관련성 분포 정보를 이용한 정보원 선택 알고리즘

김현주, 김영자, 배종민
경상대학교 컴퓨터과학과 / 정보통신연구센터

An Algorithm for Collection Selection Using Relevance Distribution

Hyun-Ju Kim, Young-Ja Kim, Jong-Min Bae
Dept. of Computer Science/Information and Communication Research Center
Gyeongsang National University

요 약

본 논문은 통합 검색에서 이질의 정보원으로부터 정보를 검색할 때 주어진 질의에 대해 가장 적합한 정보원 선택에 대한 새로운 알고리즘을 제안한다. 제안된 알고리즘은 질의어와 검색에 참여한 정보원간의 관련성 분포 정보를 사용하였다. 이때 관련성 분포 정보는 질의어와 정보원 사이의 관련성 정도를 말하며, 이에 대한 평가는 질의에 대해 정보원으로부터 임의의 크기 N 만큼 검색 문서를 수집한 후에 이들을 평가하여 추정하였다.

본 논문에서 제안한 관련성 분포 정보는 검색 문서의 재평가 값, 관련 문서의 순서 정보, 정확도 등으로 평가한다. 또한 제안된 알고리즘은 정보원 평가에서 검색 인덱스 정보가 필요 없으며, tf, df, N등의 메타 데이터로만 평가할 수 있는 장점이 있어, 동적인 환경에 적용하기가 매우 쉽다.

1. 서론

최근 인터넷은 컴퓨터 네트워크를 기반으로 정보를 교환하고, 필요한 정보를 검색할 수 있는 대표적인 예이다. 이러한 인터넷은 매우 다양하게 정보원이 생겨났으며 지금도 생성되고 있다. 또한 인터넷상에 존재하는 수많은 정보원 가운데서 자신이 원하는 정보가 어디에 있는지 찾는 것은 힘들뿐만 아니라 또한 찾았다 하더라도 정보원의 검색 엔진을 효과적으로 사용하기도 어렵다[7].

이들 정보원이 가지고 있는 검색 엔진들을 사용자가 쉽고 편리하게 이용할 수 있도록 하는 정보검색 분야의 노력 중 하나가 통합 검색 혹은 메타 검색의 등장이다[4, 5, 6, 7].

이러한 통합 검색 분야에서 질의에 대해 검색 결과에 영향을 미치는 주요 요인으로써는 세 가지 연구 분야로 분류할 수 있다. 첫 번째는 가장 좋은 정보원을 선택 문제이다. 두 번째는 질의어 자동 번역 문제이다. 마지막으로는 검색 문서의 통합 및 순위 매김하는 문제이다. 이러한 통합 검색 시스템의 세 가지 연구 분야는 통합 검색 시스템의 검색 결과에 많은 영향을 미치며, 또한 검색을 수행할 때 상호 연동되어 동작한다. 만약 검색에 참여하고 있는 정보원의 상세 정보가 많으면 많을수록 보다 양질의 검색 결과를 사용자에게 제공할 수 있다[5, 7].

그러나 통합 검색 시스템에서 검색에 참여시키고 있는 서로 다른 이질의 정보원에서 질의에 적합한 검색 문서를 추출하고, 이들을 통합하여 단일 우선 순위를 가질 수 있도록 각 문서에 대한 순위 매김을 하기란 매우 어렵다.

본 논문에서는 통합 검색 시스템에서 검색의 효율성에 영향을 미치는 분야 중 정보원 선택에 대한 새로운 모델을 제안한다. 그리고 제안된 모델을 실험을 위해 구현한 HoleInOne (=*w*HOLE IN*formation* ONE*time*) 통합 검색기의 실험 결과를 분석해 본다.

2. 관련 연구

이 장에서는 기존의 4가지 정보원 선택 모델에 대해 살펴본다. 첫 번째로, Voorhees[2]외 2명이 제안한 정보원 선택 모델은 주어진 질의어와 검색에 참여한 정보원과의 관련성 정도를 유사도를 이용하여 정보원을 평가하고 선택하는 모델이다. 이때 정보원에 대한 유사도를 추정하는 방법으로는 문서의 관련성 분포정보와 질의어 클러스트링 정보를 이용한다. 두 번째로는 Callan외 3명이 제안한 모델로 INQUERY[1, 5] 메타 검색 시스템으로 실험하였다. 이는 CORI net 검색 모델이라고도 하며, 문서와 컬렉션과 질의어와의 관련성을 df와 icf를 기반으로 평가한다. 세 번째로는 ProFusion [3] 메타 검색 시스템으로 사용자의 질의에 대하여 9개의 정보원 중에서 (1) 최상의 3개 검색 엔진, (2) 가장 빠른 검색 결과를 보여주는 3개의 검색 엔진, (3) 9개의 검색 엔진 모두다, (4) 사용자가 검색 엔진을 선택 등의 방법을 제공한다. 마지막으로 *GLOSS*[6] 모델로 Luis Gravano외 2명이 제안한 모델로 데이터 베이스에서의 단어 빈도 수 정보를 기반으로 질의에 대하여 데이터 베이스 정보원을 평가한다. 이때 단어

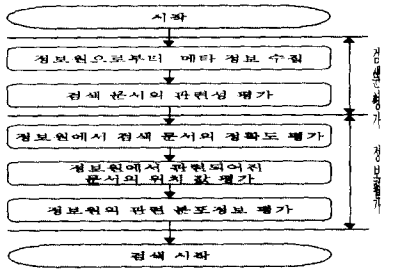
빈도 수 정보는 데이터베이스 정보원 내에서 단어가 포함된 문서의 수를 말한다. 이를 Ind_k 평가자라고 한다.

3. 제안된 메타 데이터 기반 정보원 선택 모델

이 장에서는 통합 검색 시스템에서 주어진 질의에 대해 가장 적합한 정보원을 선택하는 방법에 대해 설명한다.

3.1 정보원 선택에 대한 개괄적인 처리 과정

본 논문에서 제안한 정보원에 대한 선택 처리 과정은 <그림 1>과 같다. 이는 검색 문서의 평가와 정보원에 대한 평가 등의 두 단계로 구성되어 있다. 첫 번째는 검색 문서 평가 과정이며, 두 번째는 정보원 평가이다. 다음의 <그림 1>는 정보원 선택 처리 과정의 개괄 구조도이다.



<그림 1> 정보원 선택에 대한 개괄적인 처리 과정

3.2 정보원의 관련성 분포 정보 평가

이 절에서는 검색 문서의 평가 결과를 사용하여 정보원에 대한 관련성 분포 정보를 평가한다. 다음은 이에 대한 개괄 처리 과정이다.

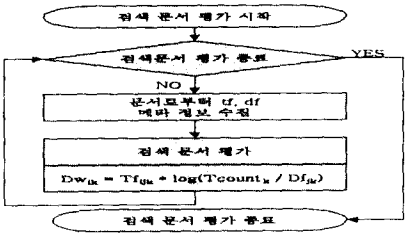
정보원에 대한 관련성 분포 정보 평가는 재평가 문서 값, 관련 문서의 위치 정보, 검색 문서의 정확도 등의 세 가지 요소의 합으로 평가한다. 이에 대한 처리 과정은 <그림 3>과 같다.

3.2.1 메타 데이터

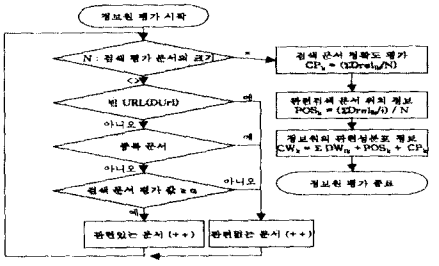
검색 문서의 평가는 다음의 메타 데이터를 기반으로 평가된다. 이때 사용되는 메타 데이터는 다음과 같다.

- Term Frequency(tf) : 이는 검색된 문서에서 질의어가 발생한 빈도 수이다.
- Document Frequency(df) : 이는 정보원 내에서 질의가 포함된 문서의 수이다.
- Total Count(N) : 이는 정보원 내에서 질의어와 관련된 문서의 전체 수이다.

위의 3가지 메타 데이터를 기반으로 검색 문서를 재평가한다. 이때 처리되는 과정은 다음의 <그림 2>와 같다.



<그림 2> 검색 문서의 평가 처리 과정



<그림 3> 정보원의 관련성 평가 처리 모델

3.2.2 관련성 분포 정보 평가 요소

다음은 관련성 분포 정보 평가 요소에 대한 설명이다.

① 검색 문서의 재평가 값

검색 문서의 재평가는 <그림 2>와 같다.

② 관련 문서의 위치 정보

관련 문서의 위치 정보는 정보원을 평가할 때, 질의어와 관련된 문서의 위치를 보상해 주기 위해 사용한다. 다음은 관련 문서 위치 정보 평가 식이다.

(수식 1)

$$POS_k = \frac{\sum_{i=1}^{N} \frac{Drel_{ik}}{i}}{N}$$

③ 문서의 정확도

정확도란 평가 모집단에서 관련 문서의 수에 따라 평가되는 값이다. 다음은 검색 문서 모집단에서 관련 문서의 정확도에 대한 수식이다.

(수식 2)

$$CP_k = \frac{\sum_{i=1}^{N} Drel_{ik}}{N}$$

4. 비교 분석

이 절에서는 기존의 정보원 선택 모델과 본 논문에서 제안한 정보원 선택 모델에서 사용된 메타 데이터와 이들의 특징들을 비교 분석해본다. 아래의 <표 1>은 정보원 선택에서 사용된 메타 정보들을 서로 비교해 보고, 그 특징들에 대해 간략히 기술하였다.

현재까지 알려진 대부분의 기존 연구들은 주어진 질의에 가장 적합한 정보원을 선택하기 위해 자신의 검색 정보를 생성하였으며, 이들을 평가 기준으로 질의가 발생할 때 사용하였다.

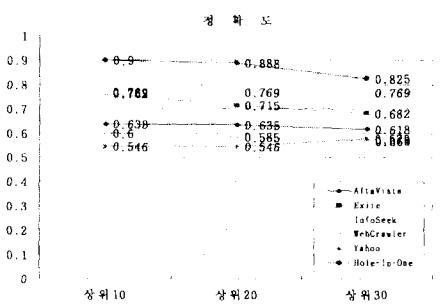
<표 1> 정보원 선택에서의 메타데이타

구 분	메타 데이타
[2]	1) df, 2) 질의어 학습 3) 검색 인덱스 정보생성
	1) df, 2)질의어 학습 3) 검색 인덱스 정보생성
[1, 5]	1) tf, icf 2) 인덱스 정보 생성
[3]	1) df, 2) 수동질의어 분석 3) 지식 DB정보 생성
[6]	1) db내에서 df 2) 인덱스 정보 생성
제안 모델	**1) tf, df, N 2) 검색인덱스 정보없음**

이는 인터넷상에 존재하는 수많은 문서들에 대해 검색 정보를 생성해야하는 단점이 있다. 또한 생성된 검색 정보들을 동적으로 변하는 인터넷 환경에서 일관성 유지를 위해 사용되는 비용도 매우 크다. 본 논문에서 제안하는 모델의 특징은 검색 정보를 생성하지 않고, 메타 정보만으로 정보원을 선택할 수 있는 장점이다.

5. 실험 결과

이 장에서는 현재 일반적으로 많이 사용되고 있는 5개의 일반 검색 엔진들과 3개의 메타 검색 엔진들과의 정확도 실험 결과는 <그림 4>,<그림 5>와 같다.

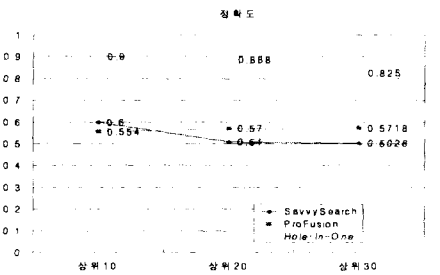


<그림 4> 일반검색 엔진과의 성능비교

본 실험에서는 제안한 관련성 분포 정보를 기반으로 검색 결과와 기존의 단일 검색엔진 결과와의 비교에서 정확성, 효율성 면에서 약 15%의 성능 향상을 확인하였다.

6. 결 론

본 논문에서는 통합 검색 시스템을 사용하여 인터넷상의 다양한 정보들을 효율적으로 검색할 수 있는 모델을 제안하고, 이 제안된 모델의 검색 성능을 평가하기 위해 HoleInOne 통합 검색 시스템을 설계 및 구현하였다. 앞으로의 연구 과제는 정보원에 대한 양질의 정보를 얻기 위해서는 질의에 적합한 정보원을 선택할 수 있도록 표준화된 메타 데이터 개발이 필요하고, 정보원에 대한 메타 정보 수집 방법과 융합 클러스터링 기법의 개발 등에 대한 연구가 필요하다.



<그림 5> 메타검색 엔진과의 성능비교

[참고문헌]

[1] J. P. Callan, Z. Lu, and W. B. Croft, "Searching Distributed Collections with Inference Networks," In Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA. pp. 21-28. 1995.

[2] E. M. Voorhees, N. K. Gupat, and B. Johnson-Laird., "The Collection Fusion Problem," In D. K. Harman, editor, The Third Text REtrieval Conference (TREC-3), Gaithersburg, MD, pp. National Institue of Standards and Technology, Special Publication 500-225., 1994.

[3] S. Gauch, G. Wang, and M. Gomez, "ProFusion: Intelligent Fusion from Multiple, Distributed Search Engines, WebNet '96", The First World Conference of the Web Society, San Francisco, October 1996.

[4] C. Baumgarten, "Probablistic Information Retrieval in a Distributed Heterogeneous Environment." PhD Thesis, Dresden Univ. of Techn., Accepted, 1999.

[5] J. C. French, A. L. Powell, J. Callan, C. L. Viles, T. Emmitt, K. J. Prey and Y. Mou, "Comparing the Performance of Database Selection Algorithms," SIGIR 99, 1999.

[6] L. Gravano, H. Garcia-Molina and A. Tomasic, "The effectiveness of GlOSS for the text database discovery problem," In Proceedings of the 1994 ACM SIGMOD Conference, May 1994.

[7] 김현주, 김상준, 배종민, "관련성 분포 정보를 이용한 컬렉션 융합",한국정보처리학회 춘계 학술 논문발표집. pp.907-910., 1999.