인터넷 문서의 자동분류 서비스 시스템에 관한 구현

A Structure on Classification Service System of Internet Documents

황성하, 최광남, 이대규, 이상호
한국과학기술정보연구원

Hwang Sung-Ha, Choi Kwang-Nam, Lee Dae-Kyu, Lee Sang-Ho
Korean Institute of Science and Technology Information

요약

인터넷 정보를 검색하고 활용하는 것은 쉽고도 어려운 일이다. 많은 정보 중에서 원하는 정보를 얻기 위한 노력은 단순히 검색뿐만 아니라 정보의 수집에서 분류 및 가공, 활용에까지 각 분야별로 그 범위와 용도에서 다양한 기술의 발전이 급속히 진행되고 있다. 특히, 이러한 발전은 다양한 용도의 에이전트와 분류, 변환 등의 가공 기술에서 더욱 두드러지게 나타나고 있다. 또한, 시스템의 자동화를 통한 편리성을 제공한다면 더욱 효과적인 정보관리가 이루어 질 것이다. 본 논문에서는 이러한 배경에서 인터넷 정보의 수집에서 자동 분류, 검색 서비스까지를 하나의 시스템에서 처리 할 수 있는 인터넷 문서 자동분류 서비스 시스템을 소개한다.

Abstract

Using for the internet information is easy or difficult. The effort to obtain the useful information is developed the various technique such as search as well as the information repository, classification, processing and the utilization. Specially, such developments are remarkable to the Agent of various uses and the classification, conversion in processing techniques. The study introduces the classification service system of internet documents which is processing from the repository of internet information to the automatic classification and search service.

I. 서 론

최근 인터넷은 급속도로 발전해 나가고 있다. 예컨대, 매일 평균 20억 이상의 웹 문서가 증가하고 있으며 우리는 다양한 정보를 인터넷상에서 수많은 HTML 문서 등을 접할 수 있게 되었다. 인터넷을 통해 정보를 검색하고 활용하는 것은 현대를 살아가는 평범한 일거리가 되었다. 또한, 정보의 홍수 속에서 보다 낳은 정보를 얻기 위한 노력은 지금도 계속되고 있다. 그러나 그 수많은 문서들을 일일이 찾아다니면서 원하는 정보를 찾게 된다면 상당히 비효율적인 일이 될 것이다. 예컨대, 인터넷상에서 정보를 검색하는데 있어서 관련성이 없는 불필요한 정보들이 많이 검색되기도 하며, 검색 결과를 체계적으로 분류하고 조직화하는데 많은 문제점이 있다. 이러한 문제점을 해결하기 위한 노력의 결실로 에이전트, 정보 분류, 정보 검색 등의 기술 분야에서 많은 발전을 해 왔으며 각각 다양하게 활용되고 있다.

특히, 필요한 정보를 수집하여 관리하고 분류하여 검색할 수 있는 자동분류 서비스 시스템은 효과적이고 편리한 정보관리 방법을 제공 할 수 있는 각각의 필요한 요소기술이 통합된 시스템이다.

본 논문에서는 이러한 요소기술을 통합하여 구현한 인터넷 문서 자동분류 서비스 시스템을 소개한다. 2장에서는 관련연구를 통해 요소기술의 발전 동향을 살펴보고, 3장에서는 시스템의 통합구조 및 기능설계 내용을 살펴보며, 4장에서는 시스템의 구현과 기능의 동작을 검증하였다. 끝으로 5장에서는 향후 발전 방향을 제시하면서 결론을 맺는다.

Ⅱ. 관련연구

1. 에이전트

에이전트란 사용자를 대신하여 원하는 작업을 정해진 스케줄에 따라 인터넷 상에서 정보수집 등의 작업을 자동적으로 해결하여 주는 소프트웨어라고 할 수 있다[1].

에이전트는 1950년대 John McCarthy로부터 시작하여 Oliver G. Selfridge에 의해 정립되었으며 초기에는 인공지능 분야에서 연구가 진행되었으나, 80년대 부터 인공지능과 분리되어 독립적인 연구 주제로 대두되기 시작하였다. 이제 에이전트는 분산 컴퓨팅, 객체지향 시스템, 소프트웨어 공학, 인공지능, 경제학, 사회학 등의 여러 분야가 결합된 영역으로 발전하고 있다[2].

최근 연구되고 있는 에이전트의 종류로는 Mobile Computing분야에서 네트워크를 돌아다니면서 일을 수행하는 모바일 에이전트(Mobile Agents), 사용자가 시스템을 사용하기 편리하도록 지원하는 사용자 에이전트(Interface Agents), 많은 양의 정보를 찾아내어 분석하고 필터링하여 필요한 정보를 제공하는 정보 에이전트(Information Agents), 에이전트 간 통신을 이용하여 자율적으로 상호 일을 수행하는 협업 에이전트(Collaborative Agents), 다양한 일을 하는 서로 다른 구조를 가지는 에이전트 간 통신과 협력을 통하여 일을 수행하는 이질 에이전트(Heterogeneous Agents) 등이 있다[3][4].

본 논문에서 활용되는 에이전트는 정보 에이전트중의 하나인 인터넷 로봇으로 여러 곳에 흩어져 있는 인터넷 정보를 지능적으로 접근할 수 있는 기능을 제공하고 원하는 정보를 찾는 역할뿐 아니라 정보의 변화상황을 감시하는 모니터링도 수행한다.

2. 문서 자동 분류

문서 분류는 전자문서 관리시스템 (Electronic Docu- ment Management System, "EDMS")과 정보검색시스템(Information Retrieval System, "IRS")에서 비롯된다[7].

EDMS는 다양한 문서의 일관되고 체계적인 저장·관리와 단일 인터페이스를 통한 정보의 접근·공유를 통해 업무에 쉽게 활용할 수 있도록 하는 시스템으로 문서 분류 체계와 관리 방법이 중요한 기능이기도 하다. 또한, EDMS에서의 문서 분류는 대부분 등록에 의한 수동 작업으로 이루어지고 있으며, 이는 자동 분류를 위한 분류기준이 명확하지 않고 정확성이 떨어지기 때문이다. 따라서 문서의 분류체계 관리와 대·중·소 등의 단순한 분류에 그치고 있으며 대부분 조직체계의 기능 및 역할 기반에 중점을 두고 있다[4][7].

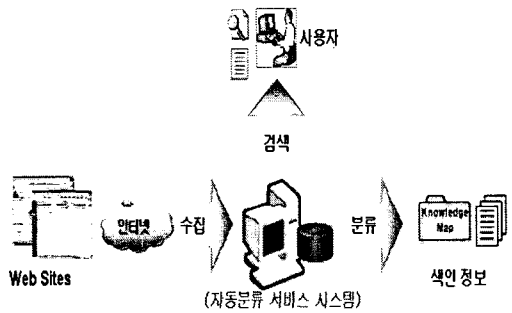
정보검색 분야에서의 IRS는 검색 속도와 결과의 정확성에 중점을 두고 있기 때문에 정확성을 높이기 위한 다양한 방법이 활발히 연구되고 있다. 특히, 영어를 기반으로 하는 솔루션은 상품화되어 활용되고 있으며 정확성이 높은 것으로 인정받고 있다. 그러나 한글의 경우에는 언어의 구조적 특성에 따른 처리 방법이 상이하여 적용하지 못하고 있으며, 최근 한글을 지원하는 검색엔진과 형태소 분석기의 발전으로 연구가 활발히 진행되고 있다. 특히, 학습기능을 도입하여 정확성을 높인 자동분류 시스템이 많이 연구되고 있다[5].

본 논문에서는 학습기능을 기반으로 하는 에이전트를 적용하여 인터넷 문서를 자동 분류하는 자동분류 시스템을 소개한다.

Ⅲ. 자동분류 서비스 시스템 구성 및 기능

1. 자동분류 서비스 시스템 개요

자동분류 서비스 시스템은 인터넷 수집로봇 에이전트를 통해 등록된 웹 사이트의 문서를 주기적으로 방문하여 수집하고 지식 카테고리 별로 수동 및 자동분류를 실행하여 사용자에게 다양한 지식 카테고리 검색 및 상세검색 서비스를 제공하며, E-mail 서비스를 원하는 사용자에게 Push Mail을 수행 등을 통합한 시스템이다. [그림 1]은 인터넷 문서의 수집에서 자동분류, 검색 서비스까지를 하나의 시스템에서 처리되는 과정을 나타낸다.

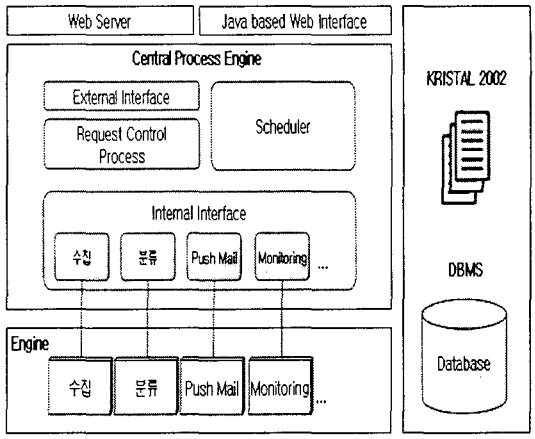


▶▶ 그림 1. 자동분류 서비스 시스템 개요

2. 시스템 구조 및 기능 설계

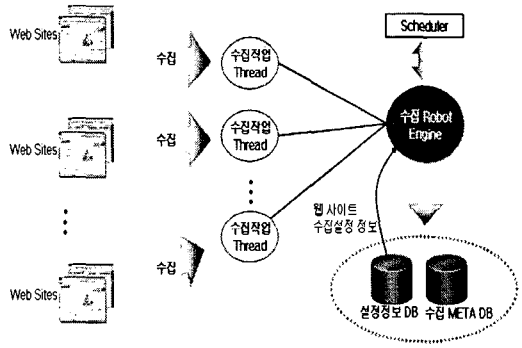
자동분류 서비스 시스템은 웹 문서 수집 및 자동분류 시스템으로 스케줄러에 의해 정해진 시간에 수집 Robot Engine이 자동 구동되어 문서를 수집하며 수집된 정보를 분류 Engine에 내장된 학습 알고리즘 및 텍스트 분석에 의한 주제어 분류 방식을 이용하여 검색엔진에 카테고리별로 분류하는 시스템이다. [그림 2]는 자동분류 서비스 시스템의 구성을 보여준다.

자동분류 서비스 시스템은 크게 4가지 기능으로 분류된다. 첫째, 수집 Robot Engine에 의한 정보수집 기능과 둘째, 분류 Engine에 의한 자동분류 기능 셋째, 분류 완료된 문서의 유효성 검사를 위한 Monitoring 기능 넷째, 사용자의 관심정보를 분류하여 검색결과를 전송해주는 Push Mailing 기능으로 분류된다.



▶▶ 그림 2. 자동분류 서비스 시스템의 구성도

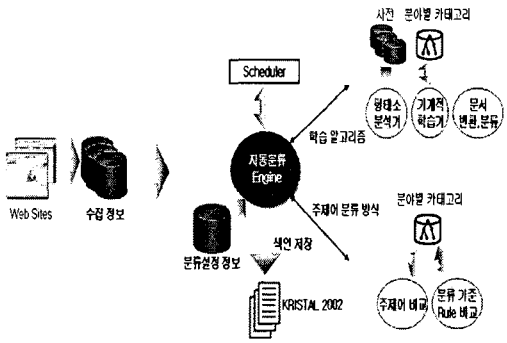
2.1 문서 수집



▶▶ 그림 3. 문서 수집

[그림 3]은 수집 Robot Engine에 의해 웹문서가 수집되는 과정을 나타내는 그림이다. 수집 Robot은 웹 사이트에서 문서 및 Meta 정보를 쓰레드(Thread)에 의한 수집 작업으로 Scheduler에 의해 수집한다. 이때, 설정정보 DB에 수집 설정 정보를 로딩(Loading)하여 수집된 정보를 수집 META DB에 저장한다.

2.2 문서 자동 분류

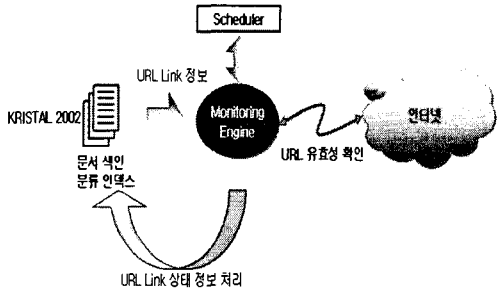


▶▶ 그림 4. 문서 자동 분류

[그림 4]는 수집된 정보를 분류 Engine에 의해 분류 및 색인되는 과정을 나타낸다. 분류 에이전트는 수집정보 DB의 문서를 형태소 분석기와 기계적 학습기 및 문서 변환, 분류를 이용하여 해당 분류에 할당하고 검색엔진(KRISTAL 2002)에 분류정보와 문서를 색인하는 기능을 한다. 분류 학습기에서는 추출된 형태소별로 각각의 기존분류 디렉토리에서의 중요도에 따른 가중치와 정확도를 계산하여 해당 분류 디렉토리에 할당하고 분류 디렉토리별 가중치를 재조정한다. 분류 Engine의 Scheduler에 의해 자동으로 분류되고 DB에서 지식 맵별 분류설정 정보를 로딩(Loading)한다. 또한, 수집된 웹 사이트 정보를 형태소 분석기, 기계적 학습기 및 문서 변환, 분류 등의 학습 알고리즘과 텍스트 분석에 의한 주제어 분류 방식으로 처리하고 분류된 정보를 검색엔진(KRISTAL 2002)에 색인하여 저장한다.

2.3 Monitoring

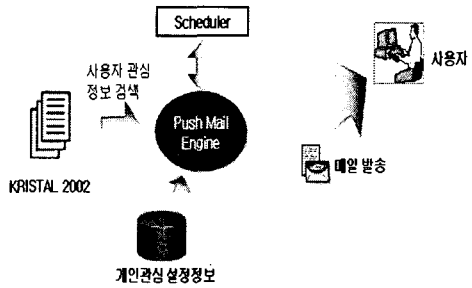
Monitoring은 Monitoring Engine의 Scheduler에 따라 분류완료 된 각 문서의 URL Link의 유효성 유무 정보를 확인하고 URL 해당 서버의 다운 및 Dead Link 시 설정된 Rule Loading 방법에 따라 상태정보 flag를 삭제한다. Monitoring 과정은 [그림 5]에서 보여준다.



▶▶ 그림 5. Monitoring

2.4 Push Mail

Push Mail 서비스를 신청한 사용자에 한하여 Push Mail Engine Scheduler에 의해 Mailing되며 사용자별로 설정된 메일 전송주기, 관심 지식 맵, 관심 키워드 등의 관심정보를 로딩(Loading)한다. 이를 통해 분류정보 내에서 검색하여, 검색 결과를 사용자에게 메일을 전송하는 방식으로 이는 [그림 6]에서 보여준다.



▶▶ 그림 6. Push Mail

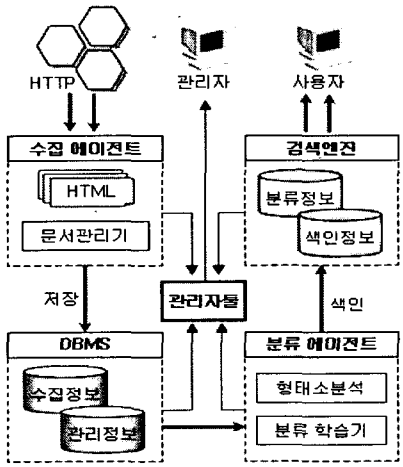
Ⅳ. 자동분류 서비스 시스템 구현

1. 자동분류 서비스 시스템 흐름

[그림 7]은 자동분류 서비스 시스템의 흐름을 보여주는 것으로 수집/분류 에이전트, DBMS와 검색엔진으로 구분된다.

수집 에이전트가 등록된 여러 개의 웹 사이트를 돌아다니며 문서를 수집해 오면 문서 관리기에 의해 중

복, 수정, 추가 등의 문서 상태를 체크하고 수집정보 DB에 저장한다.



▶▶ 그림 7. 자동분류 서비스 시스템 흐름도

만약, 수집문서가 이미 수집된 문서와 동일한 중복 문서일 경우에는 모니터링 관리 정보에 기록하고 마치며, 추가 또는 수정된 문서일 경우에는 수집정보 DB에 추가하고 분류 에이전트에 분류 해 줄 것을 통보한다. 반면, 문서가 삭제된 경우에는 분류 에이전트에 색인정보에서 삭제 할 것을 통보하고 분류 에이전트는 삭제 후, 해당 분류 디렉토리의 학습 값을 재조정한다. 따라서 수집 에이전트의 핵심은 문서의 중복성을 체크하는 작업이며, 실제로 이 부분에서 많은 처리시간이 요구된다.

분류 에이전트는 수집정보 DB의 문서를 형태소 분석기와 분류 학습기를 이용하여 해당 분류에 할당하고 검색엔진에 분류정보와 문서를 색인하는 기능을 한다. 수집된 문서는 형태소 분석기에 의해 각각의 형태소가 추출되고 빈도수가 계산된다. 분류 학습기에서는 추출된 형태소별로 각각의 기존 분류 디렉토리에서의 중요도에 따른 가중치와 정확도를 계산하여 해당 분류 디렉토리에 할당하고 분류 디렉토리별 가중치를 재조정한다.

[표 1]은 자동분류 서비스 시스템에서 수행되는 주요 기능들에 대한 설명이다.

[표 1] 자동분류 서비스 시스템 기능별 설명

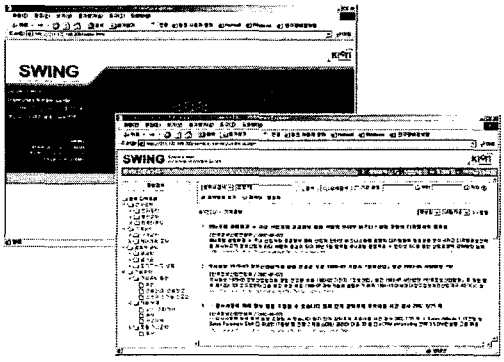
구 분	기 능	내 용
수집	수집 Site 관리	수집 Site 그룹 추가/수정/삭제 수집 Site 추가/수정/삭제 수집 포함/제외 URL 설정 수집 포함 단어 설정 즉시 수집 실행
	수집 환경 설정	다중 처리(Multi-Threading) 수집 옵션 설정
	수집 상태 보기	수집 진행/대기 상태 보기 수집 중지(진행 중 Site) 수집 취소(대기 중 Site)
	수집 현황 보기	수집 Site 현황 통계 목록
	수집 엔진	수집 실행 및 프로세스 관리 수집 스케줄링 관리 수집 자료 저장 및 관리
분류	분류 디렉토리 보기	분류 디렉토리 추가/수정/삭제
	분류 방법 설정	분류 대상 사이트 설정 분류 조건 설정 분류 키워드 관리 학습 여부 선택
	분류 관리	수집 자료 관리/검색 수동/자동 분류 실행 분류 자료 수정/삭제 분류 자료 등록
	분류 상태 보기	자동 분류 진행/대기 상태 보기 자동 분류 중지/취소
	분류 엔진	분류 실행 및 프로세스 관리 분류 작업 스케줄링 관리 수동 및 자동 분류 처리
Monitoring	Monitoring 엔진	분류 자료의 유효성 확인 및 처리
Push Mail	Push Mail 엔진	사용자 관심분야 정보 메일 발송

2. 자동분류 서비스 시스템 GUI

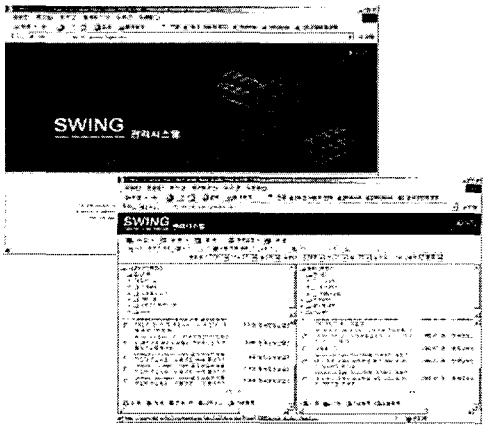
[그림 8]은 검색서비스 화면으로 수집과 분류를 끝낸 자료를 분류 디렉토리 검색, 상세검색 및 통합검색 등의 검색 기능을 제공한다. 또한, 사용자가 관심있는 분야를 설정하는 등록하여 자동 수집하여 야를 등록할 수 있도록 하여 신규정보 분류시 메일을 통한 전송서비스와 해당 분야의 정보만을 모아서 보여주는 사용자 중심 서비스도 제공한다.

[그림 9]는 관리자 화면으로 관리자의 시스템 제어

를 위한 것으로 전체 시스템의 성능 및 흐름을 파악할 수 있으며 웹 사이트 및 분류 디렉토리 관리, 사용자 및 관리자의 관리, 시스템 백업, 온라인 자동 업그레이드 등의 기능을 제공한다. 또한, 화면의 왼쪽에서는 각각의 등록된 웹 사이트 정보와 수집된 정보를 보여주고 있으며, 오른쪽은 분류 디렉토리와 분류된 결과 문서를 보여준다.



▶▶ 그림 8. 검색서비스 화면



▶▶ 그림 9. 관리자 서비스 화면

V. 결 론

본 논문에서는 인터넷 문서의 수집에서 분류, 검색 서비스까지를 하나의 시스템에서 수행할 수 있도록 통합 설계하고 구현한 인터넷 문서 자동분류 서비스 시스템을 구현하였다.

각각의 중요 기술을 통합하여 구현함으로써 효율적이고 편리한 정보관리 방법을 제시하였으며, 이는 각 분야의 정보 수집 및 관리에 있어서 시간 및 비용 절약과 업무 효율을 높이는데 효과가 있을 거라 기대된다.

향후 연구과제로는 사용자 편의가 강화된 Monitoring Engine으로 발전하기 위해 Monitoring 결과 통계 및 알람 서비스 기능, Monitoring Engine의 자율적인 판단에 의한 자료처리 기능, 자료처리에 의한 서버 부하를 감소시켜 시스템 안정화 방안을 강구할 것이다. 또한, 학습단어와 불용단어에 대한 개선된 알고리즘을 설계하여 자동분류에 대한 정확도를 높여야할 것이다.

▌참 고 문 헌▐

[1] Kalakota, R. and A. B. Whinston, Frontiers of Electronic Commerce, Addison-Wesley, 1996.

[2] Russell and Norvig, Artificial Intelligence a Modern Approach 2/E Chap 2. Prentice Hall International Co., 1994.

[3] Edmund H. Durfee and Jeffrey S. Rosenschein, "Distributed Problem Solving and Multi-Agent Systems: Comparisons and Examples," Proc. of Thirteenth International Distributed AI Workshop, pp.94-104, 1994.

[4] Oren Etzioni and Daniel Weld, "A Softbot-Based Interface to the Internet," Communications of ACM, Vol.37, No.7, pp.72-76, 1994.

[5] Gree, William B. "Introduction to Electronic Document Management Systems", Academic Press, 1993.

[6] John Bear and David Martin, Using Information Extraction to Improve Document Retrieval, Text Retrieval Conference, 1996.

[7] 한국전산원, "행정문서관리 효율화방안", Sep. 1997.