자바 기반 개인용 웹 정보 수집 에이전트의 구현

박민규, 한정기, 유태명, 김중섭, 최석민, 김준태
동국대학교 컴퓨터공학과

Implementation of Java-based Personal Web Information Gathering Agent

Minkyu Park, Junggee Han, Taemyung Yoo, Jungseob Kim, Sukmin Choi and Juntae Kim
Department of Computer Engineering, Dongguk University

요 약

본 논문에서는 웹에서 사용자의 취향에 부합하는 정보를 지속적으로 수집하여 추천해주는 지능적인 개인용 웹 정보 수집 에이전트의 구현에 대하여 기술한다. 본 논문에서 구현한 에이전트 시스템은 자바언어로 구현되었으며 인터넷에서 페이지를 모아오는 수집 단계, HTML 문서 색인 단계, 필터링 단계, 사용자 모니터링 단계 학습 단계 등 다섯 단계로 구성되어 있다. 웹 페이지는 기존의 검색 엔진으로부터 수집하도록 하였으며, 사용자의 관심에 부합되는 웹 페이지들을 추천하고, 추천된 페이지들에 대한 사용자의 행동을 모니터링하여 사용자의 취향을 학습함으로써 사용자 프로파일을 재구성한다. 본 웹 에이전트 시스템은 암시적인 피드백에 의한 학습을 수행하고 백그라운드에서 동작함으로써 사용자에게 기존의 검색 작업에 따른 시간과 수고를 덜어 주었다.

1. 서론

인터넷의 이용은 초기의 학술적인 분야에서 다양한 분야로 확대되고 있으며 인터넷 사용자의 수도 폭발적으로 증가하고 있다. 이런 상황 속에서 인터넷 사용자들은 웹에서 정보를 얻기 위해 검색 엔진(search engine)을 이용하게 되었다. 검색 엔진에서 정보를 찾기 위해서는 검색어와 연산자를 통해 사용자가 원하는 정보를 질의어로 표현하고 검색 결과인 관련 URL 리스트를 보면서 관련 있는 웹사이트들을 찾아가야 한다. 하지만 사용자가 원하는 정보를 찾기까지 시간과 인내를 필요하게 되고, 적절한 질의어를 선택하기 위하여 검색 결과에 따라 질의어를 수정하는 작업이 필요하게 된다. 이를 해결하기 위해 사용자를 대신하여 사용자의 취향에 부합하는 웹 페이지들을 찾아주는 웹 에이전트에 관한 연구가 활발히 진행되고 있다[1, 2, 3].

본 논문에서는 사용자로부터 명시적 피드백을 받지 않고도 사용자의 관심을 학습하면서 사용자의 취향에 부합하는 웹 페이지들을 지속적으로 수집하여 추천해주는 지능적인 개인용 웹 에이전트에 구현에 대하여 기술한다. 본 웹 에이전트 시스템은 기존의 검색 엔진 사용에 따른 문제점을 해결하고, 정보 검색에 있어 사용자에게 편의성을 제공한다. 본 논문에서 구현한 개인용 웹 에이전트 시스템은 자바언어로 구현되었으며, 메타 검색 엔진 형태로 페이지를 모아오는 수집단계, 수집된 HTML 문서에서 중요한 단어를 추출하는 색인 단계, 문서의 순위를 정하는 필터링 단계, 사용자의 북마크의 추가/삭제를 감시하는 단계, 사용자의 취향을 학습함으로써 사용자의 프로파일을 재구성하는 학습 단계 등 다섯 단계로 구성되어 있다.

2. 웹 에이전트

웹 에이전트는 사용자를 대신하여 인터넷에 산재되어 있는 온라인 정보들에 대하여 정보검색(information retrieval)과 정보여과(information filtering)등을 수행하는 에이전트를 말한다.

웹 에이전트는 다양한 응용 분야에서 연구되고 있는데, 사용자의 취향을 학습하여 사용자가 원하는 정보를 검색 할 수 있도록 도와주는 지능형 에이전트로 Personal Webwatcher[1], InforFinder[2], Syskill&Webert[3] 등이 있으며, 협동 여과(collaborative filtering) 방식으로 정보를 수집하는 에이전트로는 Firefly[4], Siteseer[5] 등을 들 수 있고, 전자상거래를 위한 샵로봇(shopbot)인 BargainFinder[6], Shopbot[7] 등도 웹 에이전트라고 할 수 있다.

정보 여과는 인터넷에 산재되어 있는 정보 중에서 필요한 것과 필요하지 않은 것을 구분하는 것을 말하는데, 구분 기준이 되는 것이 사용자의 프로파일(profile)이다. 사용자의 프로파일은 사용자가 관심을 보인 웹 페이지에 포함된 중요 단어와 출현 빈도수로 나타낼 수 있다.

본 에이전트 시스템은 Syskill&Webert[1]와 비슷하게 동작하지만, 사용자의 취향 분석을 위해 북마크 모니터링에 의한 암시적인 피드백(implicit feedback)을 사용하여 사용자는 에이전트의 동작에 직접 관여하지 않아도 되도록 하였고, 적합성 피드백(relevance feedback)에 의한 프로파일 갱신으로 사용자의 취향을 학습하도록 하였다.

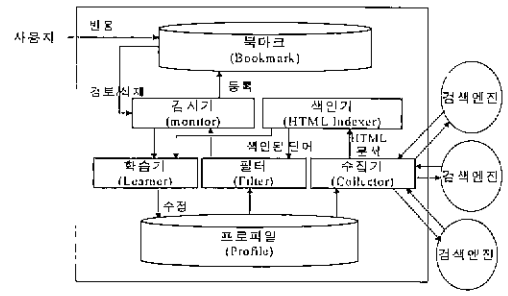
3. 개인용 웹 에이전트의 구현

본 웹 에이전트 목적은 사용자가 원하는 정보를 얻기 위해 서치엔진을 이용하는 것과 같이 직접 질의어를 작성, 수정하고 웹 페이지들을 찾아다니는 등의 작업을 할 필요 없이, 에이전트에 의해 수집되고 지속적으로 추천되는 페이지들을 볼 수 있도록 검색의 편의성을 제공하는데 있다.

본 에이전트의 동작 과정은 다음과 같다. 우선 사용자의 웹 브라우저 북마크로부터 초기 프로파일을 만들거나, 시작 URL이나 초기 질의어를 입력받아 프로파일을 만든다. 초기 프로파

일 자료를 바탕으로 해서 검색 엔진에 질의를 보내고 검색 결과 URL들을 이용하여 웹 페이지들을 가져온다. 가져온 페이지들로부터 단어들을 추출하고, 출현빈도수에 HTML 태그 정보를 이용한 가중치를 부여하여 페이지를 표현하며, 프로파일을 이용하여 사용자에 취향에 가까운 순서대로 페이지를 선정한 후 북마크의 '추천' 디렉토리에 이 페이지들을 추가한다. 이러한 과정이 완료한 후에 에이전트는 사용자에 의한 '추천' 디렉토리의 변경을 계속해서 감시하고 디렉토리의 내용이 변경되면 추가/삭제/이동된 페이지들을 이용하여 프로파일을 수정하여 다시 웹 페이지를 수집한다. 이러한 과정을 백그라운드에서 반복하면서 지속적으로 사용자의 취향에 부합하는 페이지들을 추천한다.

본 에이전트의 전체 구조는 [그림 1]과 같다.



[그림1] 개인용 웹 에이전트의 전체 구조

3.1 수집기(Collector)

수집기는 사용자 프로파일 자료를 바탕으로 대표적인 검색 엔진인 Yahoo, AltaVista, Yahoo Korea, AltaVista Korea에 OR형식으로 질의를 보낸 후, 검색결과 중 상위 100개의 웹 페이지를 가져온다.

수집기는 질의를 보내기 위하여 각 검색 엔진의 URL과 각 검색 엔진에 사용되는 연산(&, |)들로 구성된 질의 제작을 위한 검색 엔진 정보라는 데이터를 사용한다. 검색 엔진 정보를 사용하여 프로그램의 수정 없이 검색 엔진을 추가하거나 검색 엔진의 질의 형태 변화에 대처 할 수 있다. 이 부분은 또한 자바의 쓰레드 방식을 이용하여 인터넷의 시간 지연에 구애받지 않고 다른 루틴들이 작업을 할 수 있도록 구성되어졌으며, 한번 간 웹 페이지는 다시 방문하지 않도록 하였다.

3.2 HTML 색인기(Indexer)

색인기는 수집기에서 수집한 HTML 문서를 분석하여 각 웹 페이지를 문서내의 단어와 각 단어의 출현빈도수에 HTML 태그에 따른 가중치를 곱한 수치의 쌍들로 표현한다. 영문 부분은 불용어(stopword)를 제거한 후 간단한 스테밍(stemming) 작업을 하고, 사전을 비교해 색인어를 추출하였다. 한글 부분은 자바에서 유니코드를 사용하기 때문에 먼저 한글을 초성, 중성, 종성으로 분리하고 조사, 어미부분을 제거한 뒤 최장일치법으로 색인어를 추출하였다. 유니코드 한글표현은 조합될 수 있는 현대 한글 글자 모두를 가나다 순으로 정렬, 배치한 것으로[8], 코드값의 조합 방법은 다음과 같다.

코드값 = 0xAC00 + (초성값*21*28)+(중성값*28)+(종성값)

색인어들이 추출되면 각 단어에 대한 가중치를 계산하고 추출된 단어와 각 단어의 가중치를 hashtable에 저장한다. 실험적으로 HTML 태그 정보를 이용하기 위해 TITLE 태그에 대해서만 2.0의 가중치를 주고 구현하였다.

3.3 여과기(Filter)

여과기는 전 단계의 결과인 각각의 HTML 문서의 Term Frequency(TF)에 HTML 태그 가중치를 적용한 값을 가지고 사용자의 프로파일에 근거하여 사용자의 취향에 더 가까운 문서를 먼저 보여 줄 수 있게 순위를 정한다.

문서의 단어들 중에서 사용자의 프로파일에 들어 있는 단어만이 의미 있는 단어이므로 프로파일의 단어들만 추출해서 프로파일과 같은 형태의 벡터를 만든다. 그리고 프로파일과 문서와의 유사도를 다음과 같이 계산한다.

$$SIM(V_{profile}, V_{document}) = \frac{V_{profile} \cdot V_{document}}{|V_{profile}||V_{document}|}$$

$$= \frac{\sum_{di \in pv} tf(i)pw(i)}{\sum_{i \in pv} [tf(i)]^2 \sum_{di \in pv} [pw(i)]^2}$$

pw(i) profile 안에 i번째 단어의 weight

tf(i) 문서에서 profile안에 i번째 단어가 나온 빈도수

위의 식을 적용하여 수집한 각각의 HTML 문서에 대하여 프로파일과의 유사도를 계산한 다음, 유사도가 높은 20개의 문서를 선택하여 추천한다. 프로파일의 각 단어에는 가중치가 부여되어 있으므로 유사도 계산 결과에 따른 문서의 순위는 검색 엔진에서의 순위와는 차이가 있게된다.

3.4 감시기(Monitor)

감시기는 여과 단계에서 넘어온 웹 페이지들의 URL을 북마크에 '추천' 이라는 디렉토리에 추가하고, 사용자가 북마크 아이템을 추가/이동/삭제하는 행위를 모니터링하여 그 결과를 학습기에 넘겨준다. 이때 사용자가 '추천' 디렉토리에 있는 URL을 삭제했다면 그 URL을 사용자가 관심이 없는 URL로 간주하여 - 속성을 부여하고, '추천' 디렉토리에 있는 URL을 다른 곳으로 이동하거나 새로운 URL을 추가했다면 그 URL을 사용자가 관심이 있는 URL로 간주하여 + 속성을 부여한다.

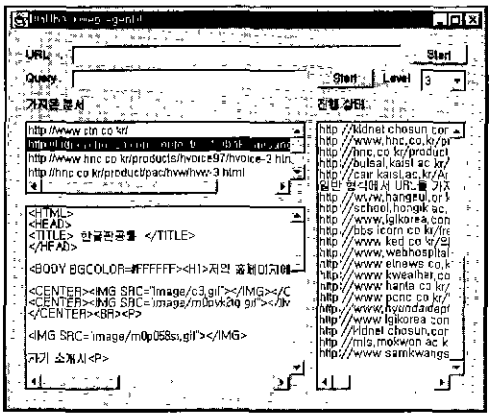
본 웹 에이전트는 백그라운드에서 사용자가 북마크 아이템을 변경하는 행위를 모니터링함으로써 사용자로부터 어떤 페이지에 관심이 있는가를 별도로 입력받는 과정이 없이 학습을 수행할 수 있다.

3.5 학습기(Learner)

학습기는 적합성 피드백을 이용한 프로파일 수정으로 프로파일에 대한 학습을 진행시킨다. 감시기로부터 전달받은 + 속성을 가진 URL들은 양성 피드백으로, - 속성을 가진 URL들은 음성 피드백으로 각각 처리하여 프로파일을 수정하게 된다. 프로파일의 수정은 사용자의 새로운 피드백이 발생하면 지금까지 발생한 모든 피드백들과 새 피드백을 합하여 평균을 내는 형식으로 진행된다.

4. 실험

본 에이전트는 윈도우즈 95 환경에서 자바 언어로 구현되었으며, 기본 사용자 인터페이스는 [그림2]와 같다. 에이전트는 초기 URL이나 질의어를 제시하여 동작시킬 수 있으며, 에이전트가 동작을 시작한 후에 사용자는 별도의 입력 없이 북마크의 추천 디렉토리만 보고 웹 페이지들을 삭제하거나 이동하면 된다. 필요한 경우에는 에이전트의 동작 상태를 볼 수 있다.



[그림2] 개인용 웹 정보 수집 에이전트 사용자 인터페이스

본 에이전트의 동작을 테스트하기 위해 사용자의 관심 분야가 '아래 한글' 관련 문서라고 가정하고 '한글'이란 단어를 초기 질의어로 하여 웹 페이지의 수집과 피드백에 의한 학습 등을 실험하였다. 본 실험에서는 Netscape Navigator를 기본 브라우저로 하여 북마크 '추천' 디렉토리에 추천된 페이지 중 관련 페이지를 사용자가 자신의 다른 디렉토리로 이동한 뒤 학습을 통한 프로파일의 변화와 추천되는 웹 페이지들의 변화를 관찰하였다.

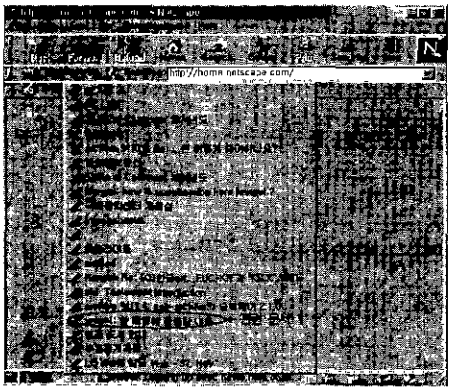
[그림3], [그림4], [그림5]가 이 실험에서의 에이전트의 동작 결과이다. 초기 프로파일은 <한글, 10.0>으로 단어와 가중치를 부여하였다. [그림3]은 초기 프로파일에 의해 수집된 웹 페이지들 중에서 추천된 페이지들을 나타낸다. 이중에서 관련 문서는 단 1개뿐이며 이 페이지를 이동하였을 때 에이전트를 이러한 변경 상황을 인식하여 학습을 수행하였다. 2회 학습 후, 그 결과 변경된 프로파일은 [그림4]와 같으며, 새로 추천된 페이지들이 추가된 북마크는 [그림5]와 같다. '한글'은 10.0에서 108.0으로 가중치가 증가되고, '아래', '한글과컴퓨터', '아래한글' 등과 같은 단어가 추가되었음을 볼 수 있다. 새로 추천된 페이지들 중에는 모두 9개의 페이지가 관련된 문서였다.

5. 결론

본 논문에서는 북마크 모니터링을 통하여 사용자의 취향을 명시적 피드백 없이 학습하고, 사용자에게 관심 있는 웹 페이지들을 지속적으로 추천하는 지능적인 개인용 웹 에이전트의 구현에 대하여 기술하였다.

간단한 실험을 통하여 본 에이전트가 암시적인 피드백에 의해 동작하며 학습에 따라 점차 관련된 페이지들을 보다 정확하게 추천할 수 있음을 보았다.

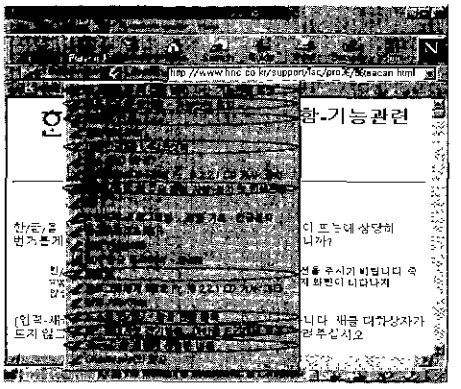
향후과제로는 복수의 주제에 대한 별도의 프로파일 관리에 대한 연구, 다양한 사용자 피드백 방법에 대한 연구, 검색 엔진의 홈페이지를 인덱싱하는 방법을 제어하는 메타 태그(META tag)인 'keywords'를 이용한 가중치 부여 방법 등에 대한 연구가 필요하며, 장기간 피드백에 따른 웹 에이전트의 학습에 대한 다양한 조건의 실험 및 분석이 수행되어야 한다.



[그림3] 초기 프로파일에 의해 추천된 웹 페이지

한글	108.0
파일	86.0
아래	73.0
한글과컴퓨터	3.0
아래한글	2.0

[그림4] 학습후 변경된 프로파일



[그림5] 학습 후 추가된 웹 페이지

참고문헌

[1] Dunja Mladenic, "Personal WebWatcher design and implementation", Carnegie Mellon University, 1996

[2] Bruce Krulwich, Chad Burkey, "The InfoFinder Agent' Learning User Interests through Heuris" IEEE Expert/Intelligent System & Their Application, Vol.12 No.5, September/October 1997

[3] Michael Pazzani, Jack Muramatsu & Daniel Billsus, "Syskill & Webert : Identifying interesting web sites",1996 AAAI

[4] Firefly, http://www firefly net

[5] James Rucker and Maros J Polanco, "Siteseer Personalized Navigation for the web", Cummunication Of ACM, March 1997/Vol 40, No 3

[6] BargainFinder, http://bf cstar ac com/bf

[7] Shopbot, http://www cs washington edu/research/shobot/

[8] Unicode, http //www unicode org/