개인화된 학술 DB 구축을 위한 에이전트 시스템
An Agent System for Constructing Personalized Research Data Base

권혁찬, 최숙영*, 유관종
충남대학교 컴퓨터과학과
우석대학교 컴퓨터교육과*
Hyeok-Chan Kwon, Sook-Young Choi*, Kwan-Jong Yoo
Dept. of Computer Science, ChungNam National University
Dept. of computer education, Woosuk university*

요 약

인터넷의 발달과 더불어 매일같이 제공되는 수많은 정보로부터 자신에게 필요한 정보만을 추출하는 데는 많은 시간과 노력이 소요된다. 이러한 정보 가운데 한가지로 메일링 리스트(mailing list)로 부터의 전자우편을 들 수 있다. 메일링 리스트의 주된 목적은 정보교류이며 많은 경우 학술적인 자원을 제공하는 URL을 소개하고 있다. 본 논문에서는 사용자를 대행하여 메일링 리스트로부터의 메일을 분석하고 메일내의 URL을 추적하여 사용자의 기호에 맞는 논문과 논문에 관한 정보를 수집하여 개인화된 학술 DB를 구축해주는 에이전트 시스템을 제안한다. 사용자는 사용자 인터페이스 에이전트를 이용하여 개인가상 도서관의 자료를 열람할 수 있으며 검색, 삭제 등의 작업과 수집된 자료들에 대한 피드백을 줄 수 있다.

1. 서론

인터넷의 발달로 매일같이 수많은 정보가 사용자에게 제공된다. 그러나 폭팔적으로 증가하는 데이터 때문에, 그 가운데서 자신에게 필요한 정보를 추출해 내는데는 많은 시간과 노력이 요구된다[1]. 매일같이 사용자에게 제공되는 정보가운데 한 가지로 전자우편(electronic mail)을 들 수 있다. 전자우편에 의한 정보 제공처로서 뉴스 그룹, 메일링 리스트, 인터넷 쇼핑 몰 등이 있다. 이러한 메일을 일일이 읽고 분류하여 필요한 정보를 생성하는 작업은 사용자에게 많은 시간과 노력을 요하는 부담이 되는 작업이다[3,5].

위의 언급된 문제들에 대한 해결 방안으로서 에이전트에 기반한 정보 필터링(information filtering), 정보 검색(information retrieval) 시스템들이 제안되었고 개발되었다.[3] 정보 필터링 시스템은 주로 전자우편에 관한 것으로서 메일을 분류하고 사용자의 기호에 따라 우선순위를 부여하는 기능을 갖는다. 정보 검색 시스템은 주로 사용자의 요구를 질의(query)로 받아 해당하는 정보를 보유하고 있는 웹 사이트(web site)를 추천해 주는 기능을 갖는다. 이러한 기능의 검색엔진은 현재 매우 보편화 되어 있다.

인터넷상에는 주제별로 많은 메일링 리스트가 존재한다. 메일링 리스트의 주된 목적은 정보교류이며 많은 경우 학술적인 자원을 제공하는 URL을 소개하고 있다. 일반적으로 사용자는 메일링 리스트로부터의 메일을 읽고 그 안의 하이퍼링크를 추적하고 자신의 관심이 있는 논문에 대한 정보를 수집하고 다운로딩(downloading) 하는 등의 작업을 수행하게 된다. 본 논문에서 제안하는 에이전트 시스템은

이러한 사용자의 작업을 대행하여 웹상에서 제공된 논문들로 이루어진 학술 DB를 구축한다. 본 논문의 에이전트 시스템은 사용자를 대행하여 메일링 리스트로 부터의 메일을 분석하고 메일내의 URL을 추적하여 사용자의 기호에 맞는 논문에 대한 정보 와 논문을 수집한다. 사용자는 사용자 인터페이스 에이전트를 이용하여 개인가상 도서관의 자료를 열람할 수 있으며 검색, 삭제등의 작업과 수집된 자료들에 대한 피드백을 줄 수 있다.

본 연구의 결과를 통해 사용자는 자신이 필요한 논문을 추출하기 위한 시간과 노력을 요하는 작업을 줄일 수 있게 된다.

본 논문의 2장에서는 관련연구로서 기존의 정보 필터링 시스템과 정보 검색 시스템을 조사하였다. 3장에서는 에이전트 시스템의 구조에 대해 다루며 4장에서는 알고리즘을 설명하고, 5장에서는 구현에 대해 설명한다. 그리고 마지막 6장에서 결론 및 향후 연구방향을 다룬다.

2. 관련 연구

2.1 정보 필터링 시스템과 정보 검색 시스템

많은 정보 필터링 시스템들이 제안되어 왔다. 그 가운데 한 가지로 Maxims[3]을 들 수 있다. Maxims는 메모리 기반 추론(memory-based reasoning) 방식의 학습방법을 이용하여 전자우편과 관련한 사용자의 행위를 분석하여 학습하는 에이전트 시스템이다. 사용자가 어떤 행위를 하면 에이전트는 모든 situation-pattern 쌍을 기억한다. 각각의 도착한 메일에 대해 에이전트는 전송자, 수신자, Cc 리스트, subject의 키워드, 메일이 읽혔는지, 답장하였는지의 여부 등을 기억한다. 이러한 정보를 갖고 학습을 하며, 사용자의 행위를 분석한 결과를 토대로, 각각의 메일에 우선순위를 부여하고, 삭제, 정렬, 전달, 보관 등의 작업을 대행한다.

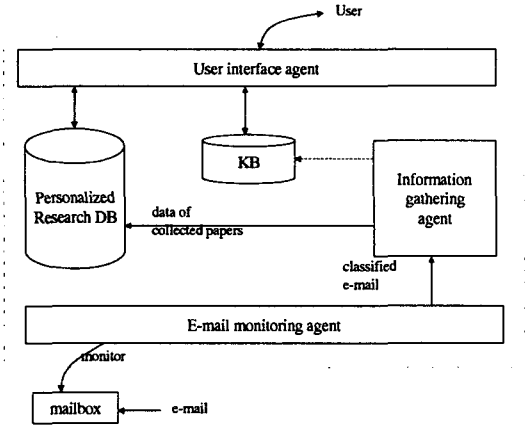
Usenet News를 필터링 해주기 위한 시스템으로 NewT[6]가 있다. 사용자가 선택한 기사와 선택하지 않은 기사를 분석하여 관련 워드를 추출한다. 사용자가 선택한 기사의 경우에는 저자, 기사가 실린 잡지 명, 인덱스, 제목 등의 정보를 보관한다. 이러한 정보들을 학습한 결과를 토대로 새로운 기사들이 도착하면, 그 가운데 사용자의 관심이 있을 만한 기사를 추천하여 주는 시스템이다.

정보 검색 시스템은 주로 사용자가 입력한 질의에 대해 웹을 검색하여 관련 사이트를 추천해 주는 시스템이 주류를 이룬다. 그러한 서비스를 제공해 주는 사이트로 국내의 네이버, 야후, 라이코스, 심마니 등과 외국의 Yahoo, Altavista, Lycos, WebCrawler 등이 있다.

Webdoggie[4]시스템은 Web을 검색하여 사용자의 기호에 맞는 site를 추천하는 기능을 갖는다.

3. 에이전트 시스템의 구조

정보 수집 에이전트 시스템의 기본 구조는 [그림 1]과 같다. 각각에 대한 설명은 다음과 같다.



[그림 45] 에이전트 시스템의 구조

▶ Email Monitoring 에이전트

데몬 형태로 메일을 기다리다가 새로운 메일이 도착하면 메일의 헤더를 분석하여, 수신자가 지정된 메일링 리스트이면 그 메일을 Information Gathering agent에게 전달한다.

▶ Information Gathering 에이전트

실제 정보 수집을 하는 에이전트로서 웹 페이지 분석, URL 추적, 논문 다운로딩 등의 기능을 갖는다. 논문이 웹에 링크로 연결된 경우엔 논문을 다운로드하여 보관하고 그렇지 않은 경우에는 논문의 정보, 즉 논문 제목, 저자, 잡지명, 출판년도, URL 등의 정보를 저장한다. 사용자의 관심도는 논문의 제목과 요약문을 분석하여 결정하며, 자세한 방법은 4장에서 설명한다. 또한 정보 수집 과정에서 이전의 방문 정보를 유지하므로 동일한 사이트를 반복해서 방문하지 않도록 하며, 전송받은 논문의 정보도 유지하여 데이터의 중복을 피한다. 웹으로부터 수집한 정보는 개인 가상도서관에 보관하며 정보수집을 위해 Knowledge base의 정보를 참조한다. 또한 사용자의 연구분야에 해당하는 CFP (Call-for-paper)에 대한 메일도 수집하여 DB에 보관한다. 또한 이전에 방문했던 사이트 중 갱신 또는 추가된 정보가 있을 수 있으므로 주기적으로 재방문하여 추가 정보를 수집한다.

▶ User Interface 에이전트

사용자와 개인 가상도서관과의 인터페이스를 담당한다. 사용자로부터 메일링 리스트를 등록받고 필요한 정보를 입력받으며, 사용자에게 수집된 논문의 정보를 보여준다. 사용자는 user interface 에이전트를 통해 논문의 검색, 삭제등의 작업과 검색된 논문에 대한 피드백을 줄 수 있다. 그리고 새로이 갱신된 - 사용자가 아직 보지 못한 - 정보를 제공한다. 또한 사용자의 기호가 변경되었을 때 그에 대한 변경 사항을 입력받아 시스템에 반영한다.

▶ 개인화된 학술 DB

개인 가상 도서관은 수집된 논문에 대한 정보, 실제 전송받은 논문, Call-for-paper 등의 기타 정보로 구성된다. Information Gathering 에이전트에 의해 자동으로 생성되고, 갱신된다.

▶ Knowledge base

메일과 웹 페이지를 분석하기 위한 정보 - 사용자의 피드백을 반영한 - 와 search rule에 대한 정보로 구성되어 있다.

4. 알고리즘

초기에 사용자로부터 새로운 메일링 리스트를 등록하기 위해 입력받는 내용은 다음과 같다.

- 메일링 리스트의 주소
- 정보 수집 에이전트가 탐색할 내부 링크의 깊이
- 정보 수집 에이전트가 탐색할 외부 링크의 깊이
- 논문을 자동으로 다운받을 것인지, 또는 논문에 대한 정보만을 보관할 것인지의 여부
- 사용자의 관심 분야를 나타낼 수 있는 두 종류의 관심 키워드
  - Required(RK.) : 사용자의 기호에 의한 키워드
  - Not(NK.) : 논문의 제목에 포함되어서는

[표 16] 알고리즘

```
1. 도착한 메일이 지정된 메일링 리스트로부터 온 것이면 Information Gathering 에이전트에게 전달한다.
   (Email Monitoring&Filtering 에이전트)
2. 전달 받은 메일을 분석한다. (Information Gathering 에이전트)
  a. 참고문헌 형식의 문장이 있으면
      사용자가 관심이 있는 문헌이면
          => 해당문장을 추출하여 보관하고 link로 연결이 되어 있는지 검사한다.
          Link로 연결이 되어 있으면
              => link로 연결된 해당 파일을 download 받아 메일링 리스트 별로 분류하여 보관한다.

      사용자가 관심이 있는 문장이 아니면
          => link는 따라가지 않고 그 정보만 보관한다.
   b. 메일 내에 URL address가 있으면
       있으면 => 해당 page를 추출하여 조사한다.
```

안되는 키워드

· 중요도(Weight) : 사용자의 기호 정도를 나타내는 것으로 각각의 RK.에 부여 된다.

· 기준 점수 : 논문이 사용자의 기호에 맞는지를 판단하기 위한 기준 점수.

4.1 논문의 선호도 결정

논문의 선호도는 논문의 제목에 포함된 관심 키워드들의 중요도의 합으로 결정한다. 이 값이 기준 점수를 초과하고, 논문의 제목에 NK. 키워드가 포함이 안된 논문이 선택되어 저장된다. 만약 초기 중요도가 (RK) agent : 15, mobile : 3, model : 3, performance : 3 이고 (NK) QoS 이고, 기준점수가 15점이라는 가정 하에 검색된 논문의 제목과 그에 대한 처리는 다음과 같다.

- "A performance model for agent system",

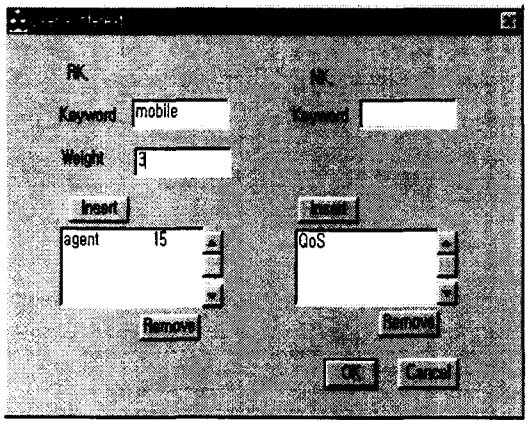
결과 : accept (선호도 : 21점)

- Tunnel Agent for Enhanced Internet QoS"

결과 : reject (NK인 QoS가 포함)

- "A Mobile code language"

결과 : reject (선호도 : 3)



[그림 46] 선호도 편집 도구

[그림 2]에서는 사용자의 선호도 편집 도구의 모습을 보인다. [표 1]은 전체 시스템의 알고리즘이며, [표 2]는 각각의 선택된 논문에 대해 저장하는 정보이다.

[표 17] 저장되는 논문의 정보

ID	고유번호
URL	논문을 받아온 곳의 URL
날자	논문을 받아온 날자
저장여부	논문을 저장하였는지 관련 정보만 저장했는지의 여부
선호도	에이전트 시스템이 부여한 선호도 점수
기타 정보	논문의 저자, 게제 잡지명 등의 정보

4.2 선호도 훈련(training)

사용자는 에이전트 시스템에 피드백(feedback)을 줌으로서 사용자의 선호도를 training할 수 있다. 다음의 세 종류로 수집된 논문에 대한 피드백을 줄 수 있다.

- good : 사용자의 선호도에 정확히 맞는 논문
- satisfactory : 선호도에 적당히 맞는 논문
- bad : 사용자 관심 밖의 논문

에이전트 시스템은 두 가지 방법의 훈련을 사용한다.

(1) Manual training

: 사용자가 에이전트 시스템의 도움 없이 직접 선호도를 변경한다. 선호도 편집 도구에서 키워드들을 추가하고, 제거하고, 변경하는 등의 작업을 수행한다.

(2) Active trainging

: 에이전트 시스템이 자동으로 사용자의 선호도를 변경한다. 먼저, 선택된 논문들에 포함된 모든 키워드 - 관사와 정관사는 제외 - 에 대해 키워드 점수를 부여한다. 키워드 점수를 부여하기 위한 함수 [표3]과 같다.

[표 18] 키워드 점수 부여 방법

$AV(k) = (F_G(P_G,k)+F_S(P_S,k)*P)*100/N$

$F_G(P_G,k)$: 'Good'으로 평가된 논문의 제목들에서 키워드 k가 포함된 횟수

$F_S(P_S,k)$: 'Satisfactory'로 평가된 논문의 제목들에서 키워드 k가 포함된 횟수

P : 'Satisfactory'로 평가된 논문의 중요도

N : 전체 수집된 논문의 수

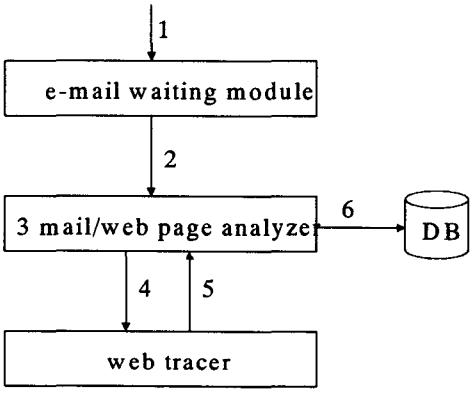
에이전트 시스템은 [표 3]의 함수를 적용해

높은 키워드 점수를 획득한 키워드를 사용자에게 관심 키워드에 추가하도록 추천하여 주거나, 또는 에이전트가 해당 키워드를 직접 RK. 키워드 리스트에 추가하고 적당한 중요도를 부가한다. 이는 사용자가 초기에 선택한 옵션에 따른다. 중요도는 획득한 키워드 값에 의해 달리 부여된다.

5장 구현

본 논문의 에이전트 시스템은 웹 페이지 분석 단계에서 먼저, hyperlink를 찾는다. html 문서의 <a>와 태그를 찾는 방법으로 이러한 작업을 수행한다. 그리고 나머지 html 코드 부분은 제거한다.

에이전트 시스템은 크게 Front end와 Back end로 구성된다. 전자우편을 모니터링 하는 부분과 사용자가 DB를 열람하면서 주는 피드백을 KB에 반영하는 부분이 Front end에 해당하며, 메일링 리스트로부터의 메일과 웹 페이지를 분석하고, 웹의 리소스를 추적하고 해당 자료를 전송받는 부분이 Back end에 해당한다. [그림 3]에서는 정보 수집 과정에서의 주된 event를 보여준다.



[그림 47] 수행과정

[그림 3]에서 각각의 수행 과정은 다음과 같다.

(1) e-mail 도착
(2) e-mail waiting module은 도착한 메일이 등록된 메일링 리스트의 것이면 mail/web page analyzer로 보낸다.
(3) mail/web page analyzer는 mail이나 web page를 분석하고
(4) web tracer에게 web page를 요청하고,
(5) web tracer는 요구된 페이지를 전송해준다.
(6) mail/web page analyzer는 사용자의 선호도에 맞는 논문을 개인화된 학술 DB에 저장한다.

6장 결론

본 논문에서는 WWW(world-wide web)으로부터 사용자의 개인화된 학술 DB 구축을 위한 에이전트 시스템을 제안하였다. 제안한 에이전트 시스템을 통하여 사용자는 메일링 리스트로 부터의 메일을 읽고 분류하고, 필요한 자료를 얻기 위해 내포된 URL을 일일이 추적하는 등의 시간과 노력을 요하는 작업을 줄일 수 있게 된다.

시스템의 구현은 현재 진행 중에 있다. 추후 과제로서 사용자의 기호를 학습하는 방법, 사용자의 피드백을 반영하기 위한 메카니즘 등에 대한 연구가 필요하다.

참고문헌

[1] Etziono, O., D.Weld, "Intelligent Agents on the Internet : Fact, Fiction, and Forecast," IEEE Expert, Aug. 1995, pp.44-49.

[2] Lieberman, H., Neil W., Van D., Adriana S. V.,"Let's Browse: A Collaborative Browsing Agent," Knowledge-Based Systems, Vol. 12, Dec. 1999, pp. 427-431.

[3] Williams, J., "Bots and Other Internet Beasties", SamsNet, pp. 257-263, 1996

[4] Webdoggie, A personalized WWW document filtering system, http://webhound.www.media.mit.edu/projects/webhound/

[5] Maes P., "Agents that reduce Work and Information Overload," CACM, vol. 37, no. 7, pp.31-40, 146, ACM Press, July 1994

[6] Lashkari, Y., M. Mentral, and P. Maes,

"Collaborative Interface Agents," In: Proceedings of the National Conference on Artificial Intellegence, 1994

[7] Shardanand U., P.Maes, "Social Information Filtering : Algorithms for Automating 'Word of Mouth'," Proceedings of the CHI-95 Conference, Denver, CO, ACM Press, May 1995

[8] Sheth B., NEWT : A Learning Approach to Personalized Information Filtering, Master Thesis 1994