일반적인 웹 검색 경로패턴 추출 알고리즘

장민석* · 하은미**

Algorithm for Extracting the General Web Search Path Pattern

MinSeok Jang* · EunMi Ha**

Dept. of Computer Information Science, Kunsan National University

E-mail : msjang@kunsan.ac.kr*, yesni77@hanmail.net**

요 약

웹 환경에서 사용자들의 정보검색 패턴을 얻어내기 위해 흔히 로그 파일의 정보검색 패턴을 분석하는 기존 연구들이 있어 왔다. 이들에서 흔히 사용하는 방법은 경로 순회 패턴(path traversal patterns)에서 효율적으로 빈번 패턴(frequent patterns)을 찾아내는 알고리즘을 제안하는 것이다. 하지만 이들의 기존 연구의 가장 일반적인 문제점들 중의 하나는 일반적인 패턴 즉, 복잡한 형태 패턴(topological patterns)에 대한 적절한 해답을 찾아주지는 못한다는 것이다. 따라서 본 논문에서는 일반적인 패턴 유형을 정의하고 이들로부터 정보검색 패턴을 알아내는 효율적인 알고리즘을 제안하고자 한다.

ABSTRACT

There have been researches about analyzing the information retrieval patterns of log file to efficiently obtain the users' information research patters in web environment. The methods frequently used in their researches is to suggest the algorithms by which the frequent one is derived from the path traversal patterns in efficient way. But one of their general problems is not to provide the proper solution in case of complex, that is, general topological patterns. Therefore this paper tries to suggest a efficient algorithm after defining the general information retrieval pattern.

키워드

Web search pattern, path traversal pattern, Data Mining

Ⅰ. 서 론

오늘날 웹에서 사용자가 원하는 정보를 얼마나 효율적으로 검색하는지에 대한 중요성이 점증하고 있다. 웹 환경에서 실제로 사용자들이 정보에 접근하는 사용자 접근 패턴들은 웹 시스템 디자인이나 효과적인 마케팅 전략을 세우는 데 도움을 줄 것이다. 이와 관련된 연구들([1]~[4])이 진행되어 왔다. 하지만 이들은 다음과 같은 점에서 한계점을 들어내고 있다. 첫째, 일반적인 경로 패턴과는 달리 간단한 경우에 대해서 적용하고 있다. 둘째, 검색의 순서를 무시하고 있다. 따라서 본 논문에서는 이러한 문제점을 극복한 사용자들의 빈번한 정보검색 패턴을 알아내는 효율적인 알고리즘을 연구한다.

웹 환경에서 사용자들이 탐색하는 패턴들이 접근하는 방법으로 두 가지 방향을 따르고 있다. 첫 번째는 법칙 기반 패턴(rule-based patterns)을 사용하는 방법이고, 두 번째는 형태기반패턴(topology-based patterns)을 구하는 방법이다. 본 논문에서는 후자의 경우에 해당된다. 사용자들이 일반적으로 접근하는 것은 특정한 형태-길(topology-path) 그래프를 제한한 사용자 접근 패턴을 효율적인 계산을 조사하는 데 첫 번째입니다. 이 논문에서는 일반적로 해결하는 새롭고 효율적인 알고리즘을 제안한다. 이 알고리즘은 효율적인 반복 패러다임(paradigm)을 기반으로 한다. 나머지의 단계는 다음과 같이 구성하였다. 두 번째 단계에서 우리는 문제를 공식적으로 정의를 한다. 세 번째 단계에서는 우리의 알고리즘뿐만 아니라 수행, 토론들을 제안한다.

II. 본 론

아래 그림 1은 웹 검색의 일반적인 경로 망을 예를 든 것이다. 각 노드(A,B,C,D,E)는 사용자가 방문하는 웹 문서(혹은 사이트)를 나타내며, A 노드로부터 검색함을 의미한다. 여기서 각 에지는 양방향성을 가질 수 있다. 따라서 방향성을 고려하지 않는다면 10(=nC2/2)개의 에지가 존재하며, 방향성까지 고려하면 20개의 방향성 에지가 있음을 알 수 있다. 기존 연구의 경우 backward traversal(역방향 통과)은 거의 고려하지 않고 있다. 예를들어 ABC와 같은 경로를 고려한 알고리즘은 존재한다. 하지만 ABCAE와 같은 경우에 대해서는 적절한 효과적인 알고리즘이 없는 실정이다. 본 논문에서는 Apriori 알고리즘[6]과 그래프 이론[5]에 기반해서 역방향 경로까지 고려한 알고리즘을 제안하고자 한다.

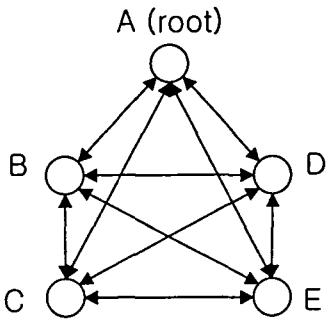

그림 1. 웹 검색 경로 망 예

위 그림에서 존재하는 웹 정보검색 경로는 역방향 경로까지 고려하면 무수히 많은 경로가 존재할 수 있다. 그 중 역방향 경로까지 포함한 아래와 같은 경로가 존재할 수 있다. 이를 기반으로 제안하고자 하는 알고리즘을 설명하고자 한다.

표 1. 경로 통과 예

Case	경로 통과 예
1	ABC
2	ABCBA
3	ABCA
4	ABCDE
5	ABCADE
...	...

III. MFTP 알고리즘

그림 2는 표 1의 경로를 그림 1의 망에 적용해 본 결과이다. 여기서 Case 2, 3, 5는 역방향 경로를 포함하고 있다. 구하고자 하는 것은 MFTP(Maximal Frequent Traversal Path; 최빈 통과 경로)이다. 알고리즘의 적용 과정을 그림 3에 도시하고 있다.

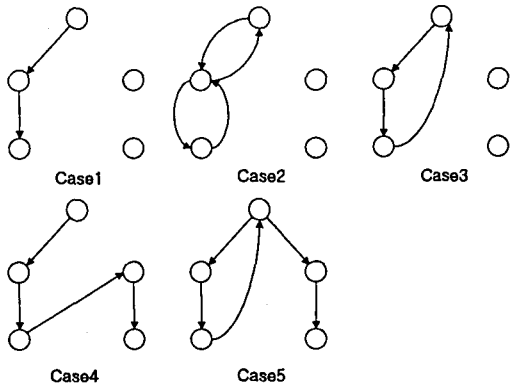

그림 2. 각 경우의 검색 경로

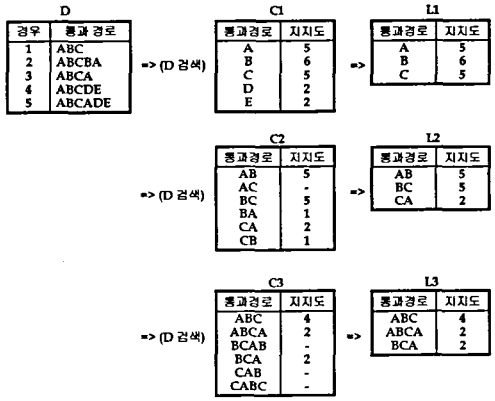

그림 3. 알고리즘 적용 절차

D(데이터베이스)에는 통과 경로들이 저장되어 있으며 MFTP를 계산하는 과정에서 이를 계속 검사하게 된다. 여기서 Lk는 k번째 빈번한 통과 경로(path traversal)을 의미하며, Ck는 Lk를 포함한 집합이다. Lk에서 일정 지지도 이상의 원소 경로를 추출한 것이 Ck이다. 최종적으로 알고리즘의 결과는 마지막 루프를 진행하고 남은 Lk를 구하는 것이다. 위 그림 3에서 C1은 최소 경로를 의미하므로 모든 노드를 열거하게 되며, 이중 D를 검사함으로써 그 빈도수를 계산하게 되며, 이중 사용자가 원하는 빈도수 이상의 지지도를 가진 통과 경로를 L1으로 선정하게 된다. 여기서는 최도 빈도수 2인 D, E 노드를 제거함으로써 L1을 산출하게 된다. 그 이후 L1의 경로를 포함한 길이가 2인 통과경로를 산출한다. 기존 연구와 다른 점은 방향성(역방향성)을 고려하기 때문에 Itemset이 아닌 순서 개념이 들어간 경로 그 자체

를 표기하게 된다. 따라서 기존의 경우 Ck로서 {A,B}, {B,C}, {A,C}를 고려하였지만, 여기서는 AB, AC, BC, BA, CA, CB의 6가지 경로를 고려하게 된다. 이의 빈도수를 D에서 계산하고자 할때도 집합으로서가 아닌 자체 경로가 존재하는 빈도수를 계산한다. 즉, CB 경로의 경우, C와 B는 모든 경로에 존재하지만, CB경로는 경우2에만 존재하게 된다. 따라서 빈도수는 1 에 불과하게 된다. C2에서 사용자 선택에 따라서 지지도를 2이상인 경로를 선택하게 되면, L2와 같이 산출된다. 이와 같은 단계를 계속하게 되면, 최종적으로 L3와 같은 MFTP가 산출되게 된다. C3에서 지지도가 2인 경로를 제거함으로써 ABC 경로를 MFTP로 정할 수 있다. 하지만 L3에서 C4로 진행하는 것은 의미가 없음을 알 수 있다.

개략적인 알고리즘을 표현하면 다음과 같다.

알고리즘 MFTP(Maximal Frequent Traversal Path)
• 입력: 웹 경로 검색 경로를 포함한 데이터베이스 D
• 출력: 지정한 지지도 이상 빈번한 경로
-1단계: C1를 계산하기 위해 D를 판독.
-2단계: k값을 증가하면서
Lk-1로부터 Ck를 계산한다.
Lk를 얻기 위해 D를 검사해서 Ck를 산출
(이때 지지도값은 사용자가 지정한다.)
마지막 Lk가 나올때까지 반복한다.

IV. 결론 및 향후 연구

이상에서 본 바와 같이 본 논문에서는 일반적인 웹 검색 경로에 대한 MFPT(빈번한 검색 경로)를 구하는 효율적인 방법에 대해 제안하였다. 특징적인 것은 역방향 경로까지 고려한다는 점과 집합 개념이 아닌 순서개념을 적용함으로써 가장 일반적인 형태 기반 패턴을 추출할 수 있다는 것이다. 기존의 연구 중에는 역방향을 고려한다고 하더라도 순방향, 역방향 경로를 따로 적용하고 있는 알고리즘이 존재한다. 본 연구는 순서개념을 적용함으로써 이러한 불편함을 해소했다.

앞으로 진행되어야 할 연구는 보다 정확한 성능 평가지수를 제공함으로써 본 알고리즘의 수치적인 우수함을 보이고자 한다.

감사의 글

본 연구는 한국과학재단 목적기초연구
(R01-2004-000-10946-0) 지원으로 수행되었음.

참고문헌

[1] M. S. Chen, J. S. Park, and P. S. Yu, Efficient Data Mining for Path Traversal Patterns, IEEE Transactions on Knowledge and Data Engineering, 10(2), 209-221, 1998.
[2] X. Lin, C. Liu, Y. Zhang, X. Zhou, "Efficiently Computing Frequent Tree-Like Topology Patterns in a Web Environment", In Proc. IEEE Conference, pp440~447, 1999.
[3] B. Mobasher, N. Jain, E. H. Han and J. Srivastava, "Web Mining, Pattern Discovery from World Wide Web Transactions", Proceeding of the 9th IEEE International Conference on Tools with Artificial Intelligence(ICTAI'97), 1997
[4] J. S. Park, M. _S. Chen, and P. S. Yu, "An effective Hash Based Algorithm for Mining Association Rules", ACMSIGMOD, 175-186, 1995
[5] J. A. Bondy and U.S.R. Murty, Graph Theory with Applications, Macmillan, 1997.
[6] Data Mining: Concepts and Techniques Chapter6 : Mining Association Rules in Large Databases , 2000