문서 수집과 필터링을 위한 개인용 에이전트의 설계와 구현

최상열*, 김진상*, 배인호*, 김윤년**, 장창수***
*계명대학교 컴퓨터공학과
**계명대학교 의과대학
***대구미래대학 모바일 컨텐츠학과
e-mail:el2idea@gmail.com

Design and Implementation of a Personal Agent for Collecting and Filtering Documents

Sang-Yeol Choi*, Jin-Sang Kim*, In-ho Bae*, Yoon-Nyun Kim**, Chang-Soo Jang***
*Dept of Computer Engineering, Keimyung University
**School of Medicine, Keimyung University
***Dept. of Mobile Contents, Daegu Mirae College

요 약

사용자가 원하는 논문 등의 문서를 웹에서 검색하여 조건에 맞지 않는 문서를 필터링하고, 문서의 제목, 저자, 출처 등의 요약 정보를 제공하며 사용자의 하드디스크로 문서를 다운로드하는 개인용 에이전트를 설계하고 구현하였다. 현재 구현된 개인용 에이전트는 사용자의 연구 분야에 대한 논문 문서를 수집하여 간략한 요약 정보를 제공하는 것을 목표로 하였으며, 따라서 PDF 형식의 논문 파일의 수집과 필터링에 한정하였다. 원하는 분야의 문서 수집을 예약할 수 있으며, 수집된 문서에 대해 사용자의 주석을 첨가할 수 있고, 또한 파일의 이름이 상이한 동일 문서를 식별하는 기능도 제공한다.

1. 배경

검색엔진을 이용한 인터넷 정보검색은 양의 방대함과 존재하는 자료의 부정확성 때문에 많은 시간과 노력이 소요된다. 특히 신속하고 정확한 정보가 필요한 대학이나 연구 기관에 종사하는 연구자들에게 필요한 학술 문서 파일의 검색과 수집은 검색엔진을 통해 수작업으로 처리하는 기존의 방식에 다음과 같은 큰 문제점이 있다. 즉, 링크 내에 존재하는 문서에 대한 내용 검색이 어렵고, 단순 키워드 검색으로 인해 검색의 정확성이 떨어지며, 방대한 양의 검색 결과를 사용자가 직접 확인해야 한다는 점이다.

이와 같은 문제점을 가진 수작업 방식의 인터넷 정보 검색과 필터링을 자동화하는 방법으로 에이전트 기술을 이용할 수 있다. 일반적으로 에이전트는 복잡한 동적 환경에 존재하면서 지각과 행동을 자율적으로 수행하면서 주어진 목표를 달성하는 계산 시스템으로 정의된다.

본 논문에서는 검색엔진을 이용해 수작업으로 이루어지는 문서 수집과 필터링의 문제점 해결과 이러한 처리의 효율성을 높이기 위해 웹상의 문서 필터링과 수집을 위한 개인용 에이전트를 설계하고 구현하였다.

웹 환경을 대상으로 개발한 에이전트는 다양한 종류가 있는데, 사용자에게 새로운 정보가 있는 사이트 정보를 찾아 보여 주는 브라우징 에이전트, 원하는 정보를 효율적으로 검색해 주는 검색 에이전트, 웹 자원을 비교 및 모니터링하여 유용한 정보를 제공하는 비교 에이전트, 동적인 웹 데이터 중 관심있는 정보만 필터링하여 제공하는 필터링 에이전트 등이 있다[1,2]. 다양한 종류의 웹 에이전트 개발은 궁

극적으로 웹 환경을 지능화시키려는 노력의 일환이며[3,4,5,6], 동시에 웹 사용자의 다양한 요구라고 볼 수 있다. 예를 들어, 음악에 관한 정보의 수집과 분석 및 관리를 위한 웹 에이전트 개발은 사용자의 다양한 요구를 수용하는 한 가지 예라고 볼 수 있다[7].

본 논문에서는 연구자들이 필요한 학술 문서를 대상하고, 수집 및 필터링할 문서 파일 종류를 PDF 파일로 한정하여 개인용 에이전트를 구현하였다.

2. 방법

본 연구에서 구현된 논문 수집 에이전트는 예약관리기, 검색기, 분석기, 데이터 관리기, 로컬 데이터 검색기의 다섯 부분으로 구성된다.

① 예약관리기는 사용자의 검색 질의를 관리하는 기능을 수행한다.

② 검색기는 예약관리기에서 수집하고자 하는 대상 정보를 받아서 검색을 수행한다.

③ 분석기는 탐색된 문서의 정보를 추출하여 검색 가능한 데이터로 변환/분석 한다.

④ 데이터 관리기는 분석기가 분석한 정보와 XML 과의 연동을 관리하여 데이터의 삽입, 삭제 등의 기능을 수행한다.

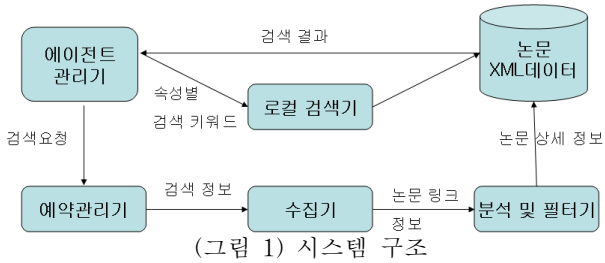
⑤ 로컬 데이터 검색 서비스는 사용자의 질의에 해당하는 문서 정보를 제공한다.

2.1. 데이터 구조

본 논문에서는 논문이라는 특수한 형식의 데이터를 대상으로 한 논문수집 에이전트에 대한 연구를 수행하였다. PDF 파일로 저장된 논문의 구조를 보면 제목, 저자, 요약, 키워드, 본문, 결론과 같은 형태로 구조화 되어 있다. 본 연구에서는 정확률 향상을 위해 문서의 구조에 대해 분석하고, 분석된 문서 데이터로부터 검색 키워드를 추출한 후 이를 이용하여 검색을 수행하였다.

2.2. 시스템 구조

본 논문에서 설계한 전체 시스템 구조는 그림 1과 같다. 사용자의 요청에 따라 예약관리기에서 사용자가 요청한 검색 정보를 예약 관리하며, 검색기에 의해 정보를 검색/수집하고, 수집된 정보를 저장하게 된다. 이 과정을 통해 생성된 데이터들을 바탕으로 이 후 사용자 요청에 대한 결과를 반환하게 된다.



(그림 1) 시스템 구조

2.2.1. 예약 관리기

예약 관리기는 사용자의 검색 요청을 관리하는 것으로 검색 옵션으로는 논문 옵션에 따른 검색, 링크 검색의 두 가지가 있다.

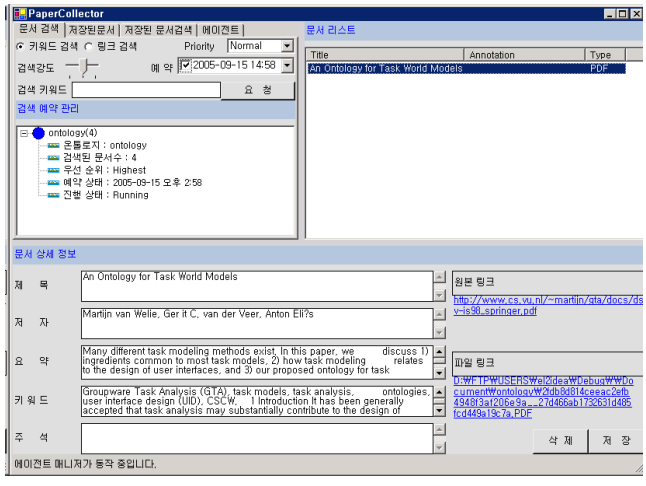
우선, 논문 옵션에 따른 검색은 검색하고자 하는 논문과 관련된 키워드를 입력 값으로 하여 논문을 검색한다.

다음으로, 링크 옵션에 따른 검색은 탐색하고자 하는 링크를 입력하면 그 링크를 탐색하여 논문을 수집한다.

부가적인 옵션으로는 검색의 강도, 검색 우선순위, 예약일시를 볼 수 있다.

그림 2는 검색 요청을 하는 화면으로서 검색 옵션에 따라서 검색 요청하고, 요청된 옵션에 따른 검색 진행 상태를 보여준다.

그리고 검색 요청 뿐 아니라 분석된 논문들의 리스트와 상세 요약 정보를 보고, 필요하다면 편집 할 수 있다. 이 과정에서 사용자가 이 논문에 부가적인 정보가 필요할 경우 주석을 추가할 수 있고 로컬 데이터에 대한 주석 검색이 가능하기 때문에 다음에 논문을 볼 때 자신이 첨가한 주석을 통해 내용이나 기타 정보를 상기할 수 있고, 주석 검색을 통해 신속한 문서 검색이 가능하다는 이점을 가지고 있다.



(그림 2) 검색 요청

2.2.2. 검색기

검색기는 예약 관리기에서 사용자의 질의를 받아 웹 검색엔진들을 통해 문서들을 수집하게 된다. 일반적인 논문 옵션으로 검색 할 경우 검색 대상은 구글(Google), 야후(Yahoo), 네이버(Naver) 등의 검색엔진을 통해 논문을 검색하게 되고, 링크 옵션으로 검색 할 경우 탐색 대상은 사용자가 요청한 링크로 한정 된다.

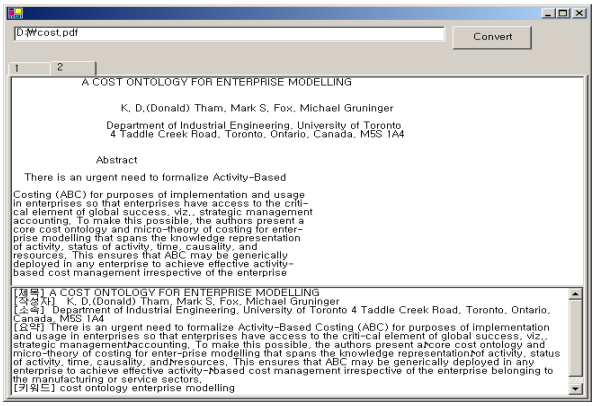
2.2.3. 분석기

분석기는 검색기에 의해 검색된 문서 데이터를 텍스트로 변환한 후 문서의 제목, 작성자, 요약, 키워드 등의 정보를 분리해내게 된다. 여기서 분리된 데이터들은 검색에 필요한 키워드로 이용한다. 이를 위한 텍스트 데이터에 대한 전처리 과정은 다음과 같다.

① 다단 제거
② 단락 나누기
③ 키워드 생성

변환된 논문 텍스트 데이터가 다단으로 편집되어 있을 경우 텍스트의 다단을 제거하는 과정을 거치게 된다. 이 과정을 통해 변환된 텍스트는 주제, 작성자, 요약, 키워드 등의 정보를 추출하기 위해 텍스트의 단락을 구분지어 잘라내게 된다. 이 데이터를 바탕으로 주제, 작성자, 요약, 키워드, 주석들의 정보를 추출하게 되며, 논문 내 키워드 정보가 없을 경우 키워드를 요약과 결론을 토대로 생성해 내게 된다.

그림 3은 원본 텍스트로부터 변환된 텍스트와 추출된 문서 정보를 보여주고 있다.



(그림 3) 분석기를 거친 후 텍스트

2.2.4. 데이터 관리기

데이터 관리기는 검색된 문서들의 정보와 파일을 관리하는 것으로 검색된 문서에 대한 추가, 삭제, 검색 등의 기능을 수행한다. 검색된 문서 정보는 XML로 저장되어 관리된다. <표 1>은 문서 정보가 저장되는 XML 파일의 형태을 보여준다.

```
<?xml version="1.0" encoding="utf-8" ?>
<Documents>
        <Paper Key="파일해시키" Type="PDF">
                <Title>제목</Title>
                <Author>저자</Author>
                <Department>소속</Department>
                <Abstract>요약</Abstract>
                <Keyword>키워드</Keyword>
                <Head></Head>
                <FilePath>파일경로</FilePath>
                <Link>원본경로</Link>
        </Paper>
</Documents>
```

<표 1> 검색된 문서 정보

2.2.5. 로컬 데이터 검색 서비스

로컬 데이터 검색 서비스는 사용자의 질의에 따라 데이터 관리기에 의해 관리되고 있는 문서 정보를 검색하여 사용자의 질의에 맞는 문서만을 보여주게 된다.

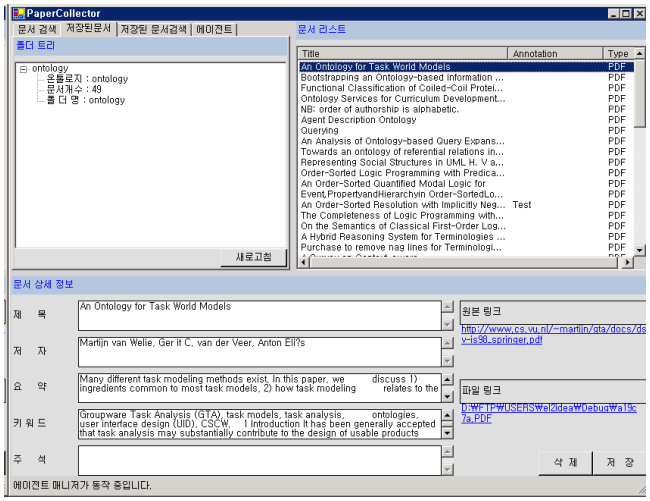
3. 결과

본 논문에서는 문서수집 에이전트를 이용하여 보다 정확하고 신속하게 논문 검색과 수집 및 필터링을 할 수 있는 시스템을 구현하였다.

1. 논문 수집 검색기

논문 수집 검색기는 사용자가 요청한 질의에 따라 검색엔진 혹은 요청한 사이트에서 논문을 검색하며, 검색된 논문을 분석하여 검색 가능한 형태로 변환하여 XML의 형태로 저장된다. 저장되는 데이터는 한글과 특수문자가 포함될 수 있기 때문에 인코딩을 하여 저장된다.

저장된 논문은 실제 검색에서 이용될 수 있도록 분석된 키워드 정보를 담고 있다. 여기서 분석된 논문 정보들이 사용자의 질의에 이용된다. 그림 5는 수집된 논문들의 리스트를 검색어 별로 브라우징하는 화면이다.

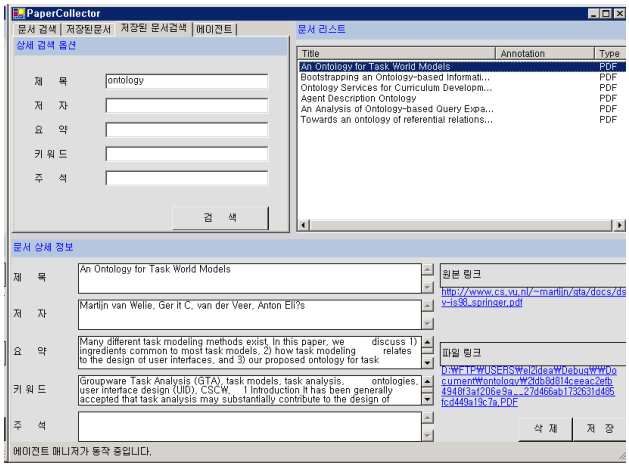


(그림 5) 검색어별 문서 수집 결과

2. 로컬 데이터 검색

로컬 데이터 검색은 수집/저장된 논문들의 풀에서 찾고자 하는 데이터를 검색하게 되는데, 검색 시 옵션으로는 주제, 저자, 요약, 키워드, 주석이 이용될 수 있다. 이러한 논문 속성에 대해 개별적으로 검색을 수행하기 때문에 단순키워드 검색에 비해 훨씬 정확한 검색 결과를 제공한다.

그림 6은 로컬 디스크에 수집된 문서 중 주제가 ontology인 것에 대한 문서들을 보여주고 있다. 검색이 끝나면 주어진 옵션들에 해당하는 논문 리스트가 오른쪽 화면에 나타나게 된다.



(그림 6) 로컬 문서 검색

3. 결론 및 고찰

본 논문에서 구현한 문서수집 및 필터링 에이전트는 논문이라는 특정 주제에 대한 구조적 분석을 통해 사용자에게 최적화된 결과를 제공할 수 있도록 설계 되었다. 웹 검색 엔진의 최대 단점은 키워드 매칭을 통해 부정확한 많은 양의 데이터를 사용자에게 제공한다는데 있다. 사용자 입장에서는 검색 엔진을 통해 제공된 방대한 양의 문서를 다시 읽고 분석하여 필터링해야 하는 불편함이 따른다. 이와 같은 비효율성을 에이전트를 통한 문서 수집과 구조적 분석을 통한 필터링으로 문서 수집의 비효율성을 극복하였다. 논문이라는 문서가 가지는 구조적 특성을 통한 검색은 사용자에게 보다 정확한 문서 검색을 가능하게 하여 검색과 분석, 그리고 필터링에 소요되는 시간과 노력을 줄 일 수 있다.

본 연구에서는 논문이라는 특수한 형식을 중심으로 연구하였다. 하지만 이 방법은 정형화 또는 부분적으로 정형화된 다른 형태의 문서들에 대해 적용 가능하다. 현재 구현된 시스템은 PDF 문서에 대한 수집만을 지원하고 있지만 다양한 포맷의 문서에 대한 확장이 가능할 것이다. 나아가 수집된 문서에 대한 자동 분류나 정보추출 등의 다양한 연구가 이루어질 수 있을 것이다.

감사의 글

본 연구는 산업자원부 지방기술혁신사업 (RTI04-01-01) 지원으로 수행되었음.

참고문헌

1. Casasola, E., Gauch, S., Intelligent Information Agents for the World Wide Web, Technical Report ITTC-FY97-TR-11100-02, 1997.

2. Luck, M., McBurney, P., Shehory, O., Willmott, M., Agent Technology Roadmap: Overview and Consultation Report, AgentLink, 2004.

3. Curran, K., Murphy, C., Annesley, S., "Web Intelligent in Information Retrieval", Information Technology Journal 3(2): 196-201, 2004.

4. Fan, Y., Gauch, S., "Adaptive Agents for Information Gathering from Multiple, Distributed Information Sources", Proc. of 1999 AAAI Symposium on Intelligents in Cyberspace, 1999.

5. Wagner, T., Phelps, J., Qian, Y., Albert, E. "A Modified Architecture for Constructing Real-Time Information Gathering Agents", Proc. of Agent-Oriented Information Systems, 2001.

6. Kushmerick, N., Thomas, B. Adaptive information extraction: Core technologies for information agents. Lecture Notes in Computer Science, 2586, 2003.

7. van Breemen, A., Feijs, L., "Architecture Evaluation of an Agent-Based Music Gathering Application", Proc. of the AAMAS-2003, 2003.