인터넷 검색결과 관리 시스템

양정훈, *정형구, 신미령, 김상철
한국외국대학교 컴퓨터공학과
e-mail : arocer@naver.com, shenmeiling13@hanmail.net, kimsa@hufs.ac.kr

A Data Management System for Internet Search Results

Jung Hoon Yang, Hyung Ku Chung, Mei Ling Shen, Sangchul Kim
Department of Computer Sci. & Eng, Hankuk University of Foreign Studies

Abstract

On the Internet we are confronted with a huge amount of information in daily lives. However, we do not have system support for storing and managing Internet search results useful for us. To solve this problem, this paper presents a data management system in which search results obtained from wikipedia and a web search engine are organized in a structured way, enabling an easy access to them later.

I. 서론

디지털 콘텐츠 시장의 성장에 따라 정보량의 증가세는 계속될 것으로 전망하고 있다. 특히, 인터넷 환경이 웹2.0으로 진화하면서 네티즌의 역할이 과거 구경꾼에서 적극적인 참여자로 변화하면서 인터넷상의 정보의 양도 급증하고 있다. 많은 사람들이 거의 매일 인터넷 검색을 통해서 원하는 정보나 지식을 획득하고 있다. 인터넷 검색의 결과 중에서 보관하고 싶은 자료를 관리하는 대표적인 방법으로는 즐겨찾기에 추가하기, 파일로 저장하기 등이 있다. 이와 같은 방법은 자료 정리에 한계가 있고 특히 같은 범주의 자료를 그룹화하여 체계적으로 정리하기가 쉽지 않다.

본 논문에서는 인터넷 검색 결과를 계층적으로 정리하는 기능을 제공하는 시스템을 제안한다. 우리들이 인터넷상에서 주제어 검색을 위해서 가장 많이 방문하는 곳은 위키피디아 [1]와 네이버 [2]와 같은 웹 검색 엔진이다. 우리의 시스템은 사용자가 원하는 주제어를 입력하면, 위키피디아와 웹 검색 엔진을 통해서 정보를 검색하고, 그 결과를 트리 형태로 분류하여 제시하고, 추가로 사용자의 의도에 맞추어 새로운 자료의 삽입, 삭제, 변경을 할 수 있는 기능을 제공한다.

기존 연구에서 지식관리시스템 [3]과 웹 검색 결과의 클러스터링 연구 [4, 5]가 본 연구와 관련된다. 일반적으로 지식관리시스템은 전문지식이나 경험지식을 수집하고 공유하는 용도를 위한 것으로서 웹 검색 결과를 위한 기능은 부족하다. 웹 검색 엔진의 결과를 클러스터링 기능은 우리 시스템에서도 채택되어 활용되고 있다.

II. 시스템 설계

본 논문에서 제안하는 시스템의 전반적인 구조는 그림 1과 같다. 사용자가 제시한 주제어를 이용해서 위키피디아 백과사전을 검색하여 그 결과를 획득한다. 위키피디아 검색의 결과는 XHTML 형식으로 되어 있으며 목차와 본문으로 정리되어 있다. 검색 결과는 필터링이라는 과정을 통함으로써 사용자가 원하지 않는 자료들을 검색 결과에서 제거한다. 사용자는 본인이 원하지 않은 주제어를 명시하거나 목차 항목을 필터로 명시할 수 있다.

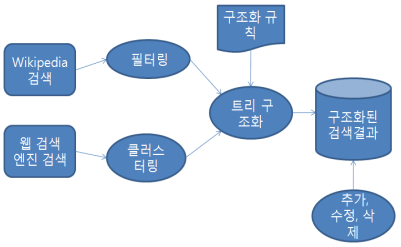

그림 1. 시스템 구조

최근 웹 검색 엔진은 외부 응용서비스에서 검색 엔진의 결과물을 사용할 수 있도록 하기 위해서 Open

API를 제공하고 있다. 사용자가 제시한 주제어에 대해서 웹 검색 엔진의 API를 이용해서 결과를 수집한 후, 클러스터링 작업을 통해서 유사 문서들을 군집화하게 된다. 본 시스템에는 기존 클러스터링 알고리즘 [4, 5]을 사용자에게 제공하고 이들 중 하나를 선택할 수 있도록 한다.

인터넷 검색결과를 후처리 (즉, 필터링이나 클러스터링) 한 후, 이들을 한 개 트리로 하나로 합치는 트리 구조화 작업을 거치게 된다. 트리 구조화 방법은 크게 두가자로서, 사용자는 이중 하나를 선택할 수 있다.

● 하나는 클러스터링 결과를 위키피디아 검색 결과에 합치는 것이다. 예를 들어 설명하면, 특정 클러스트 J가 목차 항목 I과 가장 유사하면, J는 I의 하위 항목으로 삽입된다. 문서 항목 I와 클러스트 J 간의 유사도 S(I, J)는 다음 식으로 정의된다. S_I는 I에 속한 자료 집합, S_J는 J에 속한 자료 집합, N_I는 S_I의 크기, N_J는 S_J의 크기, SF(s,t)는 자료 s와 t간의 유사도 함수를 나타낸다.

$$S(I,J) = \frac{\sum_{s \in S_I, t \in S_J} SF(s,t)}{N_I * N_J}$$

● 다른 하나는 위키피디아 검색 결과를 클러스터링 결과에 합치는 것이다. 각 클러스트를 첫 번째 레벨의 목차 항목으로 삽입한다.

트리 구조화 과정을 거친 결과는 사용자가 수작업으로 새로운 항목의 추가, 삭제, 위치 변경 등을 하여, 사용자의 의도에 맞추어 조정할 수 있다.

위키피디아는 전 세계의 사용자들이 추가 및 갱신을 계속하는 백과사전이다. 따라서, 본 시스템에서는 자동 갱신 기능을 제공한다. 사용자가 주제어와 갱신 주기를 설정해 놓으면, 본 시스템의 자료 트리의 내용을 최신 상태로 유지할 수가 있다.

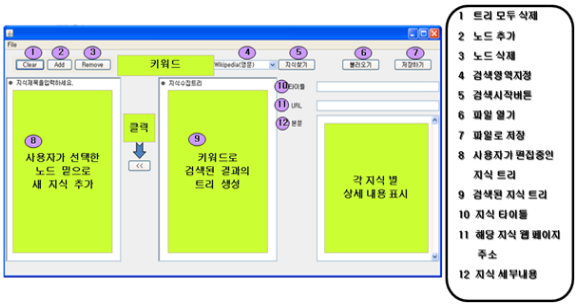

그림 2. 사용자 화면

그림 2는 사용자 화면의 구성과 사용자 기능을 정리한 것이다. 사용자가 자료 트리를 편집할 때, 편의를 위해서 임시 트리를 별도로 제공한다. 먼저 임시 트리를 원하는 트리를 완성한 후에 자료 트리에 한 개의 노드로 삽입하거나 기존 노드를 대치할 수 있다. 자료 트리를 바로 편집하면 인터넷 검색 결과와 편집한 내용이 구분이 되지 않은 경우가 자주 발생하는 어려움이 있기 때문이다.

Ⅲ. 구현

본 논문에서 제안한 시스템의 프로토타입을 자바 언어로 구현하였다. 웹 검색 엔진 API는 네이버의 도서 검색용 API를 사용하였다. 개발 환경은 자바 IDE로 널리 쓰여 지고 있는 Eclipse 3.3에 Jigloo 플러그인을 설치하여 개발하였다. 인터넷 검색 결과로 수집한 XML 코드의 내용을 분석하기 위해 JAXP (The Java™ API For XML Processing)을 이용하였고, 사용자 인터페이스 모듈은 Swing을 기반으로 하였다.

그림 3. 자료 트리 예제

Ⅳ. 결론

우리는 인터넷을 통해서 수많은 자료를 접하고, 본인에게 유용하다고 생각하는 것들은 파일 저장, 블로그에 스크랩 등의 방법으로 저장하여 관리하고 있다. 본 논문에서는 사용자에게 유용한 인터넷 검색 결과들을 보다 체계적으로 정리하고 관리하는 시스템을 기술하였다. 본 시스템은 위키피디아 검색 결과와 검색엔진 검색 결과 중에서 사용자가 원하는 자료들을 트리 형태로 표현한다. 또한, 사용자가 수작업을 통해서 자료 트리를 별도 자료를 추가하거나 수정하여 본인의 의도대로 개선할 수 있다.

참고문헌

[1] www.wikipedia.org
[2] www.naver.com
[3] A.D. Marwick, Knowledge Management Technology, IBM Systems Journal, Knowledge Management, Vol 40, No 4, 2001.
[4] J. J. Rocchio, Document Retrieval System: optimization and evaluation, Ph.D. Thesis, Havard Univ., 1966.
[5] D.R. Hill, A vector clustering technique, in:Samuelson, Mechanized Information Storage, Retrieval and Dissemination, North-Holland, Amsterdam, 1968.