사용자 중심 검색 시스템 설계 및 구현

김아용 · 이용우 · 배근호 · 정대진 · 정회경

배재대학교 컴퓨터공학과

Search for a user-centered system design and implementation

A-Yong Kim · Dae-Jin Jeong · Man-Seub Park · Jong-Moon Kim · Hoe-kyung Jung

Department of Computer Engineering, PaiChai University

E-mail : janlssary@pcu.ac.kr, leeeyongwoo@naver.com, orange0626@commu.co.kr,

jdj@realtimetech.co.kr, hkjung@pcu.ac.kr

요 약

최근 IT기술의 발전과 더불어 정보화에 대한 기술들이 이슈화 되고 있다. 웹을 사용하는 사용자들을 개인들이 필요한 정보를 찾는데 있어 검색데이터를 선별하는 방법에 대해 많은 어려움을 겪고 있다.

본 논문에서는 사용자 중심 검색 시스템을 제안한다. 제안하는 검색 시스템은 아파치 프로젝트인 Lucene과 Hadoop의 MapReduce, HDFS, Nutch, Solr를 활용하여 설계 및 구현한다.

이는 웹 검색을 이용하고자 하는 사용자의 의도에 따라 데이터를 수집하고 색인하여 원하는 정보를 제공하는 검색분야에 활용될 것이다.

ABSTRACT

addition to the advances in information technology and the latest IT technology for their issue. To enable users who are using the Web to find need the information your search data they're sifting through about how many are struggling.

In this paper, we propose a user-centered search system. Lucene search system to offer Hadoop's MapReduce with the Apache project Nutch, Solr, HDFS, utilizing design and implementation.

This is the Web search users who wish to use depending on the intentions of the data that you want to collect and index information will be utilized in the search field.

키워드

Hadoop, Lucene, Nutch, Solr, Ubuntu

Ⅰ. 서 론

IT 정보 기술이 발전하여 데이터양이 급격하게 증가하고 있다. 이로 인해 우리는 정보의 홍수시대에 살고 있으며, 수많은 정보들이 접할 수 있다. 하지만, 수많은 정보들 중에 정작 필요한 정보를 쉽게 찾을 수 없어 정보의 빈곤을 겪기도 한다. 인터넷 검색 엔진은 이러한 이유로 많은 사람들이 사용하고 있다. 하지만 인터넷 검색 엔진들은 공용 검색 엔진으로 내부 인트라넷에 있는 데이터나 특정 분야를 선택해서 검색하기에는 어려움이 따른다. 내부 인트라넷에 있는 데이터를 검색하기 위해서는 자체적으로 검색 엔진을 개발하여 사용하거나 유료 검색 엔진을 구매해서 사용해야 했었지만 Lucene 프로젝트가 시작되면서 검색 엔진 개발의 진입 장벽은 낮아짐으로 인해 개발자들이 검색 엔진을 쉽게 개발할 수 있게 되었다. Lucene[1]은 내부 네트워크 검색뿐만 아니라 외부 네트워크 검색도 할 수 있으며 Lucene 프로젝트에서 파생된 Nutch나 Solr 프레임워크를 사용하여 검색 엔진을 쉽게 구축할 수 있게 되었다[2].

본 논문에서는 오픈 소스와 오픈 소스 프레임워크를 활용하여 Hadoop 환경에서 동작하는 검색 엔진[3]을 설계하고 구현한다.

Ⅱ. 관련 연구

더그 컷팅에 의해 시작된 Lucene 프로젝트는 기존의 검색 엔진을 대체하기 위해 시작되었으며, 아파치 재단과 많은 개발자들의 참여로 인해 Lucene 프로젝트에서 파생된 Hadoop과 Nutch, Solr라는 프레임워크들도 개발하였다.

2.1 Hadoop

Hadoop은 대용량의 데이터 처리와 저장하기 위해서 자바로 개발된 오픈소스 프레임워크이며 Google에서 개발하고 검색 엔진으로 사용하고 있던 GFS(Google File System)와 MapReduce에 영감을 얻어 구현되었다. Hadoop의 주요 구성요소는 분산 파일 시스템인 HDFS(Hadoop Distributed File System)와 분산 처리 시스템인 MapReduce로 구성되어 있다. 기존의 Hadoop은 4000 노드 이상의 큰 클러스터에서 발생하던 병목현상을 개선하기 위해서 Hadoop의 다음 세대인 YARN(Yet Another Resource Negotiator)이 개발되었다[4].

2.2 Nutch

Nutch[5]는 Lucene 기반으로 개발된 오픈 소스 검색 엔진이며, 확장 가능한 웹 크롤러와 검색 엔진을 구축하기 위한 프레임 워크이다. 대량의 데이터를 크롤링 할 수 있으며 크롤링 작업은 인터넷과 인트라넷에서 수행할 수 있으며 MapReduce를 사용하여 단일 머신 또는 여러 대의 머신에서 분산 처리할 수 있다. 또한, Nutch는 Solr와 통합하여 사용할 수 있다.

2.3 Solr

Solr[6]는 Lucene 기반으로 개발된 오픈 소스 검색 엔진이며, 주요 특징은 강력한 텍스트 검색, 실시간 인덱싱, 동적 클러스터링, 데이터베이스 통합, 다양한 문서 처리 및 지리 정보 검색 등을 지원한다. Solr는 높은 신뢰성과 확장성, 내결함성, 인덱싱, 복제 및 배포를 제공하고 분산 처리와 장애 자동 복구 등을 제공한다.

Ⅲ. 시스템 설계

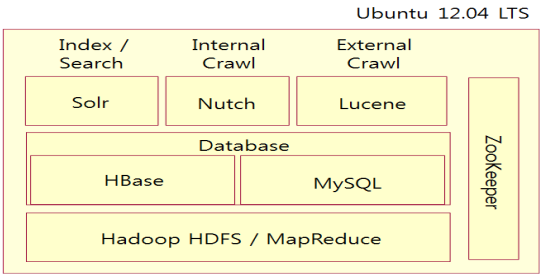

그림 1. 사용자 중심 검색 시스템 구조도

사용자 중심 검색 시스템 구조는 그림 1과 같으며, HDFS와 MapReduce 기반으로 크롤링 기능과 인덱싱 기능이 동작한다. 사용되는 데이터베이스는 MySQL과 NoSQL인 HBase를 사용한다.

Nutch 크롤링은 외부 검색이 가능하지만 사용자 중심 검색 시스템에서는 내부 검색 위주로 동작하게 설정하고, 외부 검색은 Lucene기반으로 구현된 크롤러와 특정한 주제의 문서만을 수집하기 위해서 Topical 기반으로 구현된 오픈 소스 크롤러를 사용한다. 인덱싱과 검색 기능을 Solr 기반으로 사용하며, 동작하는 흐름도는 그림 2로 표현했다.

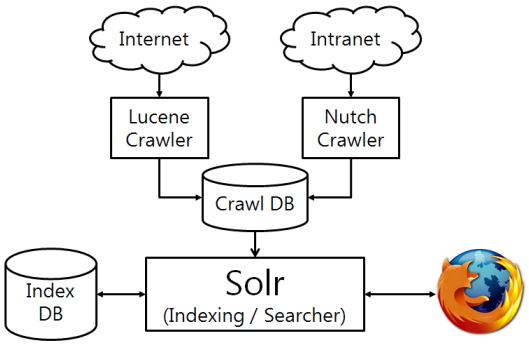

그림 2. 사용자 중심 검색 시스템 흐름도

인덱싱이 완료된 데이터는 웹 브라우저를 사용하여 원하는 데이터를 검색할 수 있다.

Ⅳ. 시스템 구현

크롤링과 인덱싱을 분산 처리하기 위해서 사용된 하드웨어는 Namenode 1대, Second Namenode 1대, Datanode 3대, Hub 1대이며, 사양은 표 1과 같다.

표 1. 사용자 중심 검색 시스템 하드웨어 사양

Namenode	
CPU	Intel i5-2400
RAM	4 GigaBytes
Second Namenode	
CPU	Intel i3-2120
RAM	4 GigaBytes
Datanode	
CPU	intel E6600
RAM	3 GigaBytes
Network Hub	
IPTime H5008	

사용자 중심 검색 엔진은 리눅스 기반으로 동작하며 자바 빌드 도구인 Ant와 Eclipse를 사용하여 세부 사항을 구현하였다. 사용된 주요 소프트웨어의 사양은 표 2와 같다.

표 2. 사용자 중심 검색 시스템 소프트웨어 사양

Operating System	Ubuntu 12.04 LTS
Hadoop	2.2.0
Nutch	2.2.1
Solr	1.4

구현한 웹 브라우저 인터페이스는 그림 3과 같으며, 특정 주제에 대한 크롤링 기능은 자바 오픈소스인 crawler4j와 WebSPHINX를 수정하여 구현했다.

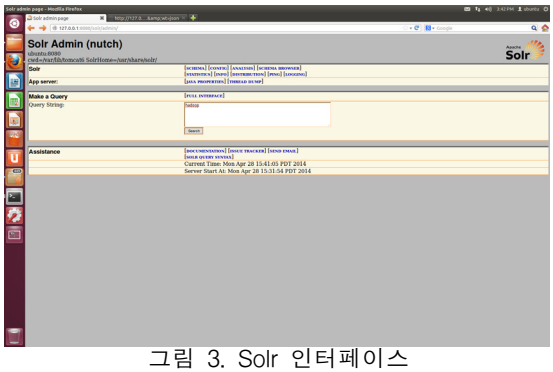

그림 3. Solr 인터페이스

V. 결 론

오픈 소스를 사용하여 구현한 검색 엔진은 인터넷과 인트라넷에 있는 데이터를 검색할 수 있게 해주며, 특정 분야만을 선택해서 크롤링할 수도 있다. 또한, 인덱싱을 분산 처리하여 대량의 데이터를 처리할 수 있다. 그러나 본 논문에서 제안하는 검색 엔진은 공용 검색 엔진에 비해 편의성과 친밀성은 저하되는 문제점을 지니고 있다.

향후 연구는 사용자가 지정한 분야에 대한 데이터들을 크롤링하는 기능을 개선하고, Google에서 제공하는 알리미 기능과 관리자에게 제공되는 메시지 알림 기능에 개발하여 검색엔진의 최적화와 관리자의 편의성, 사용자의 친밀성 등을 개선하는 연구가 필요하다.

참고 문헌

[1] Butler, Mark H, James Rutherford, "Distributed Lucene: A distributed free text index for Hadoop," HP Laboratories, 2008

[2] Maciolek, Przemyslaw, Grzegorz Dobrowolski, "Cluo: Web-scale text mining system for open source intelligence purposes," Computer Science, Vol.14, No.1, pp.45-62, 2013

[3] McCreadie, Richard, Craig Macdonald, Iadh Ounis, "MapReduce indexing strategies: Studying scalability and efficiency," Information Processing & Management, Vol.48, No.5, pp.873-888, 2012

[4] Vavilapalli, Vinod Kumar, et al. "Apache hadoop yarn: Yet another resource negotiator," Proceedings of the 4th annual Symposium on Cloud Computing, 2013

[5] Dlugolinsky, Stefan, et al. "Distributed web-scale infrastructure for crawling, indexing and search with semantic support," Computer Science, Vol.13, No.4, pp.5-19, 2012

[6] https://lucene.apache.org/solr/ 2014.4