웹 문서 변화에 관한 실험적 연구
(An Empirical Study on Changes of Web Pages)

김 성 진 † 이 상 호 ††
(Sung Jin Kim) (Sang Ho Lee)

요 약 웹 문서들은 빈번하게 생성, 소멸, 변경을 반복하고 있으며, 웹 데이타베이스는 최신의 웹 상태를 반영하여야 한다. 웹 데이타베이스의 효과적인 갱신 전략 수립을 위하여 실제 웹의 변화 성향을 파악하는 일은 매우 중요하다. 웹의 변화를 관찰한 연구들이 다양하게 발표되고 있으나 기존의 연구들은 웹 문서의 내용 변경에 주된 초점이 맞추어 있고 웹 문서의 생성과 소멸에 대한 결과가 부족하였다. 본 논문에서는 웹 문서의 변화를 표현할 수 있는 척도로서 URL의 '다운로드 성공률', '변경률', '나이 변이 계수'를 소개하고, 한국의 유명 사이트 집합과 임의(random) 사이트 집합에서 발견된 300만 개의 URL들이 2일 주기로 100일 동안 관찰한다. 본 논문에서는 '다운로드 성공률'과 '변경률'의 분포를 통해 웹 문서의 다운로드 성공과 변경이 과거 기록과 밀접한 연관이 있음을 발견하였으며, 과거 기록을 이용하여 향후 웹 문서의 다운로드 성공과 변경을 예측할 수 있는 모델을 제안한다. 또한, '나이 변이 계수'를 통해 웹 문서들이 얼마나 비주기적으로 변경되는가를 보고한다.

키워드 : 웹 데이타베이스, 웹 문서 변경, 증분 웹 로봇, 웹 통계

Abstract As web pages are created, destroyed, and updated frequently, web databases should be updated to keep up-to-date web pages. In order to keep web databases fresh effectively, we need to understand the change of real web pages. Previous researches on the change of the web pages have directed their efforts on the contents modification of web pages only, and have not taken into account the factors of creation and destruction of web pages in their research. This paper investigates the web page changes, which include contents modification, page creation, and page destruction. We introduce three metrics, namely DR (Download Rate), MR (Modification Rate), and CAV (Coefficient of Age Variation) to represent the change of the web pages. We have monitored three million web pages collected from the famous and random sites every other day for one hundred days. With the Download Rate and the Modification Rate, we learned that the download success and the modification depends on the past change of them, and proposes two estimation formulae that predict the download success and modification. With the Coefficient of Age Variation, we show how web pages do not change periodically.

Key words : web databases, change of web pages, incremental robot, web statistics

1. 서 론

웹은 실세계에 필요한 대부분의 정보를 포함하고 있으며 많은 사용자들이 웹을 통하여 정보를 공유하고 있다. 웹 검색 사이트는 웹 데이타베이스를 구축하여 사용자에게 서비스할 문서들을 유지하고 관리한다. 프록시 서버(proxy server), 메타검색엔진(meta-search engine), 웹 브라우저(web browser) 등의 다양한 응용분야에서도 캐시(cache)를 목적으로 하는 소규모의 웹 데이타베이스를 관리한다. 웹 문서 캐시는 한번 접속했던 문서를 다시 다운로드하지 않고 저장된 문서를 보여줌으로써 웹 어플리케이션(application)이 사용자에게 빠른 응답시간의 서비스를 제공할 수 있도록 한다[1].

웹 문서들은 끊임없이 생성, 변경, 소멸을 반복하고 있으며, 웹 데이타베이스는 빈번하게 갱신되어 최신 상태를 유지하여야 한다. 효과적인 웹 데이타베이스의 갱신을 위하여 실제 웹 문서들의 변화를 관찰하고 변경 모델을 수립하는 연구들이 진행되었다[2-9]. 효과적인 변경 모델은 웹 문서의 변경 가능성을 정확히 예측하여

· 본 연구는 숭실대학교 교내연구비 지원으로 이루어졌음
† 학생회원 : 서울대학교 제어계측신기술연구소 연구원
sjkim@oopsla.snu.ac.kr
†† 종신회원 : 숭실대학교 컴퓨터학부 교수
shlee@comp.ssu.ac.kr
논문접수 : 2004년 7월 27일
심사완료 : 2004년 12월 28일

변경될 확률이 높은 문서들이 우선적으로 수집되도록 한다. 즉, 불필요한 네트워크 자원의 사용을 방지하고 효과적으로 웹 데이타베이스를 최신 상태로 유지할 수 있도록 한다.

웹 문서의 변경에 관한 연구 접근 방법 중 하나[2,3]는 문서 변경이 변경률 람다(λ)의 포아송 분포(Poisson Distribution)를 따른다는 것이다. 이러한 접근 방법은 과거에 나타난 웹 문서의 변경 여부가 미래의 변경 여부와 무관하다고 가정한다. 또 다른 접근 방법[4,5]에서는 웹 문서의 변경이 과거의 변경 경향과 밀접한 관련이 있음을 지적하였다. [4]는 미국 내의 많은 웹 문서들이 업무 시간동안(월요일부터 금요일의 오전 5시부터 오후 5시)에 변경되고 있음을 보였다. [5]는 웹 문서의 변경이 변경주기를 따르기보다는 문서 각각의 과거 변경 기록에 영향을 받을 수 있음을 제시하였다.

웹 문서 변경에 대한 기존의 연구에서는 효과적인 웹 데이타베이스의 변경을 위한 주요 연구 요소들이 간과되었다. 우선, 많은 연구에서 웹 문서가 고정된 주기로 변경되고 있음이 가정되고 있으나 실제 웹에서 어느 정도의 문서들이 일정한 주기로 변경되고 있는가에 대한 연구는 부재하다. 또한 웹 문서의 변경이 문서의 내용에 대한 변경에 주된 초점이 맞추어있고, 실제 웹에서 빈번하게 발생하는 문서의 생성과 소멸에 대한 연구가 부족하였다.

본 논문에서는 [10]에서 소개된 웹 로봇을 이용하여 실제 국내 웹 문서들을 2일 주기로 50차례(100일) 수집하고, 약 300만개 URL들에 대한 관찰 결과를 제시한다. 구체적으로 다음과 같은 결과들이 기술된다. 첫째, 각 문서 수집 시에 발견된 URL들과 과거의 문서 수집에서 발견되지 않고 새롭게 생성된 문서들의 개수를 보고한다. 둘째, 웹 데이타베이스 갱신의 관점에서 일부 연구들[2,3,10]의 결과에 대한 한계를 지적한다. 일부 연구들에서는 웹 문서의 변경이 특정 카테고리별로 관찰되었다. 본 논문은 웹 문서 각각에 대한 과거의 변경 경향 기록의 필요성을 보인다. 셋째, 웹 문서 변경 모델의 기본 가정이 되는 변경 주기의 존재 여부를 관찰하고 전체 관찰 대상 문서의 약 절반에 해당하는 문서가 고정적인 변경주기 없이 불규칙적으로 변경이 발생하였음을 보인다. 넷째, 본 논문은 웹 문서의 변화를 나타내기 위한 세 가지 척도로서 '다운로드 성공률', '변경률', '나이 변이 계수'를 소개하고 수집된 문서들을 소개된 척도에 의해 분석한다. 마지막으로, 분석된 결과들에 기반을 두어 웹 문서의 변화를 통계적으로 예측하는 수식을 제시한다.

본 논문은 다음과 같이 구성되었다. 2장에서는 실험 환경과 웹 문서의 관찰 및 분석을 위한 전략적 결정사항들을 기술한다. 3장에서는 관찰 결과를 기술하고 웹 문서들의 변경 경향을 분석한다. 4장은 3장에서 관찰된 결과에 기반을 두어 향후 웹 문서 수집의 결과 상태를 통계적으로 예측하는 방법을 소개한다. 마지막으로 5장에서 결론을 맺고 향후 연구 계획을 기술한다.

2. 실험 전략

2.1 실험 환경 및 방법

웹에 존재하는 모든 문서에 대해 오랜 시간동안 빠른 주기로 변경을 관찰하는 것은 매우 큰 비용을 필요로 할뿐 아니라 전체 웹에 큰 부하를 준다. 제한된 하드웨어 환경에서 관찰 대상의 크기와 재 수집주기(또는 관찰 주기)는 비례관계에 있다. 관찰 대상 문서가 많아질수록 관찰 주기는 커지게 되며, 관찰 주기가 작아질수록 관찰 대상의 크기는 작아진다. 본 시험에는 웹의 변화 관찰을 위하여 VDSL(Very high-data rate Digital Subscriber Line) 20Mbps 네트워크 환경에서 인텔(Intel) PentiumIV-1.7GHz 프로세서(processor)와 512MB 주 기억장치를 가진 컴퓨터가 사용된다. 웹 로봇[10]은 하루 최대 150만개의 웹 문서를 수집할 수 있었다. 관찰 대상의 크기는 웹 로봇이 하루에 수집할 수 있는 양을 초과하지 않으면서 문서 수집 이후에 변화 분석에 필요한 충분한 시간이 보장되도록 약 100만개로 구성한다.

웹 문서들의 변경을 관찰하는 방법은 두 가지가 있다[2]. 첫 번째로 웹 로봇은 관찰할 문서를 정해놓고 매번 정해진 문서만을 재 수집하여 이전에 수집한 문서와 내용을 비교할 수 있다. 이러한 방법은 새로 발견된 페이지를 수집하지 못하는 단점이 있다. 두 번째로 웹 로봇은 관찰할 웹 사이트를 정해놓고 각 사이트별로 정해진 개수만큼의 문서를 수집하여 이전에 수집된 문서들과 비교할 수 있다. 본 논문에서는 두 번째 방법을 사용하여 특정 사이트를 관찰 대상 사이트로 선정하고 해당 사이트로부터 수집된 웹 문서들의 수집 상태 변경을 관찰한다.

관찰 대상은 유명 사이트 집합과 임의 사이트 집합으로 나뉘어 진다. '랭크서브' 사이트(http://www.rankserv.com/)는 국내 웹 사이트들을 15개의 카테고리(category)로 분류하고 각 카테고리별로 웹 사이트의 순위를 부여하고 있다. 유명 사이트 집합은 '랭크서브'사이트에서 높은 순위를 갖는 상위 4,000개 사이트로 구성되었다. 임의 사이트 집합은 2003년 10월 기준으로 국내 사이트 URL로 판명되었던 약 120만개 사이트 URL 중에서 임의로 추출된 3만개로 구성되었다. 하나의 사이트 URL은 유명 사이트 집합과 임의 사이트 집합에 동시에 속할 수 있다. 유명 사이트 집합에서 발견된 웹 문서들

의 변화는 임의 사이트 집합에서 발견된 웹 문서들의 변화와 비교될 수 있으며, 반대의 경우도 가능하다.

본 시험에서는 하루에 하나의 집합에 속하는 웹 문서들을 관찰한다. 즉, 시험 첫째 날에 유명사이트 집합에 속하는 웹 문서들의 변화를 관찰하고 둘째 날에는 임의 사이트 집합에 속하는 웹 문서들의 변화를 관찰하는 방식으로, 유명 사이트 집합과 임의 사이트 집합에 속한 웹 문서들을 교대로 관찰한다. 각 집합에서 관찰된 웹 문서들은 2일 주기로 재 수집되게 된다. 웹 문서는 수집 주기보다 빈번하게 변경될 수 있으며 수집 주기보다 빈번하게 변경되는 문서의 변화는 실험 결과에 반영되지 않는다. 즉, 임의의 웹 문서가 2일 동안 두 번 이상 변경이 발생했을지라도 한번만 변경하는 것으로 간주된다.

웹 로봇은 2003년 12월 중순부터 2004년 3월 말까지 100일 동안 이틀 간격으로 50차례의 문서 수집을 수행하였다. 한 사이트에는 640,000 비트(bit) 크기의 블룸-필터가 할당되었다. 시험 대상이 되는 사이트에 대한 과부하를 예방하기 위하여 사이트 당 관찰되는 문서는 [2]와 같이 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 데이타 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.

2.2 실험 척도

본 장에서는 웹 문서들의 변화 관찰을 위한 실험 척도를 제안한다. 제안되는 실험 척도의 이해를 그림 1의 웹 문서들의 변화 관찰 결과를 가정한다. 그림 1은 웹 로봇이 2일을 주기로 16번의 문서 수집을 수행한 결과이다. 16회의 수집 이후 URL A, B, C, D, E, F, G, H, I를 보유하였다. '-'는 해당 수집 차수에서 발견되지 않은 URL을 나타낸다. 1차 수집에서는 URL A, B, G, I의 4개 URL이 발견되었다. 2차 수집에서는 URL A, B, E, G, I의 5개 URL이 발견되었다. 웹 로봇은 발견된 모든 URL에 대해 문서를 요청하여 다운로드를 실패할 경우에 '●'로 표현하고 다운로드를 성공할 경우에 웹 문서의 내용을 알파벳 원문자로 표현하였다. URL A는 1차부터 16차의 수집동안 지속적으로 발견되고 있으며 매번 문서의 내용이 변하고 있음을 나타낸다. URL I는 1차부터 16차의 수집동안 지속적으로 발견되고 있으나 매번 다운로드를 실패하고 있음을 나타낸다.

URL은 네트워크의 동적인 상태 변화에 따라 발견되고 발견되지 않는 상태가 반복될 수 있다. 웹 문서의 수집 실패로 인해 실패된 문서에서 명시된 URL이 발견되지 못할 수 있다. URL의 '발견연속성'(DS, Detection Streak)은 URL의 최초 발견 이후 지속적으로 발견된 정도를 나타내는 척도로서 수식 (1)과 같이 정의된다. 임의의 URL이 높은 '발견연속성' 값을 갖기 위해서는 최초 발견 시점으로부터 매 수집 차수마다 연속으로 발견되고 발견되지 않은 시점이후부터 한번도 발견되지 않아야 한다. 임의의 URL이 최초 발견 이후 오랜 시간 발견되지 않다가 다시 발견되는 경우에 '발견연속성'은 낮게 나타난다. 본 논문에서는 '발견연속성'이 임계 값(90%) 이상인 URL들을 분석대상으로 고려한다. 그림 1의 URL D는 5, 6, 7, 9차 수집에서 발견되어 발견연속성은 4/5 = 80%이다. URL A, B, H, I는 발견 연속성은 100%이다. URL E의 발견연속성이 20%로 가장 낮다.

$$\text{발견연속성} = \frac{\text{발견횟수}}{((\text{최후발견 수집차수}) - (\text{최초발견 수집차수})) + 1} \quad (1)$$

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

그림 1 웹 문서 관찰의 예

URL의 '다운로드 성공률'(DR, Download Rate)은 웹 서버에 다운로드를 요청한 회수에 대한 다운로드 성공 회수의 비율을 나타내는 척도로서 수식 (2)와 같이 정의된다. '다운로드 성공률'은 한 URL이 안정적으로 다운로드되는 정도를 나타낸다. 반복적으로 다운로드를 실패한 URL을 제거하는 것은 관리자 입장에서는 불필요한 네트워크의 사용을 예방하고 URL 관리에 필요한 부가적인 오버헤드(overhead)를 막을 수 있으며, 사용자 입장에서는 다운로드되지 않은 URL이 질의의 결과로 조회되는 것을 방지할 수 있다. 그림 1에서 URL D는 5, 6, 7, 9차 수집에서 발견되었고 5차와 7차 수집에서 성공적으로 다운로드되었으므로 '다운로드 성공률'은 2/4 = 50%가 된다. URL A, B, H의 '다운로드 성공률'은 100%이며, URL I의 '다운로드 성공률'이 0%로 가장 낮다. '다운로드 재현률'은 총 수집횟수 중에서 URL이 발견되어 문서가 요청된 횟수로서 수식 (3)에 나타나 있다. URL C, D, E의 '다운로드 성공률'은 50%로 모두 같으나 '다운로드 재현률'은 50%, 25%, 13%로 서로 차이가 있다.

$$\text{다운로드 성공률} = \frac{\text{다운로드 성공횟수}}{\text{요청횟수 (또는발견횟수)}} \quad (2)$$

$$\text{다운로드 재현률} = \frac{\text{요청횟수}}{\text{수집횟수}} \quad (3)$$

URL의 '변경률'(MR, Modification Rate)은 문서 내용의 변경 빈도를 나타내는 척도로서 '변경률'이 높을수록 빈번하게 문서의 내용이 변경된 URL임을 나타낸다. 최초로 다운로드된 웹 문서에 대한 변경의 유무는 알 수 없기 때문에 '변경률'은 두 번 이상 성공적으로 다운로드된 문서에 대해서만 산출된다. URL의 '변경률'은 수식 (4)와 같이 정의된다. 다운로드를 실패한 URL에 대해서는 문서 변경의 여부를 파악할 수 없으므로 '변경률' 산출에서 제외된다. 그림 1의 URL H는 5차 수집부터 15차 수집까지 성공적으로 다운로드되었으며, 7, 9, 11, 13차 수집에서 문서 내용이 변경되었다. URL H의 '변경률'은 4/(11-1) = 40%가 된다. 두 문서 내용에 대한 동일성 판단은 바이트(byte)단위의 문자 비교로 이루어진다. 즉, 사용자가 웹 브라우저로 확인했을 때 동일하게 보이는 문서들이라도 HTML 문서에서 차이가 존재할 경우 두 문서는 서로 다른 내용을 가진 것으로 정의한다. 예를 들어, 6차 수집에서 발견된 HTML 문서의 태그는 모두 소문자로 작성되었으나, 7차 수집에서 발견된 HTML의 태그는 대문자로 작성되었을 때, 사용자는 웹 브라우저로 동일한 문서를 보게 되지만 이러한 경우 문서의 변경이 있었음으로 간주된다. '변경 재현률'은 수식 (5)와 같이 정의되며 '변경률'의 산출에 사용되는 문서 내용의 개수가 많고 적음을 나타낸다. 다운로드의 성공 횟수가 많을수록 변경여부를 판단하는데 사용된 문서 내용의 개수가 많아지고 '변경 재현률'은 증가한다.

$$\text{변경률} = \frac{\text{문서내용 변경횟수}}{\text{다운로드 성공횟수} - 1} \quad (4)$$

$$\text{변경 재현률} = \frac{\text{다운로드 성공횟수} - 1}{\text{총 수집횟수} - 1} \quad (5)$$

URL(또는 웹 문서)의 나이(age)는 문서의 내용이 지속적으로 유지된 기간이다. 문서의 내용이 변경되면 해당 문서의 나이는 0이 되고 시간이 지날수록 문서의 나이가 증가한다. 본 논문에서는 변경 주기의 존재 여부를 판단하기 위하여 웹 문서 나이를 정의하며, 정확한 실제 나이의 산출을 목적으로 하지 않는다. 따라서 [2]에서 웹 문서의 수명을 가정한 것과 유사하게 문서의 최초 다운로드 시점을 나이 0으로 간주하고, 관찰 기간의 마지막 시점에서 문서의 나이가 종료되는 것으로 간주한다. 예를 들어, 그림 1에서 URL H의 나이는 4일, 4일, 4일, 4일, 6일이 되고 평균 나이는 (4+4+4+4+6) / 5 = 4.4일이 된다. URL의 나이는 웹 문서의 내용이 유지된 기간을 고려하므로 다운로드되지 않거나 발견되지 않은 수집 차수도 나이에 포함이 된다. 예를 들어 URL G의 나이는 ⓧ라는 내용으로 10일, ⓨ라는 내용으로 4일, ⓩ라는 내용으로 14일 존재한 것으로 관찰되고, 평균 나이는 (10+4+14)/3=9.4일이 된다. '나이 변이 계수'(CAV, Coefficient of Age Variation)는 수식 (6)과 같이 정의된다. 일정한 주기로 변경되는 URL은 웹 문서의 나이가 일정하게 나타나고 평균에 대한 편차가 작으므로 '나이 변이 계수'가 작게 나타난다. 예를 들어, 그림 1에서 URL H의 변이계수는 0.9/4.4 = 20%이고, URL G의 '나이 변이 계수'는 5.04/9.4 = 54%이다. 내용 변경의 주기를 관찰하기 위해서는 최소 한번 이상의 내용 변경이 발생한 URL에 대해서 관찰이 가능하다. '변이 재현률'은 수식 (7)과 같이 정의된다. '나이 변이 계수'는 '변이 재현률'이 일정 수준이상이 되는 URL에 대하여 산출된다.

$$\text{나이변이계수} = \frac{\text{나이의 표준편차}}{\text{나이의 평균}} \quad (6)$$

$$\text{변이 재현률} = \frac{\text{문서내용 변경횟수}}{\text{총 수집횟수} - 1} \quad (7)$$

그림 1에 나타난 URL들에 대한 '다운로드 성공률'과 '변경률', '나이 변이 계수'를 산출한 결과는 표 1에 나타나 있다. 산출이 불가능한 값은 'N/A(Not Applicable)'로 표시되었다.

3. 문서 수집 상태 변화

임의 사이트 집합은 유명 사이트 집합에서 관찰된 내용과 비교를 위해서 사용되었으나 두 집합에서 관찰된 내용이 매우 유사하게 나타났다. 유명 사이트 집합에서

표 1 웹 문서의 관찰 결과 예

URL	발견 연속성	다운로드 성공률	다운로드 재현률	변경률	변경 재현률	평균 나이	편차	변이 계수	변이 재현률
A	100%	100%	100%	100%	100%	2	0	0%	100%
B	100%	100%	75%	0%	73%	24	0	0%	0%
C	80%	50%	50%	33%	20%	4	0	0%	7%
D	80%	50%	25%	100%	7%	2	0	0%	7%
E	20%	50%	13%	N/A	0%	2	0	0%	0%
F	100%	100%	6%	N/A	0%	2	0	0%	0%
G	75%	75%	75%	25%	53%	9.4	5.04	54%	13%
H	100%	100%	69%	40%	67%	4.4	0.9	20%	27%
I	100%	0%	100%	N/A	0%	N/A	N/A	N/A	0%

는 매 수집 차수에서 평균 80만 개의 URL들이 발견되었고 78%의 URL들이 성공적으로 다운로드되었다. 사용자(혹은 웹 클라이언트)가 웹 서버에 URL A의 문서를 요청하였을 때, 웹 서버가 URL A가 아닌 다른 URL의 문서 내용으로 응답하는 경우에 재 방향 설정(redirection)이 발생하였다고 한다. 매 수집 차수에서 재 방향 설정되는 URL들은 4%가 존재하였으며 다운로드가 성공된 문서로 간주되지 않았다. 유명 사이트 집합에서는 1차 수집에서 약 80만개의 URL을 보유하였고 50회의 문서 수집 이후 약 132만 개의 URL을 보유하게 되었다. 매 문서 수집에서 약 만개의 URL을 새롭게 발견하였다. 임의 사이트 집합에서는 매 수집마다 약 백만 개의 URL이 발견되었고 83%의 URL들이 성공적으로 다운로드되었으며 총 169만개의 URL들을 누적하여 보유하게 되었다. 그림 2와 그림 3은 유명 사이트 집합과 임의 사이트 집합에서 매 수집 차수마다 요청된 URL의 개수, 성공적으로 다운로드된 문서의 개수, 누적된 URL의 개수를 나타낸다.

그림 4는 '발견연속성'의 범위에 따른 URL들의 개수 분포를 나타낸다. 유명 사이트 집합에서 수집된 132만개의 URL들 중 발견연속성이 90%이상인 URL은 116만개(87.6%)로 나타났다. '발견연속성'이 10%미만인 URL들은 0.3%가 존재하였다. 임의 사이트 집합에서 수집된

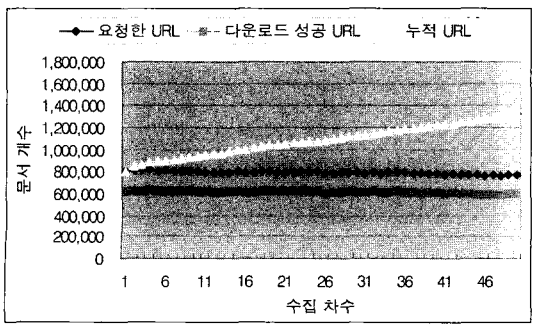

그림 2 유명 사이트 집합에서의 수집 현황

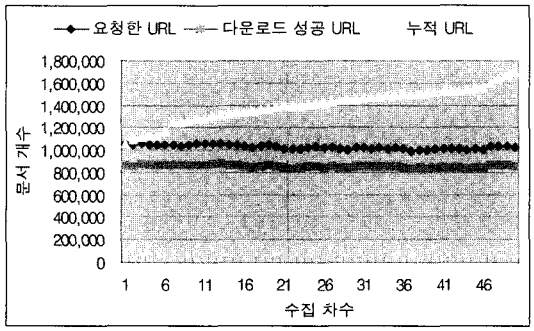

그림 3 임의 사이트 집합에서의 수집 현황

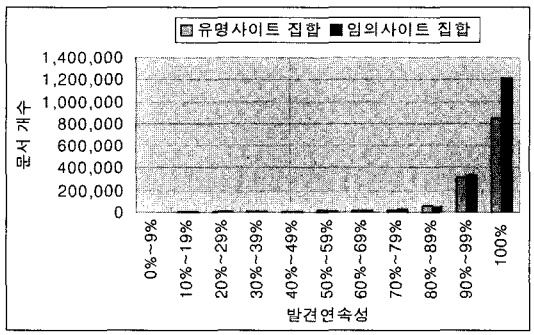

그림 4 '발견연속성' 분포

169만개 URL들에서도 155만개(91.8%)가 '발견연속성'이 90%이상으로 나타났다. 대부분의 URL들은 최초 발견이후에 연속적으로 발견되었으며 발견되지 않은 시점으로부터 지속적으로 발견되지 않는 현상이 나타났다.

'발견연속성'이 90%이상이고 '다운로드 재현률'이 20% 이상인 조건을 만족하는 URL들의 '다운로드 성공률' 분포는 그림 5에 나타내고 있다. 유명 사이트 집합의 82만개의 URL들 중 22%의 URL들은 '다운로드 성공률'이 10%미만이고, 77%는 '다운로드 성공률'이 90%이상이었다. 임의 사이트 집합에서 조건을 만족하는 111만개의 URL들 중 16%와 83%의 URL들이 '다운로드

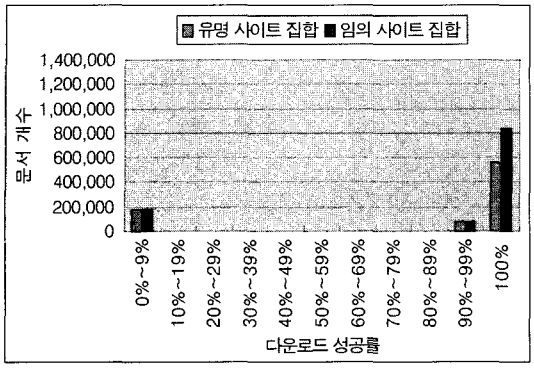

그림 5 '다운로드 성공률' 분포

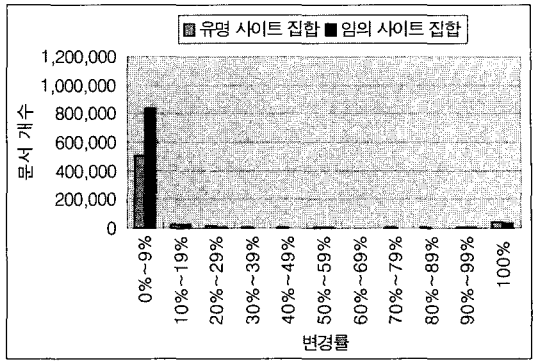

그림 6 변경률 분포

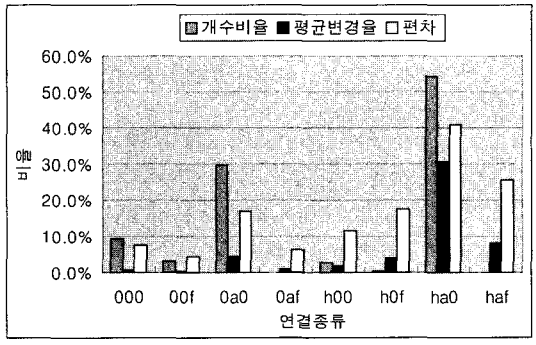

그림 7 연결종류별 '변경률'의 평균

성공률' 10%미만과 90%이상으로 나타났다. '다운로드 성공률'이 90%이상인 URL들의 비율은 임의 사이트 집합에서 상대적으로 높게 나타났다.

대부분 URL들의 '다운로드 성공률'이 10%미만과 90%이상에 분포하는 것으로 나타났다. 다운로드 실패와 성공을 반복하는 URL('다운로드 성공률'이 50% 전후인 URL)은 거의 나타나지 않았으며, 다운로드를 성공한 URL은 반복해서 다운로드를 성공하고 다운로드를 실패한 URL은 반복해서 다운로드를 실패하는 현상이 나타났다. 이러한 실험 결과는 반복적으로 다운로드가 실패된 URL이 미래에 다운로드가 성공할 것으로 기대하고 보유하면서 빈번하게 재요청하는 것은 반복적인 다운로드 실패를 유발시킬 수 있음을 나타낸다.

유명 사이트 집합에서 발견된 URL 중 '발견연속성'이 90%이상이고 '발견 재현률'이 20% 이상인 URL들은 총 64만개가 존재하였다. 64만개 URL들의 '변경률' 평균은 13%로 나타났으며, 한 웹 문서를 100번 다운로드할 때 평균적으로 13번 정도 내용이 변경되는 것으로 나타났다. 임의 사이트 집합에서는 93만개 URL들의 평균 '변경률'이 6%로 나타났다. 유명 사이트 집합에서 발견된 URL들이 빈번하게 변경되는 URL들이 상대적으로 많이 발견되었다. 그림 6은 '변경률'에 따른 문서의 개수 분포를 나타낸다. 평균 '변경률'은 매우 빈번하게 변경되는 URL들과 거의 변경되지 않는 URL들에 의해 평균적으로 산출된 수치임을 알 수 있다. 이로 인해 '변경률'의 표준 편차는 유명 사이트에서 29%이고 임의 사이트에서 22%로 평균보다 큰 값으로 나타났다.

그림 6에서 많은 URL들은 문서 내용이 변경되지 않고 있으며 일부의 URL들이 반복적으로 변경되고 있음을 알 수 있다. 즉 웹 데이타베이스의 갱신을 위하여 보유한 모든 URL들을 다시 다운로드하는 것이 불필요한 네트워크의 사용을 초래할 수 있음을 보여준다. 웹 데이타베이스의 갱신은 변경이 발생한 URL을 반복적으로 수집하고 변경이 발생되지 않았던 URL들은 그 주기를 보다 크게 설정하여 수집하여야 한다.

URL은 'f-URL', 'a-URL', 'h-URL'의 세 가지 종류로 구분된다. <FRAME> 태그에서 발견된 URL은 'f-URL'이 된다. <A HREF> 태그에서 발견된 URL은 'a-URL'이 된다. 그 외의 URL은 'h-URL'이 된다. 예를 들어, 자바 스크립트로부터 발견된 URL은 'h-URL'이 된다. 웹 문서의 "연결 종류"는 웹 문서가 포함하는 URL의 종류에 따라서 'h', 'a', 'f' 문자의 조합으로 표현된다. 예를 들어, "연결 종류"가 "0a0"인 웹 문서는 하나 이상의 'a-URL'을 포함한다. "연결종류"가 "haf"인 웹 문서는 'h-URL', 'a-URL', 'f-URL'을 각각 하나 이상 포함한다. URL을 포함하고 있지 않는 웹 문서의 "연결종류"는 "000"이 된다.

[10]에서는 20일 동안 관찰한 웹 문서에 대해 연결종류별 평균 변경 주기를 보이고 <FRAME> 태그를 포함한 문서나 하이퍼링크를 포함하지 않는 웹 문서가 상대적으로 변경주기가 길게 나타남을 보였다. 그림 7에서는 유명 사이트 집합에서 발견된 URL들을 대상으로 하여 연결유형별 문서 개수, '변경률'의 평균, 표준편차를 나타내고 있다. 연결유형이 'ha0'과 '0a0'인 웹 문서들의

'변경률'의 평균이 다른 연결종류의 문서들보다 높게 나타났으며, 연결유형이 '000'(하이퍼링크를 포함하지 않음)과 '00f'(<FRAME> 태그를 포함)인 웹 문서들의 평균 '변경률'이 상대적으로 낮게 나타났다. 편차는 URL들이 평균 '변경률'을 따르는 정도이다. 그림 7에서 각 연결 종류에 속한 URL들의 '변경률'은 평균으로부터 편차가 매우 크게 나타났다. 이러한 원인은 그림 6의 '변경률' 분포와 같이 웹 문서들은 대부분 매우 빈번하게 변경되거나 거의 변경되지 않기 때문이다.

[2,3,10]는 웹 문서들을 도메인별로 분류하여 ".com" 도메인에 속한 웹 문서들이 다른 도메인에 속한 웹 문서보다 자주 변경됨을 보였다. 본 논문에서는 도메인의 유형을 9개로 분류하고 유명 사이트 집합과 임의 사이트 집합에서 수집된 URL들 중에서 두 번 이상 다운로드된 URL에 대해 도메인별로 평균 '변경률'을 산출하였다. 그림 8은 각 도메인별로 URL의 개수, 평균 '변경률', 표준 편차를 나타낸다. [2,3,10]와 마찬가지로 ".co.kr"과 ".com"도메인에 속한 웹 문서들의 평균 '변경률'이 다른 도메인에 속한 웹 문서의 평균 '변경률'보다 높게 나타났다. 그러나 도메인별 평균 '변경률'도 연결 유형별 평균 '변경률'과 마찬가지로 편차가 매우 크며, 많은 URL들이 ".com"과 ".co.kr"도메인에 속한 URL들로 나타났다.

웹 문서들의 집합에 대한 평균적인 통계 정보는 웹 데이타베이스의 갱신을 위하여 재 수집될 URL을 개별적으로 선택하는데 무리가 있다. 예를 들어, "000"에 속하는 URL J와 "ha0"에 속하는 URL K가 있다고 하자. URL J는 "000"에 속하는 URL들의 평균 변경률을 따르지 않을 확률이 매우 높고, URL K 또한 "ha0"에 속하는 URL들의 평균 변경률을 따르지 않을 확률이 매우 높다. 따라서 URL K가 URL J보다 빈번하게 변경될 것이라고 예측하기 힘들다.

그림 9는 '발견연속성'이 90%이상이고 '변이 재현률'이 20%이상인 URL들의 '나이 변이 계수'의 분포를 나

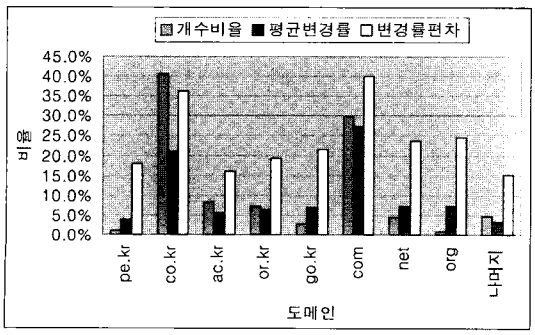

그림 8 도메인별 '변경률'의 평균

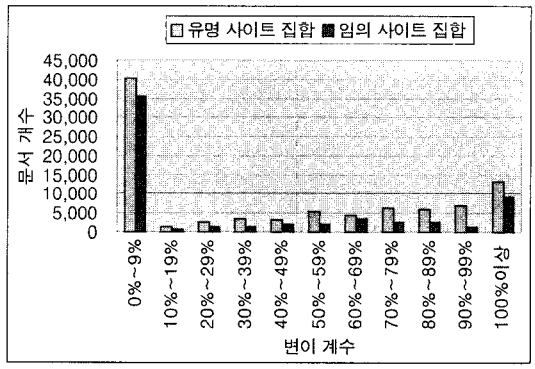

그림 9 '나이 변이 계수' 분포

타낸다. 변이계수 30%(0.3)를 기준으로 하여 30% 미만인 URL들은 변경 주기를 가지고 30% 이상인 URL들은 변경주기를 가지지 않는다고 전제할 때, 유명 사이트 집합에서 발견된 URL들 중 4만 5천개는 변경주기가 존재하고 5만 3천개는 변경주기가 존재하지 않았다. 임의 사이트 집합에서 발견된 URL들 중에서는 5만 2천개의 URL이 변경주기를 가지고 9만 1천개의 URL이 변경주기를 가지지 않는 것으로 나타났다. 그림 9에 따르면 유명 사이트 집합과 임의 사이트 집합에서 발견된 URL들 중에서 일정한 변경주기를 가지지 않는 URL들이 54%와 64%로 나타났다. 이러한 결과는 웹 문서들이 변경주기를 가지면서 변경된다는 가정의 오차를 보여준다.

4. 웹 문서 상태에 관한 예측

웹 데이타베이스 관리자는 주어진 시간 내에 웹 데이타베이스를 효과적으로 갱신하기 위하여 다운로드 성공률이 높고 변경되었을 가능성이 높은 문서를 우선적으로 수집해야 한다. 웹 데이타베이스에서 갱신을 위하여 선택된 웹 문서들의 집합이 있을 때, 선택된 집합내의 문서들 대부분이 다운로드되지 않거나 성공적으로 다운로드되었더라도 문서 내용이 이전에 다운로드된 문서 내용과 같다면, 웹 데이타베이스는 최신의 웹 상태를 효과적으로 반영하지 못하고 관리자는 불필요한 시스템 자원을 낭비하게 되는 결과가 초래된다.

본 장에서는 3장에서 관찰된 웹 문서의 상태 변화를 통해서 향후 웹 문서의 다운로드의 성공과 실패, 문서 내용의 변경과 유지에 대한 상태를 확률적으로 예측하는 방법을 소개한다. 유명 사이트 집합에서 관찰된 URL 중에서 '발견연속성'이 90%이상이고 '다운로드 재현률'과 '변경 재현률'이 20% 이상인 URL들의 개수와 비율은 표 2에 정리되어 있다. 본 장에서는 표 2에 기반을 두어 유명 사이트 집합에 속한 URL들의 문서 상태를 예측한다.

표 2 '다운로드 성공률'과 '변경률' 분포

계급구간	다운로드 성공률		변경률	
	개수	비율	개수	비율
0%	181,770	22.01%	410,325	63.85%
1% ~ 9%	489	0.06%	97,400	15.16%
10% ~ 19%	323	0.04%	28,179	4.38%
20% ~ 29%	461	0.06%	16,671	2.59%
30% ~ 39%	356	0.04%	12,448	1.94%
40% ~ 49%	396	0.05%	6,267	0.98%
50% ~ 59%	557	0.07%	6,658	1.04%
60% ~ 69%	575	0.07%	3,127	0.49%
70% ~ 79%	761	0.09%	5,553	0.86%
80% ~ 89%	2,725	0.33%	6,320	0.98%
90% ~ 99%	78,734	9.53%	8,201	1.28%
100%	558,842	67.66%	41,512	6.46%
총합	825,989	100.00%	642,661	100.00%

동일한 '다운로드 성공률'을 지닌 두 URL의 '다운로드 재현률'은 차이가 있을 수 있다. 예를 들어 하나의 URL은 20번의 문서 요청에 10번의 다운로드를 성공하였고 다른 하나의 URL은 10번의 문서 요청에 5번의 다운로드를 성공하였다면 두 URL의 '다운로드 성공률'은 50%로 같으나 '다운로드 재현률'은 전자의 URL이 더 높다. 실제 실험을 통하여 '다운로드 재현률'이 높은 URL들의 '다운로드 성공률' 분포와 '다운로드 재현률'이 낮은 URL들의 '다운로드 성공률' 분포가 유사하게 나타났다. 본 논문에서는 '다운로드 재현률'과 '다운로드 성공률'이 독립적(independent)이라고 전제한다. 즉, 20번 다운로드가 요청된 URL들을 대상으로 한 '다운로드 성공률' 분포와 10번의 다운로드가 요청된 URL들을 대상으로 한 '다운로드 성공률' 분포가 동일하다고 간주한다. '변경률'과 '변경 재현률'도 서로 독립적이고 '변경률'의 분포가 '변경 재현률'에 영향을 받지 않는다고 전제한다.

상태변화 예측에 사용되는 기호들은 표 3에 정의되어 있다. $P(DR(x\%))$, $P(MR(x\%))$은 URL의 '다운로드 성공률'이 x%일 확률과 '변경률'이 x%일 확률을 나타내며 표 2를 통해 알 수 있다. $P(DR(0\%))$ - 임의의 URL의 '다운로드 성공률'이 0%일 확률 - 은 22.01%이고 $P(DR(100\%))$는 67.66%이다. 0%와 100%를 제외한 나머지 $P(DR(x\%))$는 x%가 속하는 계급구간의 비율을 구간의 크기로 나눈 값으로 한다. $P(DR(50\%))$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 '다운로드 성공률'이 50%로 나오는 경우의 수가 ${}_{50}C_{25}=1.26*10^{14}$로 가장 많다. 즉 URL의 '다운로드 성공률'이 50%일 확률이 가장 높다. 그러나 표 2에서 $P(DR(50\%))$의 비율은 상대적으로 매우 작음을 볼 수 있다. n번의 문서 요청에서 $(n/2)$의 다운로드가 성공하는 경우(${}_nC_{n/2}$)가 발생할 확률은 $P(DR(50\%))$이다. 즉, n번의 문서 요청에서 x번의 다운로드 성공을 나타내는 경우(${}_nC_{(x/n)}$)의 확률은 $P(DR(x/n))$가 된다.

임의의 URL에 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 URL X가 있다고 하자. URL X는 다음 수집에서 다운로드가 성공할 수도 있고 실패할 수도 있다. 다음 수집에서 다운로드를 성공한다면 URL X의 '다운로드 성공률'은 25%가 되고 실패할 경우에 '다운로드 성공률'은 0%가 된다. 표 2에 따르면 URL X의 '다운로드 성공률'이 25%가 될 확률보다는 0%가 될 확률이 매우 높다. 1번의 추가 문서요청에서 URL X가 다운로드 될 확률은 $P(DR(25\%))/((P(DR(0\%)) + P(DR(25\%)))$가 된다. $P(_{Y=a,N=b}DR_{Y=c,N=d})$는 수식 (8)과 같이 정의되며 추가 문서요청에 대한 특정 다운로드 성공회수가 발생할 확률이다. 웹 데이타베이스 관리자는 두 번 연속 다운로드가 실패된 URL이 향후 3번의 추가적인 문서 요청에서 한번 이상의 다운로드가 성공될 확률을 산출할 수 있다. 그림 10은 최초 발견이후 두 번 연속 다운로드를 실패한 URL이 향후 3번의 문서수집에서 모두 다운로드를 실패할 확률을 99.9%로 산출되는 예를 나타내고 있다. 3번의 추가적인 문서 수집에서 1번이상의 다운로드 성공이 발생하는 사건은 3번 모두 다운로드를 실패하는 사건의 여사건(complement event)이다. 따라서 그 확률은 100%-99.9%=0.1%에 불과하다. $P(_{Y=a,N=b}$

표 3 문서 상태 예측에 사용되는 기호 정의

기호	정의
$P(DR(x\%))$	URL의 '다운로드 성공률'이 x%일 확률
$P(MR(x\%))$	URL의 '변경률'이 x%일 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

$MR_{Y=c,N=d}$는 $P(_{Y=a,N=b}DR_{Y=c,N=d})$와 동일한 방법으로 산출될 수 있다.

$$P(_{Y=a,N=b}DR_{Y=c,N=d}) = \frac{P\left(DR\left(\frac{a+c}{a+b+c+d}\right)\right)}{\sum_{i=0}^{c+d} P\left(DR\left(\frac{a+i}{a+b+c+d}\right)\right)} \quad (8)$$

5. 결론 및 향후 계획

본 논문에서는 국내의 유명사이트 집합과 임의사이트 집합에 속한 웹 문서들을 100일 동안 2일 간격으로 주기적으로 수집하여 웹 문서의 변경과 변경 경향을 관찰하였다. 웹 문서의 변화를 표현하기 위하여 '다운로드 성공률', '변경률', '나이 변이 계수' 척도가 정의되었다. 본 논문은 웹 문서의 다운로드 성패와 내용의 변경 여부가 쉽게 변하지 않음을 실험 결과로 제시하였다. 즉, 성공적으로 다운로드된 URL이 향후 지속적으로 다운로드가 성공되고, 내용의 변경이 없던 문서는 지속적으로 변경되지 않고 현재 상태를 유지하는 것으로 나타났다. 본 논문에서는 모든 문서들이 고정된 변경 주기로 변경이 이루어지지 않으며 약 절반의 문서들이 규칙적으로 변경되고 있음을 지적하였다. 마지막으로 본 논문은 관찰된 실험 결과에 기반을 두어 웹 문서의 변화를 통계적으로 예측할 수 있는 두 가지 수식을 소개하였다. 웹의 변화에 대한 관찰 결과는 최초에 구축된 웹 데이타베이스의 최신성(freshness)[2]을 높이기 위한 웹 문서 변경 모델의 설계에 도움을 준다[5]. 또한, 웹 로봇의 수집 빈도, 수집 순서, 웹 문서의 변경 비율에 기반을 둔 수집 전략 수립 및 운용에 도움을 줄 수 있다.

향후 다음과 같은 통계적인 관찰 연구가 필요하다. 첫째, 웹 데이타베이스는 다양한 주기로 갱신될 수 있으며 상이한 주기에서의 웹 문서의 변화 분석이 필요하다. 둘째, 관찰 기간의 확장을 통하여 본 논문에서 관찰의 시작과 끝에서 발견되는 URL들의 변화 분석에 대한 정확성을 높일 필요가 있다. 마지막으로, 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.

참 고 문 헌

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

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

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

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

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

[6] F. Douglis, A. Feldmann, and B. Krishnamurthy, "Rate of Change and Other Metrics: a Live Study of the World Wide Web," Proc. 1st USENIX Symposium on Internetworking Technologies and System, pp.147-158, 1997.

[7] S. Lawrence and C.L. Giles, "Accessibility of Information on the Web," Nature, 400(6740), pp.107-109, 1999.

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

[9] A. Ntoulas, J. Cho, C. Olston "What's New on the Web? The Evolution of the Web from a Search Engine Perspective," Proc. 13th WWW Conf., to appear, 2004.

[10] S.J. Kim and S.H. Lee, "Implementation of a Web Robot and Statistics on the Korean Web," Proc. 2nd Human.Society@Internet Conf., pp.341-350, 2003.

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

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

그림 10 $P(_{Y=0,N=2}DR_{Y=0,N=3})$의 산출 예제

김 성 진

1998년 숭실대학교 소프트웨어 공학과 졸업(학사). 2000년 숭실대학교 대학원 컴퓨터학과(석사). 2004년 숭실대학교 컴퓨터학과 대학원(박사). 2004년~현재 서울대학교 제어계측신기술연구소 연구원. 관심분야는 인터넷 데이타베이스, 데이타베이스 시스템 성능평가

이 상 호

1984년 서울대학교 전산공학과 졸업(학사). 1986년 미국 노스웨스턴대 전산학과(석사). 1989년 미국 노스웨스턴대 전산학과(박사). 1990년~1992년 한국전자통신 연구원, 선임연구원. 1999년~2000년 미국 조지 메이슨대 소프트웨어 정보 공학과 교환 교수. 1992년~현재 숭실대학교 컴퓨터학부 부교수. 관심분야는 인터넷 데이타베이스, 데이타베이스 시스템 성능 평가 및 튜닝