와는 달리 사용자들의 동선으로 인한 무작위 데이터가 생성되는데 이러한 데이터는 상당한 가치를 지니고 있다. 이런 동선의 데이터를 분석하면 사용자들의 요구사항을 파악할 수 있다. 그렇지만 동선은 단순히 항해(연결)만 해주는 페이지와 내용이 들어있는 페이지를 구별하지 못한다. 그렇기 때문에 원하는 정보와 원하지 않는 정보의 구별을 용이하게 할 수 없다.

본 논문에서는 각 분야별로 클러스터링이 되는 전자상거래의 특성을 이용하고 각 클러스터에서 머무는 시간 측정과 점유시간에 따른 페이지 구별을 통해 사용자가 원하는 페이지를 추천하는 시스템을 제안한다.

## 2. 관련연구

### 2.1 웹 로그

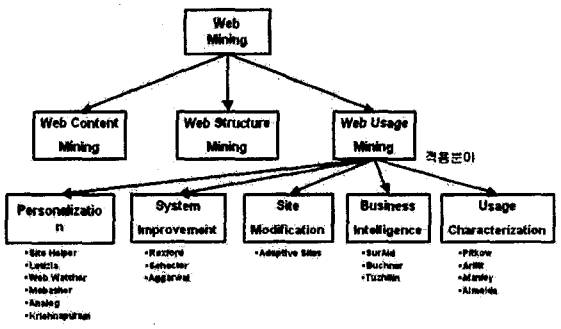
웹 서버는 사용자들의 접근을 로그파일에 기록한다. 이것은 단순히 접속이라는 데이터로, 모든 방문객과 모든 페이지에 대해 기록한다. 알 수 없는 방문객들로 인해 사이트는 움직이고 있고 웹 로그는 그러한 방문객의 위치와 ISP, 방문페이지의 종류, 시간, 횟수 등을 일목요연하게 기록한다.

이러한 로그 파일을 분석하게 되면 홈페이지에 방문한 방문자의 수와 방문자들이 접속되어 있는 서버의 도메인 또는 IP를 추적하는 과정을 통해 방문자들의 유형을 판

단할 수 있다. 또한 각 웹 페이지별로 얼마나 많이 요구되었는지 하루의 몇 시간대에 가장 방문자가 많은지 또는 요일, 계절 등을 알 수 있다.

2.2 웹 마이닝

Web Data로부터 사용 패턴을 발견하기 위한 Data Mining 기법으로 협업 필터링(Collaborative Filtering), 사례기반추론(Case-Based Reasoning), 군집분석(Clustering), 인공신경망(Newral Network), 유전적 알고리즘(Genetic Algorithm), 퍼지이론(Fuzzy Theory)등과 같은 알고리즘이 적용된다. 웹 마이닝의 적용분야에는 [그림 1]과 같은 분야에 적용되고 있다.[1]




[그림 1] Web Mining Concept Hierarchy

본 논문에서는 인터넷 쇼핑몰의 특성을 이용하여 클러스터링 기법을 사용하려 한다. 클러스터링은 주어진 관찰치 중에 유사한 몇몇의 집단으로 그룹화 하여 각 집단의 성격을 파악할 수 있다. 클러스터링의 방법 중 k-평균 군집방법은 대용량 데이터를 빠르게 처리할 수 있으며 간단한 알고리즘으로 구성되어 있어 가장 많이 사용되고 있다.[2]

## 3. 클러스터 단위의 시간측정 개념

이 기법의 기본적인 개념은 사용자가 지나온 경로를 되짚어 가면서, 사용자가 원하는 정보를 얻기 위해 지나온 페이지를 순서대로 기록하여 각 분야별로 나뉘어진 클러스터의 총 점유시간을 비교하여 다음 접속 시에 첫 페이지로 추천하는 것이다.

[그림2]를 보면, A페이지를 통해 접속을 시도하는 모든 사용자는 B, C, D등의 선택을 할 수 있다. 이러한 선택의 페이지가 사이트가 커지면 커질수록 많아지고 복잡하게 된다. B를 선택했을 경우 하위 F, G로 갈 수 있다. 따라서 B, F, G를 하나로 묶었을 경우와 묶지 않았을 경우는 다른 결과를 낳게 된다. 기존의 추천 기법들을 살펴보면 각 페이지를 단위로 Count횟수를 체크하기 때문에 연관성이 깊은 B, F, G를 특별한 연관성의 알고리즘으로 묶으려 하기 때문에 알고리즘이 어렵고 복잡해진다.

[그림 2] 사용자 경로의 예

인터넷 쇼핑몰의 특성상 상호 배반적 계층적 군집으로 나뉘게 되는 [그림 2]같은 군집은 $A_0$(B, F, G), $A_1$(C, H, I), $A_2$(D, J, K)로 분류 될 수 있다.

| 접속사이트 | 접속횟수 | 방문횟수 | 데이터량 | 방문소요시간 |
|---|---|---|---|---|
| A | 266 | 44 | 1079157 | 31701 |
| B | 234 | 1 | 750943 | 67 |
| C | 96 | 1 | 652776 | 2181 |
| D | 85 | 7 | 547080 | 435 |
| F | 81 | 8 | 53901 | 2640 |
| G | 65 | 3 | 177108 | 1769 |
| H | 48 | 3 | 719251 | 1276 |
| I | 47 | 6 | 88866 | 354 |
| J | 45 | 2 | 545706 | 1088 |
| K | 41 | 1 | 420101 | 188 |
| | 2270 | 178 | 57316001 | 62306 |

[그림 3] 방문객 현황 DB

기존의 많은 웹 로그 분석 툴을 이용하여 웹 로그 분석하면 [그림 3]에서와 같이 각 페이지별 접속 횟수, 방문횟수, 데이터 량 소요시간 등의 새로운 데이터를 알아낼 수 있다.

여기서 A사이트는 초기 접속 페이지이다. B, C, D는 A페이지에서 들어갈 수 있는 항해하는 페이지이고, F, G는 A→B→F or G로 접속이 가능하다. 여기서 B와 F, G는 유사페이지로 묶을 수 있다. 또한 C, H, I와 D, J, K도 유사 페이지로 묶을 수 있다.

이렇게 묶인 유사페이지는 사용자의 관심분야로 해석을 해도 적합하다. 따라서 B, F, G를 $A_0$, C, H, I를 $A_1$, D, J, K를 $A_3$로 표현 할 수 있다. $A_0$, $A_1$, $A_2$의 총 점유시간을 비교하면 관심분야에 접근할 수 있다.[3]

### 4. 추천 알고리즘

추천 알고리즘은 크게 사용자의 페이지를 점유하는 시간의 측정으로 우선순위를 주는 것과 단순한 항해 페이지를 구별하는 기능의 두 가지로 구성이 된다. 추천 알고리즘은 [알고리즘 1]과 같다.

페이지의 점유 시간이 최소 점유 시간보다 작을 경우 이 페이지는 항해하는 페이지로 인식될 수 있다. 항해하는 페이지는 [그림 2]에서 A→B→F로의 경우 B는 A에서 F로 가기 위한 항해 페이지이기 때문에 접속 소요 시간은 짧지만 방문횟수는 다른 페이지에 비해 많다. 이런 페이지를 항해 페이지로 인식할 수 있다. 따라서 $A_0$(4476)의 점유 시간의 합은 $A_1$(3811), $A_2$(1621)보다 크게 된다. 여기서 $A_1$의 C의 시간 점유를 보게 되면 항해 페이지의 시간 점유가 높게 나타난다. 항해 페이지에서의 무의미한 시간 점유로 인한 오류를 신뢰도의 가중치를 페이지에 주어 체크 할 수 있다.

```
Input : cur_cluster : 현재 사용자 클러스터
        last_url :사용자가 가장 최근에 요청한 URL
        δ : 최소 점유시간, α : 최소 신뢰도

Output : Recommend1 : 추천 문서 집합 1

Recommend1 = ∅ ;
{
for each T do   // T는 cur_cluster를 포함하는Time
  if ( 점유시간(T) ≥ δ )
     confidence = 신뢰도(cluster⇒url) ;
        // url 은 추천 웹 문서
     if ( confidence ≥ α )
     {
       url.score = confidence ;
       Recommend1 += url ;
         // 추천하며 추천 url을 반환(신뢰도(cluster⇒url) 포함)
     }

}
```

[알고리즘 1] 추천 알고리즘

또한 점유 시간이 최소 점유시간보다 클 경우는 다음 클러스터와 비교하여 최대치의 점유시간을 찾는다. 추천 집합의 순위를 결정하는 가중치로 시간 신뢰도를 사용한다. 신뢰도(cluster⇒url)는 최소 신뢰도를 만족하는 url들만을 추천집합에 포함시킨다.[8] 이렇게 구해진 신뢰도를 만족하는Recommend1은 다음 접속 시에 우선권을 가질 수 있다.

### 5. 결론 및 향후과제

본 논문에서는 사용자의 접속과 동적 패턴을 방문 횟수보다 방문해서 머무는 소요 시간을 통해 관심분야를 분석, 추천하는 기법을 설명하였다. 기존의 웹 페이지 추천 시스템에 대한 연구는 인터넷 상에서의 웹 페이지를 추천하는 것에 대한 연구가 주류를 이루었지만 본 논문은 한 사이트 내에서의 웹 페이지 추천을 다루었다.

이윤을 추구하는 인터넷 쇼핑몰이나 전자상거래에 있어서 사용자를 한번 들른 사이트에 다시 접속하게 하는 것은 가장 중요한 일 중의 하나이다. 이를 위해 사용자가 원하는 정보를 보다 쉽게 접근하고 얻을 수 있게 하는 것이 중요하다. 이러한 추천 시스템은 사용자에게 많은 이점을 줄 수 있다. 그렇지만 실제의 인터넷 쇼핑몰이나 전자상거래를 위한 시스템이기 때문에 이러한 시스템의 성능을 분석하고 평가하는 것은 현실적으로 많은 시간과 어려움이 있다. 앞으로 지속적인 성능평가를 통해 알고리즘을 발전시키고 최적의 방법을 제시할 것이다. 실제의 데이터를 적용해서 실제 사이트에 적용하는 것이 필요하다.

### 6. 참고문헌


[1] R. Cooley, et. al, "Data Preparation for Mining World Wide Web Browsing Patterns," Knowledge and Information Systems, Vol.1-1, 1999.

[2] 정영미, *정보검색론*, 구미무역 출판부, 1993.

[3] E.H. Han, et. al, "Clustering Based On Association Rule Hypergraphs," DMKD, 1997.

[4] Gabriela Polcicova, "Recommending HTML-documents using Feaure Guided Automated Collaborative Filtering," ACM SIGIR '99, August 1999.

[5] 문홍기, 이수원, "전자 상거래 에이전트를 위한 연관 규칙 발견 및 확장," 1999년 춘계정보과학회, August 1999.

[6] Sanford Gayle, "The Marriage of Market Basket Analysis to Predictive Modeling," The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2000.

[7] R. Agrawal and R. Srikant, "Fast Algorithm for Mining Association Rules," Proc. of the 20th VLDB Conference, pp. 487-499, 1994.

[8] 김진수, 김태용, 이정현, "웹 문서 형식과 클러스터 내의 문서 유사도를 이용한 동적 추천 시스템", 2001년 한국정보과학회 춘계 학술 발표회2001.4

[9] 강귀영, 조동섭, "사용자 프로파일을 이용한 웹페이지 추천" 2001년 한국정보과학회 춘계 학술 발표회 2001.4

[10] 이은영, 조동섭, "개인화 상품 추천을 위한 협력 필터링 에이전트" 2001년 한국정보과학회 춘계 학술 발표회2001.4

[11] Y.-G. Chong and S.-B. Cho, "A structure analysis agent for extraction, storage and visualization of web sites," Proc. Korea Information Science Society (B), Suwon, April 2001.