DOI: 10.3745/KIPSTD.2008.15-D.3.305

균등한 웹 문서 샘플링을 이용한 웹 검색 서비스들의 커버리지 측정

장 성 수[†] · 김 광 현[††] · 이 준 호[†††]

요 약

인터넷에는 유익한 정보들이 포함된 웹 문서들이 공개되고 있으며, 이러한 웹 문서들은 웹 검색 서비스를 통하여 접근할 수 있다. 따라서 웹 검색 서비스들은 보다 많은 웹 문서들을 수집하기 위하여 노력하고 있으나, 이들은 수집된 웹 문서들의 커버리지를 파악하는데 많은 어려움을 겪고 있다. 따라서 본 논문에서는 기존의 커버리지 측정 방법들을 분석하고, 효과적인 커버리지 측정 방법을 제안한다. 즉, 인터넷에서 웹 문서를 균등하게 샘플링하고, 이 웹 문서들이 웹 검색 서비스에 색인되어 있는지를 조사함으로써 웹 검색 서비스들의 절대 및 상대 커버리지를 측정한다. 그리고 본 논문에서는 제안한 방법으로 국내 웹 검색 서비스들의 커버리지를 측정하여 비교하였으며, 그 결과 구글, 네이버, 엠파스 순으로 절대 및 상대 커버리지가 높게 나타났다. 이러한 본 논문의 결과는 웹 검색 서비스들의 커버리지를 측정하는데 도움이 될 것으로 기대된다.

키워드 : 정보 검색, 웹 검색 서비스, 웹 로봇, 커버리지

Estimating Coverage of the Web Search Services Using Near-Uniform Sampling of Web Documents

Sung Soo Jang[†] · Kwang Hyun Kim[††] · Joon Ho Lee[†††]

ABSTRACT

Web documents with useful information are widely available on the internet and they are accessible with web search service. For this reason, web search services study better ways to collect more web documents, but have a difficulty figuring out the coverage of these web pages. This paper is intended to find ways to evaluate the current coverage assessment methods and suggest more effective coverage assessment technique that is, sampling internet web documents equally, monitoring how they are classified on web search services, in an attempt to assess both absolute and relative coverage of the web search engines. The paper also presents the comparison among Korean web search services using the suggested methods－the absolute and relative coverage was highest in Google followed by Naver and Empas. The result is expected to help estimating coverage of web search services.

Key Words : Information Retrieval, Web Search Service, Web Robot, Coverage

1. 서 론

인터넷의 사용이 보편화됨에 따라 수많은 정보들이 인터넷에 공개되고 있으며, 인터넷 사용자들은 원하는 정보를 찾기 위해 웹 검색 서비스를 활용하고 있다. 웹 검색 서비스는 웹 로봇을 사용하여 웹 문서들을 수집하고, 이들을 효과적으로 검색할 수 있는 검색 서비스를 제공한다. 따라서 수집된 웹 문서들의 수는 웹 검색 서비스의 품질을 측정하는 중요한 척도가 되며, 웹 검색 서비스들은 보다 많은 양질의 웹 문서들을 수집하기 위해 노력하고 있다. 그러나 급격히 증가하는 전체 웹 문서의 수를 측정할 수 없기 때문에 웹 검색 서비스들은 전체 웹 문서들에 대해 얼마나 많은 웹 문서를 수집하였는지 또는 다른 웹 검색 서비스들에 비해 얼마나 많은 웹 문서를 수집하였는지 파악하는데 많은 어려움을 겪고 있다.

일반적으로 웹 검색 서비스에서는 수집한 웹 문서들의 양을 측정하는 척도로써 커버리지(coverage)를 사용하고 있으며, 이는 절대 커버리지(absolute coverage)와 상대 커버리지(relative coverage)로 구분된다. 웹 검색 서비스 A의 절대 커버리지는 전체 웹 문서 수에 대한 A의 웹 문서 수를 의미하며, 웹 검색 서비스 B에 대한 A의 상대 커버리지 는 B의 웹 문서 수에 대한 A와 B의 중복 웹 문서 수를 의미한

† 정 회 원 : 숭실대학교 대학원 컴퓨터학과 박사과정
†† 정 회 원 : NHN (서치솔루션) 근무
††† 종신회원 : 숭실대학교 컴퓨터학부 부교수
논문접수 : 2008년 2월 22일
심사완료 : 2008년 3월 12일

다. 그러나 현실적으로 전체 웹 문서의 수나 웹 검색 서비스간의 실제 웹 문서 중복도를 정확하게 측정하기는 매우 어렵다. 현재까지 다수의 연구에서는 웹 검색 서비스들의 검색 결과에 대한 중복도를 이용 하여 상대 커버리지를 측정하였으며, 또한 이들은 상대 커버리지를 이용하여 전체 웹 문서의 수를 추정하였다[1].

Bharat는 랜덤하게 웹 문서들을 샘플링한 후 이들이 웹 검색 서비스에서 검색되는지를 확인하고, 웹 검색 서비스간 확인된 문서들의 중복도를 측정하여 웹 서비스들의 상대 커버리지를 측정하였다. 이를 위해 Bharat는 웹 검색 결과에서 웹 문서를 랜덤하게 수집하였으며, Monika는 랜덤 워크 알고리즘을 사용하여 인터넷에서 랜덤하게 웹 문서를 수집하였다. 그리고 Lawrence는 웹 문서를 추출하지 않고 사전에 선정된 질의를 웹 검색 서비스에 입력하고, 검색 결과들의 중복도를 측정하여 웹 서비스간의 상대 커버리지를 측정하였다. 그러나 현재까지 이들 측정방법의 정확도를 비교하거나 평가한 연구는 찾아보기 어렵다.

본 논문에서는 기존의 커버리지 측정 방법들을 분석하고, 효과적인 커버리지 측정 방법을 제안한다. 즉, 인터넷에서 웹 문서를 균등하게 샘플링하고, 이 웹 문서들이 웹 검색 서비스에 색인되어 있는지를 조사함으로써 웹 검색 서비스들의 절대 및 상대 커버리지를 측정한다. 그리고 본 논문에서는 제안한 방법의 정확도를 평가하였으며, 이 방법을 사용하여 국내 웹 검색 서비스들의 상대 및 절대 커버리지를 측정하여 비교하였다.

본 논문의 구성은 다음과 같다. 2장에서는 Bharat, Lawrence 그리고 Monika가 제안한 커버리지 측정 방법에 대하여 설명하고, 3장에서는 본 논문의 커버리지 측정 방법에 대하여 기술하고, 4장에서는 3장에서 기술한 방법의 성능을 비교 평가한다. 그리고 5장에서는 국내 검색 서비스들의 상대 커버리지를 분석하며, 마지막으로 6장에서 결론을 맺는다.

2. 관련 연구

2.1 Bharat의 측정 방법

Bharat는 웹 검색 서비스들의 상대 커버리지를 측정하기 위해 샘플링(sampling)과 검사(checking)의 2단계 방법을 사용하였다[2, 3]. 샘플링 단계에서는 웹 문서들을 수집하고 이들 문서에서 단어를 추출하여 질의를 작성한다. 그리고 웹 검색 서비스에 질의를 입력하고 검색 결과 중에서 웹 문서를 랜덤하게 샘플링하였다. 검사 단계에서는 샘플링 된 웹 문서들이 웹 검색 서비스에 색인이 되어 있는지를 조사하며, 이러한 웹 문서들의 중복도를 측정하여 웹 검색 서비스들의 상대 커버리지를 측정하였다. 이를 위해 Bharat는 야후에서 웹 문서를 수집하고 이들 웹 문서에서 추출된 단어들을 사용하여 35,000개의 질의를 작성하였으며, 이 질의들을 알타비스타(AlteVista), 핫봇(HotBot), 익사이트(Excite), 인포시크(Infoseek) 4개의 웹 검색 서비스에 입력하였다. 그리고 각각의 웹 검색 서비스의 검색 결과 상위 100개 중에서 랜덤하게 웹 문서를 샘플링하고 이 웹 문서들이 나머지 3개의 검색 서비스에서 검색이 되는지 확인하였다.

Bharat의 1997년 6월 실험에서 상대 커버리지는 핫봇, 알타비스타, 익사이트, 인포시크 순이었으며, 1997년 11월 실험에서 상대 커버리지는 알타비스타, 핫봇, 익사이트, 인포시크 순으로 나타났다. 이러한 실험에서 다른 검색 서비스들에 대한 알타비스타의 상대 커버리지 비율은 50%였다.

2.2 Lawrence의 측정 방법

Lawrence는 선정된 질의들을 웹 검색 서비스에 입력한 후, 검색 결과에 대한 중복도를 측정함으로써 상대 커버리지를 측정하였다[1, 4]. 이를 위해 Lawrence는 NEC 연구소 직원들이 사용하였던 질의 중 검색 결과가 50~600개인 질의 575개를 선정하였으며, 이 질의들을 알타비스타(AlteVista), 핫봇(HotBot), 노던라이트(Northern Light), 익사이트(Excite), 인포시크(Infoseek), 라이코스(Lycos)의 6개의 웹 검색 서비스에 입력하였다. 그리고 각 검색 결과들의 중복 문서를 조사하여 이들의 상대 커버리지를 계산하였으며, 그 결과 핫봇의 상대 커버리지가 가장 높았으며, 알타비스타에 대한 핫봇의 상대 커버리지는 57.5%이였다. 한편 Lawrnce는 핫봇의 상대 커버리지와 핫봇의 전체 문서 수 1억1천만개를 사용하여 전체 웹 문서 수를 추정하였으며, 그 결과로써 추정된 전체 웹 문서의 수는 약 3억2천만건이었다.

2.3 Monika의 측정 방법

Monika는 인터넷에서 웹 문서를 균등하게 샘플링한 후 이러한 웹 문서들이 웹 검색 서비스에 색인이 되어 있는지를 확인함으로써 웹 검색 서비스의 상대 커버리지를 추정하였다[5]. 이를 위해 Monika는 (그림 1)과 마코비안 랜덤워크 알고리즘으로 웹 문서들을 샘플링하였으며 이 알고리즘은 수집된 웹 문서에서 URL을 추출한 후 이 중에서 URL을 랜덤하게 선택한다. 그리고 이 URL이 지시하는 웹 문서를 수집하며 이러한 수집 과정을 반복한다[6]. 또한 웹 문서 내

```
변수의 정의 :
   HostSet        호스트 이름 집합
   Urlset(h)      호스트 이름 h의 URL집합
   Samples        샘플링 된 URL집합
   U              웹 문서 p에 포함된 URL들의 집합
RandomReset :
   HostSet에서 임의의 호스트 h를 선택
   선택된 호스트 h의 UrlSet(h)에서 URL u를 선택
   선택된 URL u가 지시하는 웹 문서 p를 다운로드
Step :
   웹 문서 p가 하나 이상의 링크를 포함하고 있다면,
         각 링크의 호스트 h, URL u를 HostSet, UrlSet(h)에 추가
   웹 문서 p에 포함된 링크들의 URL들로 U를 초기화
   While U is non-empty :
         URL 집합에서 랜덤하게 u를 선택하고 URL 집합에서 u를 삭제
         선택 URL u가 가리키는 웹 문서 p를 다운로드
         웹 문서 p를 다운로드 하는데 성공 했다면 Step으로 이동
   RandomReset으로 이동
```

(그림 1) 랜덤워크 알고리즘

에 다른 웹 문서로 링크되는 URL이 없는 경우 수집된 URL 중에서 랜덤하게 선택하여 웹 문서를 수집한다.

Monika은 랜덤 워크 알고리즘으로 웹 문서를 균등하게 샘플링하고 샘플된 웹 문서에서 단어들을 추출하고 질의를 작성하였다. 그리고 이러한 질의들을 웹 검색 서비스 알타비스타(AltaVista), 익사이트(Excite), 패스트서치(FAST Search), 구글(Google), 핫봇(HotBot), 인포씨크(Infoseek), 라이코스(Lycos), 노턴라이트(Northern Light)에 입력하였으며, 각각의 검색 결과와 샘플링 된 웹 문서들의 중복도를 조사하였다. 이러한 방법으로 각 웹 검색 서비스간의 상대 커버리지를 분석하였으며, 그 결과 알타비스타의 상대 커버리지가 가장 높았으며, 익사이트와 핫봇 순으로 상대 커버리지가 높았다[5, 7].

2. 커버리지 측정 방법

2장에서 기술한 Lawrence의 방법은 NEC 연구소 직원들이 사용한 질의 중에서 샘플 질의를 생성하였기 때문에 다른 연구원들이 반복적으로 실험을 수행하기 어려우며, 샘플 질의에 따라 상이한 결과를 나타낼 수 있다[5]. 그리고 Bharat의 방법론은 웹 문서의 크기에 따라 웹 문서가 샘플링될 확률이 상이하기 때문에 균등한 샘플링을 할 수 없다[5]. 따라서 본 논문에서는 웹 문서를 균등하게 샘플링하는 랜덤워크 방식을 사용하였으며, (그림 2)와 같이 5단계 방법으로 커버리지를 측정한다.

본 논문의 첫 번째 단계에서는 랜덤 워크 알고리즘을 사용하여 인터넷상의 웹 문서들을 편중되지 않고 균등하게 수집한다. 두 번째 단계에서는 수집된 웹 문서 중에서 웹 문서를 샘플링하였으며, 이 때 랜덤 샘플링 방법과 페이지순위 샘플링 방법을 사용한다. 랜덤 샘플링 방법은 일반적인 웹 문서들이 수집되어 있는지를 확인하기 위해 사용하며, 이는 수집된 웹 문서들 중에서 웹 문서를 랜덤하게 샘플링한다. 그리고 페이지순위 샘플링 방법은 인기도나 중요도가 높은 웹 문서들이 수집되어 있는지를 확인하기 위해 사용되며[8, 9, 10, 11], 이는 수집된 웹 문서들의 페이지순위를 계산하고 이 값이 큰 순서대로 샘플링한다.

세 번째 단계에서는 웹 검색 서비스들이 샘플링한 웹 문서들을 색인하였는지를 확인하기 위해 스트롱 질의를 생성하며, 스트롱 질의는 샘플링 된 각각의 웹 문서에서 추출된 주요 대표 단어들을 'AND' 연산자로 결합하여 작성된다. 네 번째 단계에서는 작성된 스트롱 질의들을 웹 검색 서비스에 입력한다. 다섯 번째 단계에서는 스트롱 질의들에 대한 검색 결과에 노출된 웹 문서들을 수집한다. 여섯 번째 단계에서는 검색 결과들과 두 번째 단계에서 샘플링 된 웹 문서들 간의 중복도를 측정하며[12], 이 때 웹 문서의 중복은 웹 문서의 URL 일치와 내용 일치의 2가지 방법으로 확인한다. URL 일치는 웹 문서의 URL이 동일한지를 확인하는 방법이며, 내용 일치는 URL과 상관없이 웹 문서의 내용이 동일한지를 확인하는 방법이다.

1 단계	웹 문서 수집
2 단계	웹 문서 샘플링
3 단계	스트롱 질의 생성
4 단계	질의 입력
5 단계	검색 결과 수집
6 단계	중복도 측정

(그림 2) 커버리지 측정 방법

본 논문에서는 (그림 2)와 같은 방법을 사용하여 다음과 같이 절대 커버리지와 상대 커버리지를 계산하며, 두 번째 단계에서 샘플링 된 웹 문서를 전체 웹으로 가정하고 절대 커버리지를 측정한다.

$$A\text{의 절대커버리지} \quad AC(A) = \frac{S_S \cap S_A}{S_S}$$

$$B\text{에 대한 } A\text{의 상대커버리지} \quad RC(A|B) = \frac{(S_S \cap S_A) \cap (S_S \cap S_B)}{S_S \cap S_B}$$

위의 식에서 S_s는 샘플링 된 웹 문서 집합을 의미하며, S_A와 S_A는 각각 검색 서비스 A와 B에 색인된 웹 문서 집합을 의미한다.

3. 커버리지 측정 방법의 정확도 평가

본 장에서는 3장에서 기술한 커버리지 측정 방법의 정확도를 평가한다. 즉, 본 논문에서는 절대 커버리지와 상대 커버리지의 차이가 2배인 두 개의 웹 검색 서비스를 구축하고, 3장에서 기술한 방법으로 이들의 커버리지를 측정한다. 그리고 이러한 측정 결과가 2배의 커버리지를 나타내는지 검증함으로써 3장에서 기술한 커버리지 측정 방법의 정확도를 평가 한다.

3.1 검색 서비스 구축

본 논문에서는 웹 로봇[6, 13]을 사용하여 2005년 6월 1개월 동안 1,000만 건의 정적(static) 웹 문서를 수집하였으며, 이들을 사용하여 웹 검색 서비스 S1000과 S500을 구축하였다. 검색 서비스 S1000은 1,000만 건의 웹 문서로 구축되었으며, 검색 서비스 S500은 S1000의 웹 문서 중 랜덤하게 추출된 500만 건의 웹 문서로 구축되었다. 따라서, S1000은 S500보다 2배의 절대 커버리지와 상대 커버리지를 갖는 검색 서비스가 된다. 그리고 이를 위해 검색엔진 AidSearch를 사용하였으며[14], 색인은 형태소 단위 색인 방법을 사용하였다[15]. 또한 불리언 연산자를 지원하여 'AND' 연산으로 결합된 스트롱 질의의 사용이 가능하도록 구축하였다.

3.2 커버리지 측정

본 장에서는 3장에서 기술한 방법으로 4.1절에서 구축한 2개의 검색 서비스 S1000과 S500의 커버리지를 측정하였다. 첫 번째 단계에서는 2장에 있는 (그림 1)의 랜덤워크 알고리즘을 사용하여 국내 웹 문서를 수집하였으며, 이 때 초기(seed) URL의 수가 적으면 웹 문서들이 편중되어 수집될 수 있기 때문에 5,000개의 초기 URL을 사용하였다. 그리고 유해 웹 문서나 동적(dynamic) 웹 문서는 수집에서 제외하였다[16]. 이러한 방법으로 본 논문에서는 2005년 3월에서 4월까지 2개월 동안 3회에 걸쳐 <표 1>과 같이 수집하였다. <표 1>에서 웹 문서 수는 수집된 전체 웹 문서수를 의미하고, 중복 제거는 수집된 웹 문서들 중 중복된 웹 문서들을 제거한 웹 문서수를 의미하며, 호스트 수는 수집된 웹 문서들이 포함된 호스트의 수를 의미한다.

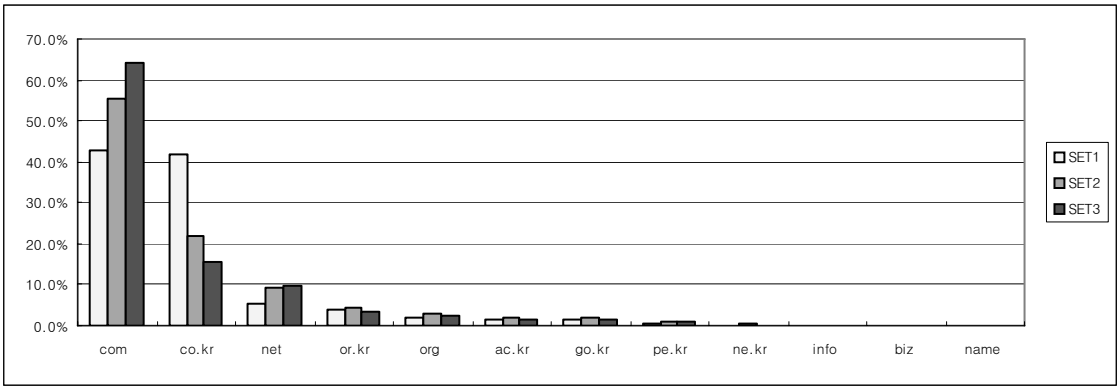
본 논문에서 첫 번째 단계에서 웹 문서들이 균등하게 수집되었는지를 분석하기 위해서 SET1, SET2, SET3 간의 중복도와 도메인별 수집 분포를 조사하였다. 각 SET간의 중복도를 살펴보면, SET1과 SET2간의 중복도는 13.7%, SET1과 SET3간의 중복도는 8.8%, SET2와 SET3간의 중복도는 8.5%로 낮았으며, 이들 3개 SET간의 중복도도 6.1%로 낮았다. 그리고 (그림 3)에서는 각 SET의 웹 문서에 대한 도메인별 수집 분포를 보여주며, 도메인별로 com, co.kr, net, or.kr, org, ac.kr, go.kr, pe.kr 등의 순으로 나타났다. 이 그림으로부터 SET1, SET2, SET3에 수집된 웹 문서들의 각 도메인별 수집 분포가 유사함을 알 수 있다. 이와 같이 SET간의 낮은 중복도와 도메인별 유사한 수집 분포는 각 SET의 웹 문서들이 균등하게 수집되었음을 나타낸다.

커버리지 측정 방법의 두 번째 단계에서는 <표 2>의 SET3에서 랜덤 샘플링 방법으로 10,000건의 웹 문서를 샘플링하였으며 인기도나 중요도가 높은 웹 문서들이 수집되어 있는지를 확인하기 위해 페이지순위 샘플링 방법으로 10,000건의 웹 문서를 샘플링하였다. 그리고 세 번째 단계로써 샘플링 된 웹 문서들이 각 웹 검색 서비스에 색인되어 있는지를 확인하기위해 샘플링 된 각각의 웹 문서가 검색될 수 있는 스트롱(strong) 질의를 생성하였다. 스트롱 질의는 각 웹 문서에서 선정된 10개의 단어를 'AND'로 결합하여 생성하였다. 예를 들어, 웹 문서에서 "정보", "검색", "시스템", "구조", "설명" 5개의 단어가 선정되었다면, 스트롱 질의는 "정보 AND 검색 AND 시스템 AND 구조 AND 설명"가 된다.

네 번째 단계에서는 세 번째 단계에서 생성된 스트롱 질의를 본 실험을 위해 구축된 웹 검색 서비스 S1000과 S500에 입력하였다. 그리고 다섯 번째 단계에서는 입력된 스트

〈표 1〉 랜덤 워크 알고리즘을 사용한 웹 문서 수집

구 분	웹 문서수	중복제거	호스트 수
SET1	3,596,245	1,092,406	139,654
SET2	3,888,541	1,010,366	118,458
SET3	2,730,131	670,530	107,947



(그림 3) 각 SET의 도메인별 웹 문서 수집 분포

〈표 2〉 샘플링 된 웹 문서들과의 중복 문서 수

구 분	랜덤 샘플링		페이지순위 샘플링	
	URL 일치	내용 일치	URL 일치	내용 일치
S500	317	387	307	338
S1000	672	819	629	682

〈표 3〉 절대 커버리지와 상대 커버리지

구 분		랜덤 샘플링		페이지순위 샘플링	
		URL 일치	내용 일치	URL 일치	내용 일치
절대 커버리지	AC(S500)	0.0317	0.0387	0.0307	0.0338
	AC(S1000)	0.0672	0.0819	0.0629	0.0682
상대 커버리지	RC(S500|S1000)	0.4717	0.4725	0.4881	0.4956
	RC(S1000|S500)	1.0000	1.0000	1.0000	1.0000

〈표 4〉 웹 검색 서비스의 절대 커버리지 비교

구 분	랜덤 샘플링		페이지순위 샘플링	
	URL 일치	내용 일치	URL 일치	내용 일치
AC(구글)	0.1395	0.1372	0.2101	0.2062
AC(네이버)	0.0719	0.0875	0.1220	0.1385
AC(엠파스)	0.0378	0.0399	0.0606	0.0657

롱 질의 검색 결과에 노출된 웹 문서들의 URL과 본문을 수집하였다. 마지막으로 여섯 번째 단계에서는 두 번째 단계에서 샘플링 된 웹 문서와의 중복 문서수를 측정하였으며, 이러한 결과를 <표 2>에서 보여 주고 있다. 이 표를 살펴보면, S1000이 S500 보다 높은 중복도가 나타난 것을 알 수 있으며, 2개의 서비스 모두 URL 일치보다 내용 일치에서, 페이지순위 샘플링 보다 랜덤 샘플링에서 높은 중복도가 나타나는 것을 알 수 있다. 그리고 이 표로부터 절대 커버리지와 상대 커버리지를 계산하면 <표 3>와 같다.

3.3 측정 방법론의 정확도 평가

본 장에서는 S1000의 커버리지에 대한 S500의 상대 커버리지는 URL 일치일 경우 2.12배, 내용 일치일 경우도 2.12배가 된다. 그리고 페이지순위 샘플링 10,000건에 대해 S1000에 대한 S500의 상대 커버리지는 URL 일치와 내용 일치에 대해 각각 2.05배, 2.02배이다. 이 결과로부터 S500에 대한 S1000의 상대 커버리지는 약 2배가 되며, 이는 S500에 대한 S1000의 실제 커버리지와 근사한 결과이다. 따라서 본 논문의 랜덤 추출을 이용한 커버리지 측정 방법론이 커버리지 측정에 우수한 성능을 제공한다고 할 수 있다.

4. 국내 웹 검색 서비스들의 커버리지 비교

본 장에서는 3장의 커버리지 측정 방법과 4장에서 수집된 웹 문서를 사용하여 국내 주요 웹 검색 서비스인 구글, 네이버, 엠파스의 절대 커버리지와 이들 간의 상대 커버리지를 측정하여 비교 분석하였다.

4.1 절대 커버리지 비교

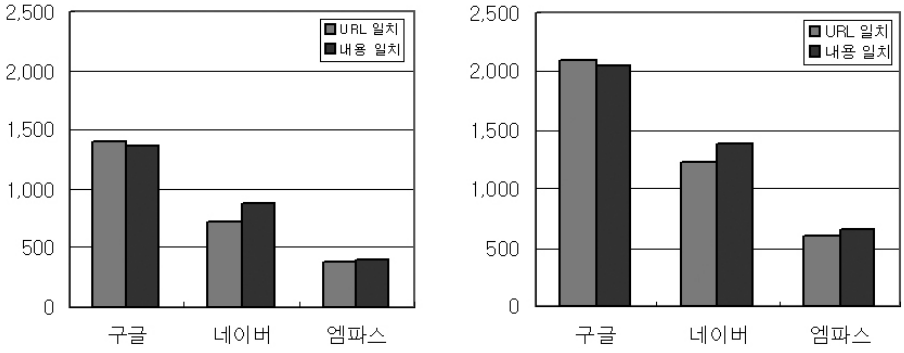
본 논문에서는 4.2절에서 랜덤 샘플링 된 10,000개의 웹문서와 페이지순위 샘플링 된 웹문서 10,000개를 각각 전체 웹으로 가정하고, 3장에서 기술한 방법으로 각 웹 검색 서비스들의 절대 커버리지를 측정하였다. 이때 각 웹 문서를 검색할 수 있는 스트롱 질의는 샘플링 된 각 웹 문서에서 추출된 단어들과 'AND' 연산을 사용하여 생성하였으며, 이를 위해 각 검색 서비스에서 제공하는 고급 검색 방법이나 불리언 질의 처리 방법을 조사하였다.

(그림 4)의 (a)와 (b)는 랜덤 샘플링 된 웹 문서와 페이지순위 샘플링 된 웹 문서 중 각 웹 검색 서비스에 색인된 웹 문서 수를 나타낸다. (그림 4)(a)에서는 랜덤 샘플링 된 웹 문서에 대해 URL 일치 방법과 내용 일치 방법으로 조사한 결과를 나타내며, URL 일치 방법의 경우 구글은 1,395개, 네이버는 719개, 엠파스는 378개의 웹 문서가 일치하였으며, 내용 일치 방법의 경우 구글은 1,372개, 네이버는 875개, 엠파스는 399개의 웹 문서가 일치하였다. 그리고 (그림 4)(b)에서는 페이지순위 샘플링 된 웹 문서에 대해 조사한 결과를 나타내며, URL 일치 방법의 경우 구글은 2,101개, 네이버는 1,220개, 엠파스는 606개가 일치하였으며, 내용 일치 방법의 경우 구글은 2,062개, 네이버는 1,385개, 엠파스는 657개가 일치하였다. 따라서 구글이 네이버나 엠파스보다 많은 웹 문서들을 색인하고 있음을 알 수 있다.

<표 4>에서는 (그림 4)를 기초로 각 검색 서비스들의 절대 커버리지를 계산하였으며, 이 표로부터 구글의 절대 커버리지가 가장 높고, 네이버와 엠파스 순으로 높음을 알 수 있다. 그리고 각 검색 서비스의 절대 커버리지는 랜덤 샘플링 방법보다 페이지순위 샘플링 방법으로 측정하였을 경우에 높게 나타났으며, 이를 통해 각 검색 서비스들은 웹 문서 수집시 페이지순위가 높은 웹 문서들을 우선적으로 수집하는 것을 알 수 있다.

4.2 상대 커버리지 비교

본 절에서는 5.1절의 실험을 바탕으로 상대 커버리지를 측정하였다. 즉, (그림 4)의 각 웹 검색 서비스에 색인된 웹 문서들이 서로 얼마나 중복되어 있는지를 비교함으로써 상

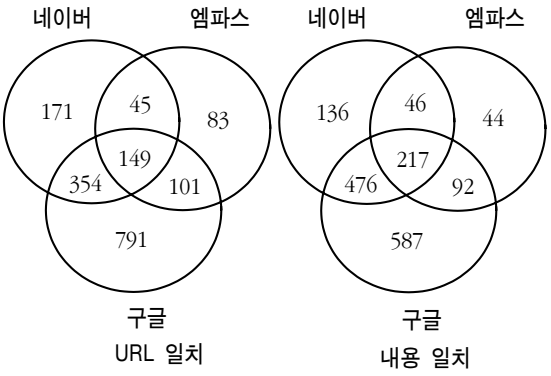


(a) 랜덤 샘플링 (b) 페이지순위 샘플링

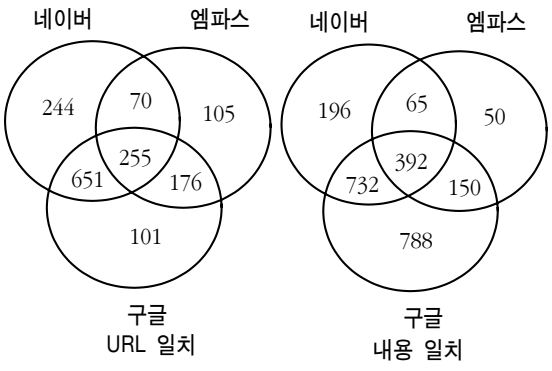
(그림 4) 샘플링 된 웹 문서 중 웹 검색 서비스에 색인된 웹 문서 수

〈표 5〉 웹 검색 서비스 간의 상대 커버리지 비교

구 분	랜덤 샘플링		페이지순위 샘플링	
	URL일치	내용일치	URL일치	내용일치
RC(구글|네이버)	0.6996	0.7920	0.7426	0.8116
RC(구글|엠파스)	0.6614	0.7744	0.7112	0.8250
RC(네이버|구글)	0.3606	0.5051	0.4312	0.5451
RC(네이버|엠파스)	0.5132	0.6591	0.5363	0.6956
RC(엠파스|구글)	0.1792	0.2252	0.2051	0.2629
RC(엠파스|네이버)	0.2698	0.3006	0.2664	0.3300



(그림 5) 랜덤 샘플링에 의한 중복도



(그림 6) 페이지순위 샘플링에 의한 중복도

대 커버리지를 측정하였다. 이를 위해 본 절에서는 웹 검색 서비스간 서로 중복된 웹 문서의 수를 조사하였으며, 그 결과를 (그림 5)와 (그림 6)에서 보여 준다. 이들 그림에서는 랜덤 샘플링과 페이지순위 샘플링에 대해 URL 일치 방법과 내용 일치 방법으로 웹 검색 서비스 간 중복된 웹 문서의 수를 나타낸다. 이 그림으로부터 구글이 다른 웹 검색 서비스의 문서를 상대적으로 많이 색인하고 있음을 알 수 있다.

<표 5>에서는 (그림 5)와 (그림 6)을 기초로 각 검색 서비스들 간의 상대 커버리지를 계산하였다. 이 표로부터 네이버나 엠파스에 대한 구글의 상대 커버리지는 66%~82% 정도로 매우 높으며, 반대로 구글에 대한 네이버나 엠파스의 상대 커버리지는 18%~54%로 낮음을 알 수 있다. 그리고 엠파스에 대한 네이버의 상대 커버리지는 51%~69%로 높으나 네이버에 대한 엠파스의 상대 커버리지는 26%~33%로 비교적 낮았다. 따라서 웹 검색 서비스들간의 상대 커버리지는 구글, 네이버, 엠파스 순임을 알 수 있다.

5. 결 론

웹 검색 서비스는 웹 문서들을 수집하고, 이들을 효과적으로 검색할 수 있는 검색 서비스를 제공한다. 따라서 수집된 웹 문서들의 수는 웹 검색 서비스의 품질을 측정하는 중요한

척도가 되기 때문에 웹 검색 서비스들은 보다 많은 양질의 웹 문서들을 수집하기 위해 노력하고 있다. 그러나 급격히 증가하는 전체 웹 문서의 수를 측정할 수 없기 때문에 웹 검색 서비스들은 전체 웹 문서들에 대해 얼마나 많은 웹 문서를 수집하였는지 또는 다른 웹 검색 서비스들에 비해 얼마나 많은 웹 문서를 수집하였는지 파악하는데 많은 어려움을 겪고 있다. 따라서 본 논문에서는 주요 웹 검색 서비스 간의 상대 및 절대 커버리지를 측정하는 방법을 제안하였다.

실험을 통하여 인터넷에서 웹 문서를 균등하게 샘플링하고, 이 웹 문서들이 웹 검색 서비스에 색인되어 있는지를 조사함으로써 웹 검색 서비스들의 절대 및 상대 커버리지를 측정하였다. 먼저 본 논문에서는 제안한 방법의 정확도를 평가하였으며, 실험 데이터의 커버리지 추정치가 실제 데이터의 커버리지와 근사한 결과 값을 얻었으므로 이 방법의 정확도는 우수하다고 판단되어진다.

이 방법을 사용하여 국내 웹 검색 서비스인 네이버, 구글, 엠파스에 대하여 상대 및 절대 커버리지를 측정하여 비교하였다. 그 결과 구글의 절대 커버리지가 가장 크게 측정되었다. 각 웹 검색 서비스간의 상대 커버리지에서 네이버나 엠파스에 대한 구글의 상대 커버리지는 66%~82% 정도로 매우 높으며, 반대로 구글에 대한 네이버나 엠파스의 상대 커버리지는 18%~54%로 낮음을 알 수 있었다. 그리고 엠파스에 대한 네이버의 상대 커버리지는 51%~69%로 높았으나 네이버에 대한 엠파스의 상대 커버리지는 26%~33%로 비교적 낮았다. 따라서 웹 검색 서비스들 간의 상대 커버리지는 구글, 네이버, 엠파스 순임을 알 수 있었다.

향후에는 본 연구에서 나타난 분석 결과를 토대로 웹 검색 서비스들의 커버리지를 향상시킬 수 있는 방법론에 관한 연구가 필요하다. 웹 검색 서비스의 기능이 향상된다면 이로인한 검색시간의 단축과 동시에 폭넓은 검색을 통한 양질의 정보를 얻을 수 있을 것이다.

참 고 문 헌

[1] S. Lawrence and C. L. Giles, “Searching the World Wide Web,” in Science 280, pp.98-100, 1998.

[2] K. Bharat and A. Broder, “A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines,” In Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, Elsevier Science, pp.379-388, April, 1998.

[3] K. Bharat, A. Broder, M. Henzinger, P. Kumar, and S. Venkatasubramanian, “The Connectivity Server: fast access to linkage information on the Web,” In Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, Elsevier Science, pp.469-477, April, 1998.

[4] S. Lawrence and C. L. Giles, “Accessibility of information on the web,” in Nature, 400, pp.107-107, 1999.

[5] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork, “On Near-Uniform Url Sampling,” in Computer Networks: The International Journal of Computer and Telecommunications Networking, pp.295-308, June 2000.

[6] 김광현, 이준호, “웹 로봇의 성능 평가를 위한 방법론,” 정보과학회논문지, 제3권 제11호, 2004.

[7] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork, “Measuring Index Quality Using Random Walks on the Web,” in Proceedings of the Eighth International World Wide Web Conference, pp.213-225, May, 1999.

[8] S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” In Proceedings of the 7th International World Wide Web Conference, Brisbane, Elsevier Science, Australia, pp.107-117, April, 1998.

[9] J. Carriere and R. Kazman, “Web query: Searching and visualizing the web through connectivity,” In Proceedings of the Sixth International World Wide Web Conference, Santa Clara, California, pp.701-711, April, 1997.

[10] J. Cho, H. Garcia-Molina, and L. Page, “Efficient Crawling Through URL Ordering,” In Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, Elsevier Science, pp.161-172, April, 1998.

[11] L. Page, S. Brin, R. Moteani, and T. Winograd, “The PageRank citation ranking: bringing order to the Web,” Manuscript in Progress, 1993.

[12] 배희진, 이진숙, 이준호, 박소연, “국내 웹 디렉토리의 커버리지 및 커버리지 중복성 분석,” 정보관리학회지, 제21권 제1호, pp. 173-186, 2004.

[13] 김성진, 이상호, “웹 로봇 구현 및 한국 웹 통계 보고,” 정보처리학회논문지, 제10-C권 제4호, pp.509-518, 2003.

[14] 이준호, 김광현, 김지승, “다양한 한글 문서 색인 방법들에 대한 평가,” 제5회 한국 과학기술 정보인프라 워크샵 학술발표논문집, 2002.

[15] 이준호, 이충식, 한선화, 김진영, “문자 인식에 의해 구축된 한글 문서 데이터베이스에 대한 정보 검색,” 한국정보처리학회 논문지 A, Vol.06, No.04, pp.833-840, 1999.

[16] 김광현, 최정미, 이준호, “웹 문서 분석에 근거한 유해 웹 문서 검출,” 정보처리학회논문지D, 제12-D권 제5호, pp.683-688, 2005.

장 성 수

e-mail : lovejang@dreamwiz.com

1985년 숭실대학교 컴퓨터학부(학사)

2000년 숭실대학교 정보과학대학원 컴퓨터학과(석사)

2001년~현 재 숭실대학교 대학원 컴퓨터학과 박사과정

관심분야 : 정보검색

김 광 현

e-mail : iamkkh@naver.com
1999년 숭실대학교 컴퓨터공학부(학사)
2001년 숭실대학교 컴퓨터학과(석사)
2006년 숭실대학교 컴퓨터학과(박사)
2000년~현 재 NHN (서치솔루션) 근무
관심분야 : 정보검색, 웹로봇, 기계학습, 유해웹문서 필터링

이 준 호

e-mail : joonho@comp.ssu.ac.kr
1987년 서울대학교 컴퓨터공학과(학사)
1989년 한국과학기술원 전산학과(석사)
1993년 한국과학기술원 전산학과(박사)
1993년~1994년 한국과학기술원 인공지능연구센터 연구원
1994년~1995년 코넬대학교 전산학과 방문연구원
1994년~1997년 연구개발정보센터, 선임연구원
1997년~현 재 숭실대학교 컴퓨터학부 부교수
관심분야 : 정보검색