웹 문서 형식과 클러스터 내의 문서 유사도를 이용한 동적 추천 시스템

김진수* 김태용** 이정현*
*인하대학교 전자계산공학과, **문경대학 컴퓨터정보과
kjspace@nlsun.inha.ac.kr tykim@munkyung.ac.kr jhlee@inha.ac.kr

Dynamic Recommendation System Using Web Document Type and Document Similarity in Cluster

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

요 약

기존의 여러 동적 추천 시스템에서 사용자들의 브라우징 패턴을 반영하려고 노력하였다. 그러나 대부분의 동적 추천 시스템들은 웹 문서들의 형식이나 웹 문서들 간의 연관성을 고려하지 않고, 사용자들의 브라우징 패턴에만 근거하기 때문에 연관성이 없거나 의미 없는 웹 문서들에 대한 추천까지 제공하는 문제점을 지니고 있다.

본 논문에서는 웹 문서들 사이의 유사도와 로그 파일 안에 들어있는 사용자들의 패턴을 이용하여 웹 문서 자체의 형식에 따라 연관된 웹 문서뿐만 아니라 순차적인 특성을 가진 웹 문서를 추천 문서로 제공한다. 이때 추천 웹 문서의 형식이 탐색 페이지이면 사용자 브라우징 순차 패턴 DB 중에서 사용자들이 자주 항해하는 순차적인 특성을 갖는 웹 문서까지 제공하는 동적 추천 시스템을 제안한다.

1. 서론

정제되지 않은 웹 데이터에는 사용자들의 축적된 경험들을 포함하는 유용한 정보들을 가지고 있다. 추천 시스템은 이러한 유용한 정보를 마이닝 기법이나 다른 측정 방법을 가지고 추출하려는 노력이 시도되고 있다. 기존의 협력적 추천 시스템에서는 사용자들에게 평가를 요구하여 축적된 평가 정보를 가지고 추천 집합을 제공하고 있다. 그러나 사용자들로부터 먼저 평가를 받아야 한다는 단점을 가지고 있다. 그리고 Letizia[5]와 같은 시스템은 사용자들의 브라우징 패턴 정보와 키워드를 가지고 링크된 웹 문서 중에서 사용자의 흥미를 만족할 만한 웹 문서를 추천하기 때문에 링크되지 않은 연관된 웹 문서를 추천 문서에 포함시키지 않고 있다. 또한 [8]과 같은 동적 링킹 시스템은 비록 사용자 브라우징 패턴과 전체적인 웹 문서에서 연관 웹 문서를 사용하지만 웹 문서 형식을 고려하지 않아 사용자에게 탐색 페이지와 같은 불필요한 웹 문서까지 추천 문서로 제공하는 문제점을 가지고 있다.

본 논문에서는 웹 문서들 사이의 유사도와 로그 파일에 포함된 사용자들의 패턴을 이용하여 웹 문서의 형식에 따라 연관된 웹 문서뿐만 아니라 순차적인 특성을 가진 웹 문서를 추천 문서로 제공한다. 이때 추천 웹 문서의 형식이 탐색 페이지이면 사용자 브라우징 순차 패턴 DB에서 사용자들이 자주 항해하는 순차적인 웹 문서까지 제공하는 동적 추천 시스템을 제안한다.

2. 관련연구

2.1 웹 마이닝(Web Mining)

웹 마이닝은 데이터 마이닝 기법을 웹에 적용시킨 기술로, 웹 자체가 지닌 리소스를 분석하는 Web Content Mining과 사용자 접근 패턴을 파악하는 Web Usage Mining, 웹 사이트와 웹 페이지의 하이퍼링크를 통해 정보를 구조화시키는 Web Structure Mining 등으로 분류할 수 있다. 본 논문에서는 웹 마이닝 기법 중 Web Usage Mining을 사용하여 사용자의 브라우징 순차 패턴을 추출한다. Web Usage Mining의 입력 자료로 주로 사용하는 것은 웹 서버에 기록된 로그 파일이다[4].

2.2 데이터 마이닝(Data Mining)

데이터 마이닝 알고리즘에는 여러 가지가 있는데, 본 논문에서는 연관 규칙, 순차 패턴, 클러스터링 등을 사용한다. 연관 규칙[1]은 통계적 방법에 의해 항목 집합들의 의존관계를 나타내는 것으로, 지지도와 신뢰도라는 측정 기준이 있다. 순차 패턴[2]은 사전에 정의된 최소 지지도를 만족하는 Large 항목 집합으로부터 생성되며, 이때 항목들 간의 시간상의 선후관계만을 고려하여 규칙을 찾아낸다. ARHP 알고리즘[3]은 Hypergraph Partitioning을 이용하여 연관 규칙에 나타나는 항목들을 가지고 클러스터링하는 방법이다.

2.3 문헌 클러스터링

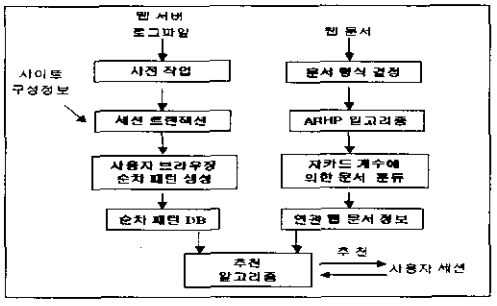
문헌 클러스터링은 문헌에 포함된 단어와 같은 식별요소를 이용하여 유사한 문헌의 클러스터를 형성하는 것이다. 클러스터링의 기준은 문헌과 문헌간 혹은 문헌과 클러스터간의 유사도를 이용하는데, 유사도의 측정을 위한 공식에는 다음과 같다.

1. 다이스계수 $\frac{2|X\cap Y|}{|X|+|Y|}$
2. 자카드계수 $\frac{|X\cap Y|}{|X\cup Y|}$
3. 코싸인계수 $\frac{|X\cap Y|}{|X|^{1/2}\times|Y|^{1/2}}$
4. 중복도계수 $\frac{|X\cap Y|}{\min(|X|,|Y|)}$
5. 타니모토계수 $\frac{|X\cap Y|}{|X|+|Y|-|X\cap Y|}$

여기서 $|X|$ 와 $|Y|$ 는 각 문헌이 갖고 있는 색인어 수이고 $|X \cap Y|$ 는 공통되는 색인어 수, $|X \cup Y|$ 는 두 문헌이 갖고 있는 다른 색인어의 합을 나타낸다[6].

3. 동적 추천 시스템

[그림 1]은 본 논문에서 제안하는 동적 추천 시스템이다.



[그림 1] 동적 추천 시스템

3.1 전처리 과정과 사용자 브라우징 순차 패턴 생성

로그 파일에서 사용자 패턴 정보를 추출하기 위해 확장자가 ".htm", ".html" 인 파일을 제외한 모든 기록을 제거하고, IP 주소, 요청 시각, 요청 URL 필드만을 가지고 트랜잭션을 결정한다[8]. 결정된 트랜잭션을 대상으로 IBM Almaden 연구소에서 개발한 AprioriAll 알고리즘[2]을 이용하여, 요청 URL을 항목으로 하는 사용자 브라우징 순차 패턴을 생성한다. 본 논문에서는 모든 사용자들이 고정 IP 주소를 사용하며, 프락시 서버를 사용하지 않는다고 가정한다.

3.2 웹 문서 형식 결정

웹 문서는 형식에 따라 Head page, Content page(내용 페이지), Navigation page(탐색 페이지), Look-up page, Personal page 등으로 나눌 수 있다[4]. Head page는 사용자가 방문하는 첫 번째 페이지이고, 내용 페이지는 웹 사이트가 제공하는 정보의 내용이 포함되어 있는 페이지이고, 탐색 페이지는 링크를 통해 내용 페이지로 안내하는 페이지이다. 그리고 Look-up page는 정의와 약어 표현을 위한 페이지이고, Personal page는 개인적 특성을 지닌 정보가 들어있는 페이지이다. 본 논문에서는 웹 문서의 형식 중 Head Page와 Look-up page는 특성상 탐색 페이지에 포함시켰고, Personal page는 실험 대상에 존재하지 않기 때문에, 웹 문서 형식을 탐색 페이지와 내용 페이지만으로 분류하였다. 웹 문서의 형식 결정 기준은 문서 내의 단어 수와 링크 수, 그리고 단어들 간의 유사도를 측정하여 판별하였다. 웹 문서 형식 결정 기준은 [표 1]과 같다.

[표 1] 웹 문서 형식 결정 기준

	단어 수(*Doc_N*)	링크 수(*Link*)	단어들 간의 유사도(*Word_Sim*)
내용 페이지	$Doc\_N \geq \alpha$	$Link < \beta$	$Word\_Sim \geq \gamma$
탐색 페이지	$Doc\_N < \alpha$	$Link \geq \beta$	$Word\_Sim < \gamma$

여기서 α, β, γ 는 한계값이고, 단어들 간의 유사도는 상호정보량을 이용한다.

3.3 연관 웹 문서 분류

연관 웹 문서 분류를 하기 위해 웹 문서 형식이 내용 페이지인 웹 문서들에서 추출된 명사들로부터 ARHP 알고리즘을 이용하여 여러 웹 문서에서 동시에 출현하는 명사들을 클러스터링한다. 연관된 웹 문서를 분류하기 위해 클러스터 내에 포함된 단어들을 가지고 벡터 모델, $V_{Cluster_i} = (w_{i1}, w_{i2}, \ldots, w_{ij}, \ldots w_{in})$ 으로 표현한다. 본 논문에서는 웹 문서의 특징을 추출하기 위해 $TF \cdot IDF$ 방법을 이용하는 것이 아니라, 처리 속도와 정확도를 높이기 위해 웹 문서에서 추출된 모든 명사들을 이용한다. 따라서 확률 모델보다는 벡터 모델이 적합하다. 벡터 모델에서 w_{ij} 의 값은 이진수로 표현되며, 단어가 $Cluster_i$ 에 포함되었으면 1로 표현하고 아니면 0으로 표현한다. 분류될 웹 문서를 위와 같이 벡터 모델로 변환시킨 후 유사도를 측정한다. 유사도 측정은 식(1)의 자카드 계수를 이용한다. 자카드 계수를 이용하는 이유는 유사도 측정 공식 중 자카드 계수가 문서 클러스터링에 사용되는 가장 보편적인 함수이기 때문이다.

$$Sim(V_i, V_j) = \frac{\sum_{k=1}^{n} (w_{ik} \cdot w_{jk})}{\sum_{k=1}^{n} w_{ik} + \sum_{k=1}^{n} w_{jk} - \sum_{k=1}^{n} w_{ik} \cdot w_{jk}} \qquad (1)$$

여기서 n 은 $Cluster_i$ 에 포함된 단어의 수이고, w_{ij} 는 벡터 V_i 의 k 번째 단어의 값이다. 이 $Sim(V_i, V_j)$ 함수는 두 번 사용된다. 하나는 클러스터 내에 포함된 단어들을 가지고 웹 문서들을 분류할 경우에 사용하고, 또 다른 하나는 사용자들이 웹 문서에 접근했을 때 클러스터 내의 웹 문서 유사도가 높은 집합을 찾을 경우에 사용한다. 이때 $V_i = (1,1,\ldots,1)$ 과 문서 V_j 의 벡터를 이용하며, $Sim(V_i, V_j)$ 값들 중 가장 큰 클러스터에 웹 문서를 할당하여 연관 웹 문서를 분류한다.

3.4 추천 알고리즘

추천 알고리즘은 순차 패턴으로부터 사용자의 세션을 가진 추천 집합과 사용자가 방문하고 있는 웹 문서와 가장 연관된 웹 문서를 제공할 추천 집합을 생성하는 알고리즘이다. 추천 알고리즘은 [알고리즘 1]과 같다.

[알고리즘 1] 추천 알고리즘

```
Input : cur_session : 현재 사용자 세션
        last_url :사용자가 가장 최근에 요청한 URL
        δ : 최소 지지도, α : 최소 신뢰도

Output : Recommend1 : 순차 패턴에 포함된 추천 문서 집합
         Recommend2 : 연관 문서에 포함된 추천 문서 집합

Recommend1 = Recommend2 = ∅ ;
if ( last_url.type == content_page ){
for each l do   // l는 cur_session을 포함,size |cur_session|+1인 large sequence 집합
  if ( 지지도(l) ≥ δ )
    confidence = 신뢰도(session=>url) ; // url 은 추천 웹 문서
    if ( confidence ≥ α ){
      url.score = confidence ;
      if ( url.type == navigation_page )
        url = Seq(url) ;     // Seq함수는 url을 포함하는 순차 패턴 중 내용 페이지를
      Recommend1 += url ;    // 추천하며 추천 url을 반환(신뢰도(session=>url) 포함)
    }
  for each url do    // url 은 last_url이 속한 클러스터의 문서들
    if (( value = Sim(last_url, url)) ≥ ξ )
      Recommend2 += url;  // 유사도가 높은 문서를 추천 문서에 포함시킴
}
else if ( last_url.type == navigation_page ){
   while (( url = Seq(last_url)) ≥ δ )
      if ( url.confidence ≥ α ){
      url.score = url.confidence ;
      if ( url.type == navigation_page )
        url = Seq(url) ;
      Recommend1 += url ;
      }
}
```

사용자 세션 중 last_url의 형식이 탐색 페이지인지 내용 페이지인지에 따라 추천 집합을 생성한다. 사용자의 last_url이

내용 페이지이면, 현재 세션을 포함하면서 1이 더 큰 large sequence 집합들 중에서 추천 집합을 생성한다. 만약 last_url이 탐색 페이지이면, 순차 패턴 DB에서 last_url을 포함하고 최소 지지도 이상을 가진 large sequence 집합에서 추천 집합을 생성한다. 이때 신뢰도(*session* ⇒*url*)는 지지도와 신뢰도 공식을 이용하여 다음과 같이 결정된다.

$$\text{신뢰도}(session \Rightarrow url) = \frac{|\ session \cap url\ |}{|\ session\ |} \qquad (2)$$

신뢰도(*session* ⇒*url*)는 추천 집합의 순위 결정을 위한 가중치로 사용하며, 최소 신뢰도를 만족하는 url들만을 추천 집합에 포함시킨다. 이때 추천될 *url*이 탐색 페이지이면 Seq 함수로부터 반환된 웹 문서를 포함시킨다. Seq 함수는 url을 포함하는 순차 패턴 DB에서 사용자의 세션을 가지고 사용자들이 자주 항해하는 순차적인 특성을 가진 웹 문서를 반환하는 함수이다. 예를 들어, A, C문서는 내용 페이지이고, B문서는 탐색 페이지라고 가정할 때, 만약 A→ B→ C로 항해하는 사용자가 많다고 하면, 동적 추천 시스템은 사용자가 A문서를 방문했을 때 B문서를 추천하는 것이 아니라 C문서를 추천 집합에 포함시켜 제공하는 것이다. 이것은 [알고리즘 1]에 의해 B문서가 탐색 페이지이기 때문에 C문서를 제공하는 것이다. 또한 last_url을 가진 클러스터에서 유사도가 높은 웹 문서들을 추천 집합에 포함시킨다.

4. 실험 및 결과

본 논문에서는 실험을 위해 인하대학교 대학원 홈페이지를 서비스하는 웹 서버의 로그 파일 가운데, 2001년 1월 18일부터 2월 2일까지의 기록을 이용하였다. 웹 문서 187개중 연관 문서 분류를 위해 168개의 내용 페이지들을 이용하였다. 웹 마이닝 알고리즘을 이용하여 로그 파일로부터 최소 지지도 30%를 만족하는 481개의 사용자 브라우징 순차 패턴이 생성되었다. [표 2]는 추출된 순차 패턴 DB를 보여준다.

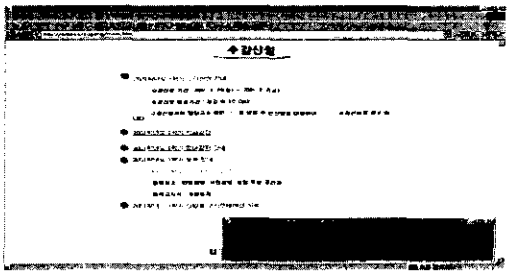
[표 2] 사용자 브라우징 순차 패턴

사용자 브라우징 순차 패턴	지지도(%)
[/~grad/] [/~grad/menu.htm]	51.78
[/~grad/] [/~grad/sugang/main.htm]	37.46
[/~grad/] [/~grad/sugang/main.html] [/~grad/]	27.52
[/~grad/] [/~grad/notice/2001-1enter/2001-1_grad.html]	23.85
[/~grad/] [/~grad/menu.htm] [/~grad/menu.htm]	23.85
...	...

[표 3]은 내용 페이지들로부터 클러스터링 한 후 포함된 단어들과 식(1)에 의해 분류된 연관 웹 문서와 웹 문서들의 벡터 표현을 보여준다. 웹 문서 형식 결정시의 한계값으로 단어 수와 링크 수는 각각 150개와 25개로 하였고, 단어들 간의 연관도는 상호 정보량의 평균값을 이용하였으며, 이때 한계값은 0.15로 하였다. 순차 패턴 DB와 연관 웹 문서 정보는 사용자가 웹 페이지에 접근했을 때의 입력 자료로 사용된다. 온라인 시스템은 사용자들의 현재 세션에 대한 기록을 저장하고, 실시간으로 추천 알고리즘을 통해 추천 집합을 생성한다. [그림 2]는 사용자가 [~grad/sugang/main.htm]이라는 페이지를 방문했을 때의 화면과 추천되는 웹 문서를 보여준다. [~grad/sugang/main.htm] 페이지의 형식이 내용 페이지이기 때문에 현재 페이지가 속한 클러스터 번호로 가서 클러스터 내에 속한 문서들과 식(1)을 이용하여 유사도를 구하여 추천 문서를 가져오며, 또한 사용자들의 순차 패턴 DB에서 최소 지지도와 최소 신뢰도를 만족하는 순차적인 특성을 갖는 페이지들까지 추천 집합에 포함시킨다.

[표 3] 분류된 연관 웹 문서의 벡터 표현과 해당 클러스터의 웹 문서 수

번호	클러스터링된 단어 및 연관 문서	문서수
1	[개설][입학][자격][전공][전형][제출][과목][교수][구술][학과][하문]... {~grad/sugang/2001~1orientation.html} (1,1,1,1,0,1,1,0,0,0,1,1,1,1,1,0,1,1,0,1,1,0,0,1,1,0,1,1,1,1,1,0,.....,0,1,1)...	72
2	[정보][정의][정책][제도][계산][서류][설명][교육][공동][공통][과학]... [~grad/about/cooperation.htm] (0,0,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,1,1,1,0,0,1,1,1,0,1,1,1,0,1,1,0,1,1)...	26
3	[경제][관리][기관][효율][모델][목표][문화][물리][반응][분류][사례]... [~grad/labs/sci.htm] (0,1,0,0,0,0,1,0,0,1,1,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,.....,0,0,1)...	13
4	[강의][개념][개발][개요][검토][제반][고찰][과정][소개][동향][교과]... [~grad/course/degree/bio.htm] (1,1,1,1,0,1,1,1,0,1,1,1,1,0,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,.....,1,1,1)...	45
5	[공정][공학][관점][세미나][기술][시스템][목적][제어][발전][효율]... [~grad/labs/eng.htm] (1,1,1,0,1,0,0,0,0,0,0,0,0,0,1,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0,1,1,0,.....,1,0,0)...	12



[그림 2] 웹 문서와 추천 집합 리스트

5. 결론

본 논문에서는 웹 문서 형식에 따라 연관된 웹 문서뿐만 아니라 연관성은 없지만 순차적인 특성을 갖는 웹 문서까지 추천하는 방법을 제공하였다. 기존의 추천 시스템과는 달리 순차 패턴과 웹 문서의 형식에 따라 연관 및 순차적인 추천 집합을 제공함으로써 사용자들이 보다 정확하고 편리하게 웹 문서 사이를 항해할 수 있도록 하였다. 향후 과제로는 각 사용자의 개별화를 통해 클러스터링하고 사용자가 웹 문서를 방문하였을 때 비슷한 흥미를 가진 사용자들로부터 정보를 얻어 재공할 필요가 있다. 또한 제공된 추천 집합이 사용자들에게 얼마나 적용이 되었는지 로그 파일을 재분석할 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 and R. Srikant, "Mining Sequential Pattern," Proc. of the Int'l Conference on Data Engineering (ICDE), Taipei, Taiwan, March 1995.

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

[6] 정영미, *정보검색론*, 구미무역 출판부, 1993.

[7] T. Tokunaga and M. Iwayama, "Text categorization based on weighted inverse document frequency," IPSJ SIG Report, NL100 (5), 1994.

[8] 박영규, 김진수, 김태용, 이정현, "연관 웹 문서 분류와 사용자 브라우징 패턴을 이용한 동적 링킹 시스템," 한국정보처리학회 추계학술발표 논문집, pp. 305-308, 2000.