정보관리연구, vol. 34, no. 2, 2002, pp. 55-66

웹 정보검색 시스템의 문서 순위 결정

Document Ranking of Web Document Retrieval Systems

안 동 언* · 강 인 호**

Dong-Un An · In-Ho Kang

차 례

초 록

인터넷의 발달로 인해 웹에서 얻을 수 있는 정보의 종류와 수는 급진적으로 증가하고 있다. 기존의 문서 위주의 구성에서 멀티미디어 서비스, 쇼핑몰 등 종류와 매체에 있어서 다양한 변화를 보이고 있다. 이에 따라 사용자가 요구하는 정보의 단위는 문서 뿐만 아니라, 사이트 그리고 서비스 단위로 확장되고 있다. 웹 환경에서 사용자의 정보 요구를 보면 크게 세가지로 볼 수 있다. 첫째는 원하는 정보를 설명하는 혹은 정보와 관련된 문서를 찾는 내용검색, 둘째는 사용자가 관심 있어 하는 개인이나 단체의 사이트 입구를 찾는 사이트 검색, 셋째는 사용자가 관심 있어 하는 서비스를 제공하는 웹 페이지를 찾는 서비스 검색을 들 수 있다. 본 논문에서는 이러한 사용자의 정보 요구 목적에 따라서 문서 순위화가 달라져야 함을 보인다. 지금까지 정보 검색에서 언급된 내용 정보, 링크 정보 그리고 URL 정보의 유용함을 사용자의 정보 요구

* 전북대학교 전자정보공학부 부교수
(Associate Professor, School of Information and Electronics Engineering, Chonbuk National University, duan@chonbuk.ac.kr)
** 한국과학기술원 전자전산학과 전산학 전공
(Department of Computer Science, KAIST, ihkang@csone.kaist.ac.kr)
· 논문접수일자 : 2003년 6월 18일
· 게재확정일자 : 2003년 6월 23일

형태에 따라서 분류한다. 내용 검색에서는 내용 정보가 유용한 반면 링크 정보와 URL 정보를 결합할 경우 성능의 저하를 초래했다. 반면 사이트 검색에서는 내용 정보만 쓰는 것 보다는 링크 정보와 URL 정보를 결합할 경우 성능의 향상을 얻을 수 있었다.

키 워 드

내용 검색, 사이트 검색, 서비스 검색, 내용 정보, 링크 정보, URL 정보

ABSTRACT

The Web is rich with various sources of information. It contains the contents of documents, multimedia data, shopping materials and so on. Due to the massive and heterogeneous web document collections, users want to find various types of target pages. We can classify user queries as three categories according to users ' intent, content search, the site search, and the service search. In this paper, we present that different strategies are needed to meet the need of a user. Also we show the properties of content information, link information and URL information according to the class of a user query. In the content search, content information showed the good result. However, we lost the performance by combining link information and URL information. In the site search, we could increase the performance by combining link information and URL information.

KEYWORDS

Content Search, Site Search, Service Search, Content Information, Link Information, URL Information

1. 서 론

다양하고 많은 양의 정보가 존재하는 웹 환경은 그 수량 또한 빠르게 증가한다. 이에 따라 방대한 규모의 웹 문서들 사이에서 사용자가 원하는 정보를 효과적으로 찾는 것은 매우 중요한 일이다. 초기 검색 시스템들은 단순히 웹 문서에 검색하는 검색어가 포함된 문서들을 사용자에게 보여주는 방식으로, 재현률(recall)을 높이기 위해 관련 문서를 모두 찾는 것에 중점을 두었다(Salton et al. 1998). 이 방식은 단순히 검색어가 포함된 모든 문서를 검색 결과로 제시하므로 사용자가 필요로 하지 않는 문서들을 많이 포함한다. 최근에는 내용 기반 검색과 웹 문서간의 링크 구조를 이용하여 검색하는 방식이 제안되어 사용되고 있다(Page et al. 1998). 이 방식은 정확률(precision)을 높이기 위해 관련성이 높고 권위있는 페이지를 먼저 제시하는 것에 중점을

두었다. 웹 문서 사이의 링크 구조를 이용하여 문서들에 대한 값을 수치로 표현한 것으로 페이지랭크(PageRank)를 들 수 있다(Brin et al. 1998). 구글(www.google.com) 검색기에서는 페이지랭크를 사용하여 많은 사람들이 추천하는 권위 있는 검증 받은 문서를 상위 결과로 제시하여 좋은 성능을 보이고 있다.

특정 상황에서 좋은 성능을 보이는 검색 기법이라 하더라도 다른 상황에서는 그렇지 못한 경우가 있다(Croft 2000). 예를 들면 내용 정보 기반의 검색이 다양한 단어와 검증된 문서를 제시하지 못한다는 단점이 있는 반면 링크 정보 기반의 검색은 불완전하고 잘못된 링크에 의한 문제가 발생할 수 있다. 이러한 각 정보의 단점을 보완하기 위해 여러 정보를 결합하는 연구가 수행되고 있다(Westerveld et al. 2001; Yang 2001). 그러나 기존의 연구는 정보를 결합하는데 있어서 사용자의 질의어를 고려하지 않은 정적인 결합 형태를 취하고 있다. 예를 들어 질의어 '*Mutual Information*'을 위하여 링크 정보를 너무 많이 사용할 경우, '*mutual funds*'와 '*information*'을 단어로 가지고 있는 유명한 투자회사가 상위 결과로 제시된다. 또한 질의어 '*Britney's Fan Club*'을 위하여 내용 정보에 너무 많이 의존할 경우, 야후(www.yahoo.com)나 라이코스(www.lycos.com)의 웹 디렉토리 문서가 상위 결과로 제시된다. 이와 같이 정보를 결합하는 것이 항상 좋은 결과를 얻는 것은 아니다.

사용자의 질의는 의도에 따라서 세 가지로 구분된다(Broder 2002).

- 내용 검색 (informational)
- 사이트 검색 (navigational)
- 서비스 검색 (transactional)

내용 검색은 사용자가 알고자 하는 정보와 관련이 있는 순서대로 순위화 된다. 예를 들어 "*What is a prime factor?*"혹은 "*prime factor*"와 같은 질의는 사용자가 '*prime factor*'에 대해서 알기를 바란다. 반면 사이트 검색은 사용자가 찾고자 하는 사이트 순서대로 순위화 된다. 예를 들어 "*Where is the site of John Hopkins Medical Institutions?*" 나 "*John Hopkins Medical Institutions*"와 같은 질의는 사용자가 *John Hopkins Medical Institution* 사이트를 방문하기를 바란다. 사이트 검색에서는 비록 그 사이트와 관련된 문서일지라도 사이트의 메인 출입 문서만을 정답으로 채택한다. 그리고 서비스 검색은 사용자가 찾고자 하는 서비스를 제공하는 순서대로 순위화 된다. 예를 들어 "*Where can I buy concert tickets?*"나 "*buy concert tickets*"와 같은 질의는 사용자가 콘서트 티켓을 구매할 수 있는 웹 문서를 찾기를 바란다. 이와 같이 검색 대상과 사용자의 의도가 다양해짐에 따라서 이에 따른 문서의 순서화 차별화가 필요하다. 본 논문에서는 위 세 가지 질의 유형 중 내용 검색과 사이트 검색에 따른 문서 순서화 방법을 살펴본다.

2. 관련연구

웹 문서 검색에서 순위화를 위하여 사용되는 정보로 크게 내용 정보, 링크 정보 그

리고 URL 정보를 들 수 있다. 내용 정보가 특정 단어에 대한 연관도를 나타내는 정보라면 링크 정보와 URL 정보는 문서 자체가 질의어와 관련 없이 가지는 정보이다.

2.1 내용 정보

질의어와 연관되는 정도를 알아보기 위해서, 질의어로 사용된 단어의 분포와 빈도 그리고 다른 문서들에서 해당 단어의 분포와 빈도를 이용한다. 이러한 단어는 발생 위치에 따라서 제목, 본문, 그리고 anchor text로 나눌 수 있다(Croft 2000). Anchor text는 하이퍼링크로 연결된 문서에 대해서 간략한 설명을 가진다. 때때로 이는 연결된 문서에 나타나지 않은 단어와 설명을 가지고 있다. Anchor text를 이용할 경우 하이퍼링크로 연결된 이미지 파일이나 아직 웹 로봇이 가지고 오지 않은 문서에 대해서도 검색이 가능하게 된다.

단어의 관련도는 tf(term frequency)와 df(document frequency)로 표현된다. tf는 웹 문서에서 특정 단어의 출현 빈도를 나타낸다. 이는 특정 단어가 얼마나 문서의 내용을 잘 표현하고 있는지를 나타낸다. df는 특정 단어가 출현한 문서의 수를 나타낸다. 이는 특정 단어가 가지는 변별성 정도를 나타낸다. df가 큰 단어는 관련 문서와 비관련 문서를 구분하는데 별로 유용하지 못한 반면 df가 작은 단어는 구분하는데 유용함을 뜻한다(Salton et al. 1983). tf와 df를 이용하여 문서의 관련도를 계산하는 기법엔 여러 가지가 있다. 주로 TFIDF 방식으로 사용되는데, tf와 df의 역수를 곱하는 형태로 각 단어의 중요도와 유사도를 계산한다. SMART 시스템에서는 tf에 따른 가중치를 계산하기 위해서 다음과 같은 수식을 제공한다(Salton 1989).

$$b: 0 or 1$$
$$n: tf$$
$$a = 0.5 + \frac{tf}{\max tf}$$
$$l: 1.0 + \log tf$$

여기에서 max tf는 그 문서에서 얻을 수 있는 최고의 tf 값을 말한다. 이외에도 더블 로그, OKAPI 등 다양한 가중치 제공 수식이 있다(Robertson et al. 1994).

$$d: 1 + \log(1 + \log(tf))$$
$$o: \frac{tf}{2 + tf}$$

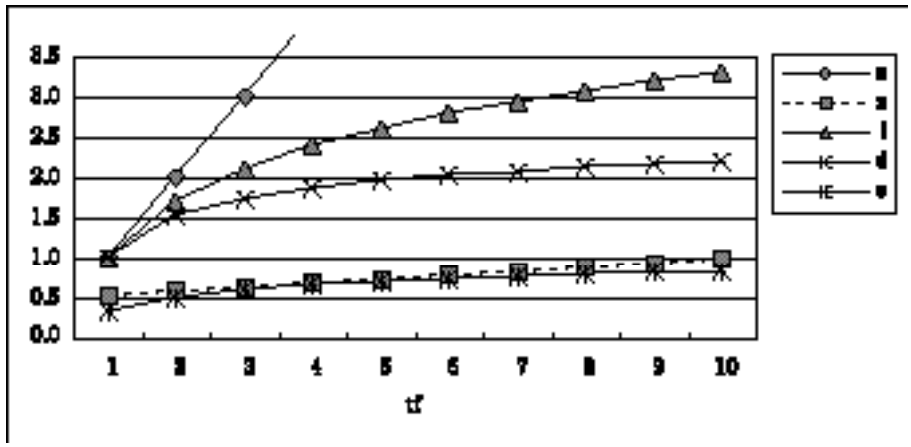
df에 따른 가중치를 계산하기 위해서는 다음과 같은 수식을 제공한다.

$$n: 1.0$$
$$t: \log\frac{N}{df}$$

여기에서 N은 총 문서 수를 나타낸다. tf에 따른 각 가중치 기법의 결과 값을 보이면 <그림 1>과 같다.

2.2 링크 정보

웹 문서들은 서로 링크로 연결되어 있



<그림 1> 가중치 기법에 따른 tf의 영향

다는 점에서 일반적인 문서와는 다른 구조적 특징을 가진다.

2.2.1 허브와 권위자

웹 문서의 구조적 특징인 링크 구조를 통해 어떤 문서가 권위적인(authoritative) 문서인지를 찾거나 유사한 문서들을 모으는 등의 정보를 알아낼 수 있다. 많은 허브(hub)가 가리키는 문서를 권위자(authority)라 하고, 많은 권위자 문서를 링크로 갖는 문서를 허브(hub)라 한다. 다음은 문서 p에 대한 권위자와 허브 값을 구하는 식이다(Kleinberg 1999).

$$authority(p) = \sum_{q,\ q \text{ Points } p} hub(q)$$

$$hub(p) = \sum_{q,\ p \text{ Points } q} authority(q)$$

권위자 값과 허브 값은 검색 결과 상위 100-200개의 웹 문서를 이용하여 계산되고, 질의어에 대한 인기도를 나타낸다. 그러나 이 값은 검색할 때마다 계산해야 하므로 실질적으로 사용하기에는 많은 제약이 따른다.

2.2.2 페이지 랭크

웹 문서들의 상대적인 중요성을 측정하기 위해 제안된 것이 페이지랭크(Page-Rank)이다. 권위자와 허브 값은 검색 시 매번 계산하는 반면 페이지랭크는 색인 작업 시 미리 계산하여 빠른 검색이 가능하다. 따라서 페이지랭크는 상업적인 정보 검색기에 더 많이 사용된다. 페이지랭크는 링크를 통해 연결되어 있는 웹 문서들을 그래프 구조로 생각하고 웹 문서들의 순위를 계산한 것이다. 많은 문서들이 가리키고 있는 문서가 더 중요하다고 생각하고 그것에 대한 값을 수치로 표현한다. 페이지랭크를 계산하는 수식은 다음과 같다(Brin et al. 1998).

$$PR(A) = (1-d) + d \times (PR(T_1)/\ C(T_1) + K + PR(T_n)\ /C(T_n))$$

*PR(A)*는 문서 *A*의 페이지랭크, *PR(Ti)*는 문서 *A*를 가리키는 문서 *Ti*의 페이지랭크, *C(Ti)*는 문서 *Ti*가 가리키는 문서 개수, *d*는 사용자가 특정 문서에서 만족하지 못하고 다른 문서로 이동할 확률을 나타낸다. 전체 웹 문서들의 페이지랭크를 구하기 위해 초기에 시작 값을 설정하고 반복적으로 페이지랭크를 계산하는 방식을 적용한다. 웹 문서의 페이지랭크는 반복 계산 과정을 통해 초기값에서 원래 페이지랭크의 근사치로 수렴한다(Page et al. 1998).

2.3 URL 정보

동일한 사이트에서 중요한 문서는 상위 디렉토리에 위치할 가능성이 크다. 또한 사이트명이나 웹 문서가 가지는 정보와 유사한 형태로 URL이 기술될 가능성이 크다. 이러한 URL 정보를 이용하는 가장 쉬운 방법은 URL의 깊이를 보거나 사용자 질의어와 URL을 일치시켜보는 것이다. URL의 깊이를 보기 위해서는 URL에 나타난 '/' 개수를 이용한다. Westerveld (2001)는 URL의 깊이를 4가지 형태로 나누어 설명했다.

- root: 호스트명만 나타나는 형태
  (예. http://trec.nist.gov)
- subroot: 호스트명에 하나의 디렉토리가 나타나는 형태
  (예. http://trec.nist.gov/pubs)
- path: 호스트명에 임의 길이의 디렉토리가 나타나는 형태
  (예. http://trec.nist.gov/pubs/trec9/papers)
- file: index.html와 같은 파일 이름으로 끝나는 형태
  (예. http://trec.nist.gov/pubs/trec9/t9proceedings.html)

Westerveld(2001)는 각 URL 형태가 사이트 출입 페이지일 확률(URLprior)을 계산하였다. Westerveld(2001)의 결과에 의하면, 확률 P(entrypage|URLtype=t)은 WT10g 테스트 컬렉션을 이용할 경우 <표 1>과 같다.

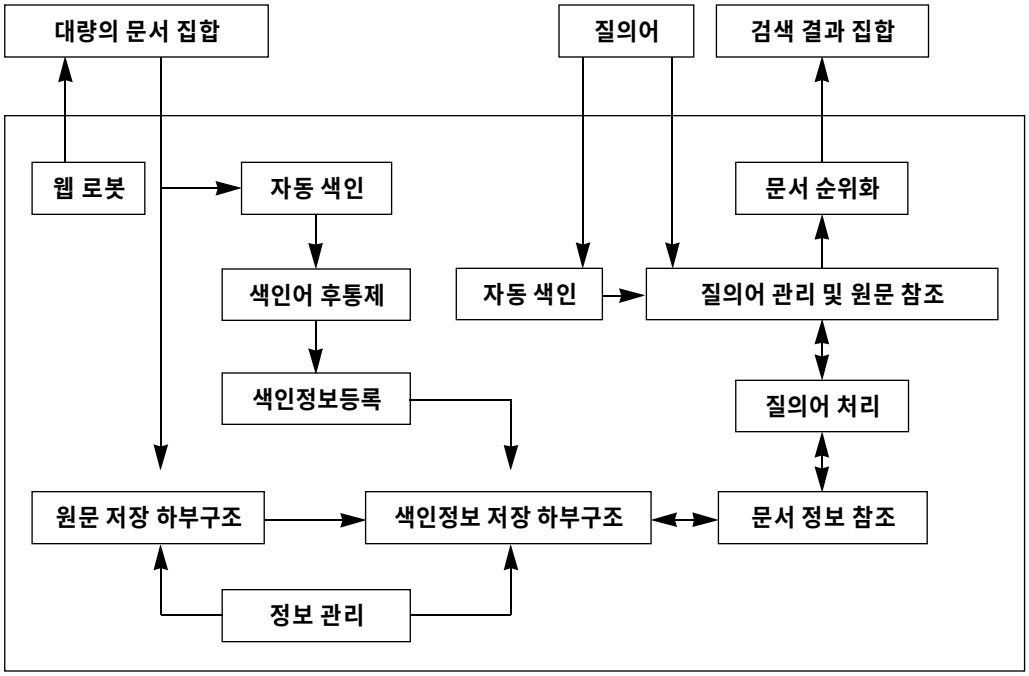
<표 1> WT10g에서 사이트 페이지 비율

URL type	# entry pages	# WT10g
root	38 (71.7%)	11,680 (0.6%)
subroot	7 (13.2%)	37,959 (2.2%)
path	3 (5.7%)	83,734 (4.9%)
file	3 (5.7%)	1,557,719 (92.1%)

본 논문에서는 질의어 유형에 따라서 내용 정보, 링크 정보 그리고 URL 정보의 특성을 보인다. 그리고 보다 나은 성능을 위해서 각 질의어 유형에 따른 정보들의 결합 방식을 제시한다.

3. 시스템 구현

본 논문에서는 검색 시스템의 다양한 문서 순위화 기법을 평가하기 위해 콘도르(Condor)검색 시스템을 사용하였다. 콘도르는 전북대학교, (주)서치라인, 그리고 카네



<그림 2> 콘도르 검색 시스템의 구조

기 멜론 대학교가 컨소시엄 형태로 개발하였다. 시스템의 질의 처리는 Inference Network을 기반으로 설계되어 있으며 문서 클러스터링 기능을 제공한다. 검색 시스템 콘도르의 구조는 <그림 2>와 같다.

각 모듈 별로 간단히 소개하면, 웹 로봇과 저장 시스템은 일반 정보 검색 시스템과 유사하다(Baeza-Yates et al. 1999; Salton et al. 1983). 질의어 처리를 위한 검색 모델은 확률 모델을 따랐다. 또한 다중어(한국어, 중국어, 일본어, 영어) 질의어 처리가 가능하다. 색인어 가중치는 검색 결과에 문서 순위를 정하는 것, 문서 요약, 문서 클러스터링 등에 사용된다. tf와 df는 변형된 OKAPI의 계산법에 따라 사용된다. 아울러 단어의 특성(P_t)을 반영하는데, 문자 폰트나 문자의 크기, 혹은 문자의 타입에 따라 정한다. 예를 들어 주변 글자와 다른 색깔을 가지거나, 글자 크기나 폰트가 변경될 경우, 변경 정도에 따라서 세 단위로 가중치를 할당한다. 가중치 계산법을 보이면 다음의 수식과 같다.

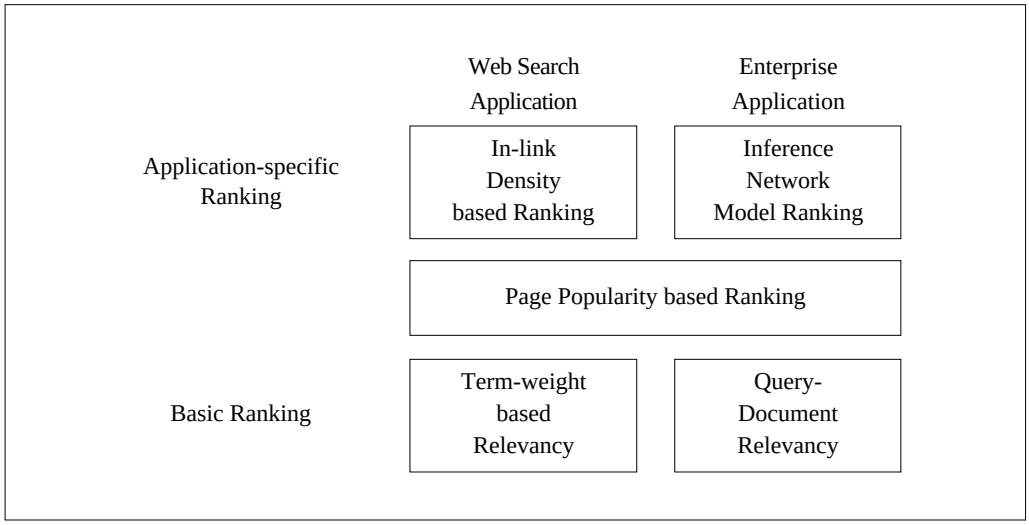
$$w_t = TF \times IDF \times P_t$$

where

$$TF = \frac{tf}{2 \times (0.25 + 0.75 \times \frac{doc\ len}{avg\ doc\ len}) + tf}$$

$$IDF = \log(\frac{N - df + 0.5}{df + 0.5})$$

$$p_t = \begin{cases} 2.0 : high \\ 1.5 : important \\ 1 : others \end{cases}$$



<그림 3> 콘도르 시스템의 문서 순위화 구조

콘도르 시스템에서는 검색의 용도에 따라서 사용하는 정보와 결합 방식을 달리한다(<그림 3>). 크게 웹 문서 검색과 기업 내 문서 및 데이터베이스 검색을 대상으로 한다. 문서의 출처 및 관리자가 정의한 중요도에 따라서 문서 고유의 가중치가 계산된다. 웹 문서의 경우에는 하이퍼텍스트를 이용하여 자동적으로 문서의 중요도를 계산한다. 이렇게 얻어진 문서의 질의 관련도와 문서 자체의 인기도와 중요도를 용도에 따라서 차별적으로 결합하여 결과 문서를 제시한다.

4. 실험 및 평가

본 논문에서는 콘도르 시스템을 기반으로 내용 검색과 사이트 검색을 대상으로 각 정보들의 유용성과 결합 방식에 대해서 살펴본다.

4.1 실험 데이터 및 평가 기준

본 논문에서는 색인을 위해서 WT10g 데이터 컬렉션을 사용한다(Bailey et al. to appear). WT10g는 10G의 웹 문서 집합으로써 CSIRO에서 배포한다(CSIRO 2001). 질의어로는 TREC-2001 Web Track의 ad-hoc task 질의어 세트(501-550)와 homepage finding task 질의어 세트(1-145)를 이용하였다(Hawking et al. 2001). Ad-hoc task 질의어 세트는 내용 검색을 위해 사용하고, homepage finding task 질의어 세트는 사이트 검색을 위해 사용하였다. 본 논문에서는 위 두 개의 질의어 세트를 *TOPIC*과 *HOME*이라고 부른다.

내용 검색의 성능을 측정하기 위해 평균 정확도(average precision)를 사용한다(Hawking et al. 2001). 평균 정확도란 검색 엔진의 결과로 나온 리스트에서 정답 문서가 나타날 때 마다 그 상위 순위까지의 정

확도를 측정하여 평균한 것이다. 정답 문서의 집합을 R이라 하고 문서 d가 검색 엔진에서 얻은 순위를 $r(d)$ 라고 하였을 때, 하나의 질의어에 대한 정확도는 다음과 같이 나타낼 수 있다.

$$P_{avg} = \frac{1}{|R|}\sum_{d \in R} \frac{|R_{\le r(d)}|}{r(d)}$$

사이트 검색의 평가를 위해서는 MRR (Mean Reciprocal Rank)을 이용한다(Hawking et al. 2001). MRR은 검색 엔진이 첫 번째 정답 문서를 몇 위로 제시하였는가를 측정한다. n개의 질의에 대해서 MRR을 측정할 때는 각 질의의 결과에 대해 첫번째 정답 문서의 순위 r의 역수를 구한 후 이 값을 평균한다.

$$MRR = \frac{1}{n}\sum_{i} \frac{1}{r_i}$$

4.2 질의어 유형에 따른 검색 기법의 비교

본 절에서는 기존의 검색 엔진에서 사용된 여러 정보들의 특성에 대해서 알아본다.

먼저 단어의 출현 위치를 크게 두 가지 경우로 나누어 생각한다. 제목이나 anchor text에 나타나는 경우와 문서 본문에 나타나는 경우로 나눈다. Anchor 방식은 문서 본문을 제외한 제목과 해당 문서를 연결하고 있는 다른 문서의 anchor text만을 가지고 색인을 수행한다. 반면 Common 방식은 해당 문서 자체만 가지고 색인을 수행한다.

정보 결합 측면을 고려하기 위해서, tf와 df 만을 이용한 내용 정보 방식과 페이지랭크와 URLprior 정보를 내용 정보에 결합(CMB)하는 두 가지 방식을 고려한다. 내용 정보와 페이지랭크 URLprior 정보를 결합하는 수식은 다음과 같다.

$$rel(d) = 0.65 \times Content\ Information + 0.25 \times URL\ Information + 0.1 \times Link\ Information$$

그리고 연산자에 대해서는 '*and*' 연산자와 '*sum*' 연산자를 고려한다. '*and*' 연산자는 질의에 사용된 모든 단어를 가지고 있는 문서만이 정답으로 채택된다. 반면 '*sum*' 연산자는 질의에 사용된 단어 중 하나라도 가지고 있는 문서가 정답으로 채택된다.

<표 2>는 위의 각 경우에 대한 성능을 나타낸다. 예를 들어 Anchor and CMB 방식은 제목과 anchor text 만을 이용해서 색인을 수행하였으며, '*and*' 연산자를 이용

<표 2> 질의어 유형에 따른 각 기법의 성능

	TOPIC	HOME
model	P_{avg}	MRR
Anchor and	0.031	0.297
Anchor and CMB	0.031	0.431
Anchor sum	0.034	0.351
Anchor sum CMB	0.034	0.583
Common and	0.131	0.294
Common and CMB	0.122	0.580
Common sum	0.182	0.355
Common sum CMB	0.169	0.673
MAX	0.226	0.774
AVG	0.145	0.432

하여 질의어와 부합 정도를 비교하며 내용 정보 외에 페이지 랭크와 *URLprior*정보를 결합하여 사용한 경우를 나타낸다. <표 2>에서 MAX와 AVG는 TREC-2001에 제출된 시스템들 중 최고 성능과 평균 성능을 뜻한다.

<표 2>를 통해 내용 검색에서는 Common 방식이 Anchor 방식에 비해 월등히 나은 성능을 보이고 있는 반면, 사이트 검색에서는 Common 방식과 Anchor 방식이 큰 차이를 보이지 않고 있음을 알 수 있다. 제목과 anchor text로는 내용 검색에 필요한 충분한 정보를 제공하지 못한다는 것을 알 수 있다. 그리고 사이트 검색에서는 페이지랭크와 URLprior를 결합하여 보다 나은 성능을 얻을 수 있는 반면, 내용 기반 검색에서는 오히려 성능이 감소함을 알 수 있다.

내용 검색에서는 사용자가 알고자 하는 개념의 중요 단어나 설명하는 단어가 질의로 사용된다. 그러나 사용자가 제시하는 설명 형태와 동일한 형태로 문서에 나타난다고 보장할 수 없다. 따라서 내용 기반 검색에 있어서 *'and'* 연산자로는 좋은 성능을 얻을 수 없다. 사이트 검색에 있어서도 *'and'* 연산자가 *'sum'* 연산자보다 성능이 다소 떨어짐을 알 수 있다. 이는 *'and'* 연산자를 이용하여 결과를 제시하지 못하는 경우가 있기 때문이다. 그러나 결과를 얻을 수 있는 경우에는 좋은 성능을 보이고 있다. 이렇게 결과가 없는 경우를 보완하기 위해 사이트 검색에서 최고 성능을 보이는 Common sum CMB와 Anchor and CMB를 결합하여 정렬할 경우 0.730의 성능을 얻을 수 있었다. 이는 Common sum CMB 보다도 나은 성능이다. 마찬가지로 내용 기반 검색에서 최고 성능인 Common sum 과 Anchor and를 결합했을 경우에는 0.173으로 성능이 떨어졌다. 이를 통해서 사이트 검색에 있어서는 *'and'* 연산자가 유용함을 알 수 있다.

위의 결과를 통해서 사용자 질의어 유형에 따라서 문서 순위화 기법이 달라져야 함을 알 수 있다. 내용 기반 검색을 위해서는 문서 전체를 사용하는 방법과 *'sum'* 연산자를 사용하는 것이 유용함을 알 수 있다. 그리고 사이트 검색을 위해서는 제목과 anchor text를 강조하는 방법과 *'and'* 연산자로 나온 결과를 우선시하며 페이지 랭크나 URLprior와 같은 링크 정보와 URL 정보가 유용함을 알 수 있다.

다음은 tf와 df를 이용한 가중치 기법에 대한 비교이다. 그 문서의 최고 tf로 보정해주는 *a* 방식과 df의 역수를 이용하는 전형적인 TFIDF와 콘도르에서 사용하는 변형된 OKAPI 기법을 비교한 결과는 <표 3>과 같다. <표 3>에서 Content Info.는 내용 정보만을 이용했을 때의 결과를 나타내며, CMB는 내용 정보에 페이지랭크와 URLprior를 결합하여 사용한 결과를 뜻한다. 링크 정보와 URL 정보를 결합할 경우 사이트 검색은 성능이 향상되어지나 내용 검색은 성능 저하를 초래한다. 그러나 OKAPI 방식은 그 저하된 성능이 TFIDF에 비해서 높으며, 사이트 검색에서 향상폭은 TFIDF 보다 큼을 알 수 있다. 이를 통해 OKAPI 가중치 기법을 이용할 경우 내용 검색과 사이트 검색에서 TFIDF 방식

<표 3> 가중치 기법에 따른 시스템의 성능

	TFIDF		OKAPI	
model	TOPIC	HOME	TOPIC	HOME
Content Info.	0.170	0.340	0.182	0.355
CMB	0.159	0.640	0.169	0.673

보다 나은 결과를 얻을 수 있음을 알 수 있다.

5. 결 론

인터넷의 발달로 인해 웹에서 얻을 수 있는 정보의 종류와 수는 급진적으로 증가하고 있다. 아울러 사용자의 요구 또한 다양해 지고 있다. 이러한 사용자의 요구는 내용 검색, 사이트 검색, 그리고 서비스 검색으로 나눌 수 있다. 내용 검색은 원하는 정보를 설명하는 혹은 정보와 관련된 문서를 찾는 것이며, 사이트 검색은 사용자가 관심 있어 하는 개인이나 단체의 사이트 입구를 찾는 것이고, 서비스 검색은 사용자가 관심 있어 하는 서비스를 제공하는 웹 페이지를 찾는 것을 말한다.

본 논문에서는 콘도르 시스템을 이용하여 사용자의 질의 유형에 적절한 문서 순위화 알고리즘을 제시했다. 기존의 연구들은 문서 순위화를 위해서 tf와 df로 대표되는 내용 정보를 보완하기 위해서 페이지 랭크와 같은 링크 정보와 URL의 깊이와 같은 URL 정보를 결합하였다. 그러나 단순한 정적인 결합은 사용자의 질의 유형에 따라서 성능이 오히려 저하되는 경우도 발생한다. 이를 보완하기 위해서는 질의 유형에 따른 차별적인 문서 순위화 방법이 요구된다. 내용 검색에서는 내용 정보와 *'sum'* 연산자를 사용하는 것이 좋으나 링크 정보와 URL 정보를 결합할 경우 오히려 성능이 저하된다. 사이트 검색에서는 문서 전체 보다는 제목과 anchor text를 이용한 정보가 유용하며 *'and'* 연산자를 이용한 결과가 이용 가능한 경우 좋은 성능을 보인다. 아울러 링크 정보나 URL 정보를 이용할 경우 월등한 성능 향상을 유도할 수 있었다. 그리고 내용 정보에 가중치를 주는 함수로 TFIDF 방식보다는 OKAPI 방식이 좋은 성능을 보였다.

앞으로의 연구로 서비스 검색에 대한 연구가 수행되어야 하며, 사용자 질의 유형을 자동으로 구분하여 문서 순위화 알고리즘을 차별하는 방법의 개발이 필요하다.

참고 문헌

Baeza-Yates, R., & Ribeiro-Neto, B. 1999. *Modern Information Retrieval.* ACM Press.

Bailey, P., Craswell, N., & Hawking, D. (to appear). "Engineering a multipurpose test collection for web retrieval experiments." *Information Processing and Management*

Brin, S., & Page, L. 1998. "The Anatomy of a Large-Scale Hypertextual Web Search Engine." *Computer Networks and ISDN Systems,* 30(1-7): 107-117.

Broder, A. 2002. "A Taxonomy of Web Search". *SIGIR Forum,* 36(2).

Croft, W. B. 2000. *Combining Approaches to Information Retrieval. In W. B. Croft (Ed.), Advances in Information Retrieval : Recent Research from the Center for Intelligent Information Retrieval*, Kluwer Academic Publishers : 1-36.

CSIRO. (2001). "Web Research Collections - Trec Web Track." <http://www.ted.cmis.csiro.au/TRECWeb/>.

Hawking, D., & Craswell, N. 2001. "Overview of the Trec-2001 Web Track". In *Text REtrieval Conference* (trec-10) : 61-67.

Kleinberg, J. M. 1999. "Authoritative Sources in a Hyperlinked Environment". *Journal of the ACM*, 46(5) : 604-632.

Page, L., Brin, S., Motwani, R. & Winograd, T. 1998. *The Pagerank citation ranking : Brining Order to the Web* (Tech. Rep.). Stanford Digital Library Technologies Project.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., & Gatford, M. 1994. "Okapi at trec-3". In *Text REtrieval Conference* (trec-2) : 109-126.

Salton, G. 1989. *Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer* Reading, MA : Addison-Wesley.

Salton, G. & McGill, M. 1983. *Introduction to Modern Information Retrieval*, New York : McGraw-Hill Press.

Westerveld, T., Kraaij, W., & Hiemstra, D. 2001. "Retrieving Web Pages using Content, Links, Urls and Anchors". In *Text REtrieval Conference* (trec-10) : 663-672.

Yang, K. 2001. "Combining Text and Link-Based Retrieval Methods for Web IR". In *Text REtrieval Conference* (trec-10) : 609-618.