메타검색에서 링크정보와 요약정보를 이용한 검색결과 통합

양명석[0*] , 이석형*, 강남규*, 윤화묵*
한국과학기술정보연구원 정보시스템연구실*
(msyang[0], skyi, ngkang, hmyoon)@kisti.re.kr

A Ranking method using link & description information in Meta searching

Myung-Seok Yang[0*] , Seok-Hyung Lee*, Nam-Kyu Kang*, Hwa-Mook Yoon*
*Dept. of Information System Research ,Korea Institute of Science and Technology Information,

요 약

본 논문은 메타검색엔진 시스템에서 다양한 검색결과를 하나의 검색결과로 통합시켜주는 통합랭킹 방법에 대한 연구이다. 검색결과 통합시 메타검색 시스템에서 실질적으로 이용할 수 있는 링크정보와 요약정보를 사용하였다. 통해 이루어져 있다. 또한 링크정보와 요약정보에 대한 가중치 값을 변화시키면서 다양한 검색결과들을 얻을 수 있었는데, 요약정보의 가중치를 높여 주었을 때 검색 효율이 좋음을 알 수 있었다.

1. 서 론

웹의 발전과 더불어 많은 검색 엔진들이 등장하게 되었다. 웹을 이용하는 사람들은 여러 검색엔진을 돌아다니면서 많은 검색 결과에 허덕이고 여러 검색 엔진들을 찾아 돌아다녀야 하는 번거로움 때문에 불편함을 상당히 느낀다. 이를 해결하기 위한 대안으로 메타 검색 엔진[1]들이 등장하고 있다.

메타 검색엔진은 하나의 검색엔진에서 여러 검색엔진들을 한꺼번에 이용할 수 있다는 점에서 상당히 매력적인 시스템이다. 사용자로부터 질의어, 검색 방법, 검색될 문서개수등을 받아서 메타 검색엔진이 연결 가능한 모든 검색엔진에 맞는 질의로 변환한 후, 검색 결과를 검색엔진으로부터 받아서 결과를 사용자에게 보여준다. 하지만 여러개의 검색엔진들로부터 검색결과를 받아서 이를 통합해 보여주는 것은 쉽지않은 일이다. 또한 각각의 질의를 각 검색엔진의 특성에 맞춰서 질의를 생성하는 것 역시 쉽지 않다. 하지만 질의를 생성하는 것은 어느 정도 검색엔진에 대한 지식을 갖고 있다면 결코 어려운 일은 아니다.

본 논문에서는 검색결과 통합에 초점을 맞추어 기존의 메타 검색엔진이 가지고 있는 통합 방법에 대해 살펴보고, 각 검색엔진의 제공정보들 중에서 실질적으로 이용가능한 링크정보와 요약정보를 사용한 결과 통합 방법을 제안한다. 또한 이러한 방법을 적용한 메타 검색엔진의 프로토타입을 구현하였으며 각각의 요인들이 검색결과에 미치는 영향에 대하여 실험을 하였다. 2절에서는 관련연구로 Lawrence와 Drelinger의 통합 방법에 대해 살펴보고, 3절에서는 본 논문에서 제안하는 결과통합방법에 대해 설명하고 4절에서는 검색결과 통합방법의 실험에 대한 것을 제시하고, 5절에서 결론을 맺는다.

2. 관련연구

2.1 Lawrence[2]

Lawrence의 결과 통합 방법은 Inquirus에서 사용하는 방법으로, 결과내의 문서를 직접 수집하고 문서내에 질의어 개수와 질의어 사이의 거리 정보등을 이용하여 직접 문서를 색인하는 것과 비슷한 효과를 얻는 방법이다. 문서를 직접 색인한다는 점에서 일반 검색엔진과 별반 차이가 없으나 문서를 전부다 수집하여 색인을 하기 때문에 상대적으로 속도가 느린 단점이 있다.

2.2 Drelinger[3]

Drelinger의 방법은 검색엔진을 문서로 보고, 일반 검색엔진의 랭킹방법처럼 질의어와 각각의 검색된 결과수등을 이용해, tf*idf과 같이 계산하고 총 검색엔진의 수로 정규화(normalize)한다. 그 결과에 최근에 사용자의 이용정보로, 최근의 클릭 정보와 검색엔진의 응답시간등을 랭킹에 반영했다. Lawrence의 방법에 비해서 속도도 빠르고, 랭킹 효율도 좋은 방법중 하나이다.

이외에도 확률모델을 이용한 결과 통합 방법[4], 클릭정보와 링크빈도를 이용한 방법[5], 또 순위 유사도[12]를 이용한 컬렉션 융합 방법등이 있다.

3. 검색결과 통합방법

메타검색에서의 검색결과 통합이란 메타검색엔진이 여러 개의 다른 일반검색엔진으로부터 얻은 검색결과를

하나로 통합하는 것을 말한다. 앞서 살펴본 lawrence 나 Drelinger의 방법이 그 대표적인 예이다.

본 논문에서 제안하는 검색결과 통합방법은 메타검색엔진의 특성과 장점을 살리면서 추가적인 장치나 도구의 사용 없이 검색결과의 특징을 분석하여 결과를 하나로 통합하는 방법이다. 검색결과를 통합하기 위하여 사용한 정보는 검색 결과 내의 링크 중복 정도와 요약정보내의 단어 빈도수이다.

링크 중복 정도는 중복된 링크의 개수를 의미하는데, 같은 검색엔진내에서의 중복 정도는 내부중복이라 하며, 다른 검색엔진과의 중복 정도는 외부중복으로 구분한다. 요약정보내의 단어 빈도수는 질의 단어가 문서의 제목과 문서 요약(description)부분에 나타난 단어 빈도수(query term frequency)를 의미한다.

그림 1은 링크정보와 요약정보를 이용한 알고리즘이다.

$$RSV_i = w_{in} * \frac{in\_dup}{Max\_in\_dup} + w_{ex} * \left(\frac{\sum_{k=1}^{N} Val_k * \frac{1}{rank_{i,k}} * exist_{i,k}}{N} \right) + w_t * \frac{wordct}{maxwt} + w_d * \frac{wordcd}{maxwd}$$

$$exist_{i,k} = \begin{cases} 0 \\ 1 \end{cases},$$

(단, $w_{in} + w_{ex} + w_t + w_d = 1$)

- i : 문서 번호 , RSV_i : 검색 상태 값
- in_dup : 검색엔진 내부에서의 중복된 문서 수
- N : 총 대상 검색엔진의 수 ,k : 검색엔진 번호
- Val_k: k번째 검색엔진의 엔진의 가치 (현재 검색된 문서 수)
- $rank_{i,k}$: 문서 i 의 k번째 검색엔진에서 랭킹
- $exist_{i,k}$: 검색엔진 k의 결과에 문서 i가 존재하는지 여부, (1 : 존재, 0 :부존재)
- wordct : 문서 i의 제목에 나타난 질의 단어의 개수
- wordcd : 문서 i의 요약정보에 나타난 질의 단어의 개수
- maxwt : 검색결과들 중에서 문서 제목에 나타난 질의 텀의 개수중 최대값
- maxwd : 검색결과들 중에서 문서 요약정보에 나타난 질의 단어의 개수중 최대값
- Max_in_dup :검색결과들 중에서 검색엔진 내부에서의 중복된 문서 수중 최대값
- w_{in}: 검색엔진 내부에서의 중복성에 대한 가중치
- w_{ex} : 외부 다른 검색엔진과의 중복성에 대한 가중치
- w_t : i번째 문서 제목에 나타난 질의 단어수에 대한 가중치
- w_d : i번째 문서 요약정보에 나타난 질의 단어수에 대한 가중치

그림 1 메타 검색엔진의 통합 랭킹 알고리즘

링크가 중복되거나 질의 단어 정보를 담고 있는 문서는 어느 정도 적합하다는 가정하에 각각의 인자 값을 이용해서 그림 1과 같은 통합 방법을 도출하였다.

검색 결과가 링크정보와 요약정보를 포함하는 정도에 따라서 통합 랭킹은 달라진다. 이 두가지 정보들을 포함하고 있지 않은 문서에 대해서는 웹 검색엔진의 성능에 따라서 이를 반영하기 위해 웹 검색엔진의 가치를 측정할 수 있는 변수를 두고(식에서는 Val_k)이 값과 앞서 관련연구에서 살펴보았던 순위 유사도와 비슷한 방법으로 검색결과가 속한 원 검색엔진의 랭킹값을 이용할 수 있도록 했다.(식에서 $rank_{i,k}$). 내부중복성의 경우 검색 결과들 중에서 내부 중복성의 최대 값으로 정규화를 하고 내부 중복성에 대한 가중치를 두었다. 외부 중복성의 경우, 외부 중복성이 발생한 문서들이 속한 검색엔진 가치를 문서의 랭킹수로 정규화 하고 이 값들을 합한 후 대상 검색 엔진수로 정규화 시킨다. 또 문서 제목에서의 질의 단어수를 최대의 질의 단어수로 정규화하고, 이값을 제어하기 위한 가중치 값을 곱한다 문서 요약에서도 문서 제목에서와 마찬가지로 계산하고 이 모든 값들을 합해서 검색 결과의 값을 계산한다.

각각의 가중치값을 조정함으로써 위 랭킹 방법에서 각각의 요인들의 관계를 조정해 줄 수 있도록 하였다. RSV_i값을 총 검색 결과 수만큼 계산하고 결과값을 내림차순으로 정렬하면 결과 통합 랭킹을 구할수 있고. 이를 인터페이스를 이용하여 통합된 랭킹 결과를 사용자에게 제공한다..

4. 실 험

4.1 링크정보에 관한 실험

실험 방법은 일반적인 웹 질의 20개에 대해 5개의 검색엔진들의 결과 중 상위 10건만을 뽑아서 이를 통합해 그 결과를 저장하여 이 검색결과들에 대해 적합성 판정을 실시했다.

중복성이 문서의 적합성과 연관관계가 있는지 판단하기 위해서 검색결과들 중에서 내부 중복성과 외부 중복성이 발생한 문서들을 대상으로 적합성 판정값을 분석하였다. 그 결과 외부 중복성의 경우에는 총 검색결과들 중에서 외부 중복성이 발생한 12개의 문서 모두가 적합한 판정 결과가 나왔으며, 내부 중복성이 발생한 13개의 문서 중에서 데드링크 3개와 부적합 1개를 제외하고 모두 적합하다는 결과가 나왔다. 이는 외부 중복성의 경우에 적합하다고 판단 할 수 있으며 내부 중복성 역시 어느 정도 적합하다고 볼 수 있음을 의미한다. 즉, 중복성이 발생한 문서들은 적합한 값을 갖으므로 이를 상위에 랭킹함으로써 검색정확도를 높힐 수 있다.

4.2 통합 랭킹방법의 정확도에 관한 실험

통합 랭킹 방법의 효과를 분석해보기 위하여 두가지 실험을 했다. 하나는 통합랭킹을 적용한 검색엔진(MSearch 이하 MSearch)과 다른 메타검색엔진에 똑 같은 질의를 넣어 얻은 검색결과를 비교해 정확도를 비교하는 방법이고, 다른 하나는 MSearch 내에서 통합 랭킹 알고리즘에 이용되는 가중치 값을 변경할 때의 정확도의 변화를 살펴보았다.

실험방법은 앞 실험에서와 마찬가지방식으로 질의 30개를 MSearch와 같은 방법으로 메타 검색을 수행하는 메타 검색엔진(편의상 D로 표기)에 넣어 검색 결과를 비교해 보고, 또한 MSearch의 결과 통합 랭킹 알고리즘에서 각각의 가중치(weight)값을 변화시키며 검색결과를 비교

해 보았다. 실험의 평가자는 앞 실험에 참여했던 2명에게 적합성 판정을 하게 하고 2명의 평균값을 이용해 평균 정확도(Average precision)와 상위 5, 10위에서의 평균 정확도를 비교해 보았다.

- MSearch A : 링크정보와 요약정보를 적절히배열
- MSearch B : 요약정보만 적용
- MSearch C : 링크정보만 적용
- D검색엔진 : MSearch와 같은 형태의 메타검색엔진

얼마나 많은 적합성 문서들이 상위에 랭킹되었는지 알아보기 위해서 상위 5위에서의 평균 정확율과 상위 10위에서의 평균 정확율을 비교해 보았다.

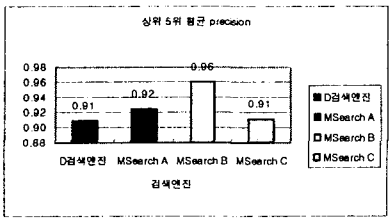

그림 2 상위 5위에서의 평균 정확률

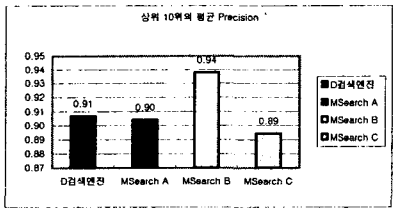

그림 3 상위 10위에서의 평균 정확률

그림 2 와 3 에서 보는 바와 같이 본 논문에서 제안하고 있는 통합 랭킹방법을 적용한 MSearch의 검색결과들이 상위결과에서 좋은 평균정확도를 보이고 있음을 알 수 있다. 특히 요약정보(질의어를 포함하는 문서)의 경우에 좀 더 높은 평균 정확도를 나타내고 있음을 알 수 있다. 또한 상위 5위에서는 D 검색엔진에 비해 차이는 크지 않지만, 약간 높은 평균 정확률을 보이고 있음을 알 수 있다. 이는 링크정보와 요약정보를 이용한 검색 결과 통합 방법이 적합한 문서들을 상위에 어느정도 적절하게 랭킹 시키고 있다는 것을 의미한다.

본 실험에서 살펴본 것처럼, 메타 검색엔진의 검색 결과값을 이용한 질의 단어 정보 추출 방법이나, 검색 결과들 사이에서 발생하는 URL 중복성 값 을 적절한 가중치를 두고 검색 결과를 통합한다면 높은 검색 정확도를 보인다는 것을 알 수 있다. 즉, 링크정보와 요약정보가 검색 결과를 통합하는데 있어서 좋은 방법이 될 수 있다는 것을 알 수 있다.

5. 결 론

검색결과 통합방법은 메타검색 시스템에서 실질적으로 이용할 수 있는 링크정보와 요약정보등을 이용하였다. 링크정보와 요약정보에 대한 가중치값을 변화시키면서 다양한 검색결과들을 얻을 수 있었으며 요약정보의 가중치를 많이 주었을 때 검색 효율이 좋음을 알 수 있었다.

향후 연구로 각 요인에 작용하는 가중치 값을 보다더 확장하여 변화값을 알아보고, 검색엔진 별로 가중치값을 적용하는 방법에 대해 더 연구를 해야할 필요가 있겠다.

6.참고문헌

[1] E. Selberg & O. Etzioni, "The MetaCrawler architecture for resource aggregation on the Web", IEEE Expert, 1997.

[2] Steve Lawrence & C. Lee Giles, "Context and Page Analysis for Improved Web Search", IEEE Internet Computing, Volume 2, Number 4, pp. 38-46, 1998.

[3] Daniel Dreilinger & Adele E. Howe, "Experiences with Selecting Search Engines Using Metasearch", ACM Transactions on Information Systems, Vol. 15, No. 3, pp. 195-222, July 1997

[4] Aslam & Montague, "Bayes Optimal Metasearch: A Probabilistic Model for Combining the Results of Multiple Retrieval Systems", SIGIR2000.

[5] 유태명, 김준태, "링크빈도와 클릭 빈도를 이용하는 메타 검색엔진의 설계", 제25회 정보과학회 춘계 학술대회. 2000

[6] 미스 다찾니, http://www.mochanni.com

[7] 와카노 , http://www.wakano.co.kr/

[8] 네이버, http://www.naver.com

[9] 한글 알타비스타, http://www.altavista.com

[10] 엠파스, http://www.empas.com

[11] Yuwono& Lee, "Search and Ranking Algorithms for Locating Resources on the World Wide Web".

[12] 이준호, 조현양, 최선희, " 정보검색에서의 다중 증거 결합에 대한 분석" , 정보과학회 논문지(B), 제 26권 제 5 호 , pp. 639-646. 1995