

웹 이용자의 접속 정보 분석을 통한 웹 활용 그래프의 구성 및 분석*

†김후곤** · 김재교***

An Analysis on the Web Usage Pattern Graph Using Web Users' Access Information*

Hu-Gon Kim** · Jae-Gyo Kim***

■ Abstract ■

There are many kinds of research on web graph, most of them are focus on the hyperlinked structure of the web graph. Well known results on the web graph are rich-get-richer phenomenon, small-world phenomenon, scale-free network, etc.

In this paper, we define a new directed web graph, so called the Web Usage Pattern Graph (WUPG), that nodes represent web sites and arcs between nodes represent a movement between two sites by users' browsing behavior. The data to constructing the WUPG, approximately 56,000 records, are gathered from some users' PCs. The results analysing the data summarized as follows :

- (i) extremely rich-get-richer phenomenon
- (ii) average path length between sites is significantly less than the previous one
- (iii) less external hyperlinks, more internal hyperlinks

Keyword : Web Graph, Small World Phenomenon, Scale-Free Network, Web Usage Pattern

논문접수일 : 2006년 07월 04일 논문게재확정일 : 2006년 10월 30일

* 이 논문은 2004년도 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2004-041-B00140)

** 경성대학교 e-비즈니스 전공

*** KT 부산본부 통신망 관리부

† 교신저자

1. 서론

웹의 급속한 확산으로 인해 약 100억개를 상회하는 것으로 추정되는 모든 웹 페이지에 대한 검색은 거의 불가능하다. Lawrence 등의 연구에서 보듯이 대표적인 검색엔진이라 하더라도 단지 전체 웹페이지의 38% 정도만을 검색할 수 있다[15]. 이처럼 검색엔진을 통한 정보검색이 가지는 한계는 이들의 정보수집 방식에 기인한다. 대부분의 검색엔진은 특정 웹 페이지에 포함된 하이퍼링크를 이용하여 새로운 웹페이지에 대한 정보 수집 및 분석을 반복하는 재귀적 과정을 통해 웹의 정보를 수집하게 된다[6, 16]. 이러한 재귀적 과정을 크롤링(crawling)이라 한다.

검색엔진들은 크롤링을 통해 수집된 웹페이지의 정보를 이용하여, 웹 페이지를 노드(node)로 보고 웹페이지간의 하이퍼링크를 아크(arc)로 하는 웹 그래프(web graph)를 만들게 된다[7]. 그렇지만 크롤링을 통해 만들어진 웹 그래프는 웹 페이지들의 하이퍼링크 구조(hyperlinked structure)에 의존적인 수밖에 없으므로, 매우 관련성이 높은 사이트들이 서로 하이퍼링크되어 있지 않다면 이들은 웹 그래프상에서 아크가 존재하지 않게 되어 서로 관련이 없는 사이트로 취급하게 된다.

이처럼 거대하게 변해가는 웹은 매우 복잡한 네트워크로 하나의 복잡계(complex system)를 형성하게 된다. 복잡계로서의 웹과 관련된 많은 연구가 최근에 활발히 진행되고 있다. 그러한 대다수 연구는 웹 페이지에 대응되는 노드(node)와 웹 페이지들을 서로 연결(hyperlink)하는 것에 대응하는 아크(arc)의 개념들을 활용하여 웹을 하나의 그래프로 다루고 있다[3, 13, 13]. Broder 등[7]은 웹의 구조적인 측면에서 무작위 네트워크(random graph)이 소수의 노드에 집중화되는 형태임을 보였고, Bharat 등[6]은 개별 웹 페이지가 아닌 웹 사이트차원에서 웹의 특징을 설명하였다. Barabasi 등은 웹이 기존의 무작위 네트워크와는 달리 비정상적으로 많은 링크를 가지고 있는 소수의 허브

(hub)가 존재하는 규모 독립적 네트워크(scale-free network)의 특성을 가지고 있음을 보였다[3-5].

본 연구에서는 기존의 하이프링크 중심의 웹 그래프와는 달리 웹사용자의 이용패턴을 반영한 웹 그래프(Web Usage Pattern Graph, WUPG)를 새롭게 정의하기로 한다. 이렇게 만들어진 웹 그래프의 특성을 분석하고, 그 결과를 바탕으로 웹의 특성 및 웹 사용자들의 웹 이용 형태 분석을 수행하기로 한다. 본 논문은 다음과 같이 구성된다. 2장에서는 웹 그래프와 관련된 연구들을 알아보고, 3장에서는 본 연구에서 제안하게 되는 WUPG의 구성 방법, 데이터의 수집 및 분석결과를 알아보고, 4장에서 연구의 결과를 요약하기로 한다.

2. 기존 연구의 고찰

하이퍼링크 구조에 의존적인 웹 그래프의 문제점을 개선하기 위해, 본 연구에서는 사용자의 웹 이용 패턴을 반영할 수 있는 새로운 웹 패턴 그래프(Web Usage Pattern Graph, WUPG)를 정의하기로 한다. 또한 WUPG의 구축 및 특성 분석을 위한 시스템을 개발하고, 이를 통해 WUPG의 특성 및 웹 이용자의 다양한 형태 이용 패턴에 대한 분석을 수행하기로 한다. WUPG와 관련된 기존의 연구로는 웹 그래프와 관련된 연구, 웹 사용자의 이용 패턴을 분석하는 웹 마이닝(web mining)에 대한 연구, 웹 검색 알고리즘들과 검색엔진에서 제공하는 브라우저 툴바(toolbar)에 대한 연구 등이 있다.

2.1 웹 그래프에 관한 연구

웹 그래프와 관련된 연구들은 전통적인 그래프 이론에 바탕을 두고, 웹의 하이퍼링크 구조로부터 웹 그래프를 도출하고 이의 특성을 분석하는데 중점을 두고 있다. 기존의 웹 그래프에 관한 연구 결과 웹에 대한 중요한 몇 가지 특성들이 파악되었는데, 이를 정리하면 다음과 같다.

① 웹 페이지의 도달가능성

웹 페이지 수의 증가속도는 매우 빠르기 때문에 어떠한 검색엔진이라도 전체 웹 페이지에 관한 정보를 모을 수 없으며, Lawrence의 연구에 의하면 가장 많은 페이지를 찾는 검색엔진도 전체 웹 페이지의 38% 정도만을 검색 할 수 있다[15].

② 스몰 월드 현상(small world phenomenon)

웹 페이지 수의 증가속도는 매우 빠르지만, 웹 상에 있는 임의의 두 페이지간의 평균거리는 항상 일정하게 유지되는 현상을 말한다. 여기에서 평균거리란 웹 그래프에서의 두 웹 페이지 간에 클릭하여 도달할 수 있는 최단 클릭 수에 대한 평균을 말하는데, Barabasi 등[3]의 연구에 의하면 두 웹 페이지간의 평균거리는 약 19인 것으로 나타나고 있다. 이처럼 네트워크의 크기에 관계없이 평균거리가 일정한 성질을 만족하는 네트워크를 규모 독립적 네트워크(scale-free network)라 한다[2, 11]. 이러한 네트워크의 예로는 인구의 수에 관계없이 임의의 두 사람 간에 최소한의 아는 사람으로 이 두 사람을 연결하는 경로의 길이는 항상 일정하게 유지되는 인적 네트워크를 예를 들 수 있다[4, 14].

③ 지수 법칙

어떤 웹 사이트가 얼마나 많이 하이퍼링크 되어 있는가를 측정해보면 다음과 같은 지수 함수(power function)의 역함수 형태를 가진다는 것이다[3].

$$y = a^{-x}$$

여기에서 a 는 상수이고 x 는 어떤 웹 사이트의 하이퍼링크된 정도를 나타내고 있다. 이를 웹 사이트의 부익부 빈익빈 현상(rich-get-richer phenomenon)이라고도 한다. 즉 포털(portal)이나 검색엔진과 같이 하이퍼링크의 빈도가 매우 높은 사이트의 수는 매우 적은 반면에, 개인홈페이지와 같이 거의 하이퍼링크가 이루어지지 않는 사이트가 대부분임을 의미한다.

이러한 웹 그래프와 관련된 연구 결과는 오직

웹의 하이퍼링크 정보만을 근거로 하고 있어서, 실제 사용자들의 웹 이용 패턴을 반영하지는 못하고 있다. 즉 하이퍼링크를 이용한 웹 그래프를 웹에 대한 물리적 구조라고 본다면, 사용자의 웹 이용 패턴을 반영하는 그래프는 웹에 대한 논리적 구조라고 볼 수 있는 것이다.

2.2 웹 마이닝에 관한 연구

웹 이용 패턴을 반영하기 위한 연구로는 웹 마이닝 분야가 있다[8, 10, 18]. 웹 마이닝과 관련된 연구는 주로 단일 또는 몇 개 서버의 로그 데이터(log data)를 분석하는 연구에 집중되어 있다. 즉 사용자의 이용 패턴을 분석하기 위하여, 웹 서버에 저장된 로그 데이터를 분석하고 이를 이용하여 '가장 즐겨 찾는 페이지'나 '오늘의 가장 많이 접속한 페이지' 등을 제공한다. 또는 사용자를 구분하여, 사용자별로 선호하는 정보를 제공하는 맞춤형 정보 제공 기능 등을 제공하기도 한다.

웹 마이닝은 마케팅 분야의 데이터 마이닝 기법을 웹의 영역에 응용한 것이다. 웹 마이닝에 대한 대부분의 연구는 특정사이트의 로그 데이터만을 대상으로 하므로, 사용자가 특정사이트를 떠나서 다른 사이트를 방문하는 것에 대한 정보는 알 수 없다는 문제점이 있다. 즉 개별사이트 단위에서 이루어지는 사용자의 웹 이용 형태는 파악할 수 있지만, 사용자들의 전체적인 웹 이용 형태에 대한 정보는 파악할 수 없는 것이다.

2.3 브라우저 툴바

일부 검색엔진들은 사용자의 웹 이용 패턴을 반영할 수 있는 툴바(toolbar)들을 설치하고, 이를 통해 사용자의 웹 이용 패턴 정보를 수집하고 있다. 다음은 검색엔진인 알렉사(www.alexa.com)의 툴바를 설치한 후 중앙일보(www.joins.com)에 접속한 경우, 이와 관련된 사이트의 정보를 알렉사의 툴바가 보여주고 있다.

주소 http://www.joins.com

A navigation bar containing a search box with the text "Search" and a magnifying glass icon, and a menu icon with the text "Menu". Below the search box, there is a list of menu items: 한국일보, 동아일보, 경향신문, 조선일보, 매일경제, 한겨레신문, 세계일보, 문화일보, 국민일보, 대한매일.

즉 중앙일보와 관련된 다른 주요 신문 사이트들을 보여주고 있는데, 실제 대부분의 신문사들 간에는 하이퍼링크 되어 있지 않으므로 웹 그래프에서는 전혀 연관성이 없다. 하지만 사용자의 웹 이용 패턴을 분석해 보면, 사용자 입장에서 보면 신문사 사이트들은 매우 연관성이 높을 수 밖에 없다. 이처럼 알렉사와 같은 기능을 제공하려면 사용자의 웹 이용 패턴을 체계적으로 수집하고 분석할 수 있는 시스템의 개발이 필요함을 알 수 있다.

3. WUPG 시스템의 구축 및 분석

3.1 WUPG 시스템의 개발

본 연구에서는 사용자의 웹 이용 패턴을 기록하고, 이를 이용하여 새로운 네트워크를 정의하게 되는데 이를 WUPG(Web Usage Pattern Graph)라 부르기로 한다. 즉 WUPG는 개별 사용자들의 사이트에 대한 접속 정보를 수집하여서, 사용자가 방문하는 사이트를 노드로 구성하고 서로 다른 사이트간의 이동이 일어나면 이를 아크로 하는 네트워크로 정의된다. 자세한 내용은 아래에서 다시 언급하기로 한다.

먼저 WUPG를 구성하기 위해 웹 사용자들의 정보를 체계적으로 수집하여야 하는데, 그 구성요소는 다음과 같다. WUPG 시스템은 사용자가 접속하게 되는 웹 브라우징 정보를 수집하여 서버로 자동 전송하는 클라이언트 에이전트(WUPG client agent, WCA), WCA가 보낸 정보를 수집하는 서버(WUPG Server, WS), WS에 모인 정보를 분석하는 분석기(WS Analyser, WA)의 세부분으로 구성된다.

WCA는 사용자의 PC에서 실행되며, 웹 브라우저에서 URL의 접속이 발생하면 이와 관련된 정보를 지정된 WS로 전송한다. 이때 WS로 보내는

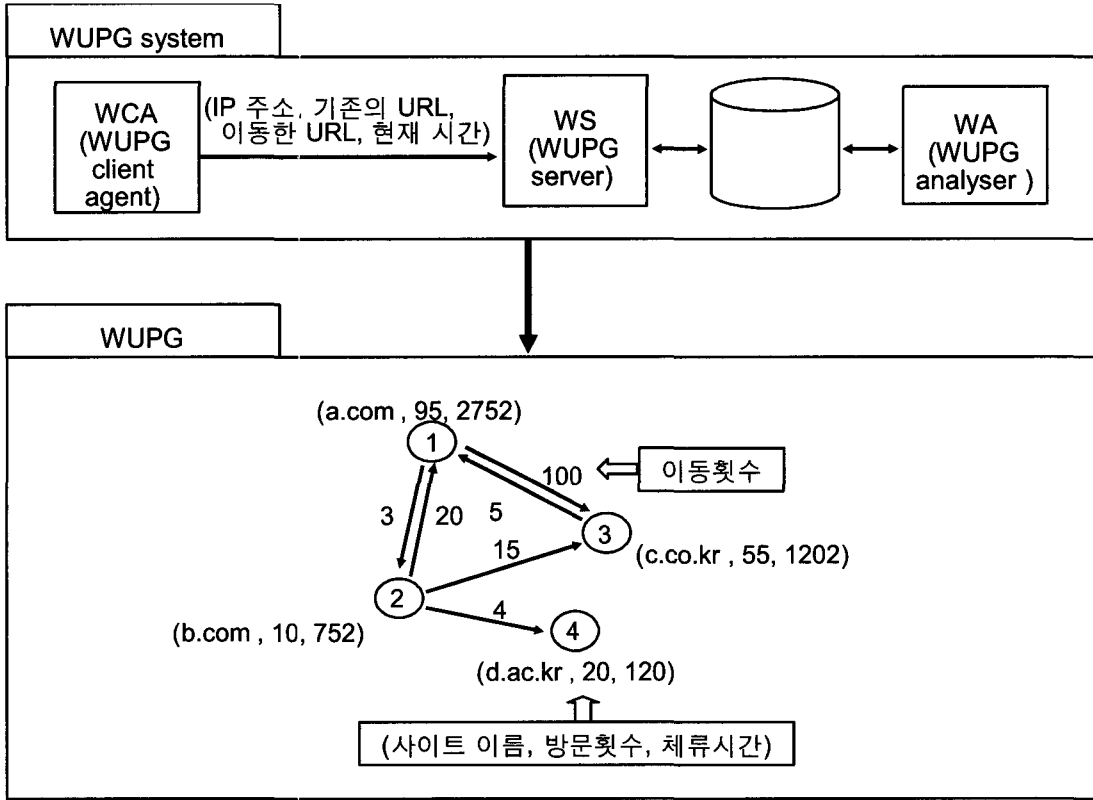
정보는(IP 주소, 접속한 URL, 접속한 시간)으로 구성된다. WS는 매우 간단한 데이터베이스를 구축하는 서버로서 단순히 WCA가 보낸 정보를 순차적으로 저장한다. 핵심이 되는 부분은 WA로 WS에서 수집한 정보를 분석하여 WUPG를 구축하게 된다.

본 연구에서 WUPG는 $G=(V,A)$ 로 나타내고, V 는 노드(node)의 집합이고 A 는 방향성이 있는 아크(directed arc)의 집합으로 정의하기로 한다. 이때 노드는 웹 사이트에 대응하는 것으로(사이트 이름, 방문횟수, 체류시간)의 3가지 레이블(label)을 가지고, 노드 i 와 노드 j 를 연결하는 아크는 (i,j) 로 표시하고 아크는 방문횟수를 값으로 가진다. <그림 1>은 WUPG 시스템과 이로부터 생성한 WUPG의 예를 보여 주고 있다.

<그림 1>에서 보듯이 WCA에서(IP 주소, 접속한 URL, 접속한 시간) 정보를 WS로 전송하면, WS는 이 정보를 데이터베이스에 저장한다. WA는 데이터베이스를 분석하여 WUPG를 생성하게 된다. <그림 1>은 4개의 노드와 6개의 아크로 이루어진 WUPG의 예를 보여 주고 있다.

각 노드는(사이트 이름, 방문횟수, 체류시간)의 3가지 값을 가지고, 아크는 노드간 이동횟수를 값으로 가지고 있다. 여기에서 아크의 이동횟수는 특정 노드에서 다른 노드로 이동한 횟수를 의미한다. 노드에서 사이트 이름은 도메인 네임을 이용하여 ccTLD(nTLD라고도 함)이면 상위 3단계 이름(x.y.z.kr이면 y.z.kr이 됨)으로 gTLD이면 상위 2단계 이름(x.y.com이면 y.com이 됨)으로 정의하기로 한다(자세한 내용은 www.icann.org 참조). 이처럼 단축시킨 도메인 네임을 편의상 SDN(Short Domain Name)이라 부르기로 한다.

새로운 웹 활용 패턴 그래프를 정의하기 위하여 자체 구축한 WUPG를 활용하여 사용자별 data를 수집하였다. 데이터의 수집기간은 2004년 3월 29



<그림 1> WUPG 시스템과 WUPG의 예

일부터 2004년 4월 30일까지이며, WCA는 모대학 실습실의 PC에 설치하였다. WCA는 새로운 웹 페이지를 접속할 때마다 (IP 주소, 접속한 URL, 접속한 시간)의 정보를 WS에 레코드 형태로 보내게 되며, 동 측정기간 동안 수집된 레코드의 수는 모두 56,578개로서 WA를 통해 분석된 결과로 WUPG 및 각종 통계자료를 추출하게 된다.

3.2 웹 패턴 그래프 분석

WS에 저장된 레코드들을 이용하여 WA를 통해 분석한 결과들을 살펴보자.

3.2.1 페이지 당 머무른 시간

웹 사용자가 특정 웹페이지에서 머무른 시간과 관련된 정보를 정리한 것으로, 1회 이상 접속이 이

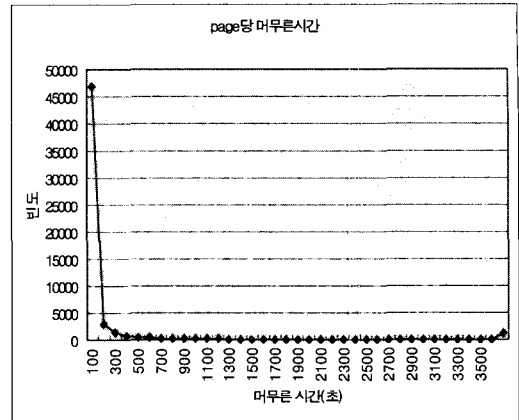
루어진 페이지의 수는 총 55,989개로 나타났다. 이에 대한 빈도 분석을 한 결과를 살펴보자. <표 1>은 0에서 10초 사이에 머무른 페이지 수에 대한 빈도를 살펴보면 다음과 같다.

아래 표를 보면 전체 접속 시간 빈도의 거의 과반수(48%)가 10초 이내에 페이지를 바꾸거나 종료하는 것으로 나타나고 있다. 이처럼 머무른 시간이 10초 이내에 집중되는 것은 “뒤로” 버튼의 실행, 브라우저 간 이동, 자주 접속하는 페이지간의 습관적 이동, 임의로 접속 끊기 등과 같이 실제 정보의 수집과는 관련이 없거나 단순히 페이지에 대한 제목 살펴보기와 같은 행위들에 기인한 것으로 보인다. 본 연구에서는 3초 이내의 페이지 간 이동 및 접속을 “단순이동”이라 정의하고, 이들은 실제 웹의 정보 수집활동과는 무관한 것으로 보고 제외하기로 한다.

<표 1> 페이지 당 머무른 시간의 빈도

초	빈도	상대 빈도	누적 빈도	초	빈도	상대 빈도	누적 빈도
0	3464	0.0619	0.0619	6	1840	0.0329	0.3791
1	4595	0.0821	0.1439	7	1697	0.0303	0.4094
2	3714	0.0663	0.2103	8	1438	0.0257	0.4351
3	3153	0.0563	0.2666	9	1303	0.0233	0.4584
4	2403	0.0429	0.3095	10	1288	0.0230	0.4814
5	2058	0.0368	0.3463	-	-	-	-

전체 페이지에서 머무른 시간의 분포를 살펴보면 <표 2>와 같이 나타났다. 아래의 표로부터 페이지 당 머무른 시간의 분포를 살펴보면, 머무른 시간이 300초(5분) 이내인 경우는 누적빈도가 91.19%이고 900초(15분) 이내인 경우가 95.38%, 1500초(25분) 이내인 경우가 96.6%를 보이고 있다. 거의 대다수 사용자들이 15분 이내동안 특정 페이지에 머무르는 것으로 나타났다.



<그림 2> 전체 페이지 당 머무른 시간의 빈도

아래 표를 그림으로 나타내면 위의 <그림 2>와 같다. 그림에서 보듯이 페이지별 머무른 시간과 빈도간의 관계는 매우 가파른 지수분포의 형태를 보이고 있다.

<그림 2>에서 3600초(1시간)를 초과하는 경우가 1.93%를 차지하고 있는데, 이는 실제 사용자가

<표 2> 전체 페이지 당 머무른 시간의 분포

초	빈도	상대빈도	누적빈도	초	빈도	상대빈도	누적빈도
100	46836	0.8365	0.8365	1900	53	0.0009	0.9706
200	2909	0.0520	0.8885	2000	50	0.0009	0.9715
300	1314	0.0235	0.9119	2100	46	0.0008	0.9723
400	734	0.0131	0.9251	2200	48	0.0009	0.9731
500	511	0.0091	0.9342	2300	46	0.0008	0.9740
600	376	0.0067	0.9409	2400	37	0.0007	0.9746
700	286	0.0051	0.9460	2500	31	0.0006	0.9752
800	236	0.0042	0.9502	2600	34	0.0006	0.9758
900	202	0.0036	0.9538	2700	39	0.0007	0.9765
1000	157	0.0028	0.9566	2800	31	0.0006	0.9770
1100	152	0.0027	0.9593	2900	25	0.0004	0.9775
1200	124	0.0022	0.9616	3000	28	0.0005	0.9780
1300	87	0.0016	0.9631	3100	25	0.0004	0.9784
1400	90	0.0016	0.9647	3200	27	0.0005	0.9789
1500	71	0.0013	0.9660	3300	27	0.0005	0.9794
1600	57	0.0010	0.9670	3400	27	0.0005	0.9799
1700	78	0.0014	0.9684	3500	30	0.0005	0.9804
1800	68	0.0012	0.9696	3600	17	0.0003	0.9807
-	-	-	-	3600 초과	1080	0.0193	1.0000

특정 페이지에 접속한 후 브라우저를 사용하지 않고 계속 방치해 둔 경우로 볼 수 있다. 이러한 부분을 효과적으로 제거할 수 있다면, 페이지별 머무른 시간에 대한 보다 정확한 정보를 획득할 수 있을 것이다. 이와 관련된 논의는 다음 절에서 자세히 다루도록 하겠다.

이런 점을 감안하더라도 페이지별 머무른 시간에 대한 빈도가 정규분포를 보이지 않고, 지수분포를 보인다는 것은 사용자들의 웹 이용형태는 적은 시간 내에 많은 정보를 획득하는 경향을 보인다고 볼 수 있다.

3.2.2 세션 분석

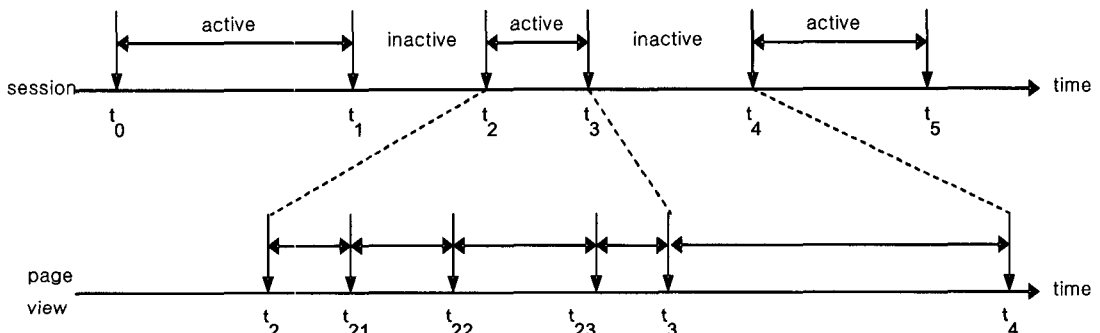
세션(session)은 일반적으로 어떠한 작업이나 서비스의 시작과 끝을 나타내는 것이다. 전화호의 경우 세션은 통화의 연결 시점부터 통화의 단절 시점까지로 정의된다. 그러나 웹 이용의 경우 웹 서핑의 시작과 끝을 하나의 세션으로 정의하려면 많은 어려움이 있다. 사용자의 경우 웹을 이용하다가 장시간 다른 작업을 하기도 하고, 웹 접속 중간 중간에 다른 작업을 수행하는 일이 많으므로 사용자 자신도 어느 시점이 웹 서핑의 시작이고 끝인지를 대답하기가 쉽지 않다. 본 연구에서는 웹 이용과 관련된 세션을 정의하기 위해 on-off 모형을 이용하기로 한다.

아래의 <그림 3>에서 두 개의 시간축인 “session”과 “page view”가 있다. “session” 시간 축에

서 active한 상태는 on을 나타내고, inactive한 상태는 off를 나타내고, active한 상태가 지속된 시간 간격을 세션으로 정의하기로 한다.

두 개의 시간 축에서 t_i 들은 사건(event)이 발생한 시간으로, 이는 새로운 웹 페이지로의 접속이 발생한 시점을 나타낸다. <그림 3>에서 t_2 와 t_4 사이의 세션을 확대한 “page view” 축을 보면, 사용자는 $t_2, t_{21}, t_{22}, t_{23}, t_3$ 의 시점에 새로운 웹 페이지에 접속하였다. 이때 $t_{21}-t_2$ 는 웹 페이지를 다운로드한 시간(d_time)과 웹 페이지 열람 시간(v_time)을 합한 것이 된다. active 상태와 inactive 상태를 구분하기 위해 본 연구에서는 d_time+v_time의 합이 정해진 임계치(threshold)를 초과하게 되면, inactive한 상태가 d_time+v_time 만큼 지속된 것으로 한다. 위의 그림에서 t_4-t_3 는 임계치를 초과하였으므로, t_2 에서 시작한 세션은 t_3 에서 끝나고, t_4 시점부터는 새로운 세션이 시작됨을 보여주고 있다. 본 연구에서는 inactive를 정의하는 임계치를 “300초”, “600초”, “900초”, “1200초”, “1500초”의 5가지로 구분하고, 이들과 관련된 분석을 수행하기로 한다. 세션이 시작되어서 끝나는 시간 간격을 “세션 지속 시간”이라 부르기로 한다.

임계치가 설정되면 WS에 저장된 레코드에서 동일한 PC에서 발생한 레코드들을 이용하여 세션을 추출할 수 있고, 세션 내에서 접속한 페이지 및 페이지에서의 머무른 시간을 쉽게 계산할 수 있다.



<그림 3> 세션 on-off 모형

〈표 3〉 임계값에 따른 세션 분포

구분	300초	600초	900초	1200초	1500초
전체 세션 수	4349	3054	2420	2043	1819
30초 미만을 제외한 세션 수	3486	2633	2148	1840	1657
최대 세션 지속 시간(초)	6307	12366	13233	15779	18071
세션 지속 시간 평균(초)	428.9	832.6	1275.8	1735.3	2130.7
최대 페이지 수	887	1383	1395	1421	1421
평균 접속 페이지 수	29.6	40.7	50.6	59.5	66.4
페이지 당 머무른 시간(초)	14.5	20.5	25.2	29.2	32.1

〈표 3〉은 세션과 관련된 통계들을 보여주고 있는데, 표에서 “전체 세션수”는 설정된 임계값에 대하여 발생한 세션의 전체수를 나타내고, “30초미만을 제외한 세션 수”는 “전체 세션수”에서 “세션 지속 시간”이 30초 미만인 세션들을 제외한 것이다. “최대 세션 지속 시간”은 주어진 임계 값에 대해서 “세션 지속 시간”이 가장 큰 값을 가지는 세션을 나타내고 있다. “세션 지속 시간 평균”은 “30초미만을 제외한 세션 수”를 대상으로 “세션 지속 시간”에 대한 산술 평균값이다. “최대 페이지 수”는 “30초미만을 제외한 세션 수”를 대상으로 가장 많은 페이지에 접속한 세션의 접속 페이지 수이다. “평균 접속 페이지 수”은 “30초미만을 제외한 세션 수”를 대상으로 세션이 접속한 페이지 수에 대한 산술평균이다. “페이지 당 머무른 시간”은 “30초미만을 제외한 세션 수”를 대상으로 세션에서 페이지당 머무른 시간에 대한 산술평균이다.

〈표 3〉에서 임계값이 300인 경우 평균 접속 페이지 수는 29.6페이지이고, 페이지당 머무른 시간은 14.5초이다.

〈표 4〉는 정의된 5가지 임계값에 대한 “세션 지속 시간”을 구하고, 이에 대한 빈도를 구한 결과를 보여주고 있다. 표에서 보듯이 5가지 경우 모두 전형적인 지수분포의 형태를 보이고 있다. 또한 임계값이 “1500초”에서 “300초”로 작아질수록 매우 가파른 지수분포의 형태를 보이고 있다. 이를 도시하면 〈그림 4〉와 같은데 “세션 지속 시간”에 대한 빈도 분석 결과는 “1500초”인 경우는 긴 꼬리 형태의 지수분포를 보이고 있고, “300초”

로 줄어들수록 매우 가파른 형태의 지수분포를 보이고 있다. 즉 세션은 임계값을 작게 설정하면 거의 지수분포에 가까워지고, 임계치가 커지면 긴꼬리 형태의 근사된 지수분포를 보이게 된다.

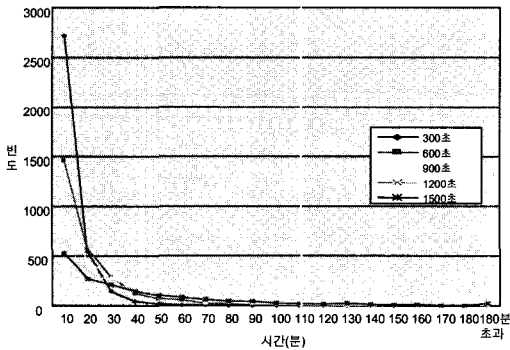
또한 〈표 5〉와 같이 5가지 임계값에 대하여 세션이 접속한 페이지 수에 대한 빈도를 보여주고 있는데 이 역시도 매우 전형적인 지수분포를 보임을 알 수 있다. 이를 그래프로 나타내면 다음 페이지의 〈그림 5〉와 같다.

〈표 4〉 임계값에 따른 세션 지속시간

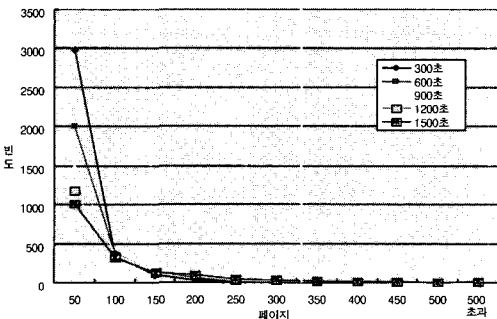
세션 지속 시간(분)	임계값				
	300초	600초	900초	1200초	1500초
10	2720	1460	908	645	531
20	544	567	448	351	270
30	143	291	288	245	214
40	43	126	168	153	142
50	19	74	98	108	102
60	7	57	82	86	84
70	5	24	36	52	64
80	1	18	40	43	46
90	1	7	28	38	47
100	2	3	15	29	27
110	1	2	12	20	18
120	0	1	7	13	17
130	0	0	4	19	23
140	0	0	6	13	17
150	0	1	4	10	10
160	0	0	1	3	9
170	0	0	1	3	6
180	0	1	1	2	7
180분 초과	0	1	1	7	23

<표 5> 임계값에 따른 접속 페이지 빈도

페이지수	300초	600초	900초	1200초	1500초
50	2985	2006	1487	1175	1007
100	349	373	362	340	315
150	92	147	141	131	125
200	28	61	86	98	93
250	12	16	29	36	38
300	10	14	19	28	35
350	2	6	8	13	21
400	3	5	8	10	8
450	0	1	2	3	8
500	1	0	1	1	1
500 초과	4	4	5	5	6



<그림 4> 임계값에 따른 세션 지속시간 분포도



<그림 5> 임계값에 대한 접속 페이지 빈도 분포도

3.2.3 네트워크의 구성 및 분석

① 네트워크의 구성

노드는 앞에서 정의한 SDN(Short Domain

Name)을 가지고 구성하게 된다. 이때 x.abc.com, y.abc.com, z.abc.com이 있으면 이로부터 한 개의 노드인 abc.com이 만들어짐에 주의하자. 56,578개의 레코드로부터 추출된 노드의 수는 총 1,638개였다. 또한 설정된 임계치에 대한 세션이 만들어지고, 세션 내에서 어떤 노드에 속한 페이지에서 다른 노드에 속한 페이지로의 이동이 발생하게 되면, 이 두 노드 간에는 아크의 방문횟수 값이 1만큼 증가하게 된다.

② 노드 정보 분석

1,638개의 노드 중에서 머무른 시간의 합이 10초를 초과하는 노드들의 수는 1,177개이다. 1,638개의 노드들에 머무른 시간의 총합(특정 페이지에 머무른 시간이 1500초를 초과하는 경우는 제외)은 3,524,421초이고, 10초 이내인 노드들에 머무른 시간의 총합은 1,844초로서 전체 시간에서 차지하는 비율은 0.523%에 불과하므로 분석에서 제외하기로 한다. 10초를 초과하는 1,177개의 노드에서 머무른 시간 순으로 상위 20위까지를 살펴보면 <표 6>과 같다.

<표 6>에서 “순위”는 “머무른 시간”이 가장 많은 순서, “SDN”은 노드의 이름, “머무른 시간”은 사이트에 머무른 시간, “inout”은 노드로 들어오고 나간 횟수, “sum in”은 외부에서 들어온 횟수, “sum out”은 외부로 나간 횟수, “first”는 세션 시작 시에 처음으로 접속하는 시작 페이지의 횟수, “last”는 세션 마지막에 선택한 횟수, “in degree”는 네트워크에서 들어오는 아크가 있는 노드의 수, “out degree”는 네트워크에서 나가는 아크가 있는 노드의 수를 각각 나타낸다. “평균”은 페이지당 머무른 시간의 평균으로 “머무른 시간”에서 “in_out”을 나눈 값, “빈도”는 “머무른 시간”을 총 접속시간(3,524,421초)으로 나눈 값, “누적 빈도”는 “빈도”를 누적시킨 값이다. “머무른 시간”은 국내에서 가장 유명한 포털인 daum.net, nate.com, naver.com의 순서를 보이고 있다. “평균”은 mocie.go.kr(산업자원부), mct.go.kr(문화관광부), dcinside.com(디

〈표 6〉 머무른 시간 기준 상위 20개 노드(사이트) 접속정보

순위	SDN	머무른 시간	inout	sumin	sum out	first	last	in deg	out deg	평균	빈도	누적 빈도
1	daum.net	940496	13389	3860	3624	326	558	427	435	70.2	0.267	0.267
2	nate.com	522351	6907	2090	2307	435	227	238	228	75.6	0.148	0.415
3	naver.com	418471	7517	3015	3001	221	258	639	663	55.7	0.119	0.534
4	damoim.net	202287	5196	667	613	26	81	57	52	38.9	0.057	0.592
5	ks.ac.kr	174504	2227	1187	1135	95	156	145	148	78.4	0.050	0.641
6	bugs.co.kr	112598	2110	529	475	40	97	80	66	53.4	0.032	0.673
7	dreamwiz.com	77856	1785	943	1320	427	40	145	157	43.6	0.022	0.695
8	yahoo.com	67255	1403	537	518	28	33	134	157	47.9	0.019	0.714
9	x-y.net	43261	450	351	277	15	100	41	38	96.1	0.012	0.726
10	sayclub.com	23528	650	199	183	4	29	34	38	36.2	0.007	0.733
11	mofat.go.kr	18292	95	83	79	0	2	21	23	192.5	0.005	0.738
12	magicn.com	17220	406	152	148	0	2	21	16	42.4	0.005	0.743
13	auktion.co.kr	16294	145	134	125	0	6	38	29	112.4	0.005	0.748
14	empas.com	16253	482	193	175	4	15	51	57	33.7	0.005	0.752
15	mym.net	13566	171	57	120	77	2	25	19	79.3	0.004	0.756
16	unikorea.go.kr	13396	74	53	50	0	3	15	13	181.0	0.004	0.760
17	dcinside.com	13210	66	60	58	3	4	14	10	200.2	0.004	0.764
18	mocie.go.kr	11868	55	50	54	0	0	22	20	215.8	0.003	0.767
19	mct.go.kr	11272	50	44	43	1	3	14	12	225.4	0.003	0.770
20	ybmsisa.com	11212	136	53	50	0	0	13	17	82.4	0.003	0.774

지털카메라), mofat.go.kr(외교통상부)의 순서를 보이고 있다. 이는 학생들이 리포트 작성을 위해 정부 부서 홈페이지를 많이 이용하는데 기인하고, dcinside.com의 경우 디지털카메라와 관련된 사이트이므로 해상도 높은 사진들이 많기 때문에 평균 머무른 시간이 높게 나오고 있다.

처음으로 접속하는 사이트를 나타내는 “first”의 경우 nate.com, dreamwiz.com, daum.net의 순서를 보이고 있는데, ks.ac.kr(경성대)와 mym.net(검색포털)이 상대적으로 높게 나타나고 있다. 이는 수집된 데이터가 본 대학이므로 시작페이지가 본 대학관련 사이트인 ks.ac.kr이 높고 mym.net의 경우 큰 용량의 무료 e-mail(100MB) 등을 제공하는 검색 포털 이어서 상대적으로 시작페이지 횟

수가 높게 나타나고 있다. “last”의 경우 daum.net, naver.com, nate.com의 순서를 보이고 있다. 검색사이트인 naver.com의 경우 머무른 시간에서는 3위이지만 in degree와 out degree가 가장 큰 값을 보이고 있다. 이러한 현상은 dreamwiz.com, yahoo.com, empas.com과 같은 검색 사이트에서도 나타나고 있다. “누적빈도”에서 상위 3개 사이트가 53.4%를 보이고 있고, 상위 20개가 77.4%로서 사이트의 집중도가 매우 높은 부익부 빈익빈 현상(rich-get-richer phenomenon)을 보이고 있음을 알 수 있다.

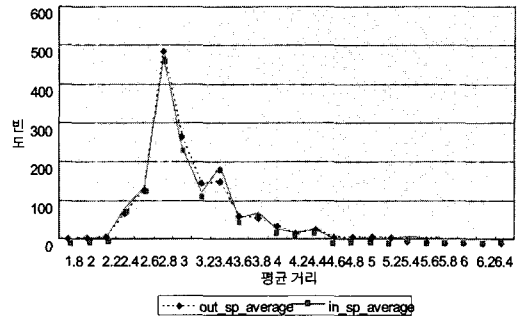
③ 네트워크 분석

1,638개의 노드로 구성된 네트워크에서 1,451개의

노드가 양방향으로 연결된 네트워크(forest)를 구성하고 있다. 나머지 187개의 노드들은 대부분 머무른 시간이 매우 적거나 중요하지 않은 노드들로 구성되어 있고, 이들 노드에 총 머무른 시간은 8,288초로 전체 분석에서 0.235%에 불과하다. 네트워크 분석에서는 1,451개의 노드로 구성된 네트워크를 대상으로 하기로 한다. 노드 간 링크의 거리는 모두 1로 가정하기로 한다. 노드들 간의 최단 경로를 알기 위해서는 “depth-first tree”를 이용하면 된다. 주어진 노드에 대한 “depth-first tree”는 모든 노드로의 최단 경로가 된다.

본 연구에서는 주어진 특정 노드와 나머지 노드간의 최단경로의 평균(out_sp_average)과 특정 노드로 들어오는 최단경로의 평균(in_sp_average)을 구하였다. 다음의 <그림 6>은 out_sp_average와 in_sp_average를 가지고 구한 빈도를 보여주고 있다. 아래의 그래프에서 보듯이 out_sp_average와 in_sp_average는 빈도에 있어서 거의 차이를 보이지 않고 있다. 1,451개의 노드에 대한 out_sp_average에 대한 평균값은 2.9761이고, in_sp_average에 대한 평균값은 2.9675로 나타나고 있다. 이는 Albert 등[2]이 URL 분석을 통한 웹 분석에서 정리한 사이트 간 평균 거리가 18.59인데 반하여, 실제 사용자들의 웹 이용 패턴으로부터 도출한 WUPG에서의 평균 거리는 3을 초과하지 않는 것으로 나타나고 있다

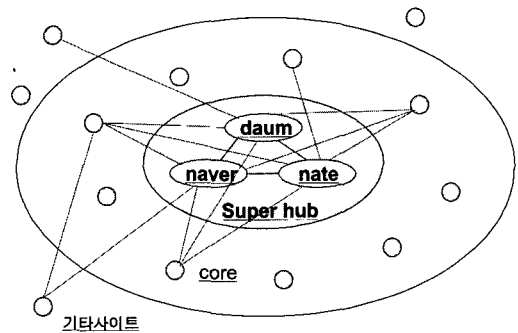
평균거리가 3을 넘지 않는다는 것은 사용자들이 특정한 몇 개의 사이트에 집중하여 웹을 사용하고 있음을 보여주는 것이다. 즉 자주가는 사이트들의 경우 하이퍼링크 정보를 이용하지 않고, URL을 직접입력하거나 아니면 즐겨찾기에 등록하여 이용한다는 것을 보여주고 있다. 특히 서로 다른 사이트간 이동시에 웹 고유의 속성인 하이퍼링크를 이용하는 경우는 매우 드물다는 것을 보여주고 있다. 물론 사이트 내에서의 이동은 대부분 하이퍼링크를 이용하겠지만, 사이트간의 이동에는 하이퍼링크를 거의 이용하지 않는다고 볼 수 있다.



<그림 6> 페이지 간 평균거리 분포

3.3 WUPG에 의한 웹의 특성

WUPG의 분석 결과를 바탕으로 사용자들의 웹 이용 패턴 그래프를 개념적으로 기술하면 <그림 7>과 같이 볼 수 있다. 즉 WUPG는 크게 세부분인 슈퍼허브(super hub), 코어(core), 기타사이트로 나누어 볼 수 있다.



<그림 7> WUPG의 구조

여기서 슈퍼허브는 daum, naver, nate와 같이 전체 트래픽의 10%이상을 차지하는 사이트이고, 코어는 트래픽 비중이 어느 정도(예를 들면 0.1%) 이상인 사이트들이라 할 수 있다. <그림 7>에서 처럼 슈퍼허브와 코어를 중심으로 거의 완전연결(full mesh) 형태를 보이면서, 나머지 사이트들은 이들을 통해 연결되는 것이다. 즉 웹은 그 자체가 하이퍼링크된 구조를 보이고 있지만, 사용자들의 이용형태는 슈퍼허브 및 코어를 주로 이용하는 형

태를 보이고 있는 것이다.

4. 결 론

본 연구는 기존에 연구되어 있는 물리적인 링크로 구성된 웹 그래프와는 달리 웹 사용자들의 논리적인 웹 이용패턴을 반영할 수 있는 웹 패턴 그래프(Web Usage Pattern Graph, WUPG)를 새롭게 정의하였다. 정의된 WUPG 시스템을 활용하여 사용자의 실질적인 웹 이용과 관련된 정보를 수집하고, 이로부터 사용자들의 웹 이용형태에 대한 분석 및 WUPG 자체와 관련된 분석을 실시하였다. 이러한 분석결과로부터, 본 연구에서 제안한 WUPG는 기존 웹 그래프에 대한 분석과는 차별되는 몇 가지 새로운 결과를 보여주고 있다.

이러한 WUPG로부터 웹의 실질적인 이용과 관련된 부분을 다음의 세 가지로 요약할 수 있을 것이다.

① 웹 이용 평균 거리의 감소

WUPG에서 보듯이 슈퍼허브와 코어들이 거의 완전연결 형태(full mesh)를 보이고 있고, 기타 사이트들은 슈퍼허브 및 코어에 직접 연결되어 있으므로 웹 이용 평균 거리는 매우 짧아짐.

② 외부 하이퍼링크 의존성의 감소

사이트간의 이동시 하이퍼링크를 이용하기 보다는 검색엔진, 자주 가는 사이트, URL의 직접 입력 등을 이용하므로, 외부 하이퍼링크의 의존성은 줄어들음.

③ 슈퍼허브의 등장

방문횟수 및 트래픽의 집중도가 극단적으로 높은 허브인 슈퍼허브는 거의 백화점식 정보제공으로 인해 집중도가 점점 높아지고 있음. 따라서 집중화의 정도가 너무 극단적이어서, 코어에 속한 사이트가 슈퍼허브가 되기는 매우 힘들 것으로 보임.

참 고 문 헌

- [1] Adamic L., "Zipf, Power-laws, and Pareto - a ranking tutorial," Xerox Palo Alto Research Center.
- [2] Albert R., H. Jeong, and A.-L. Barabasi, "Diameter of the World Wide Web," *Nature*, Vol.401(1999), p.130.
- [3] Barabasi A.-L. and R. Albert, "Emergence of scaling in random networks," *Science*, Vol.286(1999), pp.509-512.
- [4] Barabasi A.-L., R. Albert and H. Jeong, "Scale-free characteristics of random networks: the topology of the world-wide web," *Physica A*, Vol.281(2000), pp.69-77.
- [5] Barabasi A.-L., R. Albert, H. Jeong. and G. Bianconi, "Power-law distribution of the World Wide Web," *Science*, Vol.287(2000), p.2115a.
- [6] Bharat K., B.-W. Chang, M. Henzinger and M. Ruhl, "Who links to whom: Mining linkage between web sites," *Proceedings of the IEEE International Conference on Data Mining*, November 2001.
- [7] Broder A., R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph Structure in the Web," *The 9th International World Wide Web Conference*.
- [8] Chakrabarti S., B. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg "Mining the Web's link structure," *Computer*, Vol.32, No.(8), 1999.
- [9] Chakrabarti S., B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan, "Automatic resource compilation by analyzing hyperlink structure and associated text," *Proc.*

- 7th WWW, 1998.
- [10] Chen M.S., J.S. Park, and P.S. Yu, "Efficient data mining for path traversal patterns," *IEEE Transactions on Knowledge and Data Engineering*, Vol.10 No.2(1998), pp.209-221.
- [11] Kim, B.C., Yoon, S. Han, and H. Jeong, "Path finding strategies in scale-free network," *Physical Review E*, Vol.65(2002).
- [12] Kleinberg J., "Authoritative sources in a hyperlinked environment," *J. of ACM*, Vol.46, No.5(1999), pp.604-632
- [13] Kleinberg J., R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tonkins, "The Web as a Graph : Measurements, Models, and Methods," *COCOON*, (1999), pp.1-17.
- [14] Kleinberg J. and S. Lawrence, "The Structure of the Web," *Science*, Vol.294, No.30(2001), Nov.
- [15] Lawrence S. and C. Lee Giles, "Accessibility of Information on the Web," *Nature*, Vol.400, No.8(July 1999).
- [16] Lawrence S. and C. Lee Giles, "Searching the World Wide Web," *Science*, Vol.280, No.3(April 1998)
- [17] Srivastava J., R. Cooley., M. Deshpande, and P. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," *ACM SIGKDD Explorations*, (Jan 2000).