대용량 과학 기술 정보 처리를 위한 분산 시스템에 관한 연구

Study on Distributed System for Process of A Large Amount of Science Technology Information.

김광영, 강남규, 김진숙, 진두석, 정창후, 윤화묵
한국과학기술정보연구원

Kim Kwang-Young, Kang Nam-Gyu, Kim Jin-Suk, Jin Du-Seok, Jeong Chang-Hoo
Korea Institute of Science and Technology Information

요약

인터넷 기술이 급속한 발전함에 따라 인터넷은 대용량의 웹 문서, 과학기술 문서 및 DB 등으로 점점 더 복잡하게 구성되고 있다. 대용량의 웹 문서, 과학기술 문서 및 DB 등을 효율적으로 검색 및 관리 지원하기 위해서는 분산 시스템이 요구되어진다. 이러한 분산 시스템은 사용자에게는 대용량의 정보를 보다 빠르고 정확하게 검색을 지원해야한다. 또한 분산 시스템은 관리자 측면에서도 관리가 보다 용이해야한다.

본 논문에서는 과학기술정보 문서를 효율적 검색과 관리를 목적으로 시스템을 설계 및 구현하였다. 본 논문에서는 대용량의 과학기술 정보 시스템을 이용하여 구현된 분산 시스템의 검색 성능을 실험하고 그 결과를 비교 분석하였다.

Abstract

With the development of internet technologies, internet has been more complexly consisted of a large amount of Web documents, science technology documents, data-base and etc. distributed system is required to support effective retrieval and management about a large amount of Web documents, science technology documents and etc. distributed system has to support for user to search quickly and exactly. distributed system has to support for manager to manage easy.

This paper designed and made distributed system. these system effectively manages the science technology information documents. this paper made an experiment using a large mount of science technology information system. this paper analyzed the result of experimenting.

I. 서 론

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]

본문에서는 대용량의 과학기술관련 문서(현 KISTI에서 서비스 중)를 다양하게 분산하여 그 시스템의 구성 및 성능을 고찰하고자 한다.

분산 시스템은 현재 KRISTAL[1]에서 개발한 [2]dKRISTAL을 이용하여 그 실험을 측정하였다.

II. 분산 시스템 구성

본문은 분산 시스템을 실험 하기위해서 최소 3대의 서버에서 최대 8대 서버를 이용하여 실험을 하였다.

첫 번째는 최소의 서버 개수에서 색인 시스템의 테이블을 다양하게 분산하여 그 시스템을 구성하고 성능을 고찰하였다. 두 번째는 서버 개수 최대한 8대로 분산하여 대용량의 과학기술문헌 정보를 분산 색인하여 그 시스템을 구성하고 그 성능을 고찰하였다.

1. 다중 테이블 분산 시스템 구성

본 논문에서는 분산검색 시스템인 dKRISTA의 검색 속도를 측정하여 그 성능의 지표를 제시하고자 한다. 사용한 데이터베이스는 KISTI에서 서비스하고 있는 대용량 DB인 과학

1) Information Retrieval Management System

2) Distributed KRISTAL

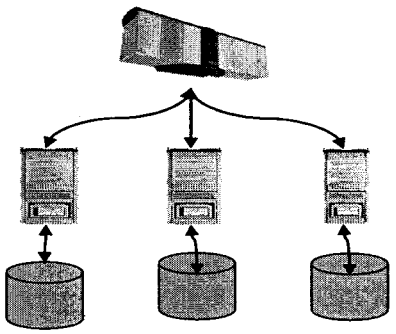
기술문헌 정보의 일부 (약 2700만 건)를 사용하였다. 이 실험의 주 목적은 대용량 DB를 하나의 서버에서 검색하는 속도와 여러 서버로 분산 시켜 놓고 dKRISTAL를 사용하여 검색하는 속도를 비교하였다.

[표 1] 다중 테이블 분산 시스템 사양

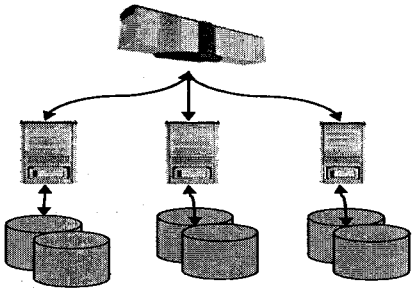
구분	사양
운영체제	Linux (RedHat 9.0)
중앙처리장치 (CPU)	Intel Pentium-IV Xeon 2.8GHz (Dual)
메모리(RAM)	8GB , 4GB, 2GB
저장장치(HDD)	SCA-type SCSI HDD (300GB)

시스템 운영체제는 리눅스를 사용하고, 메모리는 각각 8G,4G,2G인 3대의 서버로 구성을 했다.

그림 1은 각 3대에 서버에 2700만 건 문서를 3개로 나누어서 적재를 한 것이다. 그림 2는 각 서버에 2개의 Volume으로 나누어서 구성한 것이다.



▶▶ 그림 1. 3개분할(3개 서버)



▶▶ 그림 2. 6개분할(3개 서버)

2. 다중 서버 분산 시스템 구성

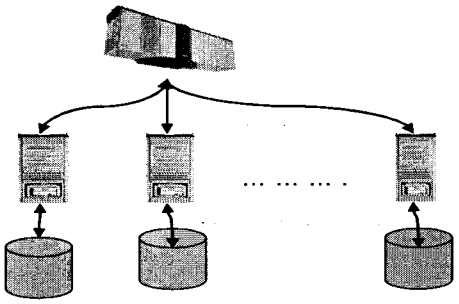
다중 서버 분산 시스템은 총 8대의 서버를 이용하여 KISTI에서 서비스하고 있는 대용량 DB인 과학기술문헌 정보의 일부 (약 5000만 건)를 사용하였다. 이 실험의 주 목적은 대용량 DB를 많은 서버에 분산하여 dKRISTAL를 사용하여 검색하는 속도를 비교 분석한다.

[표 2] 다중 서버 분산 시스템 사양

구분	사양
운영체제	Linux (RedHat 9.0)
중앙처리장치 (CPU)	AMD 2GHz (4CPU)
메모리(RAM)	8GB
저장장치(HDD)	SCA-type SCSI HDD (300GB)

시스템 운영체제는 리눅스를 사용하고, 메모리는 각각 8G인 8대의 서버로 구성을 했다.

그림 3은 8대의 서버에 DB를 나누어 구성한 시스템이다. 본 논문에서 중점적으로 실험한 시스템으로 최대한 DB를 관리하기 용이하도록 설계를 하고 또한 검색 처리 속도도 빠르게 할 수 있도록 하기위한 시스템을 구성한 것이다.



▶▶ 그림 3. 8개 서버로 분산

표 3은 그림 3의 시스템의 Volume를 구성도 기술한 것이다.

[표 4] DB 현황

	N-System	O-System
DB	해외학술지/해외회의OA	논문
	중국학술지	
	INSP	
	FSTA	
	국내학위논문	
	국내학술지/회의자료	
	특허	특허
	국내/국외 연구보고서	국내/국외 연구보고서

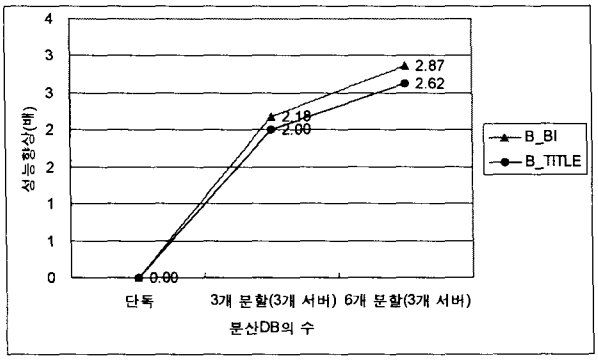
N-System은 각각 8대의 서버로 DB를 분산한 시스템이다. 반면 O-System은 3대의 서버로 분산한 시스템이다. 즉 O-System은 해외학술지, 중국, INSP, FSTA, 국내학위, 국내 학술지 등을 한 서버에서 서비스하는 시스템이다.

III. 분산 시스템 실험

1. 다중 테이블 분산 시스템 실험

본문에서는 첫 번째로 테이블을 분산하여 실험을 하였다. 실험에 사용되는 질의어는 일반 사용자들이 검색한 질의어 History를 이용하여 검색 테스트를 하였다. 실험 검색 방법은 불리언 모델을 이용하여 최대한 recall을 높여서 검색을 하였다.

그림 1과2와 같이 구성한 시스템에 대해서는 실험 결과는 그림 4와 같은 결과를 측정하였다.



▶▶ 그림 4. 다중 테이블 분할 분산 시스템

그림 4에서는 B_TI(한글/영어 제목 섹션)는 한글/영어 제목 섹션을 대상으로 검색한 것이다. 그리고 B_BI는 위 B_TI와 ABS(초록), KEYWORD를 통합한 섹션으로 검색한 것이다.

그림 4에서 볼 수 있듯이 3개로 분할하였을 때는 분할하지 않은 서버 보다 2배 이상의 검색 속도 향상을 가지고 왔다. 하지만 6개로 분할 할 때는 단독에서 3개분할 할 때와 같은 성능을 보여 주지 못하고 있다.

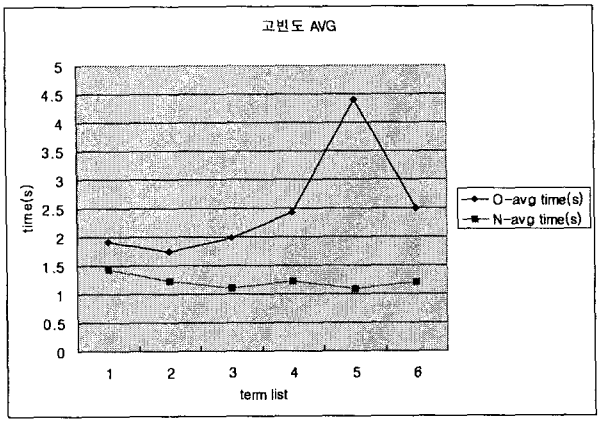
즉 한 서버에 많이 테이블을 분산하여도 한 서버가 처리할 수 있는 I/O 한계를 나타낸다고 할 수 있다. 계속 테이블을 나누어도 검색 속도 향상은 어느 수준에서 더 이상 증가하지 않은 것을 볼 수가 있다.

2. 다중 서버 분산 시스템 구성

본문에서는 다중 서버로 분산 시스템을 구성하였다. O-System과 N-System은 <표 3>과 같이 구성된 시스템이다.

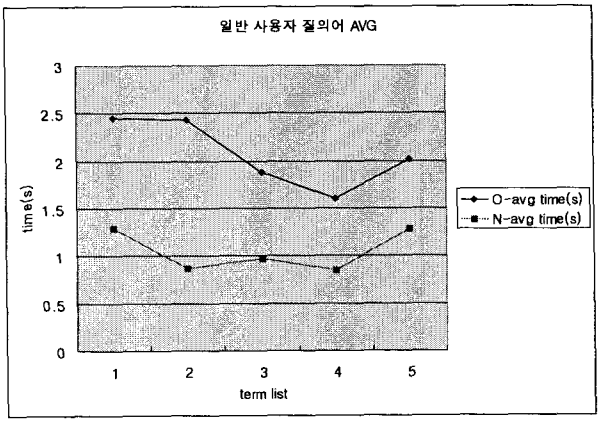
고빈도 [3] 질의어를 이용하여 각각 100개의 단어를 선정하여 검색 한 결과를 그림 5와 같이 측정되었다.

그림 5에서 볼 수 있듯이 3개의 서버로 분산된 시스템과 8개로 분산된 시스템의 검색 성능은 최대 4배 이상 차이가 난다. 그림 5에서 5번째는 O-System 4.4초, N-System은 1.08초로 최대 4배 빠른 결과를 나타낸다. 하지만 그림 5와 같이 측정된 결과를 보면 8개 서버로 분산된 시스템은 검색 성능이 일정하게 유지된다는 것을 볼 수가 있다.



▶▶ 그림 5. 고빈도 단어를 이용한 평균값 측정

그림 6은 일반 사용자들이 검색한 질의어들을 이용하야 검색한 결과이다.



▶▶ 그림 6. 일반 사용자 질의어 평균값 측정

일반 사용자들의 질의어 History를 이용하여 측정한 결과는 8대 서버로 분산하여 검색 서버가 3개도 분산한 서버에 비해서 최대 2배 이상의 성능을 보여 주었다.

IV. 결 론

본문에서는 한 서버에 다중으로 테이블을 분산한 시스템의 검색 성능을 측정하였고 그 결과 최대 2배의 검색 성능을 보여주지만 계속 테이블을 나누어도 증가 되지 않은 것을 볼 수가 있다.

다중 서버 분산 서버 실험 결과에서는 최대 8대로 서버를 분산 했을 때는 그 검색 성능이 일정하게 유지 되는 것을 볼 수가 있었다. 특히 그림 5와 같은 고빈도 단어들에 대해서 검색

3) DF가 10,000이상인 질의어

성능이 최대 4배까지 차이가 나는 것을 볼 수가 있었다. 일반 사용자들의 검색 질의어를 이용한 실험 결과로는 보통 2배 정도의 검색 성능 향상을 볼 수가 있었다.

▌참 고 문 헌▐

[1] 김광영 "단어의 위치 정보를 이용한 정보 검색 시스템 가중치 실험", KOSTI, 제10권, 제2호, pp.9-14, 2005..

[2] GIIS "dKRISTAL-2002 분산검색 시스템 속도 평가", GIIS Technical Report 20040402