論文
2-11-7

관련성 분포정보를 이용한 통합 검색 시스템의 설계 및 구현 (Design and Implementation of Meta Search using Relevance Distribution Information)

김 현 주*
(Hyun-Ju Kim)

요 약

본 논문에서는 관련성 분포 정보를 이용한 새로운 컬렉션 융합 모델을 제시한다. 이는 먼저 주어진 질의에 대하여 검색에 참여한 정보원을 평가하고 질의에 가장 적합한 정보원을 선택한다. 그리고 정보원의 평가 결과에 따라 해당 정보원으로부터 검색 문서를 차별적으로 수집하고, 검색된 문서들은 정보원의 평가 값인 관련성 분포 정보를 기반으로 최종 검색 문서의 순위 매김을 수행한다. 이렇게 순위 매김된 검색 문서는 단일 우선순위를 가지는 검색 문서의 집합으로 통합하여 사용자에게 단일 검색 결과를 제공한다.

그리고 질의어에 대하여 가장 좋은 정보원들을 분류할 수 있는 체계를 개발하여 사용자의 질의어에 대하여 최선의 정보원들을 선택할 수 있는 알고리즘의 제시하였다. 마지막으로 선택된 정보원으로부터 질의에 적합한 문서를 검색한 후에 이들을 순위 매김하고 통합하는 통합검색 시스템을 제시한다.

ABSTRACT

We design the evaluation factors to represent the relevance distribution information between a query and sources and proposes the scheme to get relevance distribution information based on evaluation factors. Then it is developed that the organization is able to classify the best source toward query, and shown the algorithms that is able to select the best source toward users query, it is developed algorithms that is decided ranking and mering these, after choose the best source to evaluate a query, Finally, it merges the result from the source, and present them to the user to the issued query.

This paper also develops the scheme to classify the best sources for query and presents the selection algorithm of the best information sources. Finally the ranking and merging Federated Retrieval System is presented.

1. 서론

인터넷상에서 전문적으로 정보 제공의 기능을 가지고 있는 것을 정보원(source) 혹은 컬렉션(collection)이라고 한다. 사용자가 인터넷상의 정보원으로부터 자신이 필요한 정보를 찾고자 할 경우에 다음과 같은 문제점이 있다.

먼저 인터넷상에 존재하는 수많은 정보원 가운데에서 자신이 원하는 정보가 어느 곳에, 어떤 정보원에 있는지 찾기가 힘들며, 다음으로는 정보원을 찾았다 하더라도 정보원의 검색 기능을 효과적으로 사용하기 어렵다는 것 등이다. 이러한 정보원 관리 방법 중의 하나로 통합 검색 혹은 메타 검색의 방법이 그 해결책 중의 하나로 제시되었다.

* 정회원 : 경남정보대학 컴퓨터정보계열 전임강사

논문접수 : 2001. 11. 6.
심사완료 : 2001. 11. 11.

최근 등장한 메타 검색 시스템은 ProFusion[8], SavvySearch[9], 미스다찾니[50] 등이 있으며, 기존의 Yahoo[51], InfoSeek[2] 등과 같은 정보 검색 포탈 싸이트도 폭발적으로 늘어나는 정보를 자신의 컴퓨터에 저장하는 중앙 집중식 정보 관리법에 한계를 느껴 InfoSeek[2]에서는 메타 검색 시스템인 InfoSeek Patent를 개발하여 실험적으로 운영하고 있다.

통합 검색 혹은 메타 검색 시스템들은 사용자에게 분산된 이질의 내용을 가진 수많은 정보원들의 존재를 숨기고 전체적으로 하나의 정보원만 있다는 관점을 제공한다. 따라서 사용자는 하나의 질의 문법만으로 분산된 이질의 정보원을 동시에 사용하여 검색을 할 수 있다.

이러한 통합 검색 분야에서 사용자들의 질의에 대하여 가장 효율적인 검색 결과를 얻기 위해서 주로 연구되고 있는 분야는 크게 세 가지로 구분할 수 있다. 첫 번째는 질의에 대해 가장 좋은 정보원을 선택하는 문제이다. 이는 통합 검색 시스템이 검색에 참여시키고 있는 수많은 이질의 정보원 중에서 사용자의 질의어를 만족시킬 수 있는 가장 좋은 정보원들을 자동으로 결정하는 방법에 대한 것이다. 두 번째는 질의어 자동 번역 문제이다. 통합 검색 시스템에서는 단일 인터페이스를 통해 질의어를 발생시킨다. 이때 발생된 질의어는 가장 적합한 정보원을 선택한 후에 자동적으로 질의된다. 그러나 검색에 참여한 이질의 정보원은 서로 다른 질의 문법을 가지고 있어서 통합 검색 시스템에서 생성된 질의어를 직접 인식하지 못한다. 따라서 이들을 자동 번역하는 질의어 번역기가 필요하다. 마지막으로는 검색 문서의 통합 및 순위 매김을 처리하는 문제이다. 통합 검색 시스템은 입력된 질의어에 대하여 분산된 이질의 정보원으로부터 검색 결과를 수집한다. 그리고 문서에 대하여 순위 매김을 수행하여 이를 기준으로 검색 문서들을 단일 검색 결과로 통합한 후에, 사용자의 질의에 대한 검색 결과로써 회신한다. 이러한 통합 검색 시스템의 세 가지 연구 분야는 통합 검색 시스템의 검색 능력에 많은 영향을 미치며, 또한 검색을 수행할 때 상호 연관되어 동작한다.

따라서 본 논문에서는 관련성 분포 정보를 기반으로 주어진 질의에 대하여 검색에 참여한 정보원을 평가하고, 평가된 값에 따라 질의에 가장 적합한 정보원을 선택하며, 마지막으로 평가된 결과를 기반으로 정보원으로부터 검색할 문서를 차등적으로 수집하고, 또한 검색된 문서들은 정보원에 대한 평가 값인 관련성 분포 정보를 적용하여 최종 순위 매김을 수행하여, 검색된 문서의 최종 순위 매김 값에 따라 단일 우선 순위를 가지는 검색 문서의 집합으로 통합하여 사용자에게 단일 검색 결과를 제공해주는 통합 검색 시스템을 제안한다.

2. 관련 연구

이 절에서는 기존의 컬렉션 융합 모델에 대하여 알아본다. 이 절에서 소개하는 3가지 컬렉션 융합 모델은 다음과 같다. ProFusion[6] 메타 검색 시스템에서 사용한 컬렉션 융합 모델, INQUERY[1, 9, 12] 메타 검색 시스템에서 사용한 컬렉션 융합 모델, Voorhees[2, 3]외 2명이 제안한 컬렉션 융합 모델 등이며, 이들이 가지고 있는 특징들을 서로 비교하여 살펴본다.

2.1 ProFusion[6] 메타 검색 시스템의 컬렉션 융합 모델

ProFusion[6]은 미국의 캔자스 대학의 Susan Gauch[6]외 2명이 개발한 메타 검색 시스템이다. 이 메타 검색 시스템의 특징은 9개의 일반 검색 엔진을 대상으로 질의를 수행하고 이들로부터 검색 결과를 수집하여 통합 검색 결과를 사용자에게 URL로 보여준다. 이때 검색 문서를 통합할 때 중복된 검색 문서의 제거, 빈 URL 제거 등의 기능도 제공한다.

ProFusion[6] 메타 검색기에서는 사용자의 질의에 대하여 9개의 정보원을 선택하는 방법으로는 최상의 3개 검색 엔진을 선택하는 방법, 가장 빠른 검색 결과를 보여주는 3개의 검색 엔진을 선택하는 방법, 9개의 검색엔진 모두 다 사용하는 방법, 사용자가 검색엔진을 선택하여 사용하는 방법 등을 제공한다.

ProFusion 메타 검색기에서의 컬렉션 융합은 질의어에 대한 신뢰도(cf)와 정규화 된 문서의 평가 값 등의 두 가지 요소를 곱으로 문서의 최종 순위 매김을 수행하며, 이들의 평가 값을 기준으로 통합을 수행을 한다. ProFusion 메타 검색기의 컬렉션 융합을 위한 수식은 다음과 같다.

$$R_{di} = cf_i * mf_{di}$$

위의 수식에서 사용된 CFi는 검색에 참여한 정보원들의 신뢰도 값이며, MFdi는 정보원으로부터 검색된 문서에 대하여 정규화 과정을 통해 생성된 문서의 평가 값이다. 이를 통해 **ProFusion** 메타 검색기에서는 컬렉션 융합을 수행하기 위해 Rdi로 검색 문서를 재평가한 값으로 평가하였다. 이는 위의 수식과 같이 CFi와 MFdi 등의 두 요소를 곱으로 생성하며, 이들 문서의 최종 순위 매김 값을 내림차순으로 컬렉션 융합을 수행한다.

2.2 INQUERY 메타 검색 시스템에서의 컬렉션 융합 모델

다음으로는 Callan[1, 9, 12]외 3명이 제안한 컬렉션 융합 모델을 INQUERY 메타 검색 시스템으로 실험 및 구현한 컬렉션 융합 모델이다. 이러한 컬렉션 융합 모델을 흔히 CORI NET 검색 모델이라고도 한다. 이 모델은 문서, 컬렉션 그리고 질의어 사이의 관련성을 df와 icf를 기반으로 평가한 후에, 이들을 방향 그래프 자료 구조를 사용하여 검색 정보를 생성한다. 또한 주어진 질의어와 문서와의 관련성 평가 정보를 문서 네트워크 부분과 질의어 네트워크 부분 등으로 분류하여 이들의 검색 정보를 표현한다.

CORI net 검색 모델에서는 주어진 질의어에 대하여 가장 적합한 정보원을 선택하기 위해 term과 검색 문서 사이를 df와 icf를 기반으로 평가하여 정보원 선택 정보를 생성한다. 이때 정보원을 평가하고 정보원에 대한 순위 매김 처리 과정은 mean squred error metric를 사용하였으며, 이때 단일 질의어에 대한 컬렉션 평가 수식은 다음과 같다.

$$\frac{1}{|C|} \cdot \Sigma_{i \in c} (O_i - R_i)^2$$

- $|C|$: 컬렉션의 수
- O_i : 컬렉션 i의 최적의 순위
- R_i : 검색 알고리즘에서 제공되는 컬렉션의 순위

위의 수식에서 O_i는 질의어가 포함되어 있는 관련 문서의 수를 기반으로 평가된다. 예를 들어 질의어와 가장 많은 관련 문서를 포함하고 있는 컬렉션은 1 순위로 매김하고, 다음으로 관련있는 컬렉션을 2 순위로 매김하는 방법이다.

CORI NET 정보 검색 모델을 기반으로 구현한 INQUERY 검색 시스템에서는 검색 문서에 대한 컬렉션 융합 방법으로 Interleaving, Raw Scores, Normalized Scores, Weighted Scores 등의 네 가지 방법을 사용하였다.

2.3 Voorhees외 2명이 제안한 컬렉션 융합 모델

Voorhees[2, 3]외 2명이 제안한 컬렉션 융합 모델은 주어진 질의어와 검색에 참여한 컬렉션과의 관련성 정도를 유사도를 이용하여 컬렉션을 평가하고 융합하는 모델이다. 이때 컬렉션에 대한 유사도를 추정하는 방법으로는 문서의 관련성 분포 방법과 질의어 클러 스트링 방법 등이 있다.

다음은 문서의 관련성 분포 정보를 이용한 처리 과정을 간략하게 표현하였다.

단계 1 : 학습된 질의어들로 컬렉션에 대한 유사도 정보 자료 구조를 생성한다.

단계 2 : 새로운 질의어가 주어지면, 질의어와 유사한 k개의 학습된 질의어로부터 이들 유사도 값의 평균값을 계산하고, 이를 새로운 질의어의 유사도 값으로 평가한다.

단계 3 : 질의어 유사도와 전체 검색 문서 수 N을 기반으로 Maximization Procedure 검색 문서 크기 추정 함수로부터 컬렉션에서 검색할 문서의 크기를 계산한다.

단계 4 : 검색 결과를 "rolling biased c-faced die" 방법으로 융합한다.

다음으로 질의어 클러스트링 정보를 이용한 방법이다. 이는 앞의 방법과 동일하게 미리 질의어들을 학습시켜 질의어와 컬렉션간의 유사도를 평가한다. 이때 질의어들을 학습할 때 질의어 상호간에 검색 결과가 공통으로 발생하는 문서의 빈도 수를 기반으로 질의어들을 클러스트링하고, 이때 클러스트링된 질의어들의 유사도를 평균 값으로 산출하여 이를 클

러스트링된 모든 질의어의 중심 값으로 한다. 이 중심 값은 클러스트링된 모든 질의어들이 컬렉션에 대한 유사도 값을 산출할 때 사용된다. 이때 평가된 유사도는 다음의 수식으로 각 컬렉션에서 검색할 문서의 크기를 결정하는데 사용된다.

$$\frac{w_i}{\sum_{i=1}^{c} w_i} * N$$

- w_i : 평가하고자하는 컬렉션의 유사도
- $\sum_{i=1}^{c} w_i$: 검색에 참여한 컬렉션 유사도의 합
- N : 전체 검색 문서의 수

위의 수식으로부터 각 컬렉션에서 검색되어질 문서의 크기가 결정되면 마지막으로 검색 문서를 통합하고 단일 우선 순위를 가지는 검색 결과로 생성하는데, 이때 적용한 컬렉션 융합 모델은 Round Robin 방법으로 검색 문서를 통합하여 이들을 최종 문서 순위 매김으로 간주한다.

3. 통합 검색 시스템의 설계

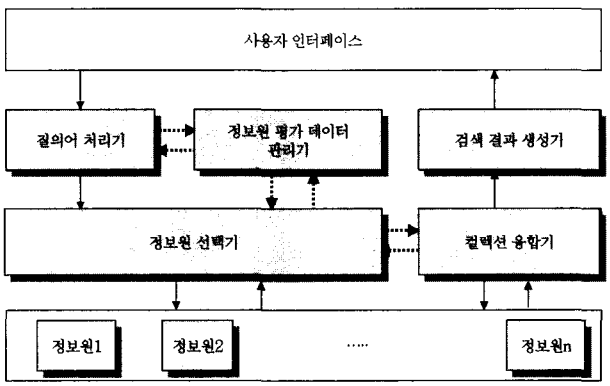
3.1 전체적인 구조도

통합 검색 혹은 메타 검색 시스템은 질의어가 주어질 때 통합 검색 시스템에서는 질의에 가장 적합한 정보원들을 평가한다. 그리고 질의어는 선택된 정보원에서 사용하는 질의 문법으로 자동 번역한 후에, 통합 검색 시스템이 각 정보원으로 질의를 대신하게 된다. 또한 질의에 대하여 검색 결과를 분산된 이질의 정보원으로부터 수집하고, 이들을 순위 매김한 후에 문서의 최종 순위 매김에 따라 단일 검색 결과로 통합하여 이를 사용자의 검색 결과로 회신한다[1, 2, 3].

통합 검색 시스템은 일반적으로 사용자 인터페이스, 질의 처리기, 정보원 선택기, 컬렉션 융합기, 검색 결과 생성기 등으로 구성되어 있다. 이들은 먼저 사용자 인터페이스로 통합 검색 시스템의 모습을 사용자에게 보여주며, 이곳에서 단일 질의 문법으로 질의할 수 있게 한다. 이때 발생된 검색 질의어는 질의어 처리기를 통해 검색에 참여한 컬렉션의 질의 문법으로 자동 변화된다. 그리고 주어진 질의에 대해 정보원의 평가 결과에 따라 차등적으로 검색 문서를 수집하고, 이를 최종 순위 매김하여 사용자에게는 마치 단일 정보원으로부터 검색 문서를 수집한 것처럼 단일 검색 결과를 문서의 목록으로 보여준다.

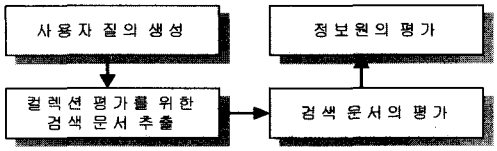
3.2 정보원 선택

정보원의 선택 과정은 사용자가 입력한 질의어에 대하여 가장 적합한 문서를 가지고 있는 정보원을 선택해 준다. 본 논문에서는 가장 양질의 정보원을 선택하기 위해 [그림 2]와 같은 과정을 수행한다.



[그림 1] 통합 검색 시스템의 구조

[Fig. 1] Architecture of Federated Retrieve System



[그림 2] 정보원의 관련성 분포 정보 평가 과정

[Fig. 2] the processing of evaluation relevance distribution information about collection

[그림 2]는 질의에 가장 적합한 정보원을 선택하기 위해서 검색 문서의 평가와 정보원에 대한 평가 등의 두 단계를 통해 질의에 대한 컬렉션의 관련성 분포 정보를 평가하고, 평가된 값을 기반으로 양질의 정보원을 구별한다. [그림 2]에서 정보원 선택의 첫 번째는 검색 문서의 평가이다. 이는 질의어와 검색 문서가 서로 관련된 문서인지를 판단하는 과정이다. 이때 문서의 관련성은 검색 문서의 재평가 값, 빈 URL, 중복 문서 등의 항목으로 판단한다. 두 번째는 정보원에 대한 관련성 판단이다. 이는 주어진 질의어와 컬렉션 사이의 관련성 정도를 평가하여 나타낸 값이다. 컬렉션의 관련성 분포 정보는 검색 문서의 재평가 값, 관련 문서의 위치 정보 값, 검색 문서의 정확도 값 등의 항목으로 평가되며, 이는 본 논문에서 제안한 질의어와 검색에 참여한 정보원 사이의 관련성 분포 정보이다.

3.2.1 관련성 판단 알고리즘

이 절에서는 질의어와 검색 문서 사이의 관련성을 검사하는 알고리즘에 대하여 설명한다. 본 논문에서 제안한 검색 문서의 관련성 유·무 평가는 검색 문서의 재평가 값, 중복 문서, 빈 URL 등의 항목으로 관련성 검사를 수행한다.

다음의 (알고리즘 1)은 질의어와 검색 문서간의 관련성 검사를 수행하는 알고리즘이다.

(알고리즘 1)에서 1번째 라인은 관련성 판단 알고리즘의 시작 부분이다. 이는 정보원을 평가하기 위해 추출된 검색 문서와 검색 문서의 수이다. 3번째 라인은 빈 URL을 검사하고, 빈 URL이 아닐 경우에는 4번째 라인에서 관련없는 문서 정보를 유지한다. 다음으로 5번째 라인은 중복된 문서를 검사하여 관련 문서와 관련없는 문서 정보를 기억한다. 마지막으로 2번 라인부터 13번 라인까지 정보원으로부터 검색된 N번을 반복 수행한 후에 관련성 평가를 위

```
1: DocumentRelevanceCheck(String Query, String content, int N)
2:   WHILE( N <= 0 ) DO
3:     IF EmptyURL(content) THEN  // 빈 URL 문서
4:         NonRelevanceValue++;  // 비관련 문서의 수
5:   ELSE IF DuplicateDocument(content) THEN //중복 문서
6:         NonRelevanceValue++; // 비관련 문서의 수
7:   ELSE IF (DocumentWeight(content) > α) THEN//문서의 재평가 값
8:         RelevanceValue++; // 관련 문서의 수
9:   ELSE
10:        NonRelevanceValue++;
11:  END IF
12:    N--;       // 평가 집단의 크기
13:  END_WHILE
14:  RETURN RelevanceValue; // 문서의 관련성 검사 결과
15: END DocumentRelevanceCheck;
```

(알고리즘 1) 관련성 판단 알고리즘

(Algorithm 1) relevance check algorithm

해 수집된 문서 중에서 관련된 문서의 정보를 되돌려준다. 이 알고리즘에서 평가된 정보는 다음 단계에서 정보원을 평가할 때 사용된다.

3.2.2 관련성 분포 정보 평가 알고리즘

이 절에서는 질의어와 검색에 참여한 정보원 사이의 관련성 정도에 대해 평가하는 알고리즘을 살펴본다. 본 논문에서 질의어와 정보원 사이의 관련성 정도를 평가하는 요소로는 검색 문서의 재평가 값, 관련 문서의 위치 정보 값, 검색 문서의 정확도 등의 항목으로 평가한다.

다음의 (알고리즘 2)은 정보원에 대한 관련성 분포 정보를 평가하는 알고리즘이다.

검색에 참여한 컬렉션의 관련성 분포 정보 평가는 평가 모집단에서 질의어와 관련된 검색 문서의 순서 정보와 관련된 문서의 수 등 두 가지 정보로 사용한다. 먼저 관련 검색 문서의 순서 정보는 관련된 문서가 정보원으로부터 부여받은 순서 정보이다. 다음은 관련된 문서의 수이다. 이는 평가 모집단 내에서 질의어에 대하여 관련 문서의 정확성을 평가할 수 있다. 이는 평가 모집단 내에서 질의어와 관련된 문서의 수를 평가를 위해 추출한 모집단의 크기 N으로 나누어서, 평가 모집단에 대한 정확도를 추정할 수 있다. (알고리즘 3)의 알고리즘에서 8번째 라인은 관련 문서의 위치 정보에 대한 가중치를 계산하며, 11번째 라인은 모집단에 대한 정확도를 계산한다. 또한 이들 8, 11번째 라인의 계산 결과를 곱하여 해당 정보원에 대한 관련성 분포 정보로 사용하였다.

3.3 검색 문서의 통합

이 절에서는 앞 절에서 추정한 정보원의 관련성 분포 정보를 사용하여 각 정보원으로부터 질의에 적합한 문서를 수집하고, 수집된 문서에 대하여 순위 매김을 수행하여 단일 검색 결과로 통합하는 알고리즘을 기술한다.

3.3.1 문서 사이의 간격 값 평가 알고리즘

다음의 (알고리즘 4)는 문서 사의의 간격 값 평가에 대한 알고리즘이다.

```
1:  CollectionToWeight(int CollectionToNum, int SelectedDocNum)
2:   FOR ALL i ∈ CollectionToNum DO     // 검색에 참여한 정보원의 수
3:       initial RelevanceDocNum, DocPositionValue;
4:       FOR ALL j ∈ SelectedDocNum DO      // 평가 모집단의 문서 수
5:          CheckValue = DocumentRelevanceCheck(String query, String content);
6:           IF(CheckValue != 0) THEN
7:                RelevanceDocNum++;        // 관련 문서 수
8:                DocPositionValue = DocPositionValue + (1/index);
9:            END-IF                  // 관련 문서의 위치 정보 보상 값
10:      END-FOR
11:      PrecisionValue = (RelevanceDocNum / SelectedDocNum);
12:      PositionValue = (DocPositionValue / SelectedDocNum)
13:      ColWeight[i] =  DW * PositionValue * PrecisionValue;
14:   END-FOR
15:  RETURN ColWeight[i]          // 정보원에 관련성 분포 정보 값
16: END CollectionToWeight
```

(알고리즘 3) 관련성 분포 정보 평가 알고리즘

(Algorithm 3) the evaluation algorithm of relevance distribution information

```
1: IntervalValue(int ColWeight, int CurrentColWeight, int ColSearchToNum)
2:  MOVE ColWeight[0] TO MinValue;
3:      FOR ALL i ∈ ColSearchToNum DO
4:          IF(MinValue < ColWeight[i+1]) THEN
5:              MinValue = ColWeight[i+1];
6:          END-IF                  // 정보원 가중 값 중에서 최소 값을 찾는 과정
7:      END-FOR
8:     InterWeight = MinValue / CurrentColWeight; // 문서간의 간격 값
9:    RETURN InterWeight;
10:  END IntervalValue
```

(알고리즘 4) 문서 사이의 간격 값 계산 알고리즘

(Algorithm 4) the evaluation algortithm of interval of docuemnt

(알고리즘 4)는 검색에 참여한 정보원의 가중치 값을 상대적인 비율 값으로 변환하여 계산한다. 즉 정보원 중에서 가장 적은 가중치 값을 분자로 하고, 평가하고자하는 정보원의 가중치 값을 분모로 하여 나눈 값을 해당 정보원에서 검색된 문서의 간격 값으로 사용한다. (알고리즘 4)에서 3번째부터 7번째 라인에서 검색에 참여한 정보원 중에서 가장 적은 가중치를 가지는 정보원의 가중치 값을 찾는다. 그리고 8번째 라인에서 평가하고자하는 정보원의 가중치 값으로 나누어서 컬렉션에서 검색된 문서간의 간격 값을 계산한다.

3.3.2 순위 매김 알고리즘

질의에 의해 검색된 문서는 관련성 분포 정보를 이용한 순위 매김에서는 검색에 참여한 모든 정보원이 동등하며, 검색 문서의 분포만 다르다고 가정하여 문서 순위 매김을 수행하였다. 그러나 검색에 참여한 정보원들 사이에서 동일한 우선 순위를 가지는 검색 문서가 서로 비교되는 경우가 있다. 이 경우에는 관련 분포 정보가 클수록 양질의 검색 문서를 가질 확률이 높다라고 가정하였으므로, 이를 위해 관련성 분포 정보를 최종 순위 매김의 가중치로 사용하였다.

다음의 (알고리즘 5)는 문서에 대한 최종 순위 매김 알고리즘이다.

(알고리즘 5)에서 5번째 라인부터 7번째 라인까지가 검색된 모든 문서에 대하여 최종 순위 매김을 수행하고 있다. 이들은 해당 정보원에 대한 가중치 값과 문서에 대한 단계별 가중치를 더하여 최종 순위

```
1: DocRankValue(int ColWeight, int ColSearchToNum, int DocMaxNum)
2:  FOR ALL i ∈ ColSearchToNum DO
3:    InterWeight = IntervalValue(ColWeight, CurrentColWeight, ColSearchToNum);
4:    DocNum = InterWeight * DocMaxNum;
5:       FOR j = 0 to (DocNum - 1) DO
6:        DocRankScore[j,i] = DWij + (DocMaxNum-(InterWeight[i] * j));
7:       END-FOR // 문서의 순위 매김
8:  END-FOR
9:  RETURN;
10: END DocRankValue
```

(알고리즘 5) 최종 순위 매김 알고리즘

(Algorithm 5) the evaluation algorithm for final ranking

매김 값으로 사용한다. 마지막으로 관련성 분포 정보를 기반으로 평가된 문서의 평가 값에 따라 내림차순으로 단일 검색 결과의 집합으로 생성되며, 사용자에게 질의에 대한 결과로써 제공한다.

4. 통합 검색 시스템의 구현

이 장에서는 본 논문에서 제안한 관련성 분포 정보 기반 컬렉션 융합 모델의 성능 평가를 위해 구현한 HoleInOne(wHOLE INformation ONEtime) 통합 검색 시스템에 대해 살펴본다. 먼저 4.1절에서는 HoleInOne 통합 검색 시스템의 전체적인 개요에 대해 살펴보며, 4.2절에서는 HoleInOne 통합 검색 시스템에서 관련성 분포 정보를 평가하기 위해 사용한 평가 요소들을 살펴본다. 이들 평가 요소는 질의어와 정보원 사이의 관련 분포 정보를 추출하는데 사용되며, 또한 질의에 대하여 가장 적합한 정보원의 선택과 검색 문서의 순위 매김 및 통합 등에 사용된다. 다음으로 4.3절에서는 질의에 대해 가장 적합한 정보원을 선택하는 정보원 선택 부 시스템에 대하여 살펴보고, 4.4절에서는 이질의 정보원으로부터 검색된 문서를 단일 검색 결과로 생성하는 컬렉션 융합 부 시스템에 대해 살펴본다.

4.1 개요

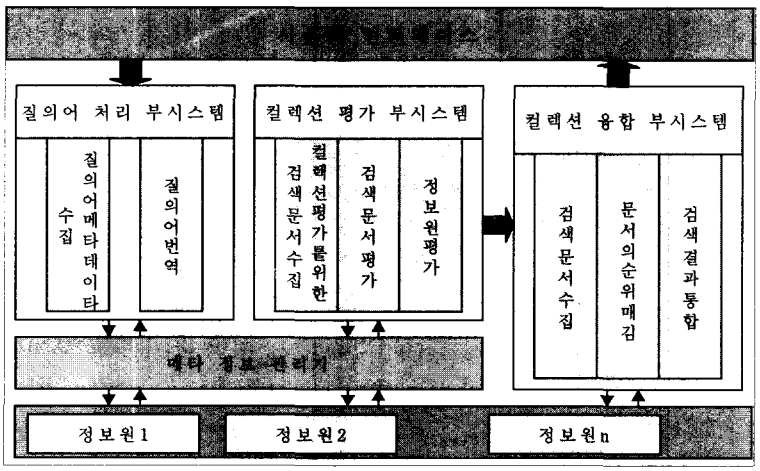
이 절에서는 본 논문에서 제안된 컬렉션 융합 모델을 평가하기 위해 구현한 HoleInOne 통합 검색 시스템의 전체적인 구조에 대해 살펴본다. 본 논문에서 구현한 HoleInOne 통합 검색 시스템은 사용자 인터페이스 부 시스템, 질의어 처리 부 시스템, 컬렉션 평가 부 시스템, 컬렉션 융합 부 시스템 등 4 개의 부 시스템으로 구성되어 있다.

[그림 3]은 실험을 위해 구현한 HoleInOne 통합 검색 시스템의 전체 구조이다.

사용자 인터페이스 부 시스템은 검색에 참여한 이질의 다양한 정보원들을 하나의 모습으로 보여주며, 사용자들은 질의 및 검색 결과 등에 관한 정보의 입력 및 수집을 할 수 있다. 이를 위해 본 논문에서는 HTML, JavaServerlet, Java 언어 등을 사용하여 구현하였다.

컬렉션 평가 부 시스템은 사용자로부터 질의어가 주어질 때 가장 양질의 문서를 가지고 있는 정보원을 선택하는 기능을 제공한다. 본 논문에서의 정보원 평가는 검색 문서의 관련성 평가와 정보원 평가 등으로 검색에 참여한 정보원을 평가한다.

컬렉션 융합 부 시스템은 이질의 정보원으로부터 검색된 문서들을 하나의 검색 결과 집합으로 통합하여 사용자에게 검색 결과로써 되돌려주는 기능을 제



[그림 3] HoleInOne 시스템의 구조

[Fig. 3] Architecture of HoleInOne System

공한다. 이는 검색된 문서의 순위 매김 및 통합 과정으로 검색 결과를 생성한다.

4.2 평가 메타데이타 HoleInOne 통합 검색 시스템에서 사용한 평가 요소

이 절에서는 본 논문에서 제안한 컬렉션 융합 모델의 실험을 위해 구현한 HoleInOne 통합 검색 시스템에서 관련성 분포 정보를 추출하기 위해 사용한 평가 요소와 이들의 기능에 대해 살펴본다. 본 논문에서 실험을 위해 사용한 평가 요소는 총 7개이며, 이들은 <표 1>과 같다. 이들은 주로 질의에 적합한 정보원의 선택과 질의에 의해 검색된 문서들의 순위 매김 정보로 사용하였다.

<표 1> 관련성 분포 정보 평가 요소

<Table 1> element of relevance distribution information

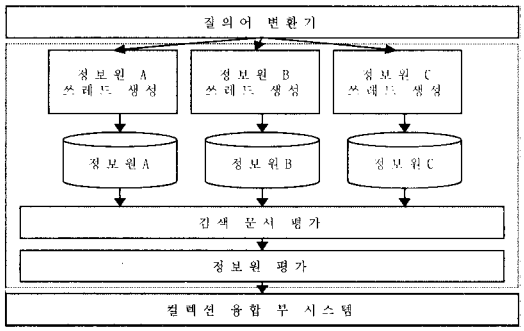
평가 요소	의 미
DocNo	검색된 문서의 순서
Re-ranking scores	검색 문서의 재평가 값
URL	검색 문서의 주소(URL)
Title	검색 문서의 Title 내용
Content	검색 문서의 Abstracts 내용
TitleCount	Title에서 질의어가 발생한 빈도 수
ContentCount	Abstract에서 질의어가 발생한 빈도 수
TCount	정보원에서 재평가된 검색 문서의 수(N)

<표 1>은 HoleInOne 통합 검색 시스템에서 컬렉션 융합을 위해 사용한 평가 요소들이다. DocNo는 검색된 문서의 순서 정보를 가지고 있다. Re-ranking scores 평가 요소는 본 논문에서 제안한 질의어 관련성 정보를 평가하기 위해 검색 문서들을 재평가한 값이다. URL 평가 요소는 검색된 문서의 인터넷 위치 정보를 가진다. Title과 Content 평가 요소는 질의에 대해 정보원으로부터 검색된 문서에서 "<title> ... </title>" 태그와 <abstract> ... </abstract> 태그 안에 있는 내용 데이터를 가진다. 이는 검색 문서와 질의어 사이의 관련성을 판단할 때 사용된다. 마지막으로 TitleCount와 ContentCount는 각 Title과 Content 평가 요소는 문서 내에서 질의어가 발생한 빈도 수 정보를 가지고 있다. 이들 평가 요소는 해당 문서가 질의어와의 관련성 분포 정보를 계산할 때 사용된다.

4.3 컬렉션 선택 부 시스템

컬렉션 평가 부 시스템은 사용자가 입력한 질의어에 대하여 가장 양질의 문서를 가지고 있는 정보원을 자동적으로 선택해 주는 기능을 제공한다. 본 논문에서는 이를 위해 검색 문서의 재평가 값, 관련 문서의 위치 정보 평가 값, 검색 문서의 정확도 등의 항목으로 검색에 참여한 정보원을 평가하였다.

다음의 [그림 4]은 컬렉션 선택 부 시스템의 구조이다.



[그림 4] 컬렉션 선택 부 시스템의 구조

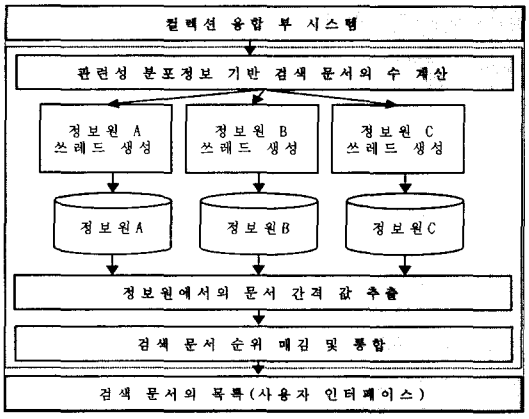
[Fig. 4] structure of Collection Selection sub-system

[그림 4]는 통합 검색 시스템에 3개의 정보원이 검색에 참여한 경우이며, 이때 컬렉션 선택 부 시스템의 처리 과정을 나타낸 것이다. 이 부 시스템에 입력되는 데이터는 통합 검색 시스템에서 생성된 질의어와 컬렉션 메타 정보 관리 부 시스템으로부터 수집된 컬렉션에 대한 메타 데이터 등이다. 수집된 메타 데이터를 컬렉션 평가를 위한 모집단이라고 하며, 검색 문서의 재평가 값, 관련 문서의 위치 정보 평가 값, 검색 문서의 정확도 등의 항목으로 검색에 참여한 컬렉션들을 평가한다.

4.4 컬렉션 융합 부 시스템

본 논문에서는 검색에 참여한 정보원으로부터 검색 문서의 수를 결정하는 것은 질의어와 정보원 사

이의 관련성 분포 정보를 추론하여 사용하였으며, 그리고 검색된 문서를 통합할 때는 관련성 분포 정보의 의미를 내포하고 있는 정보원의 문서 간격 값을 사용하였다. 다음의 [그림 5]는 컬렉션 융합 부 시스템의 구조이다.



[그림 5] 컬렉션 융합 부 시스템의 구조

[Fig. 5] structure of collection fusion sub-system

[그림 5]는 컬렉션 융합 부 시스템에서 검색된 문서를 순위 매김 후 통합하는 처리 과정에 대한 표현이다. 첫 번째 단계로 관련성 분포 정보를 기반으로 컬렉션으로부터 문서를 검색하는 과정으로 표현하였다. 두 번째는 검색된 문서들에 대해 순위 매김을 수행하고 이를 기준으로 단일 검색 결과로 통합하는 단계이다. 따라서 문서의 순위 매김에서도 분포 비율에 따라 순위 매김을 수행하여 단일 검색 결과로 통합하는 과정을 수행하였다.

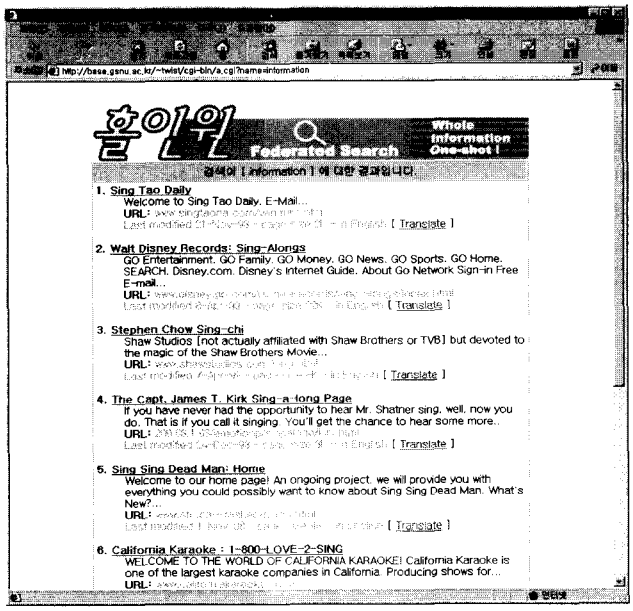
4.5 HoleInOne 시스템의 구현

이 절에서는 HoleInOne 통합 검색 시스템의 구현 화면에 대해서 살펴본다. 먼저 이 절의 구성은 HoleInOne 통합 검색 시스템의 질의 화면과 HoleIn One 통합 검색 시스템의 검색 결과 화면에 대한 내용이다. 다음의 [그림 6]은 HoleInOne 통합 검색 시스템의 질의 화면이다.

[그림 6]은 본 논문에서 제안한 관련성 분포 정보 기반 컬렉션 융합 모델의 성능을 분석하기 위해 최소의 기능으로 구현한 질의 화면이다. 질의 화면에 입력된 질의어는 HoleInOne 시스템의 검색에 참여한 이질의 컬렉션에서 사용하는 질의 문법으로 자동 번역되어 HoleInOne 시스템에서 대신 질의를 한다. 이는 사용자들에 다양한 컬렉션들의 모습을 숨기고 하나의 인터페이스만을 제공함으로써 인터넷상에 존재하는 다양한 컬렉션의 질의 문법 이해에 대한 사용자의 부담을 해결해 준다.

[그림 6] 질의 화면

[Fig. 6] The query window



[그림 7] 질의 검색 결과 화면

[Fig. 7] The result of search window from query

다음의 [그림 7]은 [그림 6]으로부터 입력된 검색 질의에 대해 HoleInOne 시스템에서 검색 결과를 생성해주는 질의 검색 결과 화면이다.

[그림 7]은 HoleInOne 시스템에 참여한 여러 개의 이질 컬렉션으로부터 검색 결과를 수집한 후에 본 논문에서 제안하는 컬렉션 융합 방법으로 단일 검색 결과를 생성한 화면이다.

5. 결론

본 논문에서 제안한 관련성 분포 정보 기반 통합 검색 시스템은 정보원 선택 문제와 검색 결과 통합 방법 등의 두 부분에서 새롭게 제안된 모델이다. 이를 실제 응용에 사용하기 위해서는 다음과 같은 문제를 보완해야 한다. 첫 번째는 질의어 자동 변환에 대한 연구이다. 두 번째는 검색 문서의 관련성 판단 알고리즘에 대한 연구이다.

앞으로의 연구 과제는 정보원에 대한 양질의 정보를 얻기 위해서는 질의에 적합한 정보원을 선택할 수 있도록 표준화된 메타데이터 개발이 필요하고, 정보원에 대한 정보 수집 방법과 융합 클러스터링 기법의 개발 등에 대한 연구가 필요하다. 또한 질의어 처리 기능의 확장이 필요하다. 즉, 불리언 모델에 바탕을 둔 질의어 처리 기능과 순위 매김 모델에 바탕을 둔 질의어 처리 기능 등의 연구이다. 이러한 정보는 본 논문에서 제시된 알고리즘의 성능을 크게 개선시킬 수 있다.

※ 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, National Institue of Standards and Tech- nology, Special Publication pp. 216-225., 1994.

[3] E. M. Voorhees, N. Gupta, and B. Johnson-Laird., "Learning Collection Fusion Strategies," *ACM SIGIR '95*, pp. 172-179, 1995.

[4] C. L. Viles and J. C. French, "Dissemination of Collection Wide Information in a Distributed Information Retrieval System," *ACM SIGIR '95*, 1995.

[5] A. Moffat and J. Zobel, "Information Retrieval Systems for Large Document Collections," *The Third Text REtrieval Conference (TREC-3)*, pp. 85-94., 1994.

[6] 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.

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

[8] C. Baumgarten, "A Probabilistic Solution to the Selection and Fusion Problem in Distributed Information Retrieval," *ACM SIGIR '99*, 1999.

[9] 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," *ACM SIGIR 99*, 1999.

[10] N. Fuhr, "Resource Discovery in Distributed Digital Libraries," *ACM SIGIR '99*, 1994.

[11] 금기문, 남세진, 신동욱, 김태균, "문서 클러스터링 정보를 이용한 컬렉션 융합," *한국정보과학회 추계학술 논문발표집*, pp. 147-149., 1998.

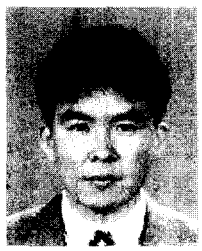
[12] 김현주, 김상준, 배종민, "관련성 분포 정보를 이용한 컬렉션 융합," *한국정보처리학회 춘계학술 논문발표집*, pp. 907-910., 1999.

[13] 김현주, 김상준, 배종민 "디지털 도서관에서 사용자 질의어와 컬렉션 사이의 관련성 분포정보를 이용한 컬렉션 융합," *한국 정보처리학회 논문지 제6권 제10호*, pp. 2728-2739., 1999.

[14] 김연곤, 엄채임, 변정용, "빈 연결을 제거하는 메타 검색 엔진의 구현," *한국멀티미디어학회 추계학술발표회*, pp. 359-364., 1998.

[15] 김현주, 배종민, "통합 검색에서 관련성 분포 정보를 이용한 정보원 선택에 관한 연구," *경상대학교 전산개발연구소 제14권*, 1999.

김 현 주



1988년 경상대학교
전산통계학과(이학사)
1990년 숭실대학교
전자계산학과(공학석사)
2000년 경상대학교
전자계산학과(공학박사)
1994년 ~ 1997년
제일정밀공업(주) 연구원
20000년 ~ 현재 경남정보대학
컴퓨터정보계열 전임강사
관심분야 : 정보검색,
디지털 도서관,
웹 프로그래밍
E-mail : khj@kit.ac.kr