제19회 한국정보처리학회 춘계학술발표대회 논문집 제10권 제1호 (2003.5)

웹문서의 구조적 정보 활용 사례에 관한 고찰

김철수, 김양범
서남대학교
e-mail:chskim@tiger.seonam.ac.kr, ybkim@tiger.seonam.ac.kr

A Survey on Example using Structural Information of Web documents

Cheol-su Kim, Yang-beom Kim
Seonam University

요약

브라우저를 통해서 보는 웹 문서는 보이는 문서 내용 이외에 많은 풍부한 정보들을 원시 문서에 포함하고 있다. 웹 문서 색인 및 분류 과정에서 이런 관련 정보를 효율적으로 사용한다면 색인어에 가중치를 부여하거나 문헌 분류과정에서 밀접하게 관련된 문헌들끼리 분류가 가능하다. 잘 색인된 색인어 집합 및 잘 분류된 문헌 집합은 검색자의 질의어에 대한 검색 결과의 문헌집합들에 대한 문헌 순위화를 효율적으로 수행하여 사용자로 하여금 검색 시간을 줄여줄 수 있다.

본 논문에서는 웹 문서를 대상으로 한 검색 시스템에서 검색 효율을 향상시키기 위하여 웹 문서의 구조적인 정보들을 이용한 방법들에 대하여 고찰해 본다. 색인 과정, 문헌 분류과정, 순위화 과정에서 활용한 방법들에 대하여 중점적으로 살펴본다.

1. 서론

웹 문서를 대상으로 하는 검색 시스템은 전통적인 검색 시스템과는 다른 성격을 가진다. 검색 대상이 될 수 있는 문헌들이 지리적으로 넓은 지역에 분산되어 있고, 웹문서의 생성과 소멸이 수시로 일어난다. 뿐만 아니라 웹 문서의 수가 폭발적으로 증가하여 웹 문서수가 2배로 증가하는데 2년이 채 걸리지 않는다[1].

정보검색 시스템의 성능은 일반적으로 재현율과 정확률로 평가한다. 웹 문서에 대한 검색 시스템의 평가는 웹 문서의 특성에 기인하여 전통적인 검색시스템의 정확률과는 조금 다른 의미를 가진다. 웹 문서 수의 폭발적인 증가에 따라 검색 결과 역시 많은 수의 문서들이 검색된다. 따라서 이용자가 검색된 결과 문헌을 모두 조사할 수 없으므로 대부분의 경우, 질의어와 유사도가 높을 것으로 예상되는 문헌들을 먼저 보여주는 문헌 순위화 과정을 거친다. 이 순위화 과정을 수행하기 위해서는 질의어와 색인어들 사이의 유사도 정도를 반영할 수 있는 색인어 가중치가 필요하다.

웹 문서에 대한 색인 과정은 전처리과정을 거쳐 얻어진 문장 및 문자열에 대하여 형태소분석과 같은 자연어 처리 과정을 거쳐 문서를 대표할 수 있는 색인어들을 얻게되며, 이 색인어들이 검색 시스템의 색인어 저장되고, 질어의 처리 과정을 거쳐 질의어와 저장된 색인어들 사이의 관련된 문서들을 검색하여 보여준다.

검색된 문헌 순위화를 위하여 검색어와 색인어 사이의 유사도 계산을 위하여 주로 빈도수와 같은 통계적인 방법들을 많이 이용하여 왔으나, 보다 효율적인 순위화를 위하여 색인어 가중치를 부여하는 방법이 사용된다.

브라우저를 통해서 보는 웹 문서는 보이는 문서 내용 이외에 많은 추가적인 정보들을 원시 문서에 포함하고 있다. 정보검색을 색인 과정에서 이런 관련 정보를 활용하여 색인어에 가중치를 부여하거나 문헌을 분류하므로 써 정확성을 향상시킬 수 있다.

본 논문에서는 웹 문서를 대상으로 한 검색 시스템에서 웹문서가 포함하고 있는 내용 정보와 내용 정보를 표현하는 과정에서의 구조적인 정보들을 검색 시스템에서 어떻게 활용하고 있는지에 대하여 살펴보고, 색인 과정 및 문헌 순위화 방법들에 대해서 고찰한다.

제 2 장 웹 문서 색인

웹 문서를 대상으로 한 정보 검색 시스템의 구조를 개략적으로 살펴보면 그림 1과 같다.

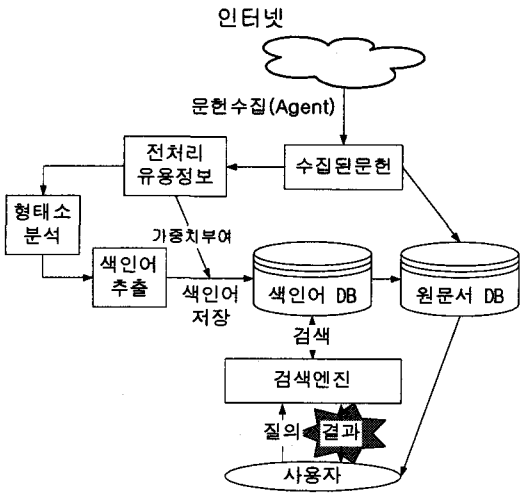

그림 1 웹 정보 검색 시스템 구조도

웹 정보검색 시스템을 이용하는 사용자의 관심사는 검색 엔진의 검색 결과이다. 웹 문서의 증대에 따라 검색된 문헌 수도 매우 많이 검색된다. 그러나 검색 된 문헌 집합의 상위 순서에 즉, 첫 번째 혹은 두 번째 검색 결과(상위 20개 정도의 문서) 화면에 관련 문헌을 먼저 보여주므로 써 검색 시간의 효율성을 증진시킬 수 있다. 문헌 순위화를 수행하기 위해서는 색인 대상인 문헌의 중요도를 반영해 줄 수 있어야 하고, 동일 문서에서 선별된 색인어라 하더라도 문헌을 대표하는 정도(중요도) 차이가 있으므로 색인된 단어가 문헌을 대표하는 정도 차이를 반영해 주므로 써 질의어에 대한 검색 모델의 처리 과정에서 문헌 중요도 및 색인어 중요도를 참조하여 문헌 순위화를 수행한다. 중요도를 반영하기 위하여 빈도수에 근거한 통계적 방법들이 많이 이용되어 왔으나, 빈도수에 근거한 방법보다는 색인어의 색인 과정에서 색인어에 대한 가중치를 부여하여 질의어 처리과정에서 사용하는 것이 더 효율적이다.

선별된 색인어에 대한 가중치를 부여하기 위해서는 문헌안에 문헌 내용의 중요도 정도를 나타내주는 구조화된 정보가 필요하다.

P_NORM 검색의 문헌 순위화에 관한 실험적 연구[2]에서는 통계적 방법보다는 비통계적 방법을 이용한 방법이 재현율과 정확율 모두 향상된 방법임을 보여주고 있으며 특히 색인어들이 나타난 필드(제목, 요약, 제목과 요약)에 따라 다른 가중치를 달리 부여한 필드 가중치 방법이 다른 가중치를 부여한 방법보다 우수한 방법임을 보여주고 있다.

웹 문서의 경우, 많은 태그들로 구성되며 구조화된 태그에 의해 표현된 문자열들은 나름대로 그 문서에서 내용의 중요도 정보를 제공해 주고 있다. 예를 들어 그림 2와 같이 동일한 임의의 웹문서 본문에서 선별된 색인어 "정보검색"과 "통계데이터", "정보처리"가 있을 때 하이퍼링크가 표시된 문자열에서 선별된 색인어 "정보검색", 진하게 표시된 문자열에서 선별된 색인어 "통계데이터", 보통 글씨의 문자열에서 선별된 색인어 "정보처리"는 문서를 대표하는 정도 즉, 색인어의 중요도가 다르다. 따라서 논문과 같은 경우 제목, 요약에서 선별된 색인어에 대하여 가중치를 다르게 주는 것 뿐만 아니라 본문 내용에 대해서도 가중치를 다르게 부여하여야 한다.

정보검색 시스템의 이해.....
.....
..... **통계데이터를 활용하는 방법** .....
.....
정보처리 과정을 거쳐 .....

그림 2 임의의 웹 문서 본문

3. 링크 정보 의 활용

링크정보의 활용 분야는 크게 3가지 분야로 나누어 볼 수 있다. 첫째는 색인과정에서 활용하는 경우로 위에서 언급한 링크 문자열에서 선별된 색인어는 다른 색인어들에 비하여 높은 가중치를 부여하는 경우이다. 둘째는 문헌 분류에서 다음과 같은 관계를 이용하여 두 개의 서로 다른 문서가 관련 있는 것으로 보고 있다[3]. ①연결 관계로 임의 문서 A에서 다른 문서 B를 링크하고 있다면 두 문서는 서로 관련이 있다. ②공동인용관계로 임의의 한 문서에서 두 개의 다른 문서를 링크하고 있다면 2개의 문서는 서로 관련 있다. ③Social Filtering 관계로 두 개의 서로 다른 문서가 공통의 한 문서를 링크하고 있다면 두 문서는 서로 관련 있다. ④전이 관계로 임의 문서 A가 문서 B를 링크하고, 문서 B가 문서 C를 링크하고 있다면 문서A와 문서 C는 서로 관련 있다. 이러한 관계와 휴리스틱 정보를 활용하여 관련 있는 문헌들끼리 문헌 클러스터링을 수행할 수 있다. 또한 다른 문서에 의해 링크되는 횟수에 의해

문헌의 중요도를 계산하여 그 문헌에서 색인되는 단어들에 대하여 가중치를 부여할 수 있다. 또한 링크의 진출 차수 에 따라 문서의 중요도를 반영할 수도 있다.

[4]에서는 하이퍼링크 정보를 이용하여 중요 문헌들을 선별하는 과정에 이용한다. 링크 관계가 있는 문헌들을 네트워크 구조로 표현한다. 링크 관계가 있는 문서들을 정점(Vertex)들의 모임에서 노드는 문서에 대응하고, 문서 p에서 문서 q를 링크한다면 방향성에지(Directed edge) (p, q) ∈ E로 정의하여 방향성 그래프 G = (V, E)로 볼 수 있다. 이 전체 그래프에 대하여 알고리즘에 의한 처리 절차를 거쳐 작은 부분그래프들로 분리해 낼 수 있다. 여기에서 분리된 작은 그래프에 포함된 노드들은 서로 관련된 문서들의 집합으로 볼 수 있으며, 질의어와 관계된 문서들 가운데 잘 정제된 소량의 정점들은 매우 유용한 문서로 볼 수 있다.

셋째는 검색된 문헌의 랭킹 과정에서 활용하는 경우이다[1]. 단순한 페이지 랭크 모델은 페이지랭크를 위해 검색 시스템에 저장된 문헌들의 번호를 1~m, 문서 번호 i의 outgoing link 수를 N(i), 문서번호 i를 가리키는 페이지들의 집합을 B(i)로 정의하면 문서 i의 페이지 랭크값 r(i)는 다음 식을 이용하여 계산한다.

$$r(i) = \sum_{j \in B(i)} \frac{r(j)}{N(j)}$$

검색엔진 구글의 경우 각 웹페이지의 품질 랭킹 계산 및 검색 결과 향상을 위하여 링크 정보를 이용한다[5]. 이 방법의 경우 직관적인 정보도 페이지 랭크에 활용한다. 예를 들면 많은 페이지들이 가리키고 있는 문서 혹은 소수의 페이지가 가리키는 문서이지만 페이지랭크가 높은 페이지가 가리키고 있다면 높은 페이지랭크를 가질 수 있다. 즉, 일반적인 경우 많이 인용되는 문서는 높은 페이지랭크를 가질 것이며 일반적인 홈페이지에서 인용되는 문서보다는 신뢰할 만한 홈페이지에서 인용하는 페이지가 상대적으로 높은 페이지랭크를 가질 것이다.

HITS(Hypertext Induced Topic Search) 방법[4]은 검색 질의어에 의존적인 방법으로 써, 문헌마다 단일 점수가 부여된 문헌 순위화 방법이 아닌 Authority 페이지(질의어와 매우 관련 있을 것 같은 페이지)와 Hub 페이지(페이지 자체가 Authority 페이지일 필요는 없지만 Authority 페이지를 가리키는 페이지)로 분류 생성한다. 검색자의 관심은 질의어와 관련이 있을 것 같은 Authority 페이지에 있지만 Hub 페이지도 Authority 페이지를 찾아내는 HITS 알고리즘에서 이용하므로 필요하다.

또한 에이전트을 이용한 문헌 수집시 링크되는 URL 정보를 활용하여 신뢰성있는 URL의 문서를 수집하는 과정에서도 활용할 수 있다.

4. 메타데이터의 활용

메타데이터 역시 브라우저에서는 보이지 않지만 유용한 정보들이다. 이 정보는 인터넷상의 정보검색은 빠르고 정확하게 하는데 도움을 준다. 웹 문서 제목, 제작자, 주제, 내용, 공헌자, 날자, 자원타입, 형식, 제작 프로그램, 제작 날자, 키워드 등 여러 가지 정보를 가지고 있다. 따라서 모든 웹 문서 생성시 메타데이터를 완벽하게 작성한다면 이 메타데이터에 포함된 정보들만을 가지고도 효율적인 검색엔진을 만들 수 있다.

그러나 몇가지 문제점들을 가지고 있다. 이 메타데이터는 반드시 작성해야 하는 의무 사항이 아닌 선택 사항으로 문서 작성시 웹 문서에 삽입하고 싶은 내용들만을 포함시킨다는 점이다. 따라서 메타 파일만을 이용한 색인은 재현율을 현저히 저하시킬 수 있는 요인들 담고 있다. 수집된 문서 전체를 관리하기 위해서는 문서 갱신 시 잡다하고 지루할 수 있다. 또한 서로 다른 문서 집합들을 결합할 때 두 문서 집단에 포함된 문서들이 가지고 있는 메타 정보들의 불일치 문제가 발생한다.

이런 문제들을 해결하기 위해서 메타데이터 코딩 구조로 RDF(Resource Description Framework)를 사용할 수 있다[6].

5. 결 론

정보검색에서 사용자가 원하는 정보에 쉽고 빠르게 접근하기 위하여 다양한 방법이 연구되고 있다. 웹문서의 급증에 따라 검색된 문헌 수 역시 매우 많다. 이처럼 검색 문헌수가 많을 경우 검색된 문헌을 대상으로 원하는 문서를 찾아내기 위해 많은 문서를 참조해야만 하는 단점이 있다. 이런 단점을 개선하기 위해 문헌 순위화 과정 제공하거나 주어진 질의어와 관련성이 많을 것으로 예상되는 소수의 문헌들만을 질의어의 결과 값으로 제공해 주기도 한다.

웹 문서는 브라우저를 통해서 보여주는 문서 내용 이외에 문서 내용을 브라우저를 보여주기 위해

필요로 하는 필수적 개념의 많은 구조적 정보(태그) 정보와 브라우저를 통해 보여주는 문서 내용에는 영향을 주지 않지만 부가적인 유용한 정보들을 많이 포함하고 있다.

본 논문에서는 웹 문서들 대상으로 한 정보 검색 시스템에서 검색 결과의 성능을 향상시키기 위하여 웹 문서 구조적 정보들을 어떤 과정에서 어떻게 사용하는지 몇가지 사례를 통해 살펴보았다.

웹문서의 실질적인 내용들에 대한 색인 과정은 필수적인 요소이며 웹 문서의 원시 파일에 포함되어 있는 구조적인 정보들을 색인과정, 문헌분류, 문헌수집, 문헌 순위화 과정 등 정보검색 시스템을 구성하는 여러 요소들에서 이용할 수 있으며 이런 각각의 과정 및 요소들은 별개의 것이 아닌 서로 밀접한 관계를 가지고 있으므로 웹 문서의 구조적인 정보를 잘 활용하므로 써 검색 시스템의 성능을 향상시킬 수 있다.

참고문헌

[1] Arvind Arasu, etc 4, Searching the Web, Acm Trans. on Internet Technology, Vol. 1, No, 1, pp. 2-43.

[2] 고미영, P-NORM 검색의 문헌 순위화 기법에 관한 실험적 연구, 연세대학교 대학원 박사논문, 1998.

[3]. Wen-Syan and K. Selcuk candan, Integrating content Search with structure Analysis for Hypermedia Retrieval and Management, ACM Computing Survey, Vol. 31, Num. 4es, pp.1-5, 1999.

[4] Jon M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, Journal of the ACM, Vol. 46, No. 5, pp. 604-632, 1999.

[5] Sergey Brin and Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm

[6] Mei Kobayashi and Koichi Takeda, Information Retrieval on the web, ACM Computing Survey, Vol. 32, No. 2, pp. 143-173, 2001.