HITS알고리즘을 적용한 개념그래프 기반검색시스템의 성능개선

배환국° ,박호성° ,이상준° ° ,김기태°
중앙대학교 컴퓨터공학과°
제주대학교 통신컴퓨터 공학부° °

Improved Concept-base Search System Using HITS algorithm on Conceptual Graph

Haen kuk Bae° ,Ho seong Park° , Sang joon Lee° ° , Ki tae Kim°
Dept. of Computer Science & Engineering, Chung-Ang University°
Cheju National Univ. Faculty of Telecommunication & Computer engineering° °

요 약

본 논문에서는 개념 그래프 기반 검색 시스템의 검색의 성능을 개선시키고자 Hits 알고리즘을 적용하였다. 기존 개념 그래프 기반 검색 시스템의 anchor text 분석을 통하여 개념을 추출하고 있는 시스템에서 더 나아가 하이퍼 링크의 선호도의 특성을 살려 하이퍼링크에 문서가 얼마나 연결되어 있는지,참조하고 있는지에 따라 해당 검색된 문서들의 중요도를 찾아서 순위를 매기는 실험을 하였다.

종래에는 해당 검색어의 빈도 순으로 개념의 결과를 나타내 주었는데, 본 시스템 구현 후에 랭킹알고리즘을 적용하여 해당검색에 유용한 정보를 가지고 있는 페이지들(authorities)과 유용한 정보를 보유하고 있는 페이지의 링크를 보유하고 있는 페이지들(hubs)를 각각 순위 순으로 보여주게 되었다.

그리하여 사용자는 실제 검색 시에 개념상으로 분류된 문서 중에 중요도가 높은 문서를 사용자에게 우선으로 접하게 되었으며,hub에 의해서 중요도가 높은 문서를 한눈에 볼 수도 있을 뿐 아니라,anchor text에서 나타나지 않은 중요한 정보를 가진 문서도 검색할 수 있었다.

1. 서론

웹 그래프는 개념기반 검색 방법을 이용한 검색엔진이다. 웹 그래프는 웹 문서의 키워드를 추출하기 위하여 하이퍼링크의 Anchor text를 이용하였다. 그리하여 웹 문서의 핵심어를 간단하고 빠르게 추출하며, 웹 문서 간의 하이퍼링크를 각 웹 문서의 핵심어 간에 링크관계로 추상화하여, 이 관계를 이용하여 핵심어의 개념 그래프를 구축하고 질의의 확장이나 영역지식을 제공하는 개념 기반이 가능한 시스템이다[최98].

이 검색시스템이 개념으로 묶은 싸이트들이 그 개념 속에서 각 싸이트들이 어떻게 상호관계를 맺는 것에 대해서 간과하고 있었다. 개념그래프상에 개념단위로 분류한 페이지들 서로간의 하이퍼 링크의 선호도에 의한 경쟁으로써 그 개념 속에서 우리가 찾고자 하는 정보에 관하여 분류되어 있는 문서들 상에서도 가중치가 높은 문서를 우선적으로 보여주어서 사용자에게 더욱 정확한 정보를 빠르게 찾도록 하는 데에 초점을 맞추었다.

이렇게 분류된 것들 안에서 더욱 정확하고 검색어에 가까운 정보를 보여주기 위하여 HITS 알고리즘을 적용하여 실험을 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문의 기반 연구 분야들을 살펴보고,3장에서 본 논문에서 제안한 Hits 알고리즘 적용한 개념그래프검색시스템에 대해서 설명한다.

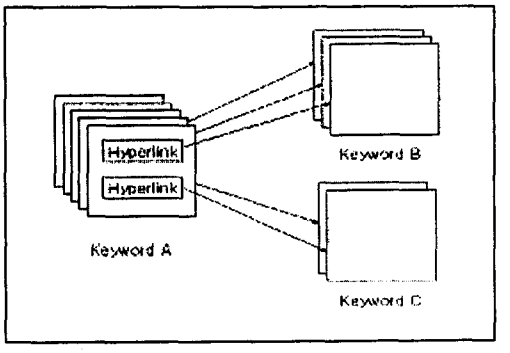
4장에서는 구현된 시스템의 대한 실험결과를 보이고 5장에서는 결론을 맺고 추후 연구 과제를 기술한다.

2. 관련 연구

2.1. 웹 그래프의 개념추출 및 생성방법

웹 그래프는 웹 문서의 핵심어 추출을 위하여 하이퍼링크 정보 중 하나인 Anchor Text 와 웹 문서의 제목(title) 태그를 이용하여 추출한다. 핵심어를 추출한 후 웹 문서마다 하나 이상의 핵심어를 가지게 된다.

각 문서의 핵심어와 하이퍼링크인 링크를 이용하여 개념을 생성한다.즉,특정 핵심어를 가진 웹 문서가 연결하고 있는 웹 문서들의 핵심어 리스트중 동일한 핵심어를 가지는 것끼리 분류를 한다.핵심어 별로 분류된 문서들은 링크로 추상화하여 웹 문서간의 관계를 핵심어 간의 관계로 바꾸어 개념을 생성한다.



[그림 1] 하이퍼링크로 생성된 웹 문서간의 관계도

2.2 하이퍼링크 정보의 특성 중 선호도

웹 문서가 다른 웹 문서에서 하이퍼링크로 얼마나 연결되었는가를 나타내는 "visibility"는 월드 와이드 웹 상에서 그 문서가 얼마나 중요한가를 나타낸다. 마치 논문 중 참고문헌으로 많이 사용된 논문이 우수하게 평가 받는 것과 마찬가지이다.

2.3 HITS (Hyperlink Induced Topic Search)

기본전제

1.quality한 문서들을 찾는 간단한 방법으로써

만약 A는 문서 B의 하이퍼링크를 가지고 있다면,그러면 문서A의 저자는 문서 B가 A에 관한 다양한 정보를 담고 있다고 생각한다.

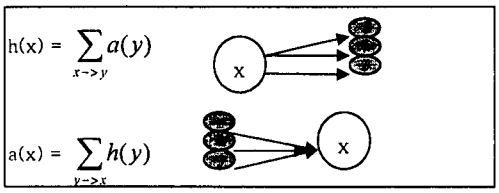
2. 만약 A가 많은 좋은 문서들을 가리 키는 것처럼 보인다면, A의 의견은 좀더 다양하고 A가 B를 가리키는 사실은 B또한 좋은 문서라는 생각을 제안한다.

쿼리에 응답하여, 각각 접하는 페이지들을 정렬하여 보여주는 대신에, 내부적으로 관련된 두 개의 집합을 찾는 것이 목표이다.

1. Authorities:: 주제에 관하여 유용한 정보를 가진 페이지들을 인식한다.

2. Hubs: 주제에 관한 유용한 정보를 가진 페이지의 많은 링크를 가진 문서들을 인식한다.

좋은 hubs는 주제에 관한 많은 authoritative 페이지들을 가리키고 있는 page를 말하며, 좋은 authoritative 페이지는 주제에 관한 많은 좋은 hubs들에 의하여 지적되고 있는 페이지를 말한다.



[그림2] authority page와 hub page

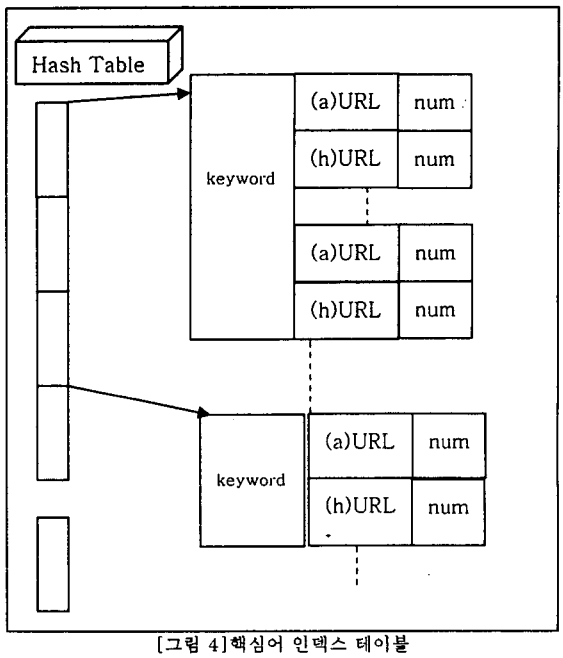
3.HITS알고리즘을 적용한 개념기반 시스템

본 시스템에서는 핵심어를 가지는 것끼리 분류 시에 같은 단어가 있을 때 그 핵심어의 url을 핵심어의 authority url저장소에 동일한 url이 들어있는지 체크한 후 동일한 url이 있으면 해당url에 가중치를 증가시킨다.
이것은 해당 단어에 대해서 해당 url이 많이 쓰인다는 것이기 때문에 authority가 증가하는 페이지라고 할 수가 있다. 또한 이 url을 지적했던 페이지를 찾아서 hub url을 저장할 수 있는 곳에서 마찬가지로 동일한 url이 있는 지 체크하게 되고, 있으면 hub url의 갯수를 증가시키게 된다.이것도 마찬가지로 가리킴을 많이 받는 갯수가 증가하게 되는 것이므로 다른 페이지보다 이 핵심어에 대해서 많은 정보를 가진 페이지라고 할 수 있다.
그리하여 핵심어 별로 hub와 authority를 내림차순으로 보여준다.
[그림3]은 해당시스템에 들어간 알고리즘이다.

```
전체 URL의 각각의 URL에 대하여
1. DB에서 제목을 가지고 온다.
2. Anchor Text를 가지고 온다.
3.  1.제목과 Anchor Text를 공백을 빼어 단어별로 m_key에 저장
    2. m_key에 저장된 각각의 단어에 대하여
        IF 인덱스 테이블이 존재하면
                (단어개수)++
            IF현재의 url이 리스트에 존재하면
                현재 authority url갯수++;
                hubRanking(현재URL);
            ELSE
                현재 URL을 추가;
                현재 authority url갯수 := 1;
                hubRanking(현재URL);
        ELSE
            테이블을 새로운 레코드로 현재의 단어를 삽입
              (단어갯수) := 1;
            현재 URL을 authority 저장소에 저장
                hubRanking(현재URL);
```

[그림3]HITS알고리즘이 적용된 테이블 작성 알고리즘

그리하여 검색결과를 보여줄 때 중가치가 가장 큰 순서로 보여주게 된다. 이때 중가치가 높을수록 해당 핵심어와 authority가 높은 문서라고 할 수가 있다. 또한 이런 높은 authority를 참조하고 있는 hub의 리스트들도 검색결과에 보여준다. hub에 관한 것도 마찬가지로 중가치가 가장 높은 순으로 보여주게 된다



[그림 4]핵심어 인덱스 테이블

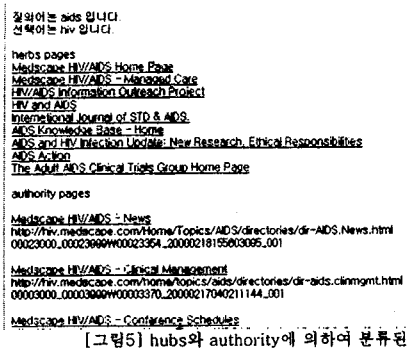
4.실험 및 평가

4.1실험환경

의학관련 초기 URL들을 시작점으로 하여 수집기를 돌려 각 URL내에 존재하는 하이퍼링크를 따라가며 웹 문서들을 수집하였다.
URL을 저장하기 위한 큐로 DB를 사용하였으며 수집된 URL총 12570개,URL에 존재하는 하이퍼링크는 34570개이다.
검색 서버 엔진은 WINDOW XP상에IIS서버상에 동작하였으며 클라이언트서버 브라우져상에 그래프를 나타내기 위하여 JAVA Applet를 사용하였다.
http://www.ama-assn.org/
http://www.medscape.com/
http://www.hc-sc.gc.ca/
초기 URL 리스트

4.2 실험평가

여러 가지 검색어에 실험을 하여 나온 결과값을 분석해본 결과 기존의 검색엔진에서는 검색어의 빈도순으로 결과값을 보여주었는데, 본 시스템에서는 hubs로써 순위가 매겨진 검색 값과 authority가 높은 순으로 되어있는 페이지들의 검색결과를 순위로 매겨 리스트로 나오게 되었다
[그림5]는 페이지에서 hubs와 authority페이지의 순위대로 나온 것을 나타낸 그림이다.
그리하여 사용자는 문서 중에 순위가 높은 것을 먼저 볼 수 있게 되었으며 보통10개 자료 중에서 본인이 원하는 정보를 찾을 수 있었다.
페이지가 두 집합으로 분류하여 보여주고 있으며, hub상에 나타난 검색 결과물의 문서를 가보면 검색어에 다양한 정보를 링크하고 있는 것을 한눈에 볼 수 있어서 사용자가 연결되어있는 문서를 능동적으로 파악하여 정보를 찾을 수 있게 되었다.
또한 authority가 높은 순으로 검색된 페이지는 검색어와 개념어가 모두 포함된 anchor text가 대부분을 차지 하였지만, 그 단어가 들어있지 않은 페이지도 순위에 올라 있었다.논문이나 참고서적에 관한 링크가 순위에 많이 있었으며,새로운 사실이나 뉴스 등에도 많이 있었다.



[그림5] hubs와 authority에 의하여 분류된 페이지들

하지만 3개의 싸이트에서 수집한 문서만으로 실험을 하여 url의 순위를 결정하는 값의 크기가 크지 않았다. 문서의 순위의 선호도를 높이기 위해서는 더욱더 많은 문서를 수집하여 실험을 해야 할 것이다.

5. Conclusion and Future Works

본 논문에서는 anchor text와 title을 이용하여 개념을 분류하던 종래의 개념기반시스템에서 hits 알고리즘을 적용한 하이퍼링크의 선호도를 첨가하여 종래에는 찾지 못했던 문서나 해당 검색어에 대하여 사용자에게 얼마나 가깝고 유용한 정보를 갖춘 검색문서를 우선 재시하여 원하는 정보를 얼마나 정확하고 빠르게 자료를 보여줄 것인가를 연구하였다.
그리하여 검색어에 분류된 문서들 중에 사용자에게 검색어에 관련된 문서를 우선 보여줌으로써 사용자는 정보에 대해서 좀 더 편리하게 정보를 찾을 수 있게 되었다. 뿐만 아니라 종래 시스템에서는 찾을수 없었던 문서까지 볼수 있게 되어서 범위가 좀 더 확대 되었다.
하지만 검색엔진의 정확성을 더욱 더 높이기 위해서 full text를 검색하여 좀더 세세한 부분까지 분석해야 하고, 개선된 hits 알고리즘을 이용하여 종래의 가중치를 1로 잡던 것을 문서의 다른 특성을 발견하여 가중치를 적절히 조절하여 문서의 정확도와 순위를 정확히 결정짓는 지표를 찾는 개발이 요구된다.
또한 순위의 경쟁을 높이기 위해서는 많은 문서들을 분류하여 경쟁시키도록 다량의 문서를 수집하는 것도 잊지 말아야 할 과제일 것이다.

참고문헌

[ht98]David Gibson, Jon Kleinberg,Prabhakar Raghavan. " Inferring Web Communities from Link Topology (1998)" ,UK Conference on Hypertext

S. Chakrabarti, B. Dom, D. Gibson, J. Keinberg, P. Raghavan, and s. Rajagopalan , and s. Rajagopalan,"Automatic resourc e compilation by analyzing hyperlink structure and associated text(1998)" ,the 7th International World Wide Web Conference

[신 01] 신일수 " 개념 그래프 기반 문맥 유지 확장을 이용한 개념기반 검색 시스템, 중앙대학교 96회 석사학위 논문,2001

[박 00] 박사준, 김상경, 황수철, 김기태, " 전문가 검색 엔진에서 개념그래프를 이용한 Web 정보 획득" , 한국정보과학회 논문집 제 27 권 1 호 2000.4

[이 00] 이권국, 신일수, 이상준, 김기태, " 전문가 검색엔진에서 데이터 마이닝을 이용한 개념관계 추출" , 한국정보과학회 논문집 제 27 권 1 호 2000.4.

[조 99] 조민재, " 웹의 개념 지식을 이용한 자동 서소러스 생성법의 설계 및 구현" , 중앙대학교 92 회 석사학위 논문, 1999

[최 98] 최준영, " 인터넷상의 하이퍼링크를 이용한 개념 그래프 기반 검색 시스템" , 중앙대학교 90 회 석사학위 논문, 1998

Jon M.Kleinberg,"Authoritative Sources in a Hyperlinked Environment(1999)",Journal of the ACM.

Ruey-Lung, Hsiao ," Web Structure Mining - PageRank and HITS lecture note".

Soumen Chakrabartiv ," Recent Results in Automatic Web Resource Discovery " , ACM Computing Survey,December, 1999.
CS276A Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Text Information Retrieval,Mining,and Exploitation Lecture 12, 14 November,2002

[Amit98]E.Amitay.Using common hypertext links to identify the best phrasal description of target web documents. In ACM SIGIR 98 Workshop
on Hypertext IR for the Web,Melbourne,1998

[Amit00]E.Amitry and C.Paris. Automatically summarizing web sites-is there a way around it? In ACM 9th International Conference on Information and Knowledge Management(CIKM 2000),Washington,DC,200

[Berry98] Michael J. A. Berry, Gordon Linoff, "Data Mining Techniques: For marketing, Sales, and Customer Support", Wiley Computer Publishing, pp 216-242, 1998

[Brin98] Sergey Brin, Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Proceeding of 7th World Wide Web Conference, 1998
[Jiawei2001] Jiawei Han, Micheline Kamber, "Data Mining - Concepts 1. and Techniques", Morgan Kaufmann 2001

[Marchiori97] Massimo Marchiori, "The Quest for Correct Information on the Web: Hyper Search Engines", Sixth International World Wide Web Conference, 1997