장르기반 분류와 주제기반 분류를 이용한 웹 로봇의 설계 및 구현

이 용 배†

요 약

특수 전문화된 정보를 자동으로 수집하기 위해서는 인터넷 상을 순회하면서 대규모 자료를 모아오는 현재의 웹 로봇의 기능만으로는 그 역할을 수행하기에 부족함이 있다. 따라서 본 논문에서는 현재의 웹 로봇의 기능과 활용도를 분석하여 보고 전문정보를 수집하는데 있어서 한계점을 알아보았다. 또한 특수화된 분야의 전문정보를 수집하기 위하여 웹 로봇이 갖추어야 할 기능들을 도출해 내고 이를 설계한 내용을 기술하였다.

웹 로봇에 접목된 주요기능은 문서를 유형기반으로 분류할 수 있는 장르기반 분류와 주제기반으로 분류하는 내용기반 분류이다. 특히 장르기반 분류는 웹 로봇이 목적 문서를 효과적으로 수집할 수 있도록 하는 주요 기능으로 작용하였다.

키워드 : 웹 로봇, 자동분류

A Design and Implementation of Web Robot by Using Genre-based Categorization and Subject-based Categorization

Yong-Bae Lee†

ABSTRACT

It still has some restrictions to collect a specialized information with only the function of existing web robot which collect an enormous of data by circulating through the internet. Therefore, in this paper the functions of the current web robot and its application areas are analyzed and the limitations of collecting a specialized information are found out. Also we define what functions are necessary for a web robot in order to collect a specialized information. Then the designed structure is described.

There are two critical functions which are applied to web robot. One is a genre-based categorization that classifies the text by the type, and the other is a content-based categorization by the subject. Most of all, genre-based categorization is used as fundamental feature which enables web robot to collect the aimed documents efficiently.

Key Words : Web Robot, Automatic Classification

1. 서 론

현재 지식정보화 사회에서 정보의 공유는 일반화되었지만 정작 필요한 특수 전문화된 정보는 공유빈도가 희박하며 공유가 된다 하더라도 매우 한정된 분야의 아주 적은 양의 정보가 비싼 값으로 유통되고 있는 상황이다. 예를 들어, 국가적 사회적으로 중요하게 인식하고 있는 전문 교육 분야의 정보는 일반인이나 교사, 학생들이 접근하기에는 쉽지 않으며 소수의 교육용 포털사이트에서 제공하는 정보 서비스는 사용자의 다양한 정보 요구를 만족시키지 못하고 정보 갱신 기간이 일반 상업용 포털사이트 보다 상대적으로 길기 때문에 신뢰성 있는 교육 정보 서비스를 할 수 없는 상황이다. 이는 일반 포털사이트에서 정보 수집을 수작업 또는 반자동으로 하고 있어 새로운 문서나 정보생성 시간이 길어지기 때문이다.

따라서, 본 연구에서는 분산환경에 산재되어 있는 디지털 콘텐츠 중에서 특수 전문화된 정보만을 선별하여 자동으로 수집할 수 있는 웹 로봇(web robot)의 기능 정의와 이를 토대로 한 기본구조 설계에 주안점을 둔다.

초기의 웹 로봇은 정보 검색엔진의 한 구성요소로서 동작하였으나 최근에는 검색엔진 뿐만 아니라 다양한 응용분야에 독립적으로 사용되어 그 중요성을 높여가고 있다. 단순한 예로 논문정보, 파일정보, 상품정보, 가격정보, 뉴스정보 등을 제공하는 특정분야에 한정된 검색 사이트에서는 이 분야에 특화된 정보 수집용 로봇들이 필요하다.

본 연구에서는 전문정보 수집용 웹 로봇의 기능을 설계하

† 정 회 원 : 전주교육대학교 컴퓨터교육과 교수
논문접수 : 2005년 2월 18일, 심사완료 : 2005년 6월 3일

기 위해 다음의 내용을 주요 연구 목표로 한다.

- 기존의 웹 로봇의 주요기능 및 웹 로봇의 활용현황 분석하여 전문정보를 수집하기에 부적절한 원인을 찾아낸다.
- 새로운 웹 환경과 전문정보를 수집하기 위한 웹 로봇의 기능을 설계하여 기존의 웹 로봇에 접목시킨다.

본 논문의 구성은 다음과 같다. 2장에서는 웹 로봇의 기능과 활용현황에 대한 분석내용을 기술하고 3장에서는 현재 웹 로봇이 전문정보를 수집하는데 있어 한계점을 알아보며 4장에서는 본 연구에서 설계한 웹 로봇의 기능에 대하여 상세히 기술하였다. 다음 5장에서는 웹 로봇의 기능평가 부분을 서술하였고 마지막 6장에서는 결론과 향후 연구 과제를 기술하였다.

2. 웹 로봇의 기능 및 활용

웹 로봇은 인터넷 상의 정보수집을 위해 사용되는 일종의 에이전트로서 서버에 상주하면서 사용자와 직접적인 상호작용 없이 사용자를 대신해서 인터넷 상에서 분산된 온라인 정보를 순회하며 정보를 수집하는 프로그램이다. 웹 로봇은 웹 서버에 접속하여 HTML 또는 XML과 같은 마크업 언어로 작성된 파일 및 다양한 형태의 파일을 가져오는 기능적인 측면으로만 볼 때에는 일반 웹 브라우저와 같은 역할을 수행[2]하지만 그 외에 해당 문서를 분석하고 문서 내에서 나타난 URL 부분을 추출하며 그 URL로 접근하여 필요한 데이터를 수집한다. 즉, 웹 로봇은 홈페이지를 자동으로 순회하면서 사람이 하기 귀찮고 어려운 다양한 정보 수집을 보다 빠르고 효율적으로 수행하는데 큰 의미가 있다고 할 수 있다.

웹이 활성화됨에 따라 사람들은 웹 로봇을 사용하여 대규모 자료에 대해 반복적이고 다양한 기능을 수행하여 왔다. 현재에는 도메인 분석을 위한 통계 자료의 수집 및 분석[23, 28], 포탈 사이트에서 링크 상태를 관리하기 위한 검사, 접속 부하 관리 또는 자료 보관을 위한 중복 사이트의 구축 및 관리(mirroring)[21, 24], 인터넷에서 필요한 정보의 자동 발견 및 수집을 위한 목적으로 웹 로봇의 그 중요성을 더해 가고 있다.

웹 로봇을 이용하는 가장 중요한 목적은 단순히 불특정 문서의 광범위한 수집으로 제한되는 것이 아니라 정보이용 목적에 적합한 문서를 수집함과 수집된 문서를 활용한 정보서비스 시스템 구축에 있다. 현재 다양한 분야에서 여러 용도로 웹 로봇이 이용되고 있는데 웹 로봇을 이용하고 있는 분야를 크게 전문정보 수집, 정보검색, 전자 상거래, 개인화된 서비스의 네 가지로 분류되며 그 활용현황을 분석한 결과는 다음과 같다.

2.1 전문정보 수집

전문정보를 수집하기 위한 로봇은 현재 범용으로 개발되어 있는 것은 없으며 특정 전문정보에 종속으로 개발된 시스템은 몇 가지 있지만 실제로 전문가들이 사용하기에는 성능이 뒷받침 되고 있지 않다. 전문정보 수집 로봇 중에서 Synap의 Brief[26]는 인터넷에 새롭게 올라오는 정보를 실시간으로 수집하여 개인 및 기업고객이 설정한 관심분야에 따라 이메일 등으로 정보를 전달하는 맞춤형 서비스를 제공하고 있다.

2.2 정보검색

웹 로봇이 검색 서비스를 지원하기 위해서는 인터넷 상의 웹 문서는 물론 문서파일이나 이미지, 오디오, 비디오, 뉴스그룹 등을 찾아 다니면서 문서를 수집해야 한다. 현재 웹 로봇이 문서를 수집하는 방법은 웹 문서에 링크되어 있는 파일의 확장자로 구분하여 다운로드 받는 형식을 취하고 있다.

2.3 전자 상거래

전자 상거래가 생활 전반으로 확대되면서 현재 웹 로봇을 가장 많이 이용하고 있는 분야 또한 전자 상거래 분야이다[3]. 이 분야의 웹 로봇은 인터넷 정보검색을 위해 이용하던 키워드를 포함하는 문서를 가져오거나 특정 범주에 포함되는 문서만을 수집하던 기존의 웹 로봇의 기능에 추가적으로 보유하고 있어야 하는 기능이 정보추출 기능이다. 정보추출은 주어진 문서로부터 내용을 요약한다든지 필요한 사실적 데이터를 자동으로 추출하는 작업을 의미한다. 현재 활성화되고 있는 온라인 전문 쇼핑몰[18, 19, 22, 27]의 웹 로봇은 정보추출을 지원하고는 있지만 기능면에서는 아직 미약한 수준에 있다.

2.4 개인용 웹 에이전트(personal web agent)

개인용 웹 에이전트로서의 로봇[1]은 사용자 프로파일을 이용한 반복 학습을 통해 사용자 개인별로 필요한 정보만을 수집하여 개인용 데이터베이스를 구축하며 사용자의 서비스 요구시에 웹 서버에 구축된 사용자 개별 데이터베이스를 참조하여 가장 적절한 정보를 사용자에게 제공하여 준다. 일부 상용 온라인 쇼핑몰[18, 25, 29]에서는 소비자의 구매패턴과 프로파일을 저장한 후 소비자가 상품을 검색할 때 적절한 제품을 추천해 주는 서비스를 제공한다.

3. 전문정보 수집을 위한 웹 로봇의 한계

2장에서 기술한 바와 같이 웹 로봇은 주로 전문정보 수집과 정보검색을 위한 문서 수집, 전자상거래 분야에서 상품정보 추출 및 개인화된 웹 에이전트의 핵심 모듈로서 사용되고 있다. 그러나 현재의 웹 로봇으로는 사용자가 원하는 목적문서를 선별적으로 수집하려는 요구사항을 만족시키기가 어렵다.

산업정보 포털사이트, 신문 포털사이트와 같은 특정 목적 분야의 정보제공은 서비스 대상 문서를 직접 만들어 제공하기도 하지만 웹에서 존재하는 문서들을 모아서 제공할 수도

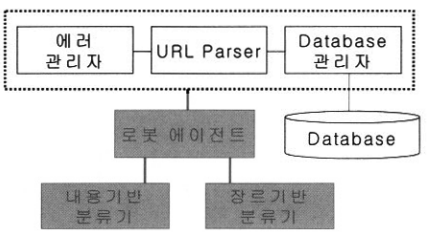
있다. 현재까지 개발되었거나 연구된 대부분의 웹 로봇들은 특정 키워드에 포함하거나 문서의 내용이 특정 범주에 해당하는 문서만을 선별하여 수집하였다. 즉, 웹 상의 수 많은 문서 중에서 그 의미가 찾고자 하는 문서와 유사한 문서만을 정제하여 저장하고 서비스를 제공하였다. 그러나 이러한 기능을 가진 웹 로봇은 문서의 유형이나 스타일에 의해 분류(예를 들어, 신문기사, 논문, 리뷰 등과 같은 특정 목적의 범주)한 문서를 찾아 올 수가 없다. 따라서, 현재 새로운 특정 목적 분야의 정보제공 서비스 구축을 위해서는 정보를 자체 제작하여 제공하거나 기존의 정보 제공 사이트를 레퍼로 연결하여 제공해야 하기 때문에 시스템 구축과 지속적 갱신을 위한 시간과 비용이 많이 소모된다.

4. 전문정보 수집을 위한 웹 로봇

전문정보를 수집하기 위한 웹 로봇은 수집되는 웹 문서를 특정 키워드에 의해 분류하는 규칙기반 분류방법 보다는 내용분석과 동시에 글의 문체나 스타일에 의해 분류할 수 있는 분류기술이 필요하다. 이를 수용하기 위해 본 연구에서 설계한 웹 로봇의 구조를 본 장에서 상세히 기술한다.

4.1 웹 로봇의 구조

본 연구에서 설계한 웹 로봇은 단순히 특정 전문 정보만을 수집할 수 있도록 한정하여 설계한 것이 아니라 소비자를 위한 제품의 평가정보 수집, 학생이나 교사를 위한 논문정보 수집, 뉴스정보 수집, 산업기술정보 수집 등 다양한 분야에 유연하게 연결할 수 있도록 설계하였다.



(그림 1) 웹 로봇의 구조

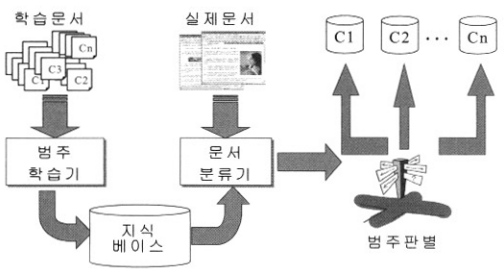
(그림 1)에서 점선 박스의 모듈은 기존의 웹 로봇에서도 공통된 주요 모듈이며 로봇 에이전트에 연결된 내용기반 분류기와 장르기반 분류기는 전문 정보 수집을 위해 새롭게 설계하여 확장한 것이다. 현재의 구조에서는 웹 로봇에 내용기반과 장르기반으로 분류하는 이차원 분류기를 확장하여 연결하였지만 향후 새로운 차원의 분류요구가 있을 경우, 로봇 에이전트에 새로운 분류기를 쉽게 연계할 수 있도록 설계에 반영하였다. 점선 박스 모듈의 기능은 다음과 같다.

- 에러관리자는 주로 웹 서버와의 통신 중에 네트워크의 과부하 또는 서비스의 장애로 인한 발생한 오류를 수정하고 관리하는 기능을 수행한다.
- URL Parser는 웹 서버에서 가져온 문서에서 하이퍼링크를 추출하며 추출된 링크가 상대경로로 되어 있으면 절대경로로 수정하여 데이터베이스에 저장한다. 또한 서로 다른 도메인 이름을 가진 문서가 동일한 아이피 주소의 문서일 경우를 확인하여 같은 아이피주소의 문서라는 표시와 함께 데이터베이스에 저장한다.
- Database 관리자의 주요 기능은 수집되는 문서를 효율적으로 저장하고 관리하는 것이다. 이와 더불어 문서 내에서 파싱된 URL을 별도로 저장하는데 URL은 방문했던 URL인지 방문예정인 URL인지를 구분하여 저장하며 URL을 저장할 때에는 실시간 검색이 가능하도록 도메인별로 최적화하여 저장한다.
- 로봇 에이전트는 에러 관리자, URL Parser, Database 관리자, 내용기반 분류기, 장르기반 분류기 모듈간의 메시지 전달과 데이터 송수신을 관리한다. 즉, 사용자로부터 문서 수집을 위한 시작 URL, 도메인제한 등의 조건을 입력 받아 문서 수집을 독립적으로 수행하며 수집된 문서를 Database 관리자가 저장할 수 있도록 전달한다. 또한 내용기반 분류기와 장르기반 분류기의 범주정보를 생성시켜 주며 문서 분류결과를 Database 관리자에게 전달하여 준다.

4.2 이차원 분류기

인터넷 사용자들은 웹 로봇이 문서를 수집할 때에 정치, 경제, 사회, 문화 등의 특정 내용이나 주제에 해당하는 문서를 수집하는 것보다는 신문기사나 논문, 상품 스펙, 상품 리뷰, 홈페이지 등과 같이 문서가 작성된 목적이나 유형에 따라 문서를 수집하기를 더 원하고 있다. 대부분의 분류기는 분류기준을 문서의 내용이나 주제를 기반으로 분류하는데 이러한 분류기법을 내용기반 분류라 하며 후자와 같은 문서의 유형에 따라 분류하는 기법을 장르기반 분류라 한다.

현재 대부분의 웹 로봇은 문서 자동 분류기를 포함하고 있지 않으며 소수 연구용 웹 로봇만이 내용기반 분류기를 내장하고 있다. 따라서 웹 로봇이 분류기에서 확장해야 하는 기술은 기존의 내용기반 분류기술에 신문기사, 논문, 광고 등과 같이 문서의 스타일 또는 문체에 의해 분류를 할 수 있는 장르기반 분류기술[8,9,10]을 결합하여 여러 차원으로 분류할 수 있는 방법을 개발하는 것이다. 본 연구에서는 이를 위해 우선 내용기반 분류기와 장르기반 분류기를 결합한 이차원 분류기를 설계하고 모형을 개발하여 웹 로봇과 결합시켰다. 다음은 이차원 분류기의 전체 구성도를 보여준다.



(그림 2) 이차원 분류기의 전체 구조

(그림 2)와 같이 이차원 분류기내의 범주 학습기는 미리 정해진 범주(category)에 적합한 학습문서를 대상으로 문서내의 용어 또는 문맥정보를 이용하여 범주학습을 수행한 후 지식베이스(knowledge base)를 생성한다. 또한 문서분류기는 지식베이스의 정보를 활용하여 실제문서가 입력되었을 때 어떤 범주와 가장 유사한지를 판단하여 범주를 할당하게 된다. 이러한 과정을 통하여 이차원 분류기는 내용기반 분류와 장르기반 분류를 복합적으로 수행할 수 있다.

4.2.1 내용기반 분류방법

글의 내용이나 주제를 기반으로 문서를 분류하는 방법은 기존에 많은 연구에서 이미 수행되어 왔다. 따라서 본 연구에서는 내용기반 분류 모듈은 분류 모델을 별도로 개발하지 않고 분류 알고리즘이 비교적 간단하고 다양한 분류 응용[11]에서 많이 사용되고 있는 베이지안(Bayesian) 모델[12,13,14,15]을 선정하여 구현하였다. 베이지안 모델은 분류 대상 문서가 각 범주에 속할 확률을 구해 가장 큰 확률값을 갖는 범주에 그 문서를 할당하는 방법이다. 베이지안 모델은 다음과 같은 수식으로 표현할 수 있다.

$$P(d \mid c_i) = P(c_i) \prod_{k=1}^{T} P(t_k \mid c_i)$$

(수식 1) 베이지안 모델

- T : 전체 문서 집합내의 용어의 수
- $P(c_i)$: 전체 문서 집합에서 범주 i의 문서가 나올 확률
- $P(t_k|c_i)$: 범주 i에서 용어 t_k가 나올 확률

4.2.2 장르기반 분류방법

기존의 장르기반 분류기[8, 9]는 장르를 구분할 수 있는 자질로 명사, 특수문자 등을 이용하였으며 분류시에 자질의 통계치를 확률함수가 아닌 규칙함수로 사용하였으므로 분류 정확률도 상대적으로 낮았다. 최근의 연구인 [10]에서도 기존의 장르기반 분류기에서 분류 자질로 사용한 명사, 특수문자 이외에 대명사, 감탄사 등도 함께 이용하였다. 그러나 분류 자질을 명사만을 이용하여 분류한 것보다 분류 정확률은 실질적으로 크게 나아지지 않았다. 본 연구에서 개발된 장르기반 분류기는 분류 정확률과 인터넷 문서의 실시간 분류의 필요성 등을 고려하여 명사만을 이용한 분류를 시도하였다.

본 연구의 장르기반 분류기는 다음과 같은 가정에서 출발하여 모델을 설계하였다. "특정 장르를 대표하는 용어들은 그 장르 내에서의 빈도수가 높으며 다른 장르에서는 상대적으로 빈도수가 낮을 것이다. 또한 장르 내에 주제별 범주가 존재한다면 특정 장르를 대표하는 용어들은 그 장르 안의 여러 주제별 범주에 걸쳐 골고루 높은 빈도로 분포되어 있을 것이다."

위와 같은 가정 하에 장르 범주 학습기는 전처리 단계, 자질선택(feature selection) 단계, 지식베이스 구축단계로 구성된다. 먼저, 전처리 단계에서는 학습 환경에 맞도록 학습문서를 전처리하고 문서에서 중요한 용어들을 추출하는 색인기능을 수행한다. 그 다음 학습과정의 핵심단계인 자질선택 과정을 거치며 그 결과 생성된 용어와 가중치를 빠르게 검색하여 분류에 이용할 수 있는 지식베이스 구축으로 학습과정을 종결한다. 자질선택과정을 단계별로 기술하면 다음과 같다.

step1 : 빈도수 높은 용어추출

자질선택의 첫번째 단계에서는 색인결과의 용어들에서 장르별로 문서출현빈도(document frequency)가 높은 용어들을 추출하고 장르 내에서도 주제별 범주가 존재하면 주제별 범주별로 문서출현빈도가 높은 용어들을 추출하여 빈도수의 내림차순으로 정렬시킨다. 빈도수를 문서출현빈도로 이용하는 이유는 본 논문에서 제안하는 방법이자 연구과정에서 분석한 결과에 의존한다.

step2 : 장르별 대표용어 계산

장르별 대표용어 계산 단계에서는 우선, 앞 단계에서 각 장르별로 문서 빈도수가 높은 순으로 추출된 용어들에 대하여 빈도수가 너무 작은 용어들은 일정 비율의 한계값(threshold)을 두어 잘라낸다. 이렇게 잘라낸 각 장르별 문서출현빈도가 높은 상위 N개의 용어들에 대하여 용어가 주제별로 어떻게 분포되어 있는지 편차를 계산한다.

주제별로 빈도수가 높으면서 편차가 작은 용어는 장르를 대표할 수 있는 용어일 확률이 높은 반면 특정 주제에서만 빈도가 높거나 편차가 큰 용어는 상대적으로 장르를 대표할 수 있는 용어가 될 확률이 낮다. 이러한 용어의 문서출현빈도와 주제별 편차가 나타내는 특성을 이용하여 장르를 대표할 수 있는 용어의 가중치를 계산해 낸다.

$$R\_Val_m(t_k) = 1 - \sqrt{\frac{\sum_{i=1}^{n_c} (DF_m(t_k) - DF_m(t_k^i))^2}{n_c}}$$

(수식 2) 장르별 대표용어 확률값

- DFm(tk) : 장르 m내에서 용어 t_k의 문서 빈도비율. 예를 들어, 장르 m의 문서 개수는 500건이고 용어 t_k의 문서출현빈도수는 100이라면 $DF_m(t_k)$는 0.2(100/500)가 된다. 즉, 문서출현빈도를 전체 장르문서 개수로 표준화(normalization) 한 값임
- $DF_m(t_k^i)$: 장르 m안의 주제별 범주 i 내에서 용어 t_k의 문서 빈도비율
- n_c : 장르 m안의 주제별 범주의 개수

(수식 2)에서 $DF_m(t_k)$과 $DF_m(t_k^i)$은 장르별로 또는 주제 범주별 문서의 개수에 따라 문서출현빈도가 상대적으로 많고 적을 수 있으므로 장르별 또는 주제 범주별 전체문서 개수로 나눈 값으로 표준화하여 장르별 대표용어 확률값 $R\_Val_m(t_k)$ 계산에 사용한다. 계산된 확률값에 장르 내에서 용어의 가중치 만큼을 곱해준 값이 용어가 특정 장르 m을 대표할 수 있는 최종값인 (수식 3) $WR_Val_m(t_k)$가 된다. 이 용어들의 값은 학습 과정의 다음 단계에서 장르간을 구분할 수 있는 변별치 계산에 이용된다.

$$WR\_Val_m(t_k) = R\_Val_m(t_k) \times DF_m(t_k)$$

(수식 3) 장르 대표용어의 최종값

step3 : 장르간 변별치 계산

여러 장르에 걸쳐 문서출현빈도가 높거나 혹은 동시에 문서출현빈도가 낮은 용어로는 어떤 장르의 문서인지를 판단하기가 어렵다. 특정 장르에만 문서출현빈도가 높거나 낮은 용어만이 장르간을 구분할 수 있는 자질이 될 수 있기 때문이다.

(수식 4)에서는 용어 t_k의 장르간 구분 변별치 $D_Val_m(t_k)$을 장르 대표용어 최종값의 장르간 편차로 계산한다. 용어의 편차가 크다는 것은 문서를 판별할 때 이 용어를 이용하면 다른 장르와 구분할 수 있는 확률이 높게 된다는 의미이고 편차가 작다는 것은 이 용어로는 타 장르와 구분할 수 있는 확률이 낮게 된다는 것을 뜻한다.

$$D\_Val_m(t_k) = \sqrt{\frac{\sum_{i=1}^{n_g} (WR\_Val_m(t_k) - WR\_Val_i(t_k))^2}{n_g}}$$

(수식 4) 용어의 장르간 변별치 계산

- n_g : 전체 장르의 개수

위와 같은 세단계의 학습과정을 거치면 지식베이스가 구축되며 이후 실제 문서가 입력되었을 때에는 지식베이스를 이용하여 문서의 장르를 판단하게 된다. 이때 판단하는 방법은 여러 가지가 있을 수 있지만 본 논문에서는 장르 대표벡터와 문서벡터와의 유사도(similarity)를 이용하여 문서 분류를 시도한다.

$$gID = \max_m [sim(G_m, D)]$$

(수식 5) 유사도 계산

- gID : 장르 식별자
- G_m : 지식베이스에 있는 장르 m의 대표벡터. 즉, 용어-변별치의 결합 테이블
- D : 장르 분류를 하기 위한 문서의 용어벡터
- sim : 장르 대표벡터와 분류할 문서의 용어벡터 간의 유사도
- max : 1~m 사이의 장르에서 가장 큰 유사도 값을 갖는 장르 번호

문서 장르 분류 과정은 먼저 지식베이스의 용어-변별치 결합테이블을 장르의 개수만큼 장르 대표벡터로 구성하고 분류하고자 하는 문서에서도 용어를 추출하여 문서 벡터를 구성한다. 다음으로 각 장르의 대표벡터와 문서 벡터와의 유사도를 계산하여 (수식 5)와 같이 가장 유사한 장르에 문서를 할당하게 된다.

위에서 살펴본 바와 같이 장르기반 분류기법을 웹 로봇에 도입하면 사용자가 문서 검색시에 찾고자 하는 장르의 문서를 빠르게 검색하여 서비스 할 수 있으며 전문정보 포털사이트를 쉽게 구축할 수 있는 장점이 있다.

5. 실험 및 평가

본 논문의 연구과정에서 설계하고 구현한 전문정보 수집용 웹 로봇은 원하는 목적 문서를 정확하게 가져오는 것이 목적이다. 이를 평가하기 위해 웹 문서를 수집하고 범주별로 분류정확도를 측정해 보았다. 수집된 문서의 유형, 즉 문서의 장르는 <표 1>에서 보는 바와 같이 신문의 사건기사와 사설, 개인 홈페이지, 리뷰, 논문, Q&A, 상품의 스펙으로 7가지이며 전체 7,828개의 문서로 구성된다.

〈표 1〉 수집된 문서

문서의 유형 (문서개수)	문서의 내용(문서개수)
사설(750)	경제(113),교육(70),국제(62),문화(53),북한(68),사회(112),스포츠(95),정보통신(55),정치(122)
사건(929)	강도(66),교통(93),사기(86),살해(142),폭력(86),약물(39),유괴(34),자살(105),절도(76),폭력(79),화재(123)
논문(1,051)	가정예술(109),공학(191),기초과학(143),농수임(101),생명의학(191),인문사회(139),전기전자(177)
리뷰(2,362)	교육(106),금융(151),문화(207),비디오(130),휴대폰(220),쇼핑(106),스포츠(242),유아(103),음식(103),의류(101),자동차(103),컴퓨터(407),통신(117),화장품(266)
개인홈(906)	교수(250),연예인(210),학생(240),회사원(206)
Q&A(960)	법률(130),소비자(102),영어(95),요리(103),유학(110),의학(156),청소년(103),컴퓨터(161)
스펙(870)	귀금속(94),스포츠(122),비디오(106),자동차(103),전자(123),컴퓨터(126),화장품(124),휴대폰(72)

<표 1>에서 문서의 내용이란 문서의 유형별로 포함된 문서의 주제를 의미한다. 예를 들어 논문 유형의 경우에는 포함된 문서들의 주제가 가정예술, 공학, 기초과학, 농수산임업 등이 있다는 의미이다.

5.1 성능 평가

실험은 크게 두 가지, 즉 장르기반과 주제기반으로 문서를 분류하였다. 장르기반 분류는 7,828건의 문서에서 절반 3,914 건은 장르학습을 위해 사용하고 나머지 절반 3,914 건의 문서를 7개 장르를 분류하는데 이용하였다. <표 2>는 장르기반 분류결과를 세부적으로 보여준다.

〈표 2〉 장르기반 분류결과

장르구분	사설	사건	리뷰	논문	프로필	Q&A	스펙
사설(375)	238	100	9	1	18	5	4
사건(464)	110	353	0	0	0	1	0
리뷰(1,181)	0	0	1,169	0	6	1	5
논문(526)	3	2	7	497	6	9	2
프로필(453)	3	5	15	22	404	4	0
Q&A(480)	0	0	15	0	3	437	25
스펙(435)	0	0	14	0	0	0	421

<표 2>에서 첫 번째 열에서 괄호 안의 숫자는 분류대상 문서의 개수를 나타내고 두 번째 열부터는 분류된 문서의 개수를 의미한다. 예를 들어, 논문으로 분류테스트를 했을 경우, 분류대상 문서의 개수는 526 건이며 이 중에서 사설로 분류된 것은 3건, 사건으로 분류된 것은 2건, 리뷰로 분류된 것은 7건, 논문으로 정확히 분류된 것은 497건, 프로필로 분류된 것은 6건, Q&A로 분류된 것은 9건, 스펙으로 분류된 것은 2건이라는 것을 뜻한다. 장르분류 실험의 전체 마이크로 평균정확도(micro average precision/recall)는 0.899로 비교적 신뢰할만한 수치로 측정되었다.

장르기반 분류의 학습방법과 분류방법을 측정해 보기 위해서 자질선택은 카이제곱 방법과 분류방법으로는 베이지안을 선택하였다. 카이제곱 방법은 용어와 범주 사이의 독립성을 계산하는 방식으로 카이제곱으로 자질선택을 한 후 문서를 분류할 때에는 비교적 높은 정확도[5, 16]를 나타내는 것으로 알려져 있다.

$$G_m(t_k) = \chi^2(t_k, m) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

(수식 6) 카이제곱 계산

	m	$\sim m$
t_k	A	B
$\sim t_k$	C	D

- A : tk 를 포함하고 장르 m에 속하는 문서의 수
- B : t_k 를 포함하고 장르 m에 속하지 않는 문서의 수
- C : t_k 를 포함하지 않고 장르 m에 속하는 문서의 수
- D : t_k 를 포함하지 않고 장르 m에도 속하지 않는 문서의 수
- N : 장르 학습 문서 전체의 개수

카이제곱으로 자질을 선택한 후 각 자질의 가중치는 (수식 6)의 값을 그대로 이용하여 분류에 사용하였다.

〈표 3〉 장르분류의 학습방법과 분류방법 비교

방법	학습 : 카이제곱 분류 : 제안방법	베이지안 방법의 학습 및 분류	제안된 방법으로 학습 및 분류
분류정확도	0.870	0.752	0.899

<표 3>에서는 카이제곱으로 학습하고 제안된 방법인 유사도를 기반으로 분류를 시도하면 약 3.3% 제안된 학습방법이 정확도가 높은 것을 알 수 있고, 베이지안 방법으로 학습하고 분류했을 경우에는 약 19.5% 제안된 방법이 정확도가 좋은 것을 볼 수 있다. 위의 실험결과로만 살펴보면 제안된 편차를 이용하여 학습하고 벡터간의 유사도를 이용하여 분류하는 장르분류 방법이 기존의 방법보다 비교적 우수한 것을 알 수 있다.

제안된 장르기반 분류는 상위계층의 장르범주를 분류하기 위하여 하위계층인 주제범주 정보를 이용하고 있다. 따라서 하위계층 정보를 이용한 효과를 측정하기 위하여 주제범주 정보를 이용한 경우와 이용하지 않은 경우의 정확도를 살펴보았다. 즉, 제안된 학습방법의 step1~step3을 이용한 것과 step1과 step3만을 이용하여 분류한 결과를 비교해 보았다.

〈표 4〉 계층정보의 이용 효과

구분	step1과 step3만을 이용	step1~step3을 모두 이용
분류정확도	0.868	0.899

<표 4>에서 볼 수 있듯이 주제범주 정보를 이용했을 경우 즉, 하위계층 정보를 이용했을 경우에 정확도가 약 3.6% 더 올라간 것을 알 수 있었다.

내용기반 분류는 각 장르내에서 <표 1>의 세부 주제별로 절반을 이용하여 내용학습을 하고 나머지를 이용하여 분류실험을 하였으며 결과는 <표 5>과 같다.

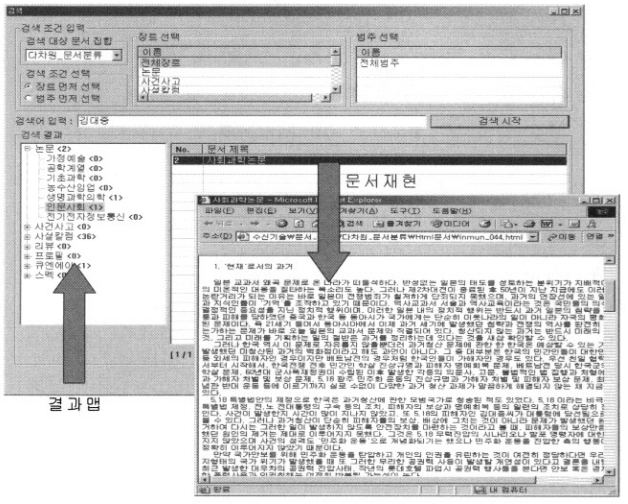
〈표 5〉 내용기반 분류결과

구분	주제 범주의 개수	분류 정확도
사설	9	0.71
사건	11	0.78
리뷰	14	0.90
논문	7	0.92
프로필	4	0.88
Q&A	8	0.89
스펙	8	0.86

5.2 정보검색에의 활용

내용기반 분류와 장르기반 분류를 접목한 웹 로봇을 정보검색에 활용하면 사용자의 검색만족도를 높일 수 있다. 현재 상용 검색사이트에서 일부 제공하는 검색결과에서의 뉴스와 이미지 등을 분류해주는 서비스는 미리 수작업으로 저장된 유형별 문서만을 대상으로 제공하는 것이다. 그러나 본 연구에서 개발된 이차원 분류기는 검색결과를 실시간으로 장르별로 주제별로 분류하여 사용자에게 검색결과를 제공할 수 있다.

(그림 3)은 웹 로봇을 검색엔진에 결합시켜 만든 검색 인터페이스이다. 가운데에는 질의입력창과 검색버튼이 있고 검색버튼을 누르면 왼쪽에 검색결과를 주제별 그리고 장르별로 실시간 분류한 결과맵이 만들어지며 결과맵의 한 목록을 선택하였을 경우 오른쪽창에 그에 해당하는 문서리스트가 나온다. 현재 인터페이스에서의 검색조건이 '장르먼저선택' 옵션으로 지정되어 결과맵에서 장르를 선택한 후 주제범주를 선택할 수 있도록 되어 있지만 옵션을 '주제먼저선택'으로 변경하면 주제범주를 선택한 후 장르를 선택할 수 있도록 재구성된다.



(그림 3) 검색결과의 분류

(그림 3)에서는 '김대중'이라는 질의어를 입력하여 왼쪽 창 결과맵 중에서 논문에서는 2건, 사설칼럼에서는 36건, 큐엔에이에서는 1건이 매칭되어 나온 결과를 보여주고 있으며 논문 장르의 인문사회 주제범주에서 걸린 1건의 문서를 클릭하여 브라우징된 모습을 나타내고 있다.

이차원 분류의 효과는 검색 사용자가 '김대중' 이라는 인물에 대하여 인문사회 분야의 논문을 원한다면 바로 결과맵을 선택하여 처음 검색결과의 39개의 문건 중에서 논문장르를 선택하여 해당 결과를 2개로 줄이고 이중에서 인문사회 주제를 선택하여 해당 문서 1건을 쉽게 찾을 수 있다. 그러나 이러한 웹 로봇이 결합되지 않았다면 전체 39개의 결과를 하나씩 읽어가면서 원하는 문서를 찾는데 시간이 많이 걸릴 것이며 검색문서의 개수가 많아질수록 본 연구에서 개발된 웹 로봇의 역할은 증가하게 될 것이다.

6. 결론 및 향후과제

최근까지 웹 로봇은 검색엔진의 핵심 구성 요소로서 검색 대상 문서 수집에 사용되었으며 접근할 수 있는 자료가 제한되기 때문에 보다 빠른 시간에 많은 자료를 수집하는 것이 중요한 관건이었다. 하지만 이후의 웹 로봇은 다양한 언어처리 기술과 결합하여 자료 수집의 정확도를 높이는 것이 중요 목적으로 변하고 있다.

본 연구에서는 현재 사용되고 있는 웹 로봇의 기능과 활용 분야을 조사하고 특수 전문화된 정보만을 선별적으로 수집하기 위해서 현재의 웹 로봇이 어떤 문제점이 있는지를 분석하여 보았다. 또한 이를 토대로 기존의 웹 로봇에 전문정보를 수집하기 위한 기능을 설계하고 구현하였으며 그 기능을 요약하면 다음과 같다.

첫째, 수집되는 웹 문서를 특정 키워드에 의해 분류하는 규칙기반 분류방법 보다는 내용분석과 학습을 통해 분류하는 학습기반 분류방법을 적용하여 웹 로봇을 개발하였다.

둘째, 문서의 주제나 의미에 따라 분류하는 내용기반 분류기와 글의 문체나 스타일에 의해 분류할 수 있는 장르기반 분류기를 개발하여 웹 로봇에 접목하였다.

본 연구에서 설계한 전문정보 수집을 위한 웹 로봇은 단지 특정 분야의 전문정보에만 제한적으로 설계한 것이 아니라 소비자를 위한 상품 평가정보, 학생이나 교사를 위한 논문정보, 뉴스정보, 산업기술정보와 같은 여러 종류의 전문분야로 확장이 가능하도록 고려하여 설계한 것이 특징이다.

향후 연구에서의 중점 부분은 우선 현재의 이차원 분류기가 내용기반 또는 장르기반으로 독립적으로 동작하는데 앞으로 이를 융합하여 하나로 보여줄 수 있는 모델개발이 필요하다. 둘째로 이차원 분류기를 평가할 수 있는 평가 데이터집합이 필요하다. 현재 수집된 문서는 주로 장르기반 분류를 평가하기 위해 구성되어 있어서 내용기반 분류와의 통합적인 평가가 어려운 상황이므로 여러 차원 분류를 종합적으로 평가할 수 있는 문서집합 구축이 필요하다. 셋째 현재의 장르기반 분류는 장르범주 정보와 장르범주 내에 있는 주제별 범주정보를 이용하는 두 계층으로 고정되어 있지만 이를 향후 다른 차원의 분류요구를 고려하여 일반적인 계층구조에 적용 가능하도록 학습방법의 확장이 필요하다. 마지막으로 차세대 의미기반의 웹 환경인 시멘틱 웹과 연동할 수 있도록 이차원 분류기를 유연성 있게 확장하는 작업이 남아 있다.

참 고 문 헌

[1] 이근배 외, "에이전트 기반 정보검색", 정보과학회지, 제16권 제8호, 1998.

[2] 마이크로소프트, 웹 로봇과 정보 추적자, 에이전트 기술/정보 찾아 3만리, 로봇 에이전트, 월간 마이크로소프트 10월, 1996.

[3] 남기범, 이건명, "전자상거래 에이전트", 정보과학회지, 제18권 제5호, 2000.

[4] 한국인터넷정보센터, URN 체계활용을 위한 메타데이터 개발, 기술보고서, 2002.

[5] 염기종, 권영식, "Suffix Tree를 이용한 웹문서 클러스터의 제목 생성 방법 성능 비교", 한국데이타마이닝학회 2002 추계학술대회 논문집, 2002.

[6] Tim Berners-Lee, James Hendler, Ora Lassila, "The Semantic Web", Scientific American, 5, 2001.

[7] W3C, Resource Description Framework (RDF) http://www.w3.org/RDF/, 2003.

[8] Andrew Dillon, Barbara Gushrowski, "Genre and the Web: Is the Personal Home Page the First Uniquely Digital Genre?", JASIS, 51(2), 2000.

[9] Jussi Karlgren, Douglass Cutting, "Recognizing Text Genres with Simple Metrics Using Discriminant Analysis", Proc. of COLING94, Kyoto, 1994.

[10] Yong-Bae Lee, Sung Hyon Myaeng, "Automatic Identification of Text Genres and Their Roles in Subject-Based Categorization", Proceedings of HICSS-37, Jan., Hawaii,

2004.

[11] Hyo-Jung Oh, Sung Hyon Myaeng, Mann-Ho Lee, "A Practical Hypertext Categorization Method using Links and Incrementally Available ClassInformation", Proc. of the 23rd ACM SIGIR Conference, Athenes, Greece, 2000.

[12] David Lewis, Marc Ringuette, "A Comparison of Two Learning Algorithm for Text Categorization", Proc. of the 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.

[13] Mehran Sahami, "Learning Limited Dependence Bayesian Classifiers", Proc. of the 2nd International Conference on KDD'96, 1996.

[14] Yiming Yang, Xin Liu, "A Re-examination of Text Categorization Methods", Proc. of the 22nd ACM SIGIR'99, 1999.

[15] Andrew McCallum, Kamal Nigram, "A Comparison of Event Models for Nave Bayes Text Classification", AAAI'98 Workshop on Learning for Text Categorization, 1998.

[16] Yiming Yang, Jan Peterson, "A comparative study on feature selection in text categorization", Proc. of 14th Int. Conf. On Machine Learning, 1997.

[17] Eberhart, "Survey of RDF data on the web", Proc. of the 6th World Multiconference on Systemics, Cybernetics and Informatics, 2002.

[18] Amazon, http://www.amazon.com.

[19] BargainFinder, http://bf.cstar.ac.com/bf.

[20] BookFinder.com, http://www.bookfinder.com.

[21] Checkbot, http://degraaff.org/checkbot.

[22] eBookExpress, http://www.ebookexpress.com.

[23] Mattew Gray, mkgray@mit.edu, http://www.mit.edu:8001/people/mkgray.

[24] MOMSpider, http://ftp.ics.uci.edu/pub/websoft.

[25] Mysimon, http://www.mysimon.com.

[26] Synaptic, http://www.synap.com.

[27] WatchPrice.com, http://www.watchprice.com.

[28] Webcrawler, http://webcrawler.com.

[29] NSTA, NSA WebWatcher Institute, http://webwatchers.nsta.org.

이 용 배

e-mail : yblee@jnue.ac.kr

1996년 충남대학교 컴퓨터과학과(학사)

1998년 충남대학교 컴퓨터과학과(이학석사)

2003년 충남대학교 컴퓨터과학과(이학박사)

2000년~2003년 (주)엔퀘스트테크놀러지 기술이사

2003년~현재 전주교육대학교 컴퓨터교육과 교수

관심분야 : 정보검색, 자연어처리, 디지털도서관, 자동분류, 지식관리시스템, 하이퍼미디어시스템