연관 웹 문서 분류와 사용자 브라우징 패턴을 이용한 동적 링킹 시스템

박영규*, 김진수*, 김태용**, 이정현*
*인하대학교 전자계산공학과, **문경대학 컴퓨터정보과
e-mail : ykpark@nlsun.inha.ac.kr

Dynamic Linking System Using Related Web Documents Classification and Users' Browsing Patterns

Young-Kyu Park*, Jin-Su Kim*, Tae-Yong Kim**, Jung-Hyun Lee*
*Dept. of Computer Science and Engineering, Inha University
**Dept. of Computer Information, Mun-Kyung College

요약

웹사이트 설계자의 주관적 판단에 의한 정적 하이퍼텍스트 링킹은 모든 사용자들에게 동일한 링크를 제공한다는 단점을 가지고 있다. 이러한 문제점을 개선하고, 각 사용자들의 브라우징 패턴에 적합한 웹 문서들을 동적 링크로 제공해주기 위한 여러 동적 링킹 시스템들이 제안되었다. 그러나 대부분의 동적 링킹 시스템들은 사용자의 현재 브라우징 패턴과 가장 유사한 패턴 정보만을 이용해 동적 링크를 제공하기 때문에 연관성이 없는 웹 문서들에 대한 링크를 수시로 제공한다는 또 다른 문제를 지니고 있다.

본 논문에서는 데이터 마이닝의 한 응용 분야인 웹 마이닝 기법을 이용하여 웹 서버의 로그파일로부터 사용자들의 브라우징 패턴을 분석해내고, 다차원 데이터 집합에 적합한 Association Rule Hypergraph Partitioning(ARHP) 알고리즘을 이용하여 서로 연관성이 있는 웹 문서들을 분류한다. 사용자 브라우징 패턴 정보로부터 사용자에게 추천해줄 1차 링크 집합을 생성하고, 연관 웹 문서 정보를 이용하여 2차 링크 집합을 생성한다. 그리고 두 링크 집합에 공통으로 포함된 링크 집합만을 사용자에게 동적으로 추천해줌으로써 사용자가 보다 편리하고 정확하게 웹사이트를 브라우징할 수 있도록 하는 동적 링킹 시스템을 제안한다.

1. 서론

대부분의 웹사이트들은 설계자의 주관적인 판단에 의해 사용자들의 브라우징 패턴과 무관한 획일적인 하이퍼텍스트 링크만을 제공해왔다. 이러한 문제를 극복하고 각각의 사용자들의 브라우징 패턴과 적합한 하이퍼텍스트 링크를 동적으로 제공하기 위해서는 사용자들의 웹 서버 액세스 기록이 저장되어있는 웹 서버 로그파일에 대한 분석이 선행되어야 한다. 그러나, 사용자 브라우징 패턴을 발견하여 동적 링크를 제공하고자하는 기존의 동적 링킹 시스템(예: MIT의 Letizia)[7]은 웹 문서들의 연관성에 대한 고려 없이 브라우징 패턴 정보 가운데 가장 유사한 패턴으로부터 동적 링크를 제공한다. 따라서 웹 사이트 설계자의 잘못된 링킹에 의해 사용자들이 브라우징을 효과적으로 하지 못할 때, 사용자 브라우징 패턴이 왜곡되는 현상이 발생하게 되므로, 사용자가 찾고자 하는 문서와 서로 연관성이 없는 문서들을 동적 링크로 제공하는 문제점을 안고 있다.

따라서 본 논문에서는 웹 마이닝 기법을 이용하여 사용자들의 브라우징 패턴을 분석하고, Association Rule Hypergraph Partitioning (ARHP) 알고리즘 [3][5][6]을 이용하여 서로 연관성이 있는 웹 문서들을 분류한다. 그리고 두 정보로부터 각각의 1차링크 집합과 2차 링크 집합을 생성한 후, 두 링크 집합에 공통으로 포함된 링크들을 가중치와 함께 동적으로 제공하는 동적 링킹 시스템을 제안한다.

2. 관련 연구

2.1 웹 마이닝

본 논문에서 채택한 기법은 웹 사용 정보 마이닝으로 하나 이상의 웹 서버에서 이루어지는 사용자들의 브라우징 패턴에 대한 마이닝 기법이다. 웹 사용 정보 마이닝을 위해서 주로 사용되는 입력 자료는 웹 서버의 로그파일이다[4]. 로그파일의 포맷은 여러 가지가 있지만, 일반적으로 로그파일에는 접속한 사용자의 IP 주소, 요청 시각, 요청 방법, 요청한 웹

문서의 URL 등이 기록되어 있다. 로그파일에 대한 분석 도구들은 많이 존재하지만, 대부분 통계적인 분석에 그치고 있을 뿐, 사용자들의 브라우징 패턴에 대한 분석은 고려하지 않고 있다.

2.2 연관 규칙 (Association Rule)

데이터 마이닝 알고리즘 가운데 연관 규칙 알고리즘이 가장 많이 이용되고 있다. 연관 규칙은 지식 표현의 하나로 항목집합 사이의 의존 관계를 나타낸다. 연관 규칙은 $X \Rightarrow Y$로 표현을 할 수 있는데, 여기서 X와 Y는 항목들(Items)의 집합이다. 이와 같은 연관 규칙은 "X를 포함하고 있는 데이터베이스 내의 트랜잭션은 Y도 함께 포함하고 있다"는 것을 의미한다. $\{X, Y\}$라는 항목집합의 지지도(Support 또는 Coverage)는 통계적 중요성을 반영하는 것으로 식 (1)과 같이 결정된다.

$$\text{지지도} = \frac{X\text{와 } Y\text{의 모든 항목들을 포함하는 트랜잭션 수}}{\text{데이터베이스 내의 전체 트랜잭션 수}} \quad (1)$$

그리고, 연관 규칙 $X \Rightarrow Y$의 결합 정도를 측정하는 신뢰도(Confidence 또는 Accuracy)는 식 (2)와 같이 결정된다[1][2].

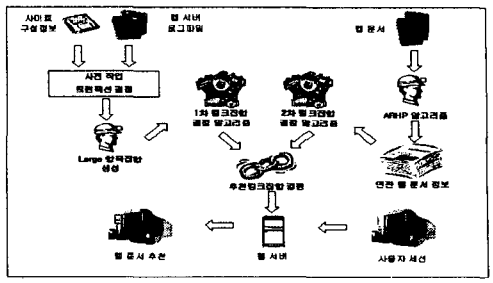
$$\text{신뢰도} = \frac{X\text{와 } Y\text{의 모든 항목을 포함하는 트랜잭션 수}}{X\text{의 항목을 포함하고 있는 트랜잭션 수}} \quad (2)$$

2.3 Association Rule Hypergraph Partitioning

ARHP 알고리즘은 Hypergraph Partitioning을 이용하여 연관 규칙에 나타나는 항목들을 클러스터링 하는 방법이다. Hypergraph $H=\{V, E\}$는 vertex들의 집합 V와 hyperedge들의 집합 E로 구성된다. Hypergraph는 각 hyperedge가 하나 이상의 vertex에 연결될 수 있다는 점에서 그래프의 확장으로 볼 수 있다. Hypergraph Partitioning 알고리즘은 항목들간 거리가 아닌 다른 가중치로 클러스터링하기 때문에 다차원 데이터 집합에 대한 클러스터링에 유용하다. ARHP 알고리즘은 클러스터링하기 위한 항목집합들의 모든 연관 규칙과 신뢰도를 구한 후, 연관 규칙에 포함되는 항목을 vertex로, 연관 관계를 hyperedge로 매핑한다. 그리고 신뢰도를 Hypergraph Partitioning을 위한 가중치로 하여, 항목들의 클러스터를 구한다[3][6].

3. 동적 링킹 시스템

아래 [그림 1]은 본 논문에서 제안하는 동적 링킹 시스템에 대한 간략한 개념도이다.



[그림 1] 전체 시스템

3.1 사용자 브라우징 패턴 분석

(1) 사전 작업

사용자 브라우징 패턴 정보를 추출해내기 위해서는 먼저 Raw Data인 로그파일에 대해 사전 작업을 해야 하는데, 요청된 파일의 확장자가 ".htm", ".html"인 파일을 제외한 모든 기록과 사용자의 IP 주소, 요청 시각, 요청 URL 필드를 제외한 모든 필드를 로그파일로부터 제거한다. 다음 단계인 트랜잭션 결정을 위한 항목 $I=\{IP, TIME, URL\}$을 생성한다. 그리고 사용자들이 사용하는 브라우저의 로컬 캐쉬와 프락시 서버의 캐쉬에 의해 요청되지 않은 기록을 복원하기 위해서 사이트 구성 정보가 필요하다. 그리고 각각의 고유한 IP 주소를 하나의 사용자로 간주한다. 본 논문에서는 제약사항으로서 모든 사용자들이 프락시 서버를 사용하지 않고, 고정 할당 IP 주소를 사용하고 있다고 가정한다[4].

(2) 트랜잭션 결정

사용자들로부터 요청된 웹 문서들 사이의 연관 규칙을 발견하기 위해서는 요청된 웹 문서들을 항목으로 하는 트랜잭션이 결정되어야 한다. 본 논문에서는 한 사용자가 한 번의 방문동안 요청한 웹 문서들에 대응되는 항목 I 들의 집합으로 구성한다. 사용자들의 사이트 방문의 시작과 끝(세션)을 결정하기 위한 여러 가지 방법이 있으나, 본 논문에서는 대부분의 상용 제품이 사용하고 있는, 시간의 임계값 30분을 사용한다[4]. 즉 30분 동안 요청 기록이 없으면, 세션의 종료로 가정한다. 트랜잭션 T는 다음과 같이 정의된다.

$$T=[\{IP, TIME_1, URL_1\}, \{IP, TIME_2, URL_2\}, \ldots, \{IP, TIME_n, URL_n\}]$$

여기서 하나의 트랜잭션 내의 모든 IP는 동일해야 하고, $TIME_{i+1} - TIME_i$ 은 30분 미만이어야 한다. 만일 30분 이상이 되면 새로운 세션의 시작으로 간주하여, 새로운 트랜잭션을 생성한다.

(3) Large 항목집합 생성

트랜잭션들이 결정되면, IBM Almaden 연구소에서

개발한 Apriori 알고리즘[1]을 이용하여 위 트랜잭션 가운데 URL만을 항목으로 하는 Large 항목집합을 결정한다. Large 항목집합이란 식 (1)에 의해 사전 정의된 최소 지지도(Minimum Support)를 만족하는 항목집합을 의미하는데, Large 항목집합에 포함된 URL들은 "최소 지지도를 보장하면서 사용자들이 한 번의 방문(세션)동안 동시에 요청한 웹 문서들"이라는 의미를 지니는 사용자 브라우징 패턴이 된다.

3.2 ARHP 알고리즘에 의한 연관 웹 문서 분류

먼저 웹 문서들로부터 HTML 태그를 제거하고 명사들을 추출한다. 이를 위해 본 연구실에서 개발한 형태소 분석기를 사용한다. 명사들로부터 문서들간의 특징을 반영하지 않는 단어들을 제거한 후, 하나의 문서에서 출현하는 단어들을 하나의 트랜잭션으로 결정한다. 그리고 Apriori 알고리즘에 의해 단어들에 대한 Large 항목집합을 구한다. Large 항목집합 사이의 연관 규칙과 신뢰도를 구한 후, 이 신뢰도를 가중치로 하여 Hypergraph Partitioning에 의해 클러스터링 하는 ARHP 알고리즘을 적용함으로써 단어들의 클러스터를 구한다. 그리고 각 웹 문서와 각 단어 클러스터에 존재하는 단어들에 대한 TFIDF 계산을 한 후, 그 값의 평균치가 가장 큰 단어 클러스터에 문서를 할당함으로써 연관 웹 문서들을 분류한다.

3.3 추천 링크 결정 알고리즘

(1) 1차 링크 집합 결정 알고리즘

아래 [알고리즘 1]은 사용자 브라우징 패턴 정보로부터 사용자에게 추천해줄 1차 링크 집합을 결정하는 알고리즘이다. 세션 내의 모든 URL을 사용하게 되면, noise가 너무 많이 들어가게 되므로, 세션 내의 모든 URL 중 최소 지지도를 만족하지 않는 URL을 제거한 후, cur_session의 크기를 Large 항목집합의 평균 크기로 한다. cur_session을 모두 포함하면서 크기가 1이 더 큰 Large 항목 집합이 존재한다면, cur_session과 조건에 만족하는 각 Large 항목집합의 마지막 URL을 이용하여 신뢰도를 측정한다. 만일 존재하지 않는다면, cur_session이 아닌 사용자가 가장 최근에 방문한 URL을 포함하는 크기 2인 Large 항목집합을 이용하여 신뢰도를 측정한다. 신뢰도(session⇒url)는 session과 url의 결합 정도를 반영하는 측정치로 식 (1)과 식 (2)를 이용하여 다음 식 (3)과 같이 결정될 수 있다.

$$\text{신뢰도}(session \Rightarrow url) = \frac{session\text{과 } url\text{을 모두 포함하는 항목 집합의 지지도}}{session \text{ 항목 집합의 지지도}} \quad (3)$$

[알고리즘 1] 브라우징 패턴 정보에 대한 1차 링크집합 결정 알고리즘

```
cur_session : 현재 사용자의 세션
last_url : 현재 사용자가 가장 최근에 요청한 URL
1차 링크 집합 = ∅
if exist(l ) /* l는 cur_session을 포함하고 size |cur_session|+1인 Large 항목집합*/
    for each l do
      recommend(l, u, cur_session)      /* u는 l의 마지막 URL */
else
    for each l2 do    /* l2는 last_url을 포함하고 size가 2인 Large 항목집합 */
      recommend(l2, u2, last_url)       /* u2는 l2의 마지막 URL */
recommend(l, url, session) {
  if 지지도(l) ≥ 최소 지지도
    confidence = 신뢰도(session⇒url)
    if confidence ≥ 최소 신뢰도 {
     url.패턴정보 신뢰도 = confidence
     1차 링크 집합 =+ url
    }
}
```

사전에 정의된 최소 신뢰도(Minimum Confidence)를 만족하는 URL만이 1차 링크 집합에 포함된다.

(2) 2차 링크 집합 결정 알고리즘

아래 [알고리즘 2]는 연관 웹 문서 분류 정보로부터 사용자에게 추천해줄 웹 문서를 결정하는 알고리즘이다.

[알고리즘 2]연관 웹 문서 정보에 대한 2차 링크집합 결정 알고리즘

```
last_doc : 현재 사용자가 가장 최근에 요청한 문서
2차 링크 집합 = ∅, word_set = ∅, word_set2 = ∅, cluster_num = 0, n = 0

word_set = 모든 단어들 in last_doc
cluster_num = last_doc가 소속된 cluster 번호
count_doc = cluster_num 클러스터내 의 모든 문서의 개수

for ( n = 1; n <= count_doc ; n++) {
 word_set2 = {cluster_num 클러스터 내의 n번째 문서의 단어들} ∩ {word_set}
 word_set2에 대한 연관 규칙 생성
 n번째 문서.연관 정보 신뢰도 = 모든 연관 규칙에 대한 신뢰도의 평균값 계산
}
2차 링크 집합 = 연관 정보 신뢰도 기준, 상위 m개 문서
```

현재 사용자가 가장 최근에 요청한 문서와 그 문서가 포함되어 있는 문서 클러스터내의 각 문서들에 대해서, 서로 동시에 포함하고 있는 공통 단어들의 집합을 구한다. 그리고 공통 단어 집합에 대한 연관 규칙을 구한 후, 각 연관 규칙에 대한 신뢰도의 평균값을 계산한다. 신뢰도의 평균값을 기준으로 상위 일정 개수의 문서를 2차 링크 집합에 포함시킨다.

(3) 최종 추천 링크 집합 결정

1차 링크 집합과 2차 링크 집합에 동시에 포함되는 링크들의 집합을 사용자들에게 제공하는데, 각 링크의 순위는 1차 링크 집합에서의 패턴 정보 신뢰도와 2차 링크 집합에서의 연관 정보 신뢰도의 평균으로 결정된다.

4. 실험 및 결과

본 논문에서는 실험을 위해 인하대학교 대학원 홈

페이지를 서비스하는 웹 서버의 Common Logfile Format(CLF) 로그파일 가운데, 2000년 7월 24일부터 8월 16일까지의 기록을 이용하였고, 웹 문서들에 대한 연관 문서 분류를 위해 대학원 홈페이지에 링크되어 있는 웹 문서 197개를 이용하였다. 웹 마이닝을 이용하여 로그파일로부터 최소 지지도 15%를 만족하는 952개의 Large 항목집합을 생성하였다.

[표 1] 사용자 브라우징 패턴의 예

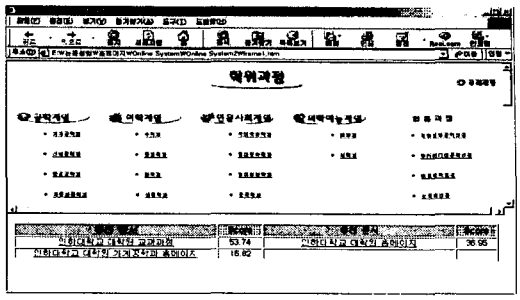
Large 항목집합	지지도(%)
{~grad/}	95.40
{~grad/sugang/index.html}	75.52
{~grad/}+{~grad/menu.htm}+{~grad/sugang/index.html}+{~grad/sugang/main.htm}	51.39
{~grad/}+{~grad/menu.htm}+{~grad/labs/main.htm}+{~grad/about/index.html}+{~grad/about/main.htm}	19.81
...	...

위 [표 1]은 사용자 브라우징 패턴 정보 생성을 위해 구한 전체 952개 Large 항목집합의 일부이다. 왼쪽 필드는 Large 항목집합들이고, 오른쪽 필드는 각 Large 항목집합에 대한 지지도이다.

[표 2] 클러스터에 포함된 단어 리스트

클러스터 번호	클러스터에 포함된 단어
1	{종합}{기준}{사항}{공동}{국가}{국제}{지원}{특강}{교류}{운영}{제도}{연구소}{인정}{협력}{학습}{강좌}{협약}{실습}{세미나} {방안}.....
2	{전공}{논문}{학과}{교수}{조교수}{입학}{학위}{일반}{필수}{개설}{학점}{이수}{과정}{내규}{사무}{업무}{학사}{조교}{과목}{성적}{합격}{학부}{졸업}..
3	{이론}{원리}{분석}{관계}{기초}{개요}{지식}{이해}{구성}{기능}{세미나}{현상}{언어}{형태}{사회}{적용}{다양}{심리}{중심}{체계}{문학}{방법론}...
4	{응용}{해석}{산업}{환경}{정보}{설계}{평가}{시스템}{처리}{자원}{실험}{작수}{기법}{활용}{컴퓨터}{반응}{측정}{효율}{생산}{통계}{제어}{전자}{진기}
5	{상호관계}{반응}{경쟁}{기관}{역학}{정치}{경제}{물성}{상태}{물리}{영향}{이동}{무역}{시대}{경영}{의미}{정론}{운동}{변환}...

그리고 웹 문서들에 대해 ARHP 알고리즘 적용 결과, 최소 지지도 30%를 만족하는 5개의 단어 클러스터가 위 [표 2]와 같이 생성되었다. 각 197개의 문서들에 대해서 위의 클러스터에 포함된 단어들과의 TFIDF를 계산한 후, 그 값의 평균치가 가장 큰 클러스터에 할당함으로써 연관 웹 문서를 분류한다. 그 결과 클러스터 1, 2, 3, 4, 5에 각각 31, 51, 40, 42, 33개의 문서가 할당되었다. 동적 링킹 시스템은 사용자들의 현재 세션에 대한 기록을 저장하고, 실시간으로 1차 링크 집합 결정 알고리즘과 2차 링크 집합 결정 알고리즘에 의해 추천 링크들을 생성한다. [그림 2]는 인하대학교 대학원 메인 페이지로부터 대학원 학위과정 페이지를 브라우징 할 때 온라인 시스템이 추천 문서를 동적인 링크로 제공하여 주는 그림이다. 여기서 Score는 1차 링크 집합의 패턴 정보 신뢰도와 2차 링크집합의 연관 정보 신뢰도의 평균값이다. [알고리즘 1]의 cur_session 크기는 Large 항목집합의 평균 크기인 3으로 하고, [알고리즘 2]에서는 연관 정보 신뢰도 기준, 상위 10개의 문서를 2차 링크 집합에 포함시켰다.



[그림 2] 동적 링킹 시스템

5. 결론

본 논문에서는 기존의 동적 링킹 시스템과는 달리 사용자 브라우징 패턴 정보뿐만 아니라 ARHP 알고리즘에 의한 연관 문서들의 분류 정보를 이용, 서로 연관성 있는 웹 문서들에 대한 동적 링킹으로 사용자들이 보다 편리하고, 정확하게 사이트를 브라우징 할 수 있도록 하였다. 향후 과제로는 본 논문에서 웹 서버의 로그파일로부터 사용자들의 브라우징 패턴을 추출할 때, 연관 규칙만을 사용하여 사용자가 한 번의 세션동안 요청했던 문서들의 Large 항목집합을 생성하였지만, 사용자들이 방문한 순서까지 고려하는 순차 패턴을 생성할 필요가 있다.

참고 문헌

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

[2] R. Agrawal, et. al, "The Quest Data Mining System," KDD, pp.244-249, 1996.

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

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

[5] C. Clifton and R. Cooley, "TopCat: Data Mining for Topic Identification in a Text Corpus," PKDD, pp.174-183, 1999.

[6] G. Karypis and V. Kumar, "Multilevel k-way Hypergraph Partitioning," DAC, pp.343-348, 1999.

[7] H. Lieberman, "Letizia : An Agent That Assist Web Browsing," http://lieber.www.media.mit.edu

[8] R. Kimball and R. Merz, *The Data Webhouse Toolkit*, Wiley Computer Publishing, pp.41-68, 2000.