웹 문서의 자동분류를 위한 과학기술 웹 정보 서비스 시스템(SWING) 소개

황 성 하, 최 광 남, 이 상 호(한국과학기술정보연구원)

차 례

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 검색되기도 하며, 검색 결과를 체계적으로 분류하고 조직화하는데 많은 문제점이 있다.

이러한 문제점을 해결하기 위한 노력의 결실로 정보 수집을 작업 스케줄링에 따라 자동으로 해결해주는 소프트웨어인 에이전트와 기계학습 등의 과정을 통해 문서를 자동으로 분류하는 문서분류 기술 분야에서 많은 발전을 해 왔으며 각각 다양하게 활용되고 있다.

특히, 필요한 정보를 수집하여 관리하고 분류하여 검색할 수 있는 자동분류 서비스 시스템은 사용자들에게 작업의 효율성과 편리한 정보관리 방법을 제공할 수 있게 되었다.

본 논문에서는 사용자 중심의 편의를 제공하는 요소기술을 통합하여 구현한 과학기술 웹 정보서비스 시스템(SWING; Science Web Information Guide)을 소개한다.

Ⅱ. 웹 문서의 자동분류 기술동향

2.1 기존의 문서분류 응용 시스템

인터넷상에 보이는 웹 문서를 대상으로 문서를 자동으로 분류하는 기법을 이용한 대표적인 시스템은 뉴스기사 분류 시스템과 검색엔진 시스템 등이 있다. 다음은 이러한 문서분류 기법을 적용하여 만들어진 문서분류 응용 시스템에 대하여 알아본다.

2.1.1 Personal Webwatcher

Personal Webwatcher는 카네기 멜론 대학에

서 개발한 시스템으로 사용자의 행동을 웹 브라우저에서 모니터링하여 사용자에게 편의를 제공하는 시스템이다. 이 시스템을 구성하는 분류기법방식은 사용자의 관심도를 학습하는 비교사 학습방식을 이용하였다. 또한, 이 시스템은 관측을 위한 모니터링 부분, 모니터링 결과에 따른 사용자의 프로파일을 만드는 부분, 사용자의 프로파일을 이용하여 사용자에게 검색하는 웹 페이지 모니터링 부분으로 구성되어 있다.[1]

2.1.2 InfoFinder

InfoFinder는 앤더슨 컨설팅 연구실에서 개발한 시스템으로 사용자의 관심을 관측함으로써 사용자의 프로파일을 만드는 시스템이다. InfoFinder는 사용자가 원하는 관심문서에 대하여 사용자의 직접적인 관심사항을 입력받아 사용자의 관심을 학습하는 교사학습방식을 이용한 시스템이다.[1]

2.1.3 NewT

인터넷의 발전으로 수많은 정보가 네트워크로 들어오는 가운데 뉴스분야의 정보는 계속적인 스트림(stream)의 형태로 웹으로 유입된다. 이러한 뉴스의 스트림 가운데 사용자가 원하는 기사의 선택을 위한 에이전트 시스템으로 NewT가 있다. NewT는 뉴스의 기사를 정치, 경제, 컴퓨터, 스포츠 4가지 클래스로 필터링 한다. 뉴스기사 문서 분석은 벡터-공간 모델을 사용한 풀-텍스트 분석으로 이루어지며, NewT 에이전트의 특징은 에이전트 협동 부분인데, 사용자는 충분히 학습된 에이전트를 복사하여 다른 사용자에게 제공할 수 있도록 유닉스 환경의 C++로 구현되었다.[1]

2.2 문서분류 및 학습 방법론

본 논문에서는 문서분류 및 학습 방법론에 관한 관련연구로 전자문서 관리시스템(EDMS), 정보검색시스템(IRS), Naive Bayesian 알고리즘과 K-NN(K-Nearest Neighbor) 알고리즘에 대해 알아본다.

2.2.1 전자문서 관리시스템(EDMS)

EDMS(Electronic Document Management System)는 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 문서분류는 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[2][5].

2.2.2 정보검색시스템(IRS)

IRS(Information Retrieval System) 정보검색 분야에서의 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. 또한, 다양한 학습기능을 도입하여 정확성을 높인 자동분류 시스템이 많이 연구되고 있다[3].

2.2.3 Naive Bayesian 알고리즘

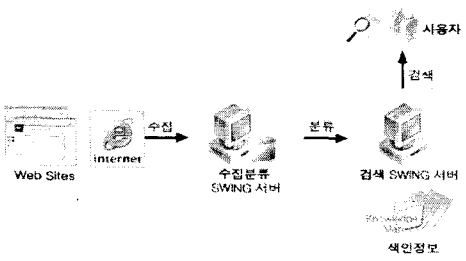
Naive Bayesian 알고리즘은 분류하고자 하는 카테고리별 학습정보와 베이지정리를 응용하여 어느 카테고리에 포함될지 확률(범주)을 구하여 확률이 높은 카테고리로 분류하는 알고리즘이다. 분류대상 문서는 형태소 분석을 통하여 명사(구)를 추출한 후, 각 명사(구)에 자질값을 추출하게 되며 측정된 자질값에 따라 자질집합(분류 카테고리)에 분류하는 알고리즘으로 단순 베이지언 정리와 퍼셉트론을 이용한 가중치 벡터구조 등에 활용되고 있다.[4]

2.2.4 K-NN(K-Nearest Neighbor) 알고리즘

K-NN 알고리즘은 메모리를 기반으로 하는 학습방법으로 분류하고자 하는 문서의 모든 단어를 각 카테고리별로 비교하여 가장 근접한 카테고리에 분류하는 알고리즘이다. 이 알고리즘을 Lazy Learing 알고리즘이라고도 부른다.[1]

Ⅲ. 과학기술 웹 정보서비스 시스템 (SWING) 구축 사례

3.1 SWING 시스템 개요

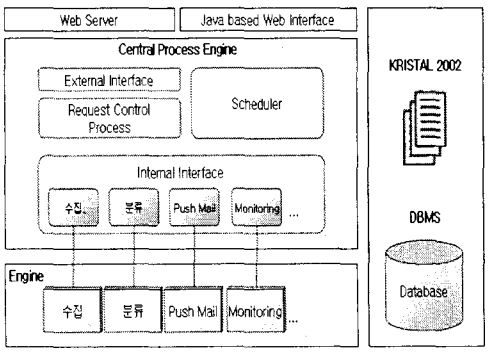


▶▶ 그림 1. SWING 시스템 개요

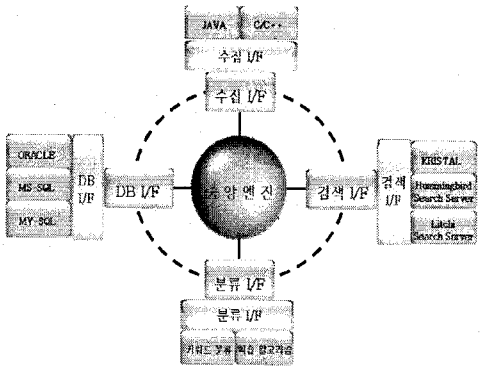
SWING 시스템은 인터넷에 산재한 광대한 정보를 효율적으로 수집, 주제별로 분류하여 정제된 지식을 사용자가 검색할 수 있도록 서비스를 제공하는 시스템이다. 즉, SWING 시스템은 인터넷 수집로봇 에이전트를 통해 등록된 웹 사이트의 문서를 주기적으로 방문하여 수집하고 지식 카테고리 별로 수동 및 자동분류를 실행하여 사용자에게 다양한 지식 카테고리 검색 및 상세검색 서비스를 제공하며, E-mail 서비스를 통해 원하는 사용자에게 Push Mail을 수행 등을 통합한 시스템이다. [그림 1]은 인터넷 문서의 수집에서 자동분류, 검색 서비스까지를 하나의 시스템에서 처리되는 과정을 나타낸다.

3.2 SWING 시스템 구조 및 기능 설계

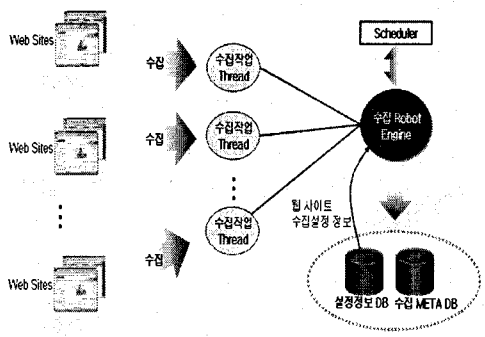
SWING 시스템은 웹 문서 수집 및 자동분류 시스템으로 스케줄러에 의해 정해진 시간에 수집 Robot Engine이 자동 구동되어 문서를 수집하며 수집된 정보를 분류 Engine에 내장된 학습 알고리즘 및 텍스트 분석에 의한 주제어 분류 방식을 이용하여 검색엔진에 카테고리별로 분류하는 시스템이다. [그림 2]는 SWING 시스템의 구조를 보여준다.



▶▶ 그림 2. SWING 시스템 구조



▶▶ 그림 3. SWING 시스템 엔진



▶▶ 그림 4. 문서 수집

SWING 시스템은 크게 네 가지 기능으로 분류된다. 첫 번째, 수집 Robot Engine에 의한 정보 수집 기능과 둘 번째, 분류 Engine에 의한 자동 분류 기능 셋 번째, 분류 완료된 문서의 유효성 검사를 위한 Monitoring 기능 넷 번째, 사용자의 관심정보를 분류하여 검색결과를 전송해주는 Push Mailing 기능으로 분류된다.

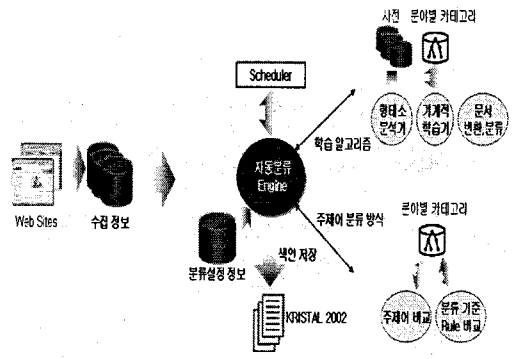
[그림 3]은 SWING 시스템의 중앙엔진을 중심으로 한 수집, 분류, DB와 검색 인터페이스를 나타낸다. 수집 인터페이스는 JAVA와 C/C++, 분류 인터페이스는 키워드 분류와 학습 알고리즘으로 구성되어 있으며, DB 인터페이스는 ORACLE, MS-SQL와 MY-SQL, 검색 인터페이스는 KRISTAL, Hummingbird Search Server와 Litchi Search Server로 구성된다.

3.2.1 문서 수집

[그림 4]는 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.

3.2.2 문서 자동 분류

[그림 5]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



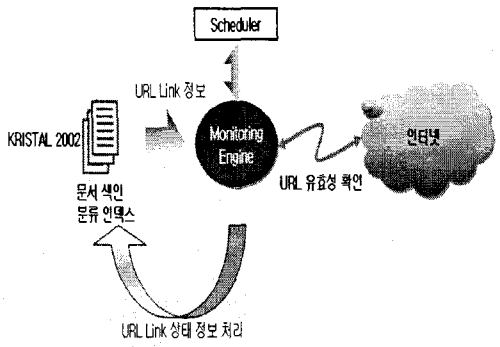
▶▶ 그림 5. 문서 자동 분류

학습기에서는 추출된 형태소별로 각각의 기존분류 디렉터리에서의 중요도에 따른 가중치와 정확도를 계산하여 해당 분류 디렉터리에 할당하고 분류 디렉터리별 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 검색엔진에 색인하여 저장한다.

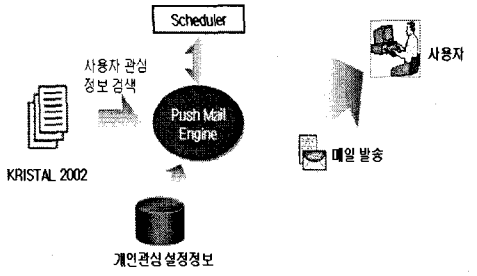
3.2.3 Monitoring

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 6]에서 보여준다.



▶▶ 그림 6. Monitoring

3.2.4 Push Mail



▶▶ 그림 7. Push Mail

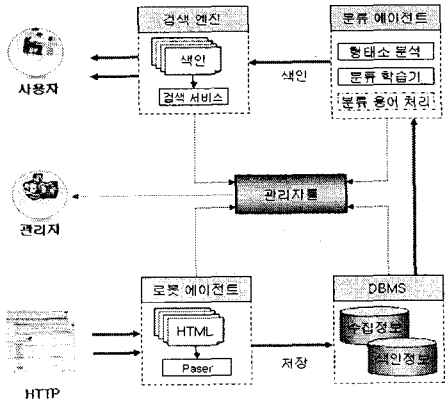
Push Mail 서비스를 신청한 사용자에 한하여 Push Mail Engine Scheduler에 의해 Mailing 되며 사용자별로 설정된 메일 전송주기, 관심 지식 맵, 관심 키워드 등의 관심정보를 로딩(Loading)한다. 이를 통해 분류정보 내에서 검색하여, 검색 결과를 사용자에게 메일을 전송하는 방식으로 이는 [그림 7]에서 보여준다.

3.3 SWING 시스템 사례

3.3.1 SWING 시스템 Flow-Chart

[그림 8]은 SWING 시스템의 흐름을 보여주는 것으로 수집 및 분류 에이전트, DBMS와 검색엔진으로 구분된다.

수집 에이전트가 등록된 여러 개의 웹 사이트를 돌아다니며 문서를 수집해 오면 문서 관리기에 의해 중복, 수정, 추가 등의 문서 상태를 체크하고 수집정보 DB에 저장한다. 만약, 수집문서가 이미 수집된 문서와 동일한 중복 문서일 경우에는 모니터링 관리 정보에 기록하고 마치며, 추가 또는 수정된 문서일 경우에는 수집정보 DB에 추가하고 분류 에이전트에 분류해 줄 것을 통보한다. 반면, 문서가 삭제된 경우에는 분류 에이전트에 색인정보에서 삭제할 것을 통보하고 분류 에이전트는 삭제 후, 해당 분류 디렉터리의 학습 값을 재조정한다.



▶▶ 그림 8. SWING 시스템 Flow-Chart

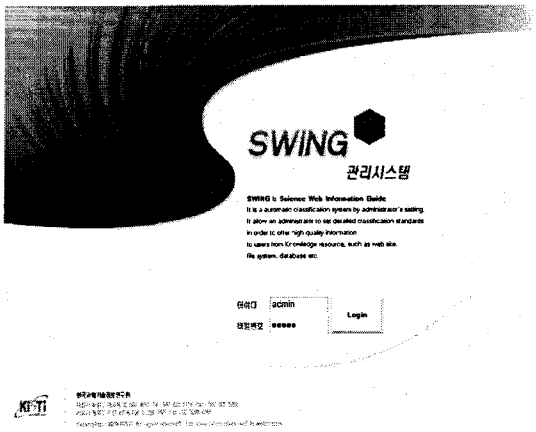
표 1. SWING 시스템 기능별 설명

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 디렉터리 보기	분류 디렉터리 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

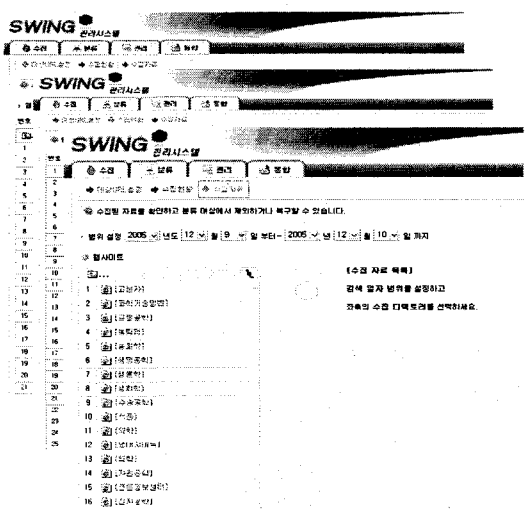
따라서 수집 에이전트의 핵심은 문서의 중복성을 체크하는 작업이며, 실제로 이 부분에서 많은 처리시간이 요구된다.

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 디렉터리에서의 중요도에 따른 가중치와 정확도를 계산하여 해당 분류 디렉터리에 할당하고 분류 디렉터리별 가중치를 재조정한다. [표 1]은 SWING 시스템에서 수행되는 주요 기능들에 대한 설명이다.

3.3.2 SWING 시스템 GUI



▶▶ 그림 9. 관리 시스템 화면



▶▶ 그림 10. 수집기능 화면

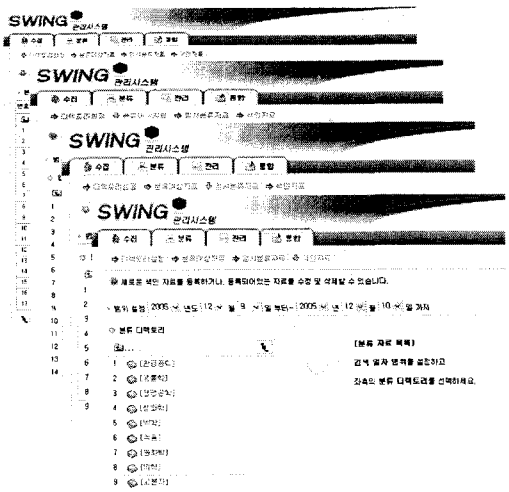
[그림 9]는 관리자의 시스템 제어를 위한 관리 시스템 화면으로 전체 시스템의 성능 및 흐름을 파악할 수 있으며 웹 사이트 및 분류 디렉터리

관리, 사용자 및 관리자의 관리, 시스템 백업, 온라인 자동 업그레이드 등의 기능을 제공한다. 예컨대, SWING 시스템은 중앙엔진과 검색엔진 상태 및 수집·분류 상태를 설정하고 확인할 수 있어 시스템의 작동상태를 한눈에 알아볼 수 있는 특징이 있다.

[그림 10]은 수집기능을 보여주는 것으로 세부기능으로는 수집하고자 하는 웹 사이트 정보의 등록, 수정, 삭제 및 그룹으로 관리할 수 있는 대상 URL 설정, 등록되어 있는 수집 대상 웹 사이트의 현황을 확인할 수 있는 수집현황과 수집된 자료를 확인하고 분류 대상에서 제외하거나 복구할 수 있는 수집자료 기능으로 구분된다.

[그림 11]은 분류기능을 나타내며, 세부기능으로는 수집된 웹 사이트의 정보를 정리하기 위해서 분류 디렉터리를 등록하거나 수정 및 삭제할 수 있는 디렉터리 설정, 분류 대상 자료를 확인하고 수동 또는 자동으로 분류할 수 있고 분류 대상에서 제외할 수 있는 분류대상자료, 색인 등록 대기 자료를 수정 및 삭제하거나 색인에 개별등록 및 일괄등록을 할 수 있는 임시분류자료와 새로운 색인 자료를 등록하거나 등록되어있는 자료를 수정 및 삭제할 수 있는 색인 자료 기능으로 구분된다.

기타 기능으로는 엔진상태와 사용자 및 관리자 관리를 위한 접속관리, 다양한 검색조건으로 색인된 자료를 검색하여 관리하는 자료관리, 색인된 자료의 유효성 여부를 확인할 수 있는 설정, 예약에 의한 DB의 주기적 백업 또는 즉시 백업 및 복구처리를 할 수 있는 백업, 시스템의 수집, 분류 및 검색에 관련된 통계를 보여주는 관리기능과 통합 검색에 참여하는 시스템 서버의 등록, 수정, 삭제할 수 있는 사이트 정보, 개별 시스템의 통합검색 참여여부 설정과 분류 디렉터리 수



▶▶ 그림 11. 분류기능 화면

집을 할 수 있는 사이트 관리, 통합검색에 참여하는 시스템들의 전체 통계를 볼 수 있는 사이트 통계, 통합검색에 참여하는 시스템들의 엔진상태를 확인하고 관리할 수 있는 사이트엔진상태, 주기적으로 통합 대상 사이트의 분류자료를 수집할 수 있도록 예약을 설정하고 관리하는 자료수집 설정과 수집된 분류 자료를 관리할 수 있는 수집자료를 관리하는 통합기능이 있다.

[그림 12]는 검색서비스 화면으로 수집과 분류를 끝낸 자료를 분류 디렉터리 검색, 상세검색 및 통합검색 등의 검색 기능을 제공한다. 예컨대, 검색한 결과는 제목, 서지정보, 출처와 등록일, 수집기관과 자료 위치를 알려주는 URL로 표시된다. 또한, 사용자가 관심있는 분야를 관심분야 목록에 등록하여 자동 수집된 자료를 신규정보 분류 시 메일을 통한 전송서비스와 해당 분야의 정보만을 모아서 보여주는 사용자 중심 서비스도 제공한다. [표 2]는 SWING 시스템의 특징을 요약한 것이다.

▶▶ 그림 12. 통합검색 화면

표 2. SWING 시스템 특징

구분	특 징
관리	• 웹 인터페이스를 통한 시스템 관리의 편의성 - 시스템 운영을 위한 기본 설정 - 중앙엔진과 검색엔진의 관리 - 수집에서 분류 및 검색 서비스까지의 순차적 정보 관리
수집	• 편리한 설정을 통한 웹 문서 수집 - 수집단어 설정 기능 - 수집 포함 또는 제외 URL 설정 기능 - 문서 File URL 수집 기능
분류	• 다양한 문서분류 처리 - 분류 키워드에 의한 자동 분류 - 분류 학습에 의한 자동 분류 - 분류 키워드 및 학습 분류 혼합 적용을 통한 분류 - 시스템 관리자에 의한 자료 분류
통합	• SWING 시스템 간의 연계 통합 기능 - 통합 시스템 설정을 통한 분산된 개별 시스템의 통합검색 용이 - 분산된 개별 SWING 시스템의 엔진 관리 - 자동화된 메타 통합 엔진
검색	• 사용자 중심의 검색 서비스 - 키워드 검색 - 분류별 검색 - 상세조건 검색 - 사용자 맞춤 서비스(관심분야 검색/ Push mail 서비스)
기타	• 표준 인터페이스에 의한 모듈화 된 시스템 구성 - 각 엔진 인터페이스 모듈화 - DBMS 인터페이스 모듈화 - 검색엔진 인터페이스 모듈화

V. 결 론

요즘 대부분의 현대인들은 인터넷을 통해서 원하는 정보를 검색한다. 예컨대, 인터넷에서 원하는 정보를 용이하게 수집하고 자동으로 분류하여 사용자에게 품질 좋은 맞춤형정보 서비스를 제공한다면 인터넷의 활용도는 더욱 커질 수 있다. SWING은 인터넷에 산재한 다양한 정보를 효율적으로 수집, 주제별로 분류하여 정제된 지식을 사용자가 검색할 수 있도록 서비스를 제공하는 시스템이다. 즉, SWING은 일반 포털 서비스와는 차별화된 다양한 정보 수집 및 분류 정제를 통한 양질의 개인 맞춤형 정보를 제공한다.

향후 연구과제로는 사용자 편의가 강화된 Monitoring Engine으로 발전하기 위해 Monitoring 결과 통계 및 알람 서비스 기능, Monitoring Engine의 자율적인 판단에 의한 자료처리 기능과 학습단어와 불용단어에 대한 개선된 알고리즘을 설계하여 자동분류에 대한 정확도를 높여야 할 것이다.

참고문헌

[1] 최정민, 진훈, 김인철, "웹 문서 분류법의 실험적 비교", 한국인터넷정보학회 학술발표논문집, 1권, 1호, May. 2000.

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

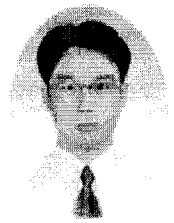
[3] Gree, William B. "Introduction to Electronic Document Management Systems", Academic Press, 1993.

[4] T. Mitchell. Machine Learning. McGraw-Hill, 1997.

[5] 한국전산원, "행정문서관리 효율화방안", Sep. 1997.

저 자 소 개

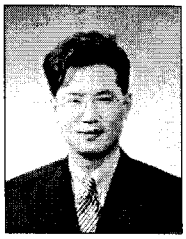
황 성 하(Sung-Ha Hwang) 정회원



- 2001년 2월 : 한남대학교 컴퓨터공학과(공학사)
- 2003년 2월 : 한남대학교 컴퓨터공학과(공학석사)
- 2003년 5월~현재 : 한국과학기술정보연구원 연구원

<관심분야> : 소프트웨어공학, 웹공학, 정보검색 및 위험분석

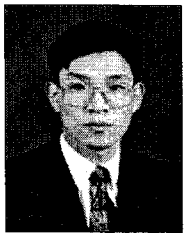
최 광 남(Kwang-Nam Choi) 정회원



- 1992년 2월 : 충남대학교 컴퓨터공학과(공학사)
- 1994년 2월 : 충남대학교 컴퓨터공학과(공학석사)
- 2004년 2월 : 연세대학교 문헌정보학과(박사수료)
- 1994년 7월~현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 정보검색, 전자도서관, 계량정보학

이 상 호(Sang-Ho Lee) 정회원



- 1982년 2월 : 충북대학교 화학공학과(공학사)
- 1984년 2월 : 충북대학교 화학공학과(공학석사)
- 1993년 3월 : 동경농공대학 물질생물공학과 (공학박사)
- 1983년 7월~현재 : 한국과학기술정보연구원 책임연구원

<관심분야> : 사실정보, 물성정보, 데이터베이스, 화학정보