학습방법을 이용한 분산통합검색시스템의 설계

강무영[0] 이민호 주원균
한국과학기술정보연구원
(kmy[0], cokeman, joo)@kisti.re.kr

A Design of Distributed Information Retrieval System using Learning Method

Mu-Yeong Kang[0], Min-Ho Lee, Won-Kyun Joo
Dept. Information System Research, Korea Institute of Science & Technology Information

요 약

본 논문에서는 여러 가지 분산통합 검색 방법중 학습을 이용한 분산통합 검색 시스템을 설계한다. 분산통합 검색시스템의 여러가지 이슈중 결과통합 문제에 주안점을 두었으며, 설계목적은 다양한 학습방법을 적용한 검색 결과 통합 실험을 위함이다. 이러한 목적을 달성하기 위하여 확장성을 고려한 모듈화를 통한 설계를 적용하여 다양한 실험과 향후 컬렉션 선택모듈, 질의변환 모듈도 삽입이 가능하도록 설계하였다.

1. 서 론

분산 통합 검색은 사용자가 원하는 결과를 포함하는 문서집합(컬렉션)을 선택하는 컬렉션 선택문제, 원천검색서버에 적합한 질의형태로 변환하고 얻어진 결과를 변환하는 변환문제, 검색 서버들로부터 얻어온 결과들을 하나의 통합된 결과로 만드는 결과통합 문제등을 주 논의점으로 하며 많은 연구가 진행되었다. 그 중 결과통합은 검색된 많은 정보에서 중복된 정보들을 제거하고 질의에 더욱 가까운 문서를 우선적으로 사용자에게 제시한다는 중요성 때문에 추론네트워크를 사용한 방법[1], 원문 재색인 방법[2], 링크 정보를 이용한 방법[3], 프로토콜을 이용한 방법[4], 학습을 이용한 방법[5][6]등 다양한 방법이 제시되었다. 이 중 이민호의 방법[6]은 기존의 학습을 이용한 분산통합 방법보다 더 효과적인 방법을 제시한다. 하지만 대용량의 데이터와 다양한 지역 검색엔진의 사용을 하지 않은 실험이기 때문에 더 많은 실험이 필요하다. 본 논문에서는 이민호의 학습방법을 이용한 분산통합 검색의 실험을 위하여 분산통합 검색 시스템을 설계한다. 본 시스템은 다양한 학습방법의 적용, 다양한 지역 검색엔진의 적용등의 확장성을 고려하여 모듈화를 통한 설계를 하여 다양한 실험이 가능하도록 한다. 다음 장에서는 이민호의 방법[6]을 개략적으로 설명하며, 3장에서는 본 논문에서 제시한 분산통합 검색 시스템의 구조를 설계한다. 마지막으로 4장에서는 결론과 향후 발전 방향에 대하여 제시한다.

2. 학습을 이용한 분산 검색 결과 통합

학습을 이용한 분산 검색 결과 통합 방법은 여러가지가 있다. 이 중 적합성 판단 정보를 이용한 순위 결정 방법[6]은 검색 정확도와 속도면에서 기존 방법보다 유용한 방법이다. 이는 임의의 질의 Q1가 I1 문서집합에서 검색되어 나온 결과 문서들의 순위별 적합 여부가 존재할 때, Q1 질의와 충분히 유사한 질의 Q2를 같은 I1 문서집합에서 검색하여 나온 결과문서들의 순위별 적합 여부도 Q1의 결과문서들의 순위별 적합여부와 비슷할 것이라는 개념으로 만들어진 방법이다. 그림 1은 이 개념을 설명하고 있다.

좀 더 자세히 살펴보면, 학습된 적합성 판단 정보를 이용한 결과통합은 학습단계와 실제 질의처리 단계로 나누어진다.

학습단계부터 살펴보면,

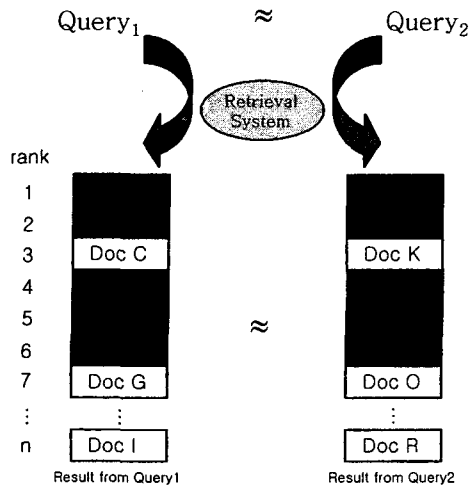

그림 1. 적합성 판정정보를 이용한 결과 통합 개념

학습질의는 벡터 방식의 색인단계를 거쳐 학습질의벡터를 구성한다. 학습질의를 지역 검색 서버에 넣어 나온 결과를 적합성 판정 정보와 비교하여 각 컬렉션 별, 학습질의별, 순위별로 판정 정보를 저장한다. 학습질의에 대한 모든 컬렉션의 문서의 적합 여부는 이미 판정되어 있어야 한다. 학습질의는 많은 수록 좋으며, 학습이 어느 정도 되었으면 실제 질의를 처리할 수 있다.

실제 질의 처리 단계에서는 사용자가 질의한 실제 질의를 각 지역 검색 서버에 넣는다. 또한, 실제 질의 역시 벡터를 구성하여 학습질의 벡터들과 비교한다. 실제 질의와 유사하다고 생각되는 미리 지정된 개수의 학습 질의를 선택하면 그 학습질의들에 대응하는 적합문서 판정 정보를 얻을 수 있다. 이 적합문서 판정 정보에 나온 순위, 컬렉션 별 적합여부를 그대로 적용하여 실제 질의가 각 원천 검색 서버로부터 얻어온 결과 문서들의 적합 여부를 판정하여 통합된 순위를 얻어낸다.

3. 분산 통합시스템 (LIRE) 구조

본 논문에서 설계한 분산 통합검색 시스템은 Learning-Based Distributed Information Retrieval Engine (LIRE)라 명명하였다. LIRE는 앞 절에서 설명한 학습방법을 사용하였으며, C++로 작성되어 있고 분산 검색 과정의 각 단계가 모듈화되어 있어서, 수정이 쉽고, 검색엔진 혹은 저장시스템의 변경이 용이하다. 현재는 학습질의의 색인 및 검색은 KRISTAL-2000 검색엔진[7]을 사용하였고, 적합성 판단은 문서 검색 컨퍼런스 TREC[8]에서 사용하는 검색결과 판정 유틸리티인 Trec_eval 을 수정하여 사용하였고, 적합성 판단 정보의 저장은 체계적인 관리를 위하여 데이터베이스 관리 라이브러리인 GDBM [9]을 사용하였다.

질의는 학습 질의인지 실제 질의인지 구분하여, 각각 필요한 모듈들로 돌어가며 , 원천검색 서버로의 검색은 multi-thread 방식으로 동시에 수행한다.

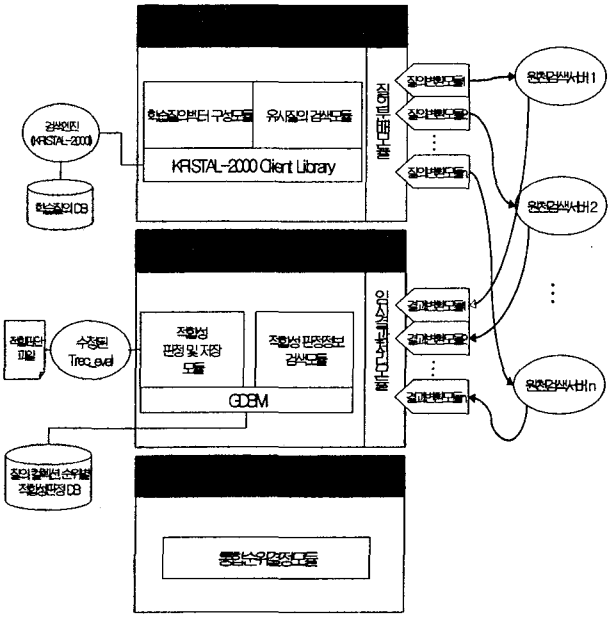

그림 2. LIRE 구조

학습단계는 학습질의가 들어오면 학습질의벡터 구성모듈에서 질의 벡터 구성을 위한 전처리 작업을 한다. KRISTAL2000 Client Library를 통하여 KRISTAL2000에서 색인을 한 후 DB에 넣는다. 질의분배 모듈과 질의변환 모듈을 거쳐 각 지역(원천)검색서버에 질의를 한다. 검색 결과로 나온 문서들은 결과변환 모듈과 임시 결과 처리 모듈을 거쳐 임시로 저장되며, 적합성 판정 모듈에서 적합성 판정이 이루어진다. 이때 적합성 판정 파일에서 각 문서의 적합여부를 입력받아 수정된 Trec_eval을 통하여 판정된다. 판정 정보는 GDBM을 거쳐 적합성 판정 DB에 들어가게 된다.

위와 같이 학습질의가 처리되며, 충분한 학습질의가 처리되면 실제 질

의를 넣게 된다.

실제 질의처리 단계는 실제 질의가 들어오면 먼저 유사질의 검색 모듈에서 KRISTAL2000 Client를 통하여 들어온 실제 질의와 유사한 학습 질의를 학습 질의DB에서 꺼내게 된다. 이것은 KRISTAL2000의 유사 문서 검색 기능을 통하여 이루어진다. 유사 학습질의의 ID는 우선 기억해두고, 실제질의는 질의 분배모듈과 질의 변한 모듈을 거쳐 지역(원천)검색 서버로 돌어간다. 검색 결과는 결과변환 모듈과 임시결과처리 모듈을 통하여 임시로 저장된다. 적합성 판정 정보검색 모듈에서는 기억된 유사 학습질의 ID에 해당되는 질의,컬렉션, 순위별 적합성 판정 정보를 GDBM을 통하여 DB에서 추출하며 이렇게 추출된 다수의 적합성 판정 정보의 순위별 적합 여부를 계산하여 통합 순위 결정모듈에서 통합 결과의 순위를 결정하게 된다.

- KRISTAL-2000

대용량 데이터의 분산검색 기능과 데이터 관리기능을 강화한 검색 엔진이다. 사용자의 입력 및 로드 밸런싱을 처리하는 Job Scheduler, 온라인 문서 관리를 처리하는 Data Manager, 실제 검색을 수행하는 Fire, 검색 결과의 Cache기능을 하는 Set Manager등 검색 엔진을 구성하는 중요 구성요소들이 독립적인 데몬 프로세스의 형태를 취하고 있어 별도의 시스템에 이식 가능한 형태로 설계되어있다. 또한, 데이터의 삽입, 수정, 삭제 등의 기능 수행 시 트랙잭션 처리와 회복 기법을 도입하여 제공함으로써, 보다 신뢰성 있는 데이터 관리 기능을 제공한다. 그리고 불리언과 벡터 모델 검색을 기본으로 제공하며, 유사 문서 검색등의 확장된 검색방법도 제공한다.

4. 결론

본 논문은 다양한 학습방법을 이용한 분산 검색 통합 결과의 성능 평가를 위한 테스트 베드 구축 설계에 그 목적이 있다. 학습방법으로는 []에서 제시한 적합성 판단 정보를 이용한 방법을 사용하였으며, 학습 질의의 색인 및 검색으로는 KRISTAL-2000을 사용하였다. 현재는 검색 결과 통합만 설계되어 있으나, 앞으로 컬렉션 선택 부분을 추가하여 사용자 질의에 적합한 컬렉션 선택, 질의 변환, 결과 통합등 분산 검색 시스템이 수행하는 모든 일을 처리할 수 있게 수정되어야 한다. 또한 KRISTAL-2000은 불리언 및 벡터 기반 검색 엔진으로서 다양한 파라미터와 검색 방법을 사용할 수 있게 설계되어 있으므로, 여러 종류의 원천 검색 서버로 변형하여 다양한 환경을 시뮬레이션 하기 적합하다. 앞으로 KRISTAL-2000을 다양하게 변형시켜 실제 분산환경과 유사한 조건에서 실험하는 과정이 필요하다.

[참고 문헌]

[1] James P. Callan, Zhihong Lu and W. Brice Croft, "Searching Distributed Collections With Inference Networks", ACM SIGIR '95, 1995.

[2] Nikolaus Walczuch, Norbert Fuhr, Michael Pollman and Birgit Sievers, "Routing and Ad-hoc Retrieval with TREC-3 Collection in a Distributed Looseley Federated Environment", TREC-3, 1994.

[3] Michail Salampasis and John Tait, "A link-based collection fusion strategy", Information Processing & Management, Vol.35, Issue 5, 1999

[4] Luis Gravano, Chen-Chuan K. Chang, Hector Garcia-Molina and Andreas Paepcke, "STARTS: Stanford Proposal for Internet Meta-Searching", ACM SIGMOD Record, Vol.26, No.2, Pages 207-218, 1997.

[5] Ellen M. Voorhees, Narenda K. Gupta, and Ben Johnson-Laird, "The Collection fusion problem", Proceeding of the 3rd Text Retrieval Conference (TREC-3). NIST Special Publication 500-225, 1994.

[6] 이민호, "복수문서 집합 검색결과의 효율적인 병합 방법", 한국어정보처리학회, 1999.

[7] KISTI , "KRISTAL2000 사용자 매뉴얼", 2002.

[8] TREC, http://trec.nist.gov/

[9] GDBM, http://www.gnu.org/directory/Database/Administration/gdbm.html