웹 문서 변경 예측☆

Estimation of Web Page Change Behavior

김 성 진*
Sung-Jin Kim

요 약

본 논문은 웹 문서의 다운로드 가능 여부와 내용 변경 여부를 예측하는 도구를 기술한다. 웹 데이터베이스 관리자는 자신이 관리하는 웹 문서 집합을 최신 상태로 유지하려고 할 때, 예측 도구를 통하여 다운로드되지 않거나 변경되지 않았을 웹 문서에 대한 불필요한 요청을 감소시킬 수 있다. 본 논문에서는 웹 문서들의 과거 변경이 미래 변경과 매우 밀접한 관련이 있음을 가정한다. 본 논문에서는 약 300만개의 웹 문서들을 2일 주기로 100일 동안 관찰하여 변경 경향을 분석하고, 관찰된 문서들의 다운로드 가능 여부와 내용 변경 여부를 예측한다. 예측 결과는 실제의 변경 사실과 비교 평가되었다.

Abstract

This paper presents the estimation methods computing the probabilities of how many times web pages are downloaded and modified, respectively, in the future crawls. The methods can make web database administrators avoid unnecessarily requesting undownloadable and unmodified web pages in a page group. We postulate that the change behavior of web pages is strongly related to the past change behavior. We gather the change histories of approximately three million web pages at two-day intervals for 100 days, and estimate the future change behavior of those pages. Our estimation, which was evaluated by actual change behavior of the pages, worked well.

☞ Keyword : 웹 데이터베이스(web database), 웹 문서 변경(web page change), 웹 진화(web evolution)

1. 서론

많은 웹 어플리케이션들은 - 예를 들어, 웹 검색 엔진(web search engine), 메타 검색 엔진(meta search engine), 프록시 서버(proxy server) 등 - 내부적으로 웹 데이터베이스를 생성하여 사용자들이 웹 문서를 조회하고 검색 가능 하도록 한다. 웹은 매우 동적이며, 웹 데이터베이스 관리자는 자신이 관리하는 웹 문서들이 최신 상태로 유지되도록 하기 위해, 변경된 웹 문서를 새롭게 다운로드하고자 한다. 그러나 웹 문서의 변경 여부는 실제 다운로드하여 저장된 문서와 비교하기 전에는 알 수가 없다. 따라서 변경되지 않은 웹 문서를 반복적으로 다운로드 하거나 다운로드되지 않을 웹 문서를 반복적으로 요청하는 경우가 빈번하게 발생한다.

효과적 변경 예측 모델은 웹 데이터베이스의 갱신 전략 수립에 매우 중요하다. [1, 2, 3]은 포아송(Poisson) 프로세스 모델에 근거한 예측 방법을 기술하였다. 포아송 프로세스 모델은 기억없는(memoryless) 프로세스 모델이다. 즉, 웹 문서들의 과거 변경 사실을 현재의 변경 사실과 독립적 사건으로 가정한다. 그러나 [1]은 웹 문서의 변경이 과거의 변경 사실과 관련이 있음을 실험적으로 보이고 인정하였다. [4]는 포아송 프로세스의 변경 가정을 따르지 않고 웹 문서들의 과거 변경 기록을 기반으로 하여 실제 웹과 웹 데이터베이스의 내용이 다른 웹 문서의 수를 최소화 하는 방법을 제안하였다.

* 정 회 원 : UCLA 박사후과정연구원
sjkim@cs.ucla.edu
[2007/07/15 투고 - 2007/07/18 심사 - 2007/08/07 완료]
☆ 이 논문은 2006년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF-2006-214-D00136)

본 논문에서는 웹 관리자들이 웹 데이터베이스를 다수의 그룹으로 나누어 관리할 때, 각 그룹의 효과적 유지 관리에 필요한 변경 예측 방법을 기술한다. 웹 데이터베이스 관리자는 문서를 지역성(locality), 도메인(domain), 주제(theme) 등의 기준으로 하여 다수의 그룹(group or category)으로 나누고, 각 그룹을 서로 다른 변경 전략으로 유지 관리할 수 있다. 본 예측의 주요 목적은 관리자들이 다운로드되지 않을 웹 문서나 변경되지 않았을 웹 문서에 대한 불필요한 요청을 최소화하는 것이다. 본 논문에서는 웹 문서의 현재 변경이 과거의 변경과 밀접하게 관련이 있음을 가정한다. 본 논문에서는 기존의 연구에서 다루어 지지 않은 웹 문서의 다운로드 상태 변경을 관찰하고 예측한다. 웹 문서의 다운로드 불가능 상태는 매우 빈번한 현상임에도 불구하고[5, 6, 7, 8] 현재까지 이에 대한 연구는 매우 미비하다.

본 논문에서 두 가지 예측 도구-$P(_{Y=a,N=b}DR_{Y=c,N=d})$과 $P(_{Y=a,N=b}MR_{Y=c,N=d})$ -가 기술된다. $P(_{Y=a,N=b}DR_{Y=c,N=d})$은 $(a + b)$번의 다운로드 요청으로 부터 a번의 성공적인 다운로드와 b번의 다운로드 실패가 있었던 웹 문서가 향후 $(c + d)$번의 다운로드 요청에서 c번 다운로드 성공과 d번의 다운로드 실패가 발생할 확률이다. $P(_{Y=a,N=b}MR_{Y=c,N=d})$은 과거 $(a + b)$번의 성공적 다운로드로부터 a번의 문서 내용 변경이 있었던 웹 문서가 향후 $(c + d)$번의 성공적인 다운로드에서 c번의 문서 변경이 발생할 확률이다. 본 논문에서는 웹 문서의 다운로드율과 변경율을 정의한다. 다운로드율은 하나의 웹 문서가 일관적으로 다운로드 되는 정도를 나타낸다. 변경율은 한 웹 문서가 빈번하게 변경되는 정도를 나타낸다. 다운로드율의 분포와 변경율의 분포는 $P(_{Y=a,N=b}DR_{Y=c,N=d})$과 $P(_{Y=a,N=b}MR_{Y=c,N=d})$의 계산에 사용된다.

예측 방법 평가를 위하여 34,000개 국내 웹 사이트(web site)가 선정되었다. 각 사이트는 100일간 2일 주기로 모니터링(monitoring)되고, 그룹 각각의 다운로드율 분포와 내용변경율 분포가 산출되었다. 산출된 분포에 근거하여 향후 5번을 2일 주기로 수집할 때 발생할 수 있는 변경 여부를 예측하고 예측된 값과 실제 값이 비교 평가된다.

본 논문은 다음과 같이 구성되었다. 2장에서는 웹 문서들의 변경을 표현하는 도구로서 다운로드율과 내용 변경율을 정의한다. 3장에서는 실제 웹 사이트들을 관찰하여 다운로드율과 내용 변경율의 분포를 수집한다. 4장에서는 웹 문서의 다운로드 성공과 내용 변경을 예측하는 도구를 제시하고, 5장에서 결론을 맺는다.

2장. 다운로드율과 변경율

본 장에서는 웹 문서의 다운로드 가능 여부와 내용 변경의 관점에서 웹 문서 변경을 표현하는 데에 사용되는 다운로드율과 내용 변경율을 정의한다. (그림 1)은 간단한 문서 수집 예를 나타낸다. 4개의 URL A, B, C, D에 대해 총 16번의 문서 수집이 수행된다. '-'은 다운로드 요청이 없었음을 의미한다. 예를 들어, 웹 관리자는 자신의 고유한 문서 요청 정책에 해당 URL을 요청하지 않을 수 있다. '●'은 URL에 해당하는 웹 문서에 대해 다운로드 요청이 있었으나 다운로드가 실패되었음을 의미한다. 웹 문서가 성공적으로 다운로드 되었을 때, 다운로드된 문서의 내용은 알파벳 원문자로 표현된다. 예를 들어, 4번째 수집을 살펴보자. URL A, C, D에 해당하는 3개의 웹 문서를 요청하고, 그 중 URL A와 C에 대한 웹 문서를 성공적으로 다운로드 하였으며, 문서 내용은 ⓓ와 ⓢ로 표현되었다. URL D에 대한 문서는 다운로드를 시도하였으나 실패하였다. 본 논문에서 각각의 URL들은 개별적인 웹 문서를 나타내는 것으로 간주한다.

정의 1. 웹 문서의 "최초 요청 번호"는 해당 웹 문서에 대한 다운로드 요청이 최초로 발생한 수집 번호이다. 웹 문서의 "최후 요청 번호"는 해당 웹 문서에 대한 다운로드 요청이 마지막으로 발

	수집 번호															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	ⓐ	ⓑ	ⓒ	ⓓ	ⓔ	ⓕ	ⓖ	ⓗ	ⓘ	ⓙ	ⓚ	ⓛ	ⓜ	ⓝ	ⓞ	ⓟ
B	-	-	-	-	ⓠ	●	ⓡ	-	●	-	-	-	-	-	-	-
C	-	-	ⓢ	ⓢ	-	●	ⓣ	ⓣ	●	-	●	●	-	-	-	-
D	●	ⓤ	-	●	ⓤ	ⓤ	ⓥ	ⓥ	-	ⓦ	●	ⓦ	ⓦ	-	-	ⓦ

(그림 1) 웹 문서 변경 관찰 예

생한 수집 번호이다. 한 문서의 요청율은 "(요청 회수) / ((최후 요청 번호) - (최초 요청 번호) + 1)"로 정의된다.

예를 들어, URL B에 해당하는 웹 문서는 5, 6, 7, 9번째 수집에서 다운로드 요청이 있었다. 따라서 요청율은 (4 / (9 - 5 + 1)) = 0.8이다. 요청율이 낮은 URL은 변경 경향을 예측에 사용되기에 충분치 않은 양의 정보를 포함하고 있다. 따라서 임계값(threshold) 이상의 요청율을 가지는 문서를 대상으로 변경 예측이 고려된다.

정의 2. 웹 문서의 다운로드율은 "(다운로드 성공회수) / (다운로드 요청회수)"로 정의된다. 다운로드 재현율은 "(다운로드 요청회수) / (총 수집 회수)"이다. 다운로드율은 실패 없이 안정적으로 다운로드가 성공되는 정도를 나타낸다. 다운로드 재현율은 다운로드 요청의 빈번한 정도를 나타낸다.

예를 들어, URL C에 해당하는 웹 문서는 3, 4, 6, 7, 8, 9, 11, 12번째 수집에서 요청되었고, 이중 3, 4, 7, 9번째 수집에서 성공적으로 다운로드되었다. 따라서 다운로드율은 (4 / 8) = 0.5이다. 총 16번의 문서 수집 중 8번의 다운로드 요청이 있었으므로, 다운로드 재현율은 (8 / 16) = 0.5이다. 낮은 다운로드 재현율을 가지는 웹 문서들의 다운로드율은 다운로드 성공을 예측하기에 충분한 정보를 포함하고 있지 않다. 따라서 임계값 이상의 다운로드 재현율을 가지는 웹 문서들의 다운로드율이 예측에 고려된다.

정의 3. URL u의 웹 문서 P_u가 있다고 하자. *i*번째 문서수집에서 P_u가 성공적으로 다운로드되었고, 첫 번째부터 (*i*-1)번째 수집 중 P_u가 최소 한 번 이상 성공적으로 다운로드되었다고 하자. 이때, *i*번째 문서 수집에서 P_u의 "현재 내용"은 i번째 수집에서 다운로드된 P_u의 문서 내용이다. i번째 문서 수집에서 P_u의 "이전 내용"은 첫 번째부터 (*i*-1)번째 수집 중 성공적으로 다운로드된 문서 내용 중 *i*의 값이 가장 큰 수집 번호에서 다운로드된 문서 내용이다. *i*번째 수집에서 P_u의 "현재 내용"과 "이전 내용"이 서로 다를 때 P_u는 *i*번째 수집에서 변경되었다고 한다.

정의 4. 웹 문서의 내용 변경율은 "(문서 내용 변경회수) / ((다운로드 성공회수) - 1)"로 정의된다. 웹 문서의 내용 변경 재현율은 "((다운로드 성공회수) - 1) / ((총 수집회수) - 1)"로 정의된다. 변경율은 웹 문서의 내용이 빈번하게 변경되는 정도를 나타낸다. 변경율은 주어진 관찰 기간 동안 최소한 두 번 이상 성공적으로 다운로드된 문서들을 대상으로 계산된다. 해당 웹 문서에 대한 다운로드가 성공적으로 이루어졌을 때, 내용 변경 재현율은 변경 여부 비교 연산의 빈도(frequency)를 나타낸다.

예를 들어, 10번째 수집에서 URL D에 해당하

는 웹 문서의 현재 내용은 ⓦ이고, 이전 내용은 ⓥ이다. 즉, URL D의 웹 문서는 10번째 수집에서 변경이 발생하였다. 12번째 수집에서는 현재 내용과 이전 내용이 서로 같으므로, 해당 웹 문서는 12번째 수집에서 변경되지 않았다. URL D의 웹 문서는 총 9번이 성공적으로 다운로드 되었으며, 이는 다운로드된 문서 내용 간에 8번의 변경 여부 비교 연산 - 2번째와 5번째, 5번째와 6번째, ..., 13번째와 16번째 문서 내용 간의 비교 - 이 있었음을 의미한다. 이 중 7번째와 10번째 수집 번호에서 변경이 발생하였다. 따라서 내용 변경율은 (2 / (9 - 1)) = 0.25이고 변경 재현율은 (8 / 15) = 0.53이다. 낮은 변경 재현율을 가지는 웹 문서들의 내용 변경율은 변경 여부를 예측하기에 충분한 정보를 포함하고 있지 않다. 따라서 임계값 이상의 내용 변경 재현율을 가지는 웹 문서들의 내용 변경율이 예측에 고려된다.

3. 문서 변경 기록

본 장에서는 문서 변경 예측에 필요한 다운로드율과 내용 변경율의 분포를 얻는다. 우선 34,000개의 국내 웹 사이트를 실험 대상으로 선정하였다. 자체 개발된 웹 로봇[7]을 통하여 100일간 2일 주기로 관찰하였다. 선정된 34,000개의 웹 사이트는 4,000개의 유명 사이트와 30,000개의 임의 사이트로 구분되었다. 유명 사이트와 임의 사이트들에 대해 총 50회의 문서 수집이 이루어 졌다. 본 시험에서는 하루에 하나의 그룹에 속하는 웹 문서들을 관찰되었다. 즉, 시험 첫째 날에 유명 사이트 그룹에 속하는 웹 문서들의 변화를 관찰하고 둘째 날에는 임의 사이트 그룹에 속하는 웹 문서들의 변화를 관찰하는 방식으로, 유명 사이트 그룹과 임의 사이트 그룹에 속한 웹 문서들을 교대로 관찰하였다.

시험 대상이 되는 사이트에 대한 과부하를 예방하기 위하여 사이트 당 관찰되는 문서는 3,000개로 제한되고 최대 수집 깊이는 9로 제한되었다. 관찰 대상 사이트에서 로봇 배제 규칙으로 명시된 웹 문서는 수집하지 않았으나, 최상위 문서는 로봇 배제 규칙과 상관없이 관찰 대상에 포함하였다. 각 문서에는 5초의 타임아웃(timeout)이 설정되어 웹 서버와 웹 로봇 사이에 5초 이상의 데이터 전송이 없을 경우 다운로드를 실패한 문서로 간주되었다. 파라미터 값의 포함 여부는 '?' 문자의 유무로 결정된다. 웹 문서의 URL이 파라미터 값을 포함할 경우에 해당 웹 문서는 관찰대상에서 제외되었다.

URL의 요청 전략은 다음과 같다. 34,000개 웹 사이트의 루트 문서(root page) - 혹은 홈페이지(home page), 메인페이지(main page) -가 요청된 이후 다운로드된 루트 문서에 존재하는 URL들을 추출한다. 추출된 URL들이 관찰 대상 웹 사이트의 웹 문서일 경우, 이를 다시 요청하여 다운로드하고 다운로드된 문서에서 새로운 URL들을 또 다시 추출한다. 즉 문서요청, 다운로드, 새로운 URL 추출의 과정을 반복한다. 한 문서 수집 번호에서 동일한 URL에 해당하는 웹 문서가 두 번 이상 요청되지 않는다.

각 문서 수집 단계에서 평균적으로 180만 개의 - 유명 사이트에서 80만개, 임의 사이트에서 100만개의 URL - URL들이 수집되었다. 수집된 URL들에 해당하는 모든 문서들을 요청하였으며, 이 경우 요청된 웹 문서의 22%와 17%에 해당하는 웹 문서를 다운로드하는데 실패하였다. 50회의 문서 수집이후에 총 300만개의 - 유명 사이트들에서 130만 개, 임의 사이트들에서 170만 개 - 유일한(unique) URL을 확보하였다. 확보된 URL 각각에 대하여 요청율, 다운로드율, 다운로드 재현율, 내용 변경율, 내용 변경 재현율을 계산하였다.

관찰 주기 내에 두 번 이상 발생한 변경 사항은 고려되지 않는다. 웹 문서의 다운로드 여부와

(표 1) 다운로드율 분포

유명 사이트				임의 사이트			
다운로드율	비율	다운로드율	비율	다운로드율	비율	다운로드율	비율
0	22.01%	0.50~0.59	0.07%	0	15.92%	0.50 ~ 0.59	0.03%
0.01 ~ 0.09	0.06%	0.60~0.69	0.07%	0.01 ~ 0.09	0.20%	0.60 ~ 0.69	0.08%
0.10 ~ 0.19	0.04%	0.70~0.79	0.09%	0.10 ~ 0.19	0.02%	0.70 ~ 0.79	0.07%
0.20 ~ 0.29	0.06%	0.80~0.89	0.33%	0.20 ~ 0.29	0.03%	0.80 ~ 0.89	0.35%
0.30 ~ 0.39	0.04%	0.90~0.99	9.53%	0.30 ~ 0.39	0.28%	0.90 ~ 0.99	7.34%
0.40 ~ 0.49	0.05%	1	67.66%	0.40 ~ 0.49	0.03%	1	75.65%
총 100.00%				총 100.00%			

문서 내용의 변경은 수집주기보다 빈번하게 발생할 수 있으며 수집 주기보다 빈번하게 변경되는 문서의 변경은 한 번의 변경으로 간주된다. 예를 들어, 본 실험에서는 각 사이트를 2일 주기로 관찰하였으므로, 임의의 웹 문서가 2일 동안 두 번 이상의 변경이 발생하였을 경우에도 한 번의 변경이 있었던 것으로 간주한다.

(표 1)은 확보된 URL중에서 다운로드 재현율이 0.2 이상이고 요청율이 0.9이상인 웹 문서들의 다운로드율 분포를 나타낸다. (표 2)는 확보된 URL중에서 변경 재현율이 최소 0.2 이상이고 요청율이 0.9 이상인 웹 문서들의 변경율 분포를 나타낸다.

문서 내용의 비교 방법은 다양하다. 바이트(byte)-기반 비교, 쉥글링(shingling) 비교[5], TF.IDF 코사인(cosine)[9] 비교 등이 문서 변경 여부 판단에 사용될 수 있다. 본 논문에서는 바이트-기반 비교 방법을 사용한다. 바이트-기반 비교는 두 웹 문서의 내용을 문자 단위로 비교하는 방법으로서 가장 엄격한 웹 문서 비교 방법이다. 즉, 태그를 포함하여 웹 문서 상에서 발생하는 어떠한 사소한 변경까지도 모두 웹 문서 변경으로 간주된다. 예를 들어, 웹 문서에서 광고를 보여주는 하이퍼링크의 내용이 변경될 경우, 웹 문서가 변경되는 것으로 간주된다. 바이트-기반 비교 외의 다른 변경 비교 방법이 사용될 경우 각 문서의 내용 변경율은 (표 2)에 수치보다 전반적으로 낮게 나타날 것이다.

(표 2) 변경율 분포

유명 사이트				임의 사이트			
변경율	비율	변경율	비율	변경율	비율	변경율	비율
0	22.01%	0.50 ~ 0.59	0.07%	0	15.92%	0.50 ~ 0.59	0.03%
0.01 ~ 0.09	0.06%	0.60 ~ 0.69	0.07%	0.01 ~ 0.09	0.20%	0.60 ~ 0.69	0.08%
0.10 ~ 0.19	0.04%	0.70 ~ 0.79	0.09%	0.10 ~ 0.19	0.02%	0.70 ~ 0.79	0.07%
0.20 ~ 0.29	0.06%	0.80 ~ 0.89	0.33%	0.20 ~ 0.29	0.03%	0.80 ~ 0.89	0.35%
0.30 ~ 0.39	0.04%	0.90 ~ 0.99	9.53%	0.30 ~ 0.39	0.28%	0.90 ~ 0.99	7.34%
0.40 ~ 0.49	0.05%	1	67.66%	0.40 ~ 0.49	0.03%	1	75.65%
총 100.00%				총 100.00%			

(표 3) 관련 기호 요약

기 호	설 명
$P(DR(x))$	URL의 '다운로드 성공률'이 x일 확률
$P(MR(x))$	URL의 '변경률' x일이 확률
$P\left(_{Y=a,N=b}DR_{Y=c,N=d}\right)$	a번 다운로드의 성공과 b번의 다운로드 실패가 있었던 URL이 향후 c번의 다운로드 성공과 d번의 다운로드 실패가 있을 확률
$P\left(_{Y=a,N=b}MR_{Y=c,N=d}\right)$	a번 문서 내용이 변경되었고 b번은 내용 변경이 없었던 URL이 향후 c번의 내용변경이 발생하고 d번은 내용 변경이 없을 확률

4. 웹 문서 변경 예측

본 장에서는 3장에서 관찰된 웹 문서의 다운로드율과 내용 변경율에 근거하여, 향후 웹 문서 수집에서 발생하게 될 문서들의 다운로드의 성공/실패와 문서 내용의 변경/유지를 예측한다. 본 논문에서는 다음과 같은 사항들을 전제한다. 첫째, 웹 문서의 현재 변경 여부는 과거의 변경 여부와 밀접한 관련이 있다. 둘째, 다운로드 가능 사건과 내용 변경 사건은 독립된 사건이다. 비록 한 웹 문서가 어느 한 시점에 변경되었다고 하는 것은 다운로드가 성공적으로 이루어 졌다는 것을 의미하지만, 이는 다운로드 성공률이 높은 웹 문서의 내용 변경율이 높거나 혹은 낮을 수 있다는 것을 의미하지 않는다. 셋째, 예측 가능한 기간은 변경 관찰 기간을 초과하지 않는다. 넷째, 관찰 기간 내에 발생한 변경 내용은 동일한 중요도를 가진다. 다섯째, 웹 문서 변경의 관찰 주기와 예측 주기는 동일하다. 즉, 2일을 주기로 웹 문서가 관찰되었을 경우 2일 주기로 예측이 이루어진다.

웹 문서 변경 예측에 사용되는 기호들은 (표 3)에 정의되어 있다. $P(DR(x))$, $P(MR(x))$은 웹 문서의 다운로드율이 x일 확률과 변경율이 x일 확률을 나타내며 (표 1)과 (표 2)를 통해 계산된다. 즉, $P(DR(0))$ - 임의로 선택한 한 웹 문서의 다운로드율이 0%일 확률 - 은 22.01%이고 $P(DR(1))$는 67.66%이다. 0와 1을 제외한 나머지 $P(DR(x))$는 x가 속하는 계급구간의 비율을 구간의 크기로 나눈 값으로 한다. $P(DR(0.5))$는 0.07%/10 = 0.007%가 된다. $P(MR(x))$도 $P(DR(x))$와 동일한 방식으로 산출된다. 다운로드의 성공과 실패가 독립사건이라면 50번의 문서 요청에서 다운로드율이 50%로 나오는 경우의 수가 $_{50}C_{25}=1.26*10^{14}$로 가장 많아야 한다. 즉 URL의 '다운로드 성공률'이 0.5일 확률이 가장 높다. 그러나 (표 2)에서 $P(DR(0.5))$의 비율은 상대적으로 매우 작음을 볼 수 있다. n번의 문서 요청에서 $(n/2)$의 다운로드가 성공하는 경우($_nC_{n/2}$)가 발생할 확률은 $P(DR(0.5))$이다. 즉, n번의 문서 요청에서 x번의 다운로드 성공을 나타내는 경우($_nC_{(x/n)}$)의 확률은 $P(DR(x/n))$가 된다.

웹 문서에 대한 $P(_{Y=a,N=b}DR_{Y=c,N=d})$는 $(c+d)$번의 추가 문서요청에서 다운로드 성공이 c번 발생할 확률이다. $P(_{Y=a,N=b}DR_{Y=c,N=d})$는 $(c+d)$번의 문서 요청에서 나타날 수 있는 모든 확률들의 합에서 다운로드 성공이 c번 발생할 조건부 확률이 된다. 3번 연속으로 다운로드를 실패한 웹 문서 P_u가 있다고 하자. P_u는 다음 수집에서 다운로드가 성공할 수도 있고 실패할 수도 있다. 다음 수집에서 다운로드를 성공한다면 P_u의 다운로드율은 25%가 되고 실패할 경우에 다운로드율은 0%가 된다. (표 1)에 따르면 P_u의 '다운로드 성공률'이 0.25가 될 확률보다는 0이 될 확률이 매우 높다. 1번의 추가 문서요청에서 P_u가 다운로드될 확률은 $P(DR(0.25)) / ((P(DR(0)) + P(DR(0.25)))$가 된다.

$P_{(Y=a,N=b}DR_{Y=c,N=d)}$는 수식 (1)과 같이 정의되며 추가 문서요청에 대한 특정 다운로드 성공회수가 발생할 확률이다. 웹 데이터베이스 관리자는 두 번 연속 다운로드가 실패된 URL이 향후 3번의 추가적인 문서 요청에서 한번 이상의 다운로드가 성공될 확률을 산출할 수 있다. (그림 2)는 최초 발견 이후 두 번 연속 다운로드를 실패한 URL이 향후 3번의 문서수집에서 모두 다운로드를 실패할 확률을 99.9%로 산출되는 예를 나타내고 있다. 3번의 추가적인 문서 수집에서 한 번 이상의 다운로드 성공이 발생하는 사건은 세 번 모두 다운로드를 실패하는 사건의 여사건(complement event)이다. 따라서 그 확률은 (100% - 99.9%) = 0.1%에 불과하다.

$$P(_{Y=a,N=b}DR_{Y=c,N=d}) = \frac{P\left(DR\left(\frac{a+c}{a+b+c+d}\right)\right)}{\sum_{i=0}^{c+d} P\left(DR\left(\frac{a+i}{a+b+c+d}\right)\right)} \quad (1)$$

웹 문서의 내용변경을 예측하는 것은 웹 문서의 성공적인 다운로드를 예측하는 것과 동일한 방법으로 계산된다. 웹 문서들의 변경율 분포를 보이는 표가 사용되어 $P(MR(x))$이 $P(DR(x))$와 같은 방법으로 계산된다. $P_{(Y=a,N=b}MR_{Y=c,N=d)}$는 수식 (2)와 같이 정의된다.

$$P(_{Y=a,N=b}MR_{Y=c,N=d}) = \frac{P\left(MR\left(\frac{a+c}{a+b+c+d}\right)\right)}{\sum_{i=0}^{c+d} P\left(MR\left(\frac{a+i}{a+b+c+d}\right)\right)} \quad (2)$$

$P_{(Y=a,N=b}DR_{Y=c,N=d)}$의 평가를 위하여 우리는 유명 사이트 그룹으로 부터 581,608개의 웹 문서를 선택하고 임의 사이트 그룹으로 부터 838,035개의 웹 문서를 선택하였다. 선택된 웹 문서들은 46번째부터 50번째까지의 수집에서 5번이 모두 발견된 URL들에 해당하는 문서들이다. 선택된 웹 문서들 각각에 대해 5번의 추가적인 요청을 수행할 경우, 나타날 수 있는 현상의 예측과 실제 5번의 추가적인 요청을 통해서 얻어진 결과를 비교한다.

5번의 추가적인 요청으로 발생할 수 있는 다운로드 성공 회수는 총 6가지 (0부터 5회의 다운로드)가 있다. 즉, 웹 문서는 0회 또는 1회 또는 2회 또는 3회 또는 4회 또는 5회가 성공적으로 다운로드 될 수 있을 것이다. 본 실험에서는 선택된 모든 웹 문서에 대해 $P_{(Y=a,N=(5-a)}DR_{Y=0,N=5)}$, $P_{(Y=a,N=(5-a)}DR_{Y=1,N=4)}$, $P_{(Y=a,N=(5-a)}DR_{Y=2,N=3)}$, $P_{(Y=a,N=(5-a)}DR_{Y=3,N=2)}$, $P_{(Y=a,N=(5-a)}DR_{Y=4,N=1)}$, $P_{(Y=a,N=(5-a)}DR_{Y=5,N=1)}$의 6 가지 값을 산출한다. a는 과거에 성공적으로 다운로드된 회수를 나타내며 a의 값은 각각의 웹 문서에 따라 정해진다. $P_{(Y=a,N=(5-a)}DR_{Y=0,N=5)}$의 의미는 46번째 수집부터 50번째 수집에서 a번 다운로드된 성공한 문서가 향후 5번의 수집에서 한 번도 성공적으로 다운로드 되지 못할 확률을 나타낸다. 6 가지 예측 값의 합은 항상 1이다.

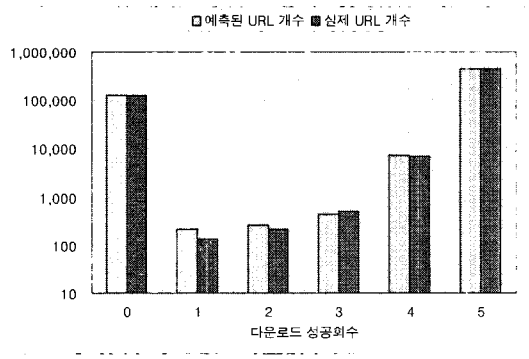
한 웹 문서의 미래 변경 가능성에 대한 예측 값은 확률 분포로 나타나게 된다. 그러나 실제 발생하는 사건은 그 중의 한 가지 경우만 나타나게 된다. 따라서 우리는 동일한 사이트 그룹에 속한 웹 문서들을 대상으로 하여 분포 비교로

$$P(_{Y=0,N=2}DR_{Y=0,N=3}) = \frac{P(DR(0))}{P\left(DR\left(\frac{0}{5}\right)\right)+P\left(DR\left(\frac{1}{5}\right)\right)+P\left(DR\left(\frac{2}{5}\right)\right)+P\left(DR\left(\frac{3}{5}\right)\right)}$$

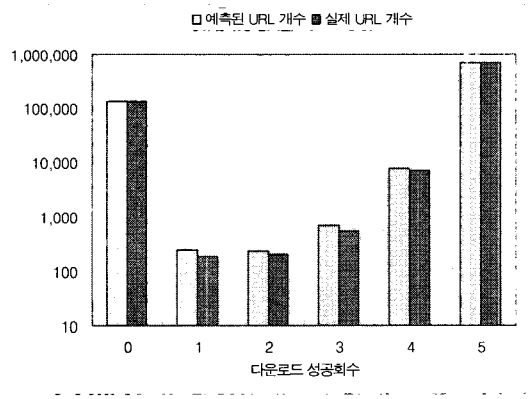
$$= \frac{22.01}{22.01+0.006+0.005+0.007} = 99.9\%$$

(그림 2) $P_{(Y=0,N=2}DR_{Y=0,N=3)}$의 산출 예

$P_{(Y=a,N=b}DR_{Y=c,N=d)}$를 평가한다. 예를 들어, 한 웹 문서 P_u의 $P_{(Y=a,N=(5-a)}DR_{Y=0,N=5)}$, $P_{(Y=a,N=(5-a)}DR_{Y=1,N=4)}$, $P_{(Y=a,N=(5-a)}DR_{Y=2,N=3)}$, $P_{(Y=a,N=(5-a)}DR_{Y=3,N=2)}$, $P_{(Y=a,N=(5-a)}DR_{Y=4,N=1)}$, $P_{(Y=a,N=(5-a)}DR_{Y=5,N=1)}$의 값이 각각 0.25, 0.15, 0.1, 0.1, 0.15, 0.25라고 하자. 그러나 실제 P_u를 요청할 경우 한 번도 다운로드되지 않을 수도 있고, 5 차례 모두 성공적으로 다운로드 될 수도 있다.



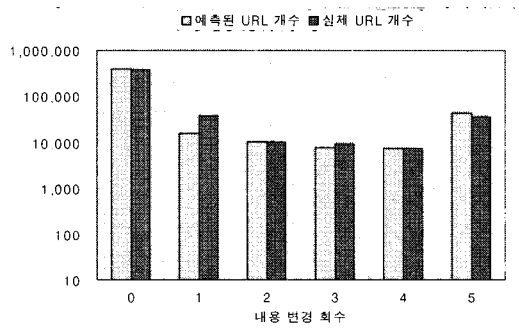
(가) 유명 사이트



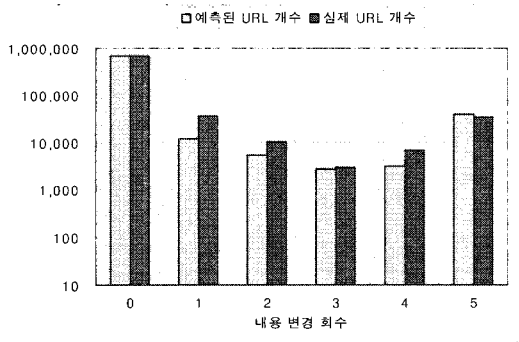
(나) 임의 사이트

(그림 3) $P_{(Y=a,N=b}DR_{Y=c,N=d)}$의 평가 결과

(그림 3)은 $P_{(Y=a,N=b}DR_{Y=c,N=d)}$의 평가 결과를 보여준다. 왼쪽 막대 바(bar)는 성공적으로 다운로드된 웹 페이지의 개수를 나타낸다. 오른쪽 막대 바는 실제로 다운로드된 웹 문서들의 개수를 나타낸다. 가로축은 성공적으로 다운로드될 회수를 나타내고, 세로축은 성공적으로 다운로드된 회수가 동일한 웹 문서들의 개수를 나타낸다. (그림 3 (가))에서 두 막대 바의 차이는 324, 86, 39, 64, 175, 92개로 나타났다. 또한 (그림 3(나))에서 두 막대 바의 차이는 129, 55, 35, 120, 732, 813개로 각각 나타났다. 선택된 URL의 개수가 유명 사이트에서 581,608개 임의 사이트에서 838,035개라는 것을 감안할 때, 잘못된 예측 비율은 각 문서 그룹에서 0.07과 0.11%로 무척 작았다.



(가) 유명 사이트



(나) 임의 사이트

(그림 4) $P_{(Y=a,N=b}MR_{Y=c,N=d)}$의 평가 결과

$P_{(Y=a,N=b}MR_{Y=c,N=d)}$의 평가를 위하여, 유명사이트 그룹으로 부터 486,603개의 웹 문서를 선택하고 임의 사이트 그룹으로 부터 766,328개의 웹 문서를 선택하였다. 선택된 웹 문서들은 45번째부터 50번째까지의 수집에서 성공적으로 다운로드된

웹 문서들이다. 추가적인 5번(51 번째부터 55 번째까지)의 수집에 대한 예측과 실제 다운로드로드 요청 후의 결과가 비교된다. (그림 4)는 문서 내용의 변경 예측과 실제 결과의 차이를 보여준다. (그림 4(가))에서 예측값과 실제값의 차이는 16,360, 22,057, 250, 1,347, 194, 6,988 (총 47,196)개로 나타났다. (그림 4(나))에서 막대 바의 차이는 26,759, 24,659, 5,039, 191, 3,630, 6,760 (총 23,598)개로 각각 나타났다. 23,598(4.8%)개와 33,518 (4.4%)의 웹 문서가 잘못 예측되었다. 잘못 예측된 웹 문서의 비율이 다운로드 성공의 예측에 비해서 높게 나타났으나, 이는 여전히 사용자가 불필요하게 변경되지 않았을 웹 문서를 요청하는 불필요한 요청을 상당히 줄일 수 있음을 보여준다.

5. 결론 및 향후 계획

본 논문에서는 웹 문서가 향후 웹 문서 수집에서 발생하는 다운로드 변경/실패와 내용 변경/유지 여부를 예측하는 도구를 기술하였다. 본 예측 도구는 미래의 변경이 과거의 변경과 매우 밀접한 관련이 있다고 가정한다. 예측도구의 평가를 위하여 약 300백만개의 URL들을 100일간 2일 주기로 관찰하여 그들의 실제 변경 기록을 분석하여 변경 예측에 사용하였다. 다운로드 예측 평가에서 유명사이트와 임의 사이트 그룹의 0.07%와 0.11% 문서들에서 예측 오류가 나타났다. 변경 예측에 대한 평가는 4.8%와 4.2%의 웹 문서들에서 예측 오류가 나타났다. 제안된 예측도구는 웹 관리자들이 다운로드되지 않거나 변경되지 않았을 웹 문서를 불필요하게 요청하여 다운로드하는 것을 상당히 예방할 수 있다.

향후 다음과 같은 통계적인 관찰 연구가 필요하다. 첫째, 웹 문서 변화에 대한 잘못된 예측이나 판단으로 인하여 발생할 수 있는 커버리지 손실에 대한 연구가 필요하다. 실제 변경되었으나 변경되지 않았을 것으로 예측된 웹 문서가 존재할 수 있으며, 실제 변경된 문서로부터 얻어질 수 있는 웹 문서들이 수집되지 못할 수 있다. 둘째, 현재의 변경 사실이 얼마나 오랜 미래의 변경 사실에 영향을 미칠 수 있는 가에 대한 연구가 필요하다.

참 고 문 헌

[1] Brewington, B. and Cybenko, G., "How Dynamic is the Web?", Proc. 9th WWW Conference, pp.257-276, 2000.

[2] Cho, J. and Garcia-Molina, H., "The Evolution of the Web and Implications for an Incremental Crawler", Proc. 26th VLDB Conference, pp.200-209, 2000.

[3] Cho, J. and Garcia-Molina, H., "Synchronizing a Database to Improve Freshness", Proc. 26th SIGMOD Conference, pp.117-128, 2000.

[4] Edwards, J., McCurley, K., and Tomlin, J., "Adaptive Model from Optimizing Performance of an Incremental Web Crawler", Proc. 10th WWW Conference., pp.106-113, 2001.

[5] Fetterly, D., Manasse, M., Najork, M., Wiener, J.L., "A large-scale study of the evolution of web pages", Proc. 12th World Wide Web conference. pp.669-678, 2003.

[6] Heydon, A. and Najork, M., "Mercator: A Scalable, Extensible Web Crawler", International Journal of WWW, Vol.2, No.4, pp.219-229, 1999.

[7] Kim, S.J. and Lee, S.H., "Implementation of a Web Robot and Statistics on the Korean Web", Springer-Verlag Lecture Notes in Computer Science Vol.2713, pp.341-350, 2003.

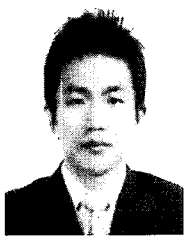
[8] V. Shkapenyuk and T. Suel, "Design and Implementation of a High-performance Distributed Web Crawler", Proc. 18th Data Engineering Conference, pp.357-368, 2002.

[9] Salton, G., Mcgill, M.J., "Intorduction to Modern Information Retrieval", MCGraw-Hill, 1983.

[10] Toyoda, M. and Kitsuregawa, M., "What's Really New on the Web? Identifying New Pages from a Series of Unstable Web Snapshots", Proc. 15th World Wide Web Conference, pp.233-241, 2006.

[11] Wills, C. and Mikhailov, M., "Towards a Better Understanding of Web Resources and Server Responses for Improved Caching", Proc. 8th WWW Conference, 1999.

◐ 저 자 소 개 ◑



김 성 진(Sung-Jin Kim)
1998년 숭실대학교 소프트웨어공학과 졸업(학사)
2000년 숭실대학교 대학원 컴퓨터 학과 졸업(석사)
2004년 숭실대학교 대학원 컴퓨터 학과 졸업(박사)
2004년~2006년 서울대학교 박사후과정연구원
2006년~현재 UCLA 박사후과정연구원
관심분야 : 데이터베이스, 웹, XML
E-mail : sjkim@cs.ucla.edu