웹 문서 변경 측정 방법의 효과성 평가†

권신영[0] 이상호
숭실대학교 대학원 컴퓨터학과
숭실대학교 컴퓨터학부
{sykwon[0], shlee}@comp.ssu.ac.kr

Effectiveness Evaluation of the Metrics Measuring the Change Degree of Web Pages

Shin Young Kwon[0] Sang Ho Lee
Department of Computing Graduate School, Soongsil University
School of Computing, Soongsil University

요 약

웹의 진화를 연구하기 위해 다수의 문서 비교 방법들이 웹 문서 변경 측정 도구로서 사용되어 왔다. 웹의 진화 연구는 웹 데이터베이스의 효율적 관리를 위해 필수적이기 때문이다. 그러나 같은 웹 문서의 변경에 대하여 어떠한 방법으로 측정하였는지에 따라 상이한 결과를 보일 수 있음에도 불구하고, 각 측정 방법의 비교 평가는 연구되지 않았다. 본 논문에서는 웹 문서 변경 측정 방법의 효과성 평가 척도를 제안한다. 그리고 수집된 실제 웹 문서를 통해 기존에 사용되어온 다섯 가지 측정 방법들의 결과 차이를 보인다. 아울러 정의한 평가 척도에 따라 각 측정 방법을 비교 평가한다.

1. 서론

웹 데이터베이스(웹 문서들의 집합)는 많은 웹 어플리케이션 분야에서 구축되고 관리되어진다. Google, Yahoo 등과 같은 웹 검색 서비스는 자체적으로 웹 데이터베이스를 구축하여 사용자들이 원하는 정보를 검색할 수 있도록 제공한다. 웹 문서는 끊임없이 변화한다. 따라서 웹 데이터베이스는 원격지에 있는 원본 웹 문서들의 변경을 반영할 수 있도록 갱신되어야 한다. 그러나 웹 데이터베이스 전체를 갱신하는 것은 웹 로봇의 불필요한 문서 수집과 불필요한 데이터베이스 갱신 작업, 네트워크 부하 증가 등 여러 자원의 낭비를 야기하기 때문에, 변경된 웹 문서만 갱신하기 위한 시도가 필요하다. 그런데 어떤 웹 문서의 변경 여부는 그 문서를 다운로드 받아 확인하기 전에는 알 수 없다. 따라서 웹 데이터베이스의 효율적 관리를 위해 변경된 웹 문서의 예측이 요구된다. 이러한 목적으로 웹 문서의 변경 규칙을 찾기 위한 웹의 진화 연구가 여러 문헌에서 진행되었다.

웹의 진화 연구에 있어 웹 문서 변경 측정 방법의 선택은 매우 중요하다. 같은 웹 문서 변경에 대하여 측정 방법에 따라 다른 결과를 나타내기 때문이다. 그럼에도 불구하고 지금까지 연구된 문헌에서는 각기 다른 측정 방법을 사용해왔으며, 이들을 서로 비교 평가하는 연구는 수행되지 않았다. 본 논문에서는 웹 문서 변경 측정 방법을 평가하기 위한 척도를 제시한다. 정의된 척도는 웹 문서의 변경 종류를 여섯 가지로 분류하며, 각 변경 종류에 따라 기준이 되는 변경 측정 값을 제시한다. 또한 정의된 척도를 통해 바이트 단위 비교(byte-wise comparison)[1, 2, 3], TF·IDF 코사인 거리(TF·IDF cosine distance)[4], 단어 거리(word distance)[4], 편집 거리(edit distance)[5], 쉥글링(shingling)[6, 7] 방법의 효과성을 평가한다.

본 논문의 구성은 다음과 같다. 2장에서는 실제 수집된 웹 문서를 사용하여 기존 측정 방법들의 결과 차이를 보인다. 3장에서는 웹 문서 변경을 여섯 가지 종류로 분류하고, 변경 측정 방법의 효과성 평가를 위한 척도를 정의한다. 4장에서는 웹 문서와 유사한 변경을 보이도록 생성된 데이터를 이용하여, 정의된 척도에 따라 기존에 사용된 다섯 가지 측정 방법들의 효과성을 비교 평가한다. 끝으로 5장에서는 결론과 함께 향후 계획을 기술한다.

2. 연구 동기

†본 연구는 한국학술진흥재단의 지원에 의하여 수행되었음.
(KRF-2004-005-D00172)

먼저 기존에 사용된 측정 방법들의 결과 차이를 보이기 위해 다음과 같은 실험을 진행하였다. 국내 웹 문서로부터 임의로 50,000개의 문서를 선택하여 하루 간격으로 두 번의 수집을 수행하였다. 확보된 웹 문서 50,000 쌍을 대상으로 각 문서의 변경도를 측정하였다. 그림 1은 그 중 변경도가 0보다 큰 문서를 대상으로 쉥글링(k=10)과 TF·IDF 코사인 거리 방법을 비교한 결과이다. 두 방법의 차이를 가시화하기 위해 쉥글링 방법의 결과값에 따라 오름차순으로 정렬하였다.

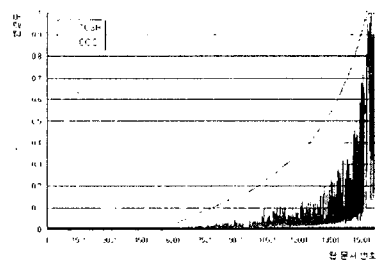


그림 1. 웹 문서의 변경도 차이 (10SH vs. COS)

결과를 통해 알 수 있듯이, 같은 변경 문서에 대하여 사용된 측정 방법에 따라 결과가 상이하며, 그 차이가 심각하게 큰 경우도 존재한다(쉥글링 방법에서는 변경도가 0.9 이상이고 TF·IDF 코사인 거리 방법에서는 0.1 미만인 문서가 존재하는 것을 알 수 있다). 이는 적절한 웹 문서 변경 측정 방법의 선택이 얼마나 중요한지를 잘 보여준다.

3. 평가 척도

평가 척도를 제안하기에 앞서 웹 문서 변경의 종류를 다음과 같이 여섯 가지로 분류하여 정의한다. 정의되는 각 변경 종류의 예는 그림 2에서 보인다.

정의1. 변경 전 문서에 존재하지 않는 새로운 단어가 삽입되는 변경을 "추가(add)" 라 하며, 변경 전에 존재하는 단어가 다시 삽입되는 변경을 "복사(copy)" 라 한다. "추가" 와 "복사" 변경 후 문서의 크기(단어 수)는 삽입된 단어의 수만큼 증가한다.

정의2. 변경 전 문서에 존재하던 단어가 삭제되어 변경 후 문서에 존재하지 않게 되는 변경을 "제거(drop)" 라 하며, 삭제된 단어가 변경

후 문서에 여전히 존재하게 되는 변경을 "축소(shrink)" 라 한다. "제거" 와 "축소" 변경 후 문서의 크기(단어 수)는 삭제된 단어의 수만큼 감소한다.

정의3. 변경 전 문서의 단어가 다른 단어로 바뀌는 변경을 "대체(replace)" 라 하며, 문서 내 단어의 위치가 바뀌는 변경을 "이동(move)" 이라 한다. "대체" 와 "이동" 변경 후 문서의 크기(단어 수)에는 변화가 없다.

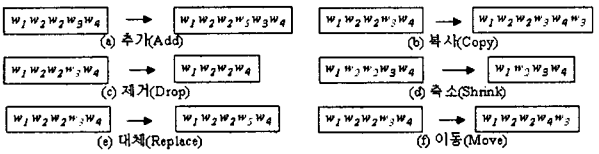


그림 2. 웹 문서 변경의 분류

위에서 정의한 여섯 가지 웹 문서 변경의 종류는 실제 웹에서 빈번히 발생한다. 예를 들어, 어떤 온라인 서점 사이트의 웹 문서를 가정해 보자. 해당 문서는 다양한 책의 정보(이미지, 이름, 출판사, 가격, 요약, 서평 등)를 포함하고 있으며, 책의 인기 순서에 따라 나열되어 있다. 이러한 웹 문서에서 다음과 같은 변경들은 빈번히 발생한다. 새로운 책 정보가 삽입된다("추가"). 기존의 책 정보가 삭제된다("제거"). 기존의 책 정보가 다른 책으로 변경된다("대체"). 책의 인기도 변경에 의해 책 정보 순서가 바뀐다("이동"). 기존에 등록되어 있던 서평과 유사한 서평이 게시된다("복사"). 기존에 등록된 유사 서평 중 일부가 제거된다("축소"). "복사" 변경을 "추가" 변경과 구별한 이유는, 기존에 있던 내용과 유사한 내용이 삽입되는 것이 새로운 내용의 삽입보다 의미적으로 더 작은 변경이 될 수 있기 때문이다. "축소" 변경과 "제거" 변경 역시 같은 이유로 구별되었다.

제안되는 척도는 다음 네 가지 속성을 가진다.

1. 웹 문서의 변경도는 0에서 1사이 값으로 표현된다.
2. 웹 문서 내에 변경이 많을수록 변경도 값은 커진다.
3. 모든 단어는 의미적으로 동일한 중요도를 갖는다.
4. 어떤 웹 문서 A가 B로 변경되었을 때의 변경도와 B가 A로 변경되었을 때의 변경도는 같다.

그림 3에서 5는 여섯 가지 변경 종류에 대한 평가 척도를 나타낸다. 변경 전 웹 문서 내의 단어 수가 n일 때, x축은 변경된 단어의 개수를 의미하며, y축은 변경도를 나타낸다.

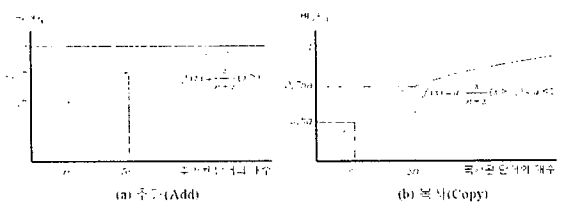


그림 3. 웹 문서 변경 척도 : "추가" , "복사" 변경

그림 3(a)에서 점 (n, 0.5)는 n개의 단어가 웹 문서에 추가되면 변경도는 0.5가 됨을 의미한다. 문서 내 단어 수가 2n이 되므로, 전체 2n개 중 n개의 변경으로 n/2n = 0.5가 되는 것이다. 점 (3n, 0.75)는 3n개의 단어가 추가된 경우이다. 변경 후 단어 수는 4n이 되므로, 변경도는 3n/4n = 0.75가 된다. 그림 3(b)의 "복사" 변경은 변수 a를 제외하고 "추가" 와 동일한 원리이다. 변수 a는 "추가" 변경에 대한 "복사" 변경의 가중치를 의미한다. 예를 들어, 어떤 어플리케이션에서 한 단어의 추가를 두 단어의 복사와 동일하게 여긴다면 a의 값은 0.5가 된다. 두 변경 모두 삽입되는 단어가 무한히 증가할수록 변경도 1에 가까워진다.

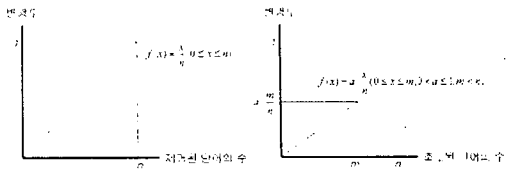


그림 4. 웹 문서 변경 척도 : "제거" , "축소" 변경

그림 4(a)에서 제거될 수 있는 단어 수는 n으로 제한된다. 물론 n개의 단어가 제거될 경우 웹 문서 전체가 변경된 것이므로 변경도는 1이 된다. 그림 4(b)에서 축소 가능 최대 단어 수는 m이다. m은 웹 문서 내에 중복되는 단어의 수를 의미한다. 웹 문서에 따라 다를 수 있으나 n보다 항상 작다. 변수 a는 "복사" 변경에서의 가중치와 동일하게 설정된다.

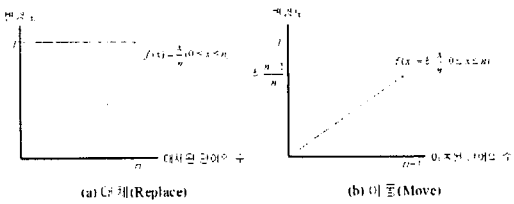


그림 5. 웹 문서 변경 척도 : "대체" , "이동" 변경

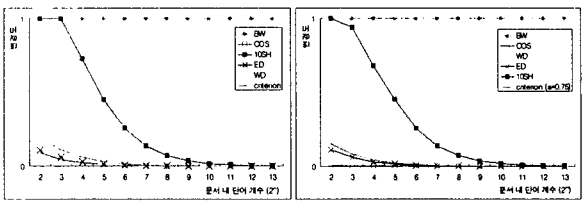
그림 5(a) 역시 대체될 수 있는 단어 수는 n으로 제한된다. n개의 단어가 대체된 후 웹 문서의 단어 수에는 변함이 없지만, 모든 단어가 변경된 것이므로 변경도 1을 갖게 된다. 그림 5(b)에서는 이동 가능 최대 단어 수가 n-1(문서 내 모든 단어가 역순으로 변경되는 경우)이다. 변수 b는 변수 a와 유사하게 "제거" 또는 "대체" 변경에 대한 "이동" 변경의 가중치를 의미한다. 만약 어떤 어플리케이션에서 단어 순서의 변경이 전혀 중요하지 않다면, 평가 척도 사용 시 b의 값을 0으로 설정한다.

4. 효과성 평가

본 장에서는 다섯 가지 측정 방법의 효과성을 세 종류의 실험을 통해 평가한다. 비교 대상 방법은 바이트 단위 비교, TF-IDF 코사인 거리, 단어 거리, 편집 거리, 쉥글링(k=10) 방법이며, 이하 BW, COS, WD, ED, 10SH로 기술한다. 실험에 사용된 데이터는 실제 웹의 변경 상태를 반영하도록 생성된 데이터이다. 첫번째 실험은 다양한 크기의 웹 문서에서 하나의 단어가 변경될 때 각 측정 방법의 효과성을 평가한다. 사용된 데이터의 크기(단어 수)는 2^2, 2^3, ..., 2^{13}이다. 이는 실제 웹 문서의 95%는 2^2~2^{13}범위 개수의 단어를 포함한다는 연구 결과를 반영한 것이다[7]. 두번째 실험은 하나의 웹 문서에서 다양한 수의 단어가 변경될 때 각 측정 방법의 효과성을 평가한다. 사용된 데이터는 1,000개의 단어를 가지며 변경되는 단어수의 비율은 5%, 10%, ..., 95%이다. 이때 변경되는 단어들은 한 부분에 군집되어 존재하도록 구성하였다. 실제 웹에는 약 1,000개 단어를 포함하는 문서가 가장 많이 존재하며(전체의 약 25%)[7], 웹 문서의 변경은 일반적으로 군집되어 발생한다는 특징을 반영한 것이다[5]. 끝으로, 앞서 기술된 변수 a, b 값의 변화에 따른 효과성 평가를 수행하였다(앞의 두 실험에서는 두 변수 모두 0.75로 설정하였다). 사용된 데이터 내의 단어 수는 역시 1,000개 이며 그 중 100개의 단어가 군집되어 변경되었을 경우를 대상으로 실험하였다.

그림 6은 첫 번째 실험의 결과이다. x축은 변경 전 데이터 내의 단어 수를 2에 대한 지수 값으로 나타낸다. 즉, x축의 값이 n이면 2^n개의 단어를 가진 데이터를 말한다. y축은 해당 변경에 따른 데이터의 변경도를 나타낸다. 어떤 측정 방법의 그래프가 평가 척도보다 상위

에 위치할수록 그 방법은 웹 문서 변경에 지나치게 민감하다고 할 수 있고, 하위에 위치할수록 둔감한 방법이라 할 수 있다. BW는 문서의 변경 여부를 판단하기 위해 사용되는 방법이며 변경된 정도는 표현할 수 없기 때문에 (0 또는 1 리턴) 결과 비교 분석에서 생략하기로 한다. 모든 변경 종류에 대해 10SH은 항상 민감한 결과를 보였으며, 척도와의 차이 역시 (특히 작은 문서일수록) 매우 높게 나타났다. 이와 달리 COS은 항상 변경에 둔감한 결과를 보였으며, WD와 함께 "이동" 변경을 전혀 고려하지 못하는 결과를 보였다. 전체적으로 ED가 가장 효과적인 결과를 나타냈으며, "이동" 변경 외의 변경에서는 WD도 효과적인 결과를 보였다. "제거" , "축소" 변경은 각각 "추가" , "복사" 변경과 유사하므로 생략하였다.

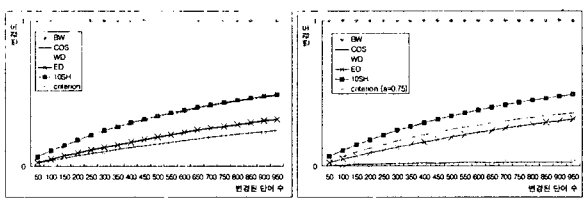


(a) 추가(add) (b) 복사(copy)

(c) 대체(replace) (d) 이동(move)

그림 6. 문서 크기에 따른 효과성 평가

그림 7은 두 번째 실험의 결과이다. x축은 1,000개 단어 중 변경된 단어 수의 비율을 나타낸다. 10SH는 여전히 모든 변경에 대해 민감한 결과를 보였지만 척도와의 차이는 상당히 작아진 것을 알 수 있다. COS 역시 여전히 둔감한 결과를 보였다. 특히 "복사" 변경에서 척도와 매우 큰 차이를 보였으며, "이동" 변경은 전혀 고려하지 못하였다. WD와 ED 역시 "이동" 변경을 제외한 모든 변경에서 척도와 유사한 결과를 나타냈다. "이동" 변경에서 ED는 다소 민감한 결과를 보였으며 WD는 항상 0을 리턴하였다.

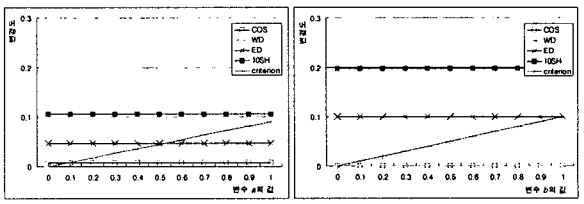


(a) 추가(add) (b) 복사(copy)

(c) 대체(replace) (d) 이동(move)

그림 7. 문서 내 변경 단어 비율에 따른 효과성 평가

그림 8은 세 번째 실험의 결과를 나타낸다. 이 결과는 특정 어플리케이션에 적절한 측정 방법을 선택할 때 지침이 될 수 있다. 예를 들어, 어떤 어플리케이션에서 두 단어의 복사를 한 단어의 추가와 동일한 변경으로 간주한다고 가정하자. 이 경우 WD나 ED 방법이 "복사" 변경을 효과적으로 표현할 수 있다는 것을 그림 8(a)를 통해 알 수 있다. 그림 8(b)에서, 10SH은 "이동" 변경에 대해 "제거" 또는 "대체" 변경 이상으로 민감하게 나타난 반면, COS과 WD는 여전히 "이동" 변경을 전혀 고려하지 못하였다. ED는 "이동" 변경을 "제거" 또는 "대체" 변경과 동일하게 간주함을 알 수 있다.



(a) 복사 (copy) (b) 이동(move)

그림 8. 변수 a, b에 따른 효과성 평가

지금까지의 실험 결과를 정리하면 다음과 같다.

1. 싱글링 방법은 웹 문서 변경에 지나치게 민감한 방법이다. 민감하게 반응하는 정도는 특히 대상 문서가 작을수록, 그리고 변경 종류가 "대체" 또는 "이동" 일 때 더욱 커진다.
2. COS은 웹 문서 변경에 둔감한 방법이다. 특히 "복사" 또는 "축소" 변경에서 둔감한 정도는 더욱 커진다. 또한 "이동" 변경을 전혀 고려하지 못한다.
3. WD는 "이동" 변경을 전혀 고려하지 못하지만 다른 변경에 대해서는 비교적 효과적인 방법이다. "이동" 변경을 고려할 필요가 없고 "복사" 또는 "축소" 변경도를 "추가" 또는 "제거" 변경도의 1/2 정도로 여기는 어플리케이션에서 적당한 방법이라 할 수 있다.
4. ED는 "이동" 변경을 "제거" 또는 "대체" 변경과 동일하게 여기는 어플리케이션에서 가장 효과적인 방법이다.

5. 결론 및 향후계획

본 논문에서는 웹 문서의 변경 종류를 여섯 가지로 분류하여, 각각에 대한 척도를 정의하였다. 이를 통해 기존에 사용된 다섯 가지 웹 문서 변경 측정 방법의 효과성을 비교 평가하였다. 이러한 평가는 특정 어플리케이션에서 웹 문서의 변경을 조사할 필요가 있을 때 적합한 측정 방법 선택을 위한 지침이 될 수 있으며, 또한 기존 방법보다 더 효과적인 측정 방법의 연구를 가능케 할 수 있다.

향후, 평가 척도의 정확성을 위해 웹 문서 변경 종류를 더 세부적으로 분류하기 위한 연구가 필요하다. 또한 정의된 척도를 잘 반영하는 효과적 웹 문서 측정 방법을 연구할 계획이다.

참고 문헌

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

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

[3] S. J. Kim and S. H. Lee, "An Empirical Study on the Change of Web Pages", Proc. the 7th APWeb Conf. pp.632-642, 2005

[4] A. Ntoulas, J. Cho, and C. Olston, "What's New on the Web? The Evolution of the Web from a Search Engine Perspective", Proc. the 13th WWW Conf. pp.1-12, 2004

[5] L. Lim, M. Wang, S. Padmanabhan, J. S. Vitter, and R. Agarwal, "Characterizing Web Document Change" Proc. the 2nd WAIM Conf. pp.133-144, 2001

[6] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, "Syntactic Clustering of the Web", Computer Network and ISDN Systems, Vol.29, No.8-13, pp.1157-1166, 1997

[7] D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener, "A Large-Scale Study of the Evolution of Web Pages" Proc. the 12th WWW Conf. pp.669-678, 2003