제20회 한국정보처리학회 추계학술발표대회 논문집 제10권 제2호 (2003. 11)

# KRISTAL2002 상의 검색 성능 비교 연구

김광영*, 진두석*, 최성필*, 서정현*
*한국과학기술정보 연구원
e-mail: kykim, dsjin, spchoi, jerryseo@kisti.re.kr

# A Comparative Study on KRISTAL2002 Retrieval Efficient

Kwang-Young Kim*, Du-Seok Jin*, Sung-Pil Choi*, Jerry-Seo*
*Group for Intelligent Information System, Korea Institute of Science and Technology Information

요 약

본 논문에서는 KRISTAL[1]2000을 성능 향상시킨 KRISTAL2002의 특징을 간략히 소개하고 KRISTAL2002와 성능을 비교 분석한다. 개선된 부분 중에서 KRISTAL2000 시스템의 단점이었든 하나의 데몬으로 하나의 테이블 밖에 지원하지 못하던 기능을 KRISTAL2002에서는 하나의 데몬이 여러 개의 테이블들을 검색 지원한다. 본 논문에서는 다중 테이블을 이용하여 KRISTAL2002의 성능을 실험하고 그 결과를 비교 분석하였다.

## 1. 서론

오늘날의 인터넷과 네트워크는 거대한 정보의 집합체로 바뀌고 있다. 널리 확산된 각 종 IT 인프라를 통해 웹, DB, 비정형문서 등 매년 새로이 생성되는 데이터는 전 세계적으로 1~2 exabyte (1exabyte = $10^{18}$)에 이르고, 인터넷 정보가 급증하고 있다, 만약 효율적인 Access 방법을 제공해주는 적절한 검색 엔진이 없다면 정보의 생산, 유통, 소비에 이르는 정보 사이클 자체가 불가능 하다. 또한 인터넷 사용자들은 원하는 정보를 정확하게 찾기가 점점 어려워질 것이다. 또한 검색 대상 문서의 수가 급격히 증가함에 따라 검색 결과 또한 상당한 양으로 사용자가 원하는 정보인지를 쉽게 판단하고 확인하기가 어렵다.[1]

KRISTAL2002 검색 시스템은 오늘날의 거대한 정보 집합체를 가공하여 사용자에게 정확한 정보 검색 결과 제공한다. 거대한 정보 집합체를 관리하는 관리자 측면에서도 KRISTAL2002에서는 관리도구(GUI)를 제공하고 있다.

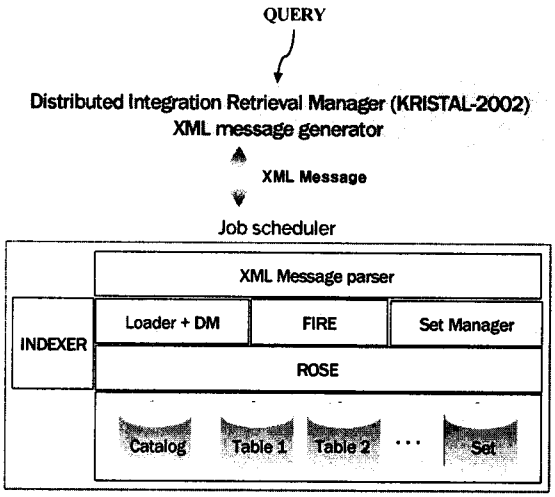
KRISTAL2002는 KRISTAL2000 보다 많은 성능 개선을 하고 있으며 또한 개선된 부분도 있다. 본 논문에서는 KRISTAL2002 개선된 부분인 멀티스키마지원 및 검색 방법을 소개하고 그 검색 성능을 실험을 통한 비교 분석 및 설명하고자 한다. KRISTAL2002 검색시스템은 다중 테이블 검색 지원을 위해서 Multi-Thread를 사용하여 사용자가 원하는 정보를 검색한다.

## 2. KRISTAL2002 시스템 구성도

KRISTAL2000에서 여러 개의 데몬을 사용하는 반면에 KRISTAL2002 시스템은 하나의 데몬으로

1) KRISTAL: IRMS

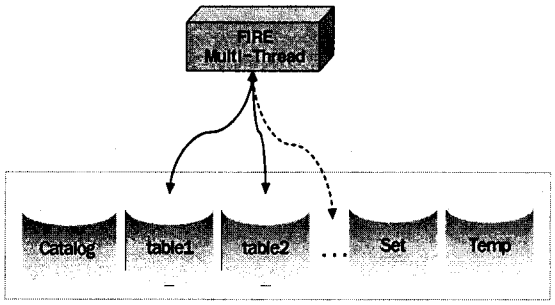
동작을 한다. KRISTAL2002 시스템은 [그림1]과 같이 크게 사용자 질의처리기, 검색기(FIRE), 저장관리기(ROSE), 색인기(INDEXER), 검색결과 관리기(Set Manager) 및 온라인/오프라인 적재기(DM/Loader)로 구성되어있다. KRISTAL2002에서는 Client/Server간의 메시지는 XML를 이용하여 통신을 처리하고 있다. KRISTAL2002는 멀티 스키마를 지원하고, 가변길이 섹션과 고정길이 섹션을 지원한다. KRISTAL2002에서는 관리도구(GUI)를 제공함으로서 DB생성에서 수정 및 편집 등의 모든 관리 기능을 제공한다. KRISTAL2002에서 포팅 및 차후 관리 등에서도 많이 개선되었다.




[그림 1 ] KRISTAL2002 시스템 구성도

본 논문에서는 KRISTAL2002의 개선된 많은 기능들 중에서도 멀티 테이블 검색 기능을 처리하는 검색기(FIRE)의 성능을 KRISTAL2000과 비교 분석하고자 한다.

[그림2]는 멀티 테이블 검색을 검색기(FIRE)에서 Multi-Thread를 이용하여 검색 처리함으로서 각 테이블 당 검색한 그 결과를 취합한다.




[그림 2] 멀티 테이블 검색

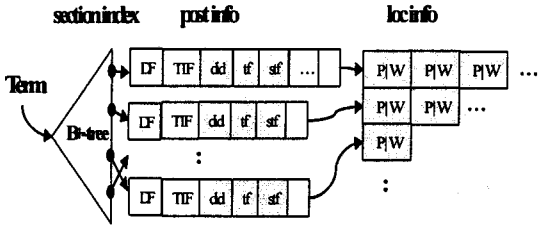
취합한 결과를 이용하여 문서의 가중치를 계산하고 문서의 랭킹을 처리하게 된다.

## 3. KRISTAL2002 포스팅 파일 구조

KRISTAL2002는 KRISTAL2000과 다른 포스팅 파일 구조를 사용한다. [그림3]과 같이 KRISTAL2002에서 가장 크게 다른 부분은 KRISTAL2002에서는 위치 정보를 따로 저장하는 구조를 사용하고 있다. KRISTAL2002에서는 사용자가 입력한 질의어를 질의 처리기의 최적화를 통하여 위치 정보사용 여부를 결정한다.

이러한 구조를 사용하는 가장 큰 이유는 검색 속도 및 검색 메모리 최적화를 향상시키기 위한 구조이다. 즉 위치 정보가 필요할 때만 위치 정보를 제공해 주기 위해서이다.

KRISTAL2002와 KRISTAL2000에서 근접도 연산을 사용할 때 위치 정보를 사용하고 있다. 또한 KRISTAL2002에서는 SumTF/MinTF/MaxTF 정보를 제공하여 준다. 위의 정보를 데이터베이스를 적재할 때 선택하여 적재를 할 수 있는 유연한 구조를 가지고 있다.




[그림 3] KRISTAL2002 Posting File 구조

포스트정보(postinfo)에는 문서ID, TF, STF(sumTF) 정보를 저장하고 있고, 위치정보(locinfo)에서는 WordNum 및 PsgNum 정보를 저장하고 있다.

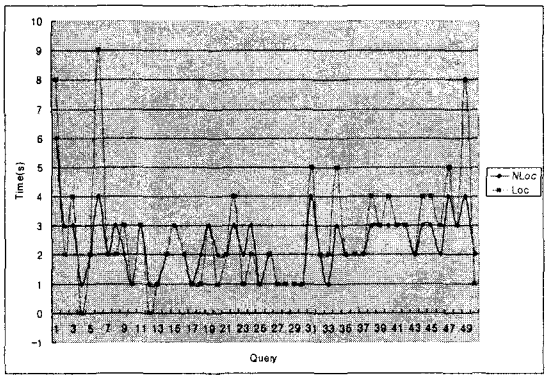
## 4. 실험

본 논문에서는 HANTEC V2.0을 이용하여 실험을 하였다. 이 컬렉션은 120,000 건 문서와 TREC 형식의 50 개 질의, 그리고 각 질의에 대해 적합성 정도에 의한 8종류의 적합문서 집합으로 구성되어 있다.[3] TREC 형식의 질의에서 <quer> 필드만을 사용하여 성능 평가 하였다. 평가는 G2.rel, L2.rel를 이용하여 50개의 질의에 대한 Average Precision값을 구하였다. KRISTAL2002에서는 8종류의 적합문서 집합을 각각 하나의 테이블로 구성하여 적재하였다. 반면 KRISTAL2000은 하나의 테이블에 8종류의 적합문서 집합을 적재하였다.

KRISTAL2002에서는 위치 정보를 자동으로 사용할 때와 사용하지 않을 때의 검색 속도를 측정하였다. 본 논문에서는 위치 정보에 따른 검색 속도를 측정하기 위해서 HANTEC V2.0에서 제공하는 <quer> 필드만을 사용하여 근접도 연산을 수행하였다. KRISTAL2002에서는 근접도 (WITHIN/NEAR) 연산에서 위치 정보를 사용하기 때문이다.[2] KRISTAL2000에서 검색 속도 측정은 가장 긴 위치 정보(WordNum)를 이용한다.[1]

[표 1] KRISTAL2002 성능 비교 분석

| 50개 질의 <quer> | KRISTAL2002 (8개 테이블) | | KRISTAL2000 (1개 테이블) | | 비고 |
|---|---|---|---|---|---|
| | 자동위치 정보사용 | 위치정보 사용 | | | |
| 속도(s) | 118(s) | 133(s) | 82(s) | | |
| 성능 | SumTF | | Max(WordNum) | | |
| | G2 | L2 | G2 | L2 | |
| | 0.1908 | 0.1749 | 0.173 | 0.1673 | |




[그림 4] 위치정보 비사용/사용에 따른 검색 속도

검색 속도 측면에서 보면 KRISTAL2000이 더 빠르게 나오는 이유는 다음과 같다. HANTEC V2.0을 이용하여 데이터베이스를 만들 때 KRISTAL2000에서는 하나의 테이블에 모두 적재된 것이고, KRISTAL2002에서는 HANTEC V2.0을 8종류의 적합문서 각각 테이블을 만들었다. 그러므로 테이블 여러 개를 열고 검색하였기 때문에 속도의 차가 발생했다. 실제 하나의 테이블로 구성하였을 때 속도 차이가 나지 않았다.

본 논문에서는 KRISTAL2002에서는 위치 정보 자동으로 선택하여 사용할 때와 하지 않을 때 15초 정도의 차이가 나타남을 볼 수가 있었다. 실제 복합명사 확장할 때만 근접도 연산이 반영된다.

KRISTAL2002에서의 검색 성능은 SumTF 이용하여 벡터 가중치에 정규화 처리를 한다.

KRISTAL2002에서의 벡터 가중치 및 정규화 식은 아래와 같다.

$$\text{Weight } Doc_j = \text{Weight}_{Vector}(Doc_j) / \log(\text{SumTF} + 1.0)$$

Weight $Doc_j$는 Term에 대한 가중치 계산인 Vector 모델로 가중치를 계산 처리한 것이다. [1]

KRISTAL2002에서는 기본적으로 SumTF를 이용한 정규화 방식을 사용한다. KRISTAL2000에서는

Max(WordNum)를 이용한 정규화 방식을 사용하고 있다.[1] KRISTAL2000에서는 검색 속도를 향상시키기 위해서 Max(WordNum)를 이용하여 검색 처리하고 있다.[1] 그러나 KRISTAL2002에서는 포스팅 파일 안에 정규화 값을 저장하고 있으므로 로 SumTF를 이용하여 정규화를 처리하는 방식을 사용하고 있다. 그러므로 KRISTAL2002가 KRISTAL2000보다 검색 속도 및 성능 면에서 우수함을 볼 수가 있다.

## 5. 결론

KRISTAL2002는 KRISTAL2000보다 검색 성능과 속도 측면에서 많은 향상을 가지고 왔다. 실험결과 검색 속도 측면에서 위치정보를 분리하고 질의처리기의 최적화를 통하여 자동적으로 위치정보를 선택할 수 있는 구조를 사용함으로 보다 빠르게 검색 결과를 제공해 주고 있다. 그러나 모든 질의어가 근접도 연산이 필요할 때는 위치 정보를 읽어서 처리해야 함으로 비슷한 속도를 보여 주고 있다.

KRISTAL2002는 KRISTAL2000보다 검색 성능 측면에서도 정규화를 SumTF를 이용하여 처리함으로서 보다 높은 검색 성능을 제공해 주고 있다.

## 참고문헌

[1] 김광영, 서정현, 이민호, 주원균, 정창후, 류범종 "위치 정보를 이용한 확장 벡터 문서 길이 정규화에 관한 연구", 정보처리학회 춘계 2003년 5월 p1623

[2] 김광영, 서정현, 최성필, "이웃한 어절의 위치정보를 이용한 KRISTAL2000 검색 성능 향상", 정보과학회 2001년 10월 p121~123

[3] 이석훈, 맹성현, 김지영 "정보 검색 평가체제 구축을 위한 HANTEC 테스트 컬렉션의 패키징" KOSTI2000 p31~48 Roger S. Pressman "Software Engineering A Practltiners' Approach" 3rd Ed. McGraw Hill