사용자 프로파일 기반 개인 웹 에이전트
(User Profile based Personalized Web Agent)

소 영 준 † 박 영 택 ††
(Young-Jun So) (Young-Tack Park)

요 약 본 논문은 웹을 이용해 정보를 검색하는 사용자의 관심도를 사용자 프로파일로 구축하여 구체적이고 정확한 사용자 관심 정보를 제공하는 개인 웹 에이전트를 구축하는데 목적을 두고 있다. 사용자에게 웹 검색 행위를 감시하는 모니터 에이전트에 자신의 관심도를 직접 기술하여 관심문서 정보를 구축하고 이에 대한 정확도를 향상시키기 위한 여러 키워드 추출작업을 수행한다. 추출된 키워드는 학습서버의 작업에 의해 사용자별 프로파일을 생성하여 이를 사용자가 확인 및 편집할 수 있게 하였다. 본 논문에서 구현하고자 하는 웹 에이전트의 사용자 프로파일 구축작업에는 사용자 관심 문서 정보의 정확한 키워드 추출작업과 학습 작업이 매우 중요하다. 이에 본 논문에서는 키워드 추출에 적용되는 여러 가중치 설정작업에 대하여 중점적으로 다루며 적용된 귀납적 기계학습에 대하여 알아본다. 이로써 구축된 사용자 프로파일은 관심 문서를 검색하는데 적절한 정보를 제시한다. 이에 따라 사용자 프로파일을 본 웹 에이전트에서 구현한 사용자 적응형 웹 검색 에이전트와 사용자 적응형 푸쉬 에이전트에 적용하여 사용자에게 적합한 서비스를 제공한다.

Abstract This paper presents a personalized web agent that constructs user profile which consists of user preferences on the web and recommends his/her relevant information to the user. The personalized web agent consists of monitor agent, user profile construction agent, and user profile refinement agent. The monitor agent makes a user describe his/her preferences directly and it creates the database of preference document, finally performs several keyword extraction to increase the accuracy of the DB. The user profile construction agent transforms the extracted keywords into user profile that could be confirmed and edited by the user. and the refinement agent refines user profile by recursively learning and processing user feedback. In this paper, we describe the several keyword weighting and inductive learning techniques in detail. Finally, we describe the adaptive web retrieval and push agent that perform adaptive services to the user.

1. 서 론

WWW(World Wide Web)의 폭발적인 팽창으로 인하여 웹 상에서 제공되는 정보가 기하급수적으로 증대됨에 따라 웹 사용자가 자신이 원하는 정보를 정확하고 신속하게 검색하기에 많은 어려움이 있다. 이에 사용자에게 관심 정보를 제공해주는 시스템에 대한 요구가 있어왔다. 최근에 들어서 이 문제에 대해 정보검색과 기계학습 분야로의 접근이 시도되면서 지능형 웹 에이전트 시스템에 관한 연구가 진행되고 있는 추세이다[1].

지능형 웹 에이전트는 사용자에게 관심 문서를 수집, 제공한다[2][3][4]. 본 논문에서는 사용자 관심 영역의 특성을 추출하여 사용자 프로파일을 생성하고 이를 토대로 사용자에게 관심 문서를 제공하는 개인 웹 에이전트에 대해 서술한다. 사용자가 브라우징한 문서들을 대상으로 관심도를 표시하면 키워드 추출기법과 특성 추출 기법을 적용하여 중요 키워드를 추출한다. 이 과정에서 관심영역에 속하는 키워드에 가중치를 설정하여 중요도를 높였다. 추출된 다량의 키워드는 귀납적 기계 학습 시스템을 이용하여 결정 트리 형태로 최적화 되고 소수의 중요 키워드를 추출하는데 이용되어 사용자 프

· 본 연구는 한국과학재단 핵심전문연구(971-0901-009-2)지원으로 수행되었음.
† 학생회원 : 숭실대학교 컴퓨터학부
so@multi.soongsil.ac.kr
†† 종신회원 : 숭실대학교 컴퓨터학부 교수
park@multi.soongsil.ac.kr
논문접수 : 1999년 5월 3일
심사완료 : 2000년 1월 6일

로파일을 구성한다. 이 작업에서 각 관심영역의 결정 트리를 구성하는 여러 키워드 구조를 추출하기 위하여 재귀적 학습 작업을 반복한다.

이를 통해 생성된 사용자 프로파일은 개별적인 서비스 제공을 목적으로 하는 사용자 적응형 검색 에이전트와 사용자 적응형 푸쉬 에이전트에 이용되어 검색 질의어 확장과 관심문서 푸쉬 서비스를 제공한다. 특히 푸쉬된 관심 정보중 사용자의 관심도에 벗어난 정보는 사용자 피드백에 의한 재학습으로 프로파일의 여러 번에 걸쳐 수정 보완되도록 하였다.

2장에서는 관련연구에 대하여 알아본다. 3장에서는 시스템을 구성하고 있는 각 모듈에 대해 소개한다. 4장에서는 사용자가 설정한 각 관심영역별 관심 문서 내에서 중요 키워드를 추출하는 작업에 대하여 설명하며, 5장에서는 추출된 키워드를 대상으로 수행되는 학습작업과 그 결과 구축되는 사용자 프로파일에 대해 살펴본다. 6장에서는 제공된 푸쉬 정보에 대한 사용자 피드백정보로 인하여 사용자 프로파일이 재 구축되는 작업을 설명한 후, 7장에서는 생성된 프로파일의 정확도 실험 및 평가를 하고 마지막으로 8장에서는 결론을 맺는다.

2. 관련연구

현재 웹을 통해서 정보를 찾고자하는 사용자를 위한 정보검색 서비스를 위해서 여러 가지 형태의 에이전트 시스템이 개발, 상용화되고 있다. 그중 본 논문에서 구현하고자 하는 에이전트 시스템과 유사한 구조를 갖는 2개의 웹 에이전트에 대해서 알아보고 이것들에 적용된 기술들을 중심으로 시스템 구조를 알아보기로 한다.

카네기 멜론 대학(CMU)에서 구축한 웹 에이전트인 Personal Webwatcher는 자동으로 모니터링된 사용자 브라우징 행위 정보를 기반으로 관심 문서를 추측한다. 이는 비 감독 학습(Unsupervised Learning)방식의 에이전트로 사용자 관심문서의 전처리 작업과 학습 작업에는 추출된 키워드를 벡터 테이블로 생성 이를 기반으로 TFIDF 및 베이지안 확률(Bayesian Probability)을 적용하여 사용자 프로파일이 구축된다[6]. 이는 사용자 관심 문서를 시스템이 자동으로 추측하기 때문에 사용자에게 편리성을 준다는 이점이 있지만 학습에 입력되는 관심 정보의 신뢰성이 낮다는 단점이 있다. 이에 비해 본 논문에서는 사용자의 직접적인 관심문서 입력에 의해 보다 정확한 학습 입력 값이 구축된다. 또한 학습 과정에서도 특성 추출 방식과 결정 트리를 이용한 소량의 대표 키워드 추출 작업으로 인해 더욱 정확한 사용자 프로파일을 구축한다.

이밖에도 앤더슨 컨설팅 랩에서 개발한 웹 에이전트인 InfoFinder는 사용자의 브라우징 과정과 이에 대한 관심도를 기반으로 학습이 이루어진다는 점에서는 Personal Webwatcher와 동일한 구조를 갖는다[12]. 그러나 사용자의 관심을 학습하는 방식에서 큰 차이점을 갖고있다. 즉, Personal Webwatcher는 사용자의 관심도를 사용자는 보는 문서에 대한 반복횟수에 따라 일방적으로 결정한다. 이에 반해, InfoFinder는 사용자가 직접 자신의 관심도를 표현할 수 있는 감독 학습(Supervised Learning)을 사용하고 있다[11].

사용자는 InfoFinder 자체 브라우저를 사용하게 되고 이를 이용해서 사용자 자신의 관심문서를 직접 에이전트 시스템에 표시하게 된다. InfoFinder는 관심 영역내의 중요 문서에 대한 키워드를 단순한 빈번도 위주로 추출하였다. 이에 비해 본 논문에서는 문서 특성 추출작업으로 관심영역의 특성을 고려한 키워드 추출작업을 적용, 해당 관심영역을 대표하는 중요 키워드의 정확도를 높였으며 영역 특성에 따른 가중치 설정 작업으로 추출되는 키워드의 중요도를 더욱 높였다. InfoFinder는 본 논문에서 적용한 것과 마찬가지로 귀납적 기계학습을 통해 관심문서를 학습하여 결정 트리를 생성하고 이를 토대로 사용자 프로파일을 생성한다.

InfoFinder와 위에서 소개한 Personal Webwatcher는 모두 관심문서에 대해 학습을 한번만 수행하는데 비해 본 논문에서는 재귀적 방식으로 학습이 이루어진다. 이는 영역 분류에 커다란 비중을 가지는 키워드로 인하여 결정 트리에 나타나지 않을 수 있는 중요 키워드가 추출될 수 있게 한다.

3. 시스템 구성

본 논문에서 구현한 사용자 적응형 웹 에이전트는 모니터 모듈과 문서 전처리를 통한 키워드 추출 모듈, 관심영역 별 키워드를 대상으로 사용자 프로파일을 구축하는 학습 에이전트, 푸쉬 정보의 비관심 피드백을 수용하여 프로파일을 수정하는 재학습 모듈, 프로파일을 적용하여 사용자에게 적합한 서비스를 제공하는 적응형 에이전트 모듈로 구성되어 있다. [그림1]은 웹 에이전트의 전체적인 시스템을 나타낸 것이다.

먼저 웹 에이전트는 클라이언트의 사용자 브라우징 행위를 모니터한 결과를 서버에 위치한 학습 모듈에 전달한다. 본 논문에서는 사용자가 브라우저를 통해 행하는 모든 요청/응답 정보를 인식할 수 있게 하는 기능을 갖고 있는 프락시 서버를 이용하여 사용자의 행위를 모니터한다. 사용자는 먼저 자신의 관심 정보를 특정 관심

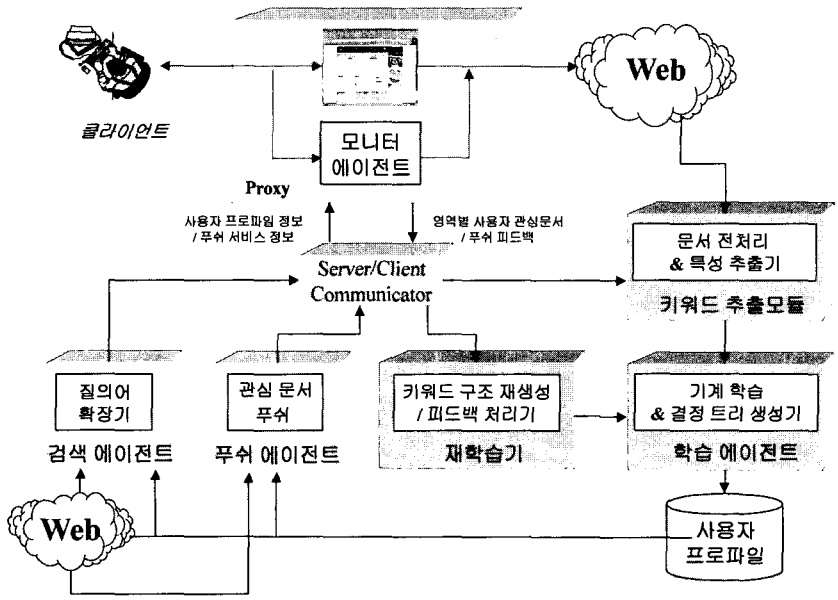

그림 1 웹 에이전트 시스템 구조도

영역을 생성한 후 이에 저장한다. 생성된 관심 영역들은 학습될 정보가 저장된 공간이며 사용자가 임의로 생성 삭제할 수 있다.

문서 전처리 과정을 통해 사용자가 정의한 각 관심영역들을 구성하는 관심문서에서 중요 키워드만을 먼저 추출한다. 추출된 키워드들은 학습 모듈을 통해 각 관심영역의 문서들을 대상으로 특성 추출 작업을 통해 영역을 분류하는 특성을 표현하는 중요 키워드들을 추출한다. 다량의 해당 키워드들은 학습 작업을 통하여 관심영역을 표현하는 각 키워드간의 연관 관계를 갖고있는 사용자 프로파일을 구성한다. 중요 키워드간의 관계를 나타내기 위해 본 논문에서는 학습작업에 귀납적 기계학습 시스템을 적용하였다. 이를 통해 사용자 프로파일은 중요 키워드들을 결정 트리 구조로써 표현하게 된다. 결정 트리의 표현으로 인해 단순히 다량으로 나열된 키워드들로 구성되었던 초기 중요 키워드들은 결정 트리를 구성하는 최상위 노드를 중심으로 일정한 규칙 형태로 관심영역을 표현하게 된다.

그러나 최상위 노드가 갖는 영역의 분류도가 아주 높을 경우 해당 노드의 키워드만을 중심으로 하는 편중된 키워드 구조가 나올 가능성이 있다. 그러므로 본 논문에서는 다양한 키워드 결정 트리 구조를 얻기 위해 학습의 입력 값을 변화시켜 학습작업을 반복 수행한다. 이 작업은 재학습기를 통해 수행되며 영역 결정의 커다란 비중을 차지하는 키워드로 인하여 추출되지 않는 중요 키워드들이 본 작업을 통해 새로운 키워드 구조를 이루며 추출된다. 이를 통해 사용자의 특정 관심 정보를 더욱 구체적이고 정확한 사용자 프로파일로 표현할 수 있게 한다[6]. 구축된 사용자 프로파일은 이로써 개별적인 서비스를 제공하는 시스템에 적용되어 각 사용자의 관심도와 관련된 정보를 제공할 수 있을 것이다.

4. 사용자 프로파일 구축을 위한 문서 표현

기계학습을 이용하여 사용자의 관심도를 구체화하는 프로파일을 구축하기 위해서는 문서의 표현이 정확해야만 한다. 이를 위해서는 사용자가 특정 관심영역에 저장해 놓은 관심문서의 표현이 매우 중요하다. 본 논문에서는 이를 위해 특정 영역을 구성하고 있는 문서들의 특징을 추출해내는 특성 추출기법과 이의 결과로 추출되는 특성에 해당되는 키워드들에 가중치를 주는 가중치 설정 기법, 그리고 보다 세부적인 의미를 갖는 키워드 추출을 위한 N-그램 단위의 키워드 추출 기법 등을 사

용한다.

4.1 가중치 기반 키워드 추출

실질적으로 학습 작업에 중요한 비중을 갖는 키워드는 관심영역내의 문서를 구성하고 있는 모든 키워드 중 소수에 불과하다. 그러므로 문서에서 추출된 키워드들을 걸러내어 양질의 키워드만을 선정하는 작업이 필요하다. 이를 위해 본 논문에서는 영역을 구성하는 문서의 특성을 대표적으로 표현하는 키워드들만을 추출해내기 위해 특성 추출 기법을 적용한다. 이를 위해 관심 문서를 구성하는 단어들의 출현 확률 값을 가중치로 설정하여 관심 영역의 키워드에 적용함으로써 관심 영역의 특성 키워드가 갖는 중요도를 높이는 작업을 수행한다.

이러한 확률 값을 이용한 가중치 설정 기법으로 정보검색 분야에서 응용되고 있는 Odds ratio기법이 있다[9]. 본 논문에서는 이러한 OddsRatio기법을 기초로 한 Exp기법을 사용하였다[9]. Exp의 수식은 다음과 같다.

$$\mathrm{Exp}\,P(A) = e^{P(WC1)\cdot P(WC2)}$$

$$WeighedKeyword = P(A) \times W(Cn)$$

위의 식에서 W(Cn)는 문서에 나타나는 빈도 수이다. 이로써 특정 관심 영역 C1의 문서들에 출현하는 빈도수가 높고 사용자의 비관심 영역 C2에 출현하는 빈도수가 낮은 키워드들이 관심영역 C1을 대표하는 것으로 표현되어 더욱 높은 가중치가 설정된다. 이로써 특정 관심 영역의 특성을 추출하는데 가중치가 설정된 대표 키워드가 추출된다.

문서	Term Frequency								관심도
	키워드1	키워드2	키워드3	키워드4	키워드5	키워드6	키워드7	..	
문서1	3	15	0	0	0	0	0	0	Interest
문서2	9	11	5	0	1	0	0	0	Interest
문서3	11	16	7	2	1	0	0	0	Interest
문서4	7	2	1	0	0	0	0	0	Interest
문서5	4	3	0	5	1	0	0	0	Interest
..	0	0	0	0	0	5	4	0	Disinterest
문서n	3	0	0	0	0	6	3	0	Disinterest

A : 관심 문서 B : 비관심 문서 A B

키워드1	키워드2	키워드3	키워드4	키워드5	키워드6	키워드7
Agent	Intelligent agent	System	neural	operation	Travel Agent	Tour

그림 2 추출된 키워드들의 positive문서와 negative문서 표현

본 논문에서는 텍스트 기반의 학습 작업을 위한 입력 값을 추출하는데 각 문서의 단어 출현 빈도 수를 고려하였다. 이 과정에서 추출되는 단어의 세부적인 정보를 얻기 위해 N-그램이라 일컫는 단어 집합 개념을 적용하였다. 이로 인해 예를 들어 "intelligent soft agent"의 세 음절로 구성된 키워드의 경우 "intelligent", "soft", "agent"의 3개의 단어 이외에도 "soft agent" "intelligent soft agent" 등의 더욱 추가적인 의미의 단어를 추출할 수 있다.

이렇게 추출된 세부 단어들은 정보 검색과 텍스트 기반의 학습 시스템에서 주로 사용되는 문서 처리 기법인 단어의 TF와 DF 기법을 적용 받게 된다[7][8]. 여기서 TF는 각 단어의 해당 문서 당 추출 빈도 수(term frequency)를 의미하고 DF는 해당 단어가 특정 영역 당 나타나는 문서의 수(document frequency)를 의미한다. 이러한 빈도 수를 고려한 키워드 추출과정에서 의미 없는 접속사 등은 삭제되고 동일한 단어의 여러 변화형을 스테밍(stemming) 알고리즘을 사용하여 같은 단어 형태로 추출하였다.

4.2 문서 표현에 적용되는 추출 키워드군

특성 추출 기법을 적용하여 선정된 중요 키워드들은 그 자체로 관심 영역의 특성을 표현할 수 있지만 문서 내의 모든 단어들을 대상으로 추출하였으므로 다량으로 존재한다. 이러한 다량의 키워드들을 본 논문에서는 관심영역의 각 문서들에 따른 키워드 벡터로 표현하여 이를 학습 시스템의 입력 값으로 적용하게 되는데 [그림2]는 생성된 키워드 벡터를 나타내었다. 관심영역에는 각 문서마다 사용자의 관심도에 따라 두종류의 문서들로 구성되었다. 이는 [그림2]에서 보는 바와 같이 "Interest" 와 "disinterest" 정보이다.

위에서 표현된 문서들은 특정한 한 개의 관심 영역을 나타내는 것이다. 본 논문에서는 여러 개의 관심영역이 있을 수 있다. 각 영역에서 수행되는 학습 작업은 동일하므로 한 개의 영역만을 가정하여 설명하였다.

5. 사용자 프로파일 학습

문서의 특성 추출 기법을 적용하여 추출된 키워드들을 대상으로 일정 빈도수 이상의 TF와 DF값을 갖는 단어들을 선별하여 중요 키워드 군이 설정되었다. 이 결과 관심 영역을 대표하는 다량의 키워드가 추출된다. 이 경우 사용자가 지정한 관심 문서를 정확히 구분 지을 수 있는 중요 키워드는 다량으로 나열된 키워드가 아니라 새로운 문서를 분류하는데 적합한 정확한 키워드여야 할 것이다. 본 논문에서는 이를 위해 귀납적 기계학습 방식을 적용하고 이의 결과 값인 결정 트리 형태로 중요 키워드를 추출한다. 또한 학습된 높은 분류도의 키워드로 인해 추출되지 않을 수도 있는 상위 분류도 다른

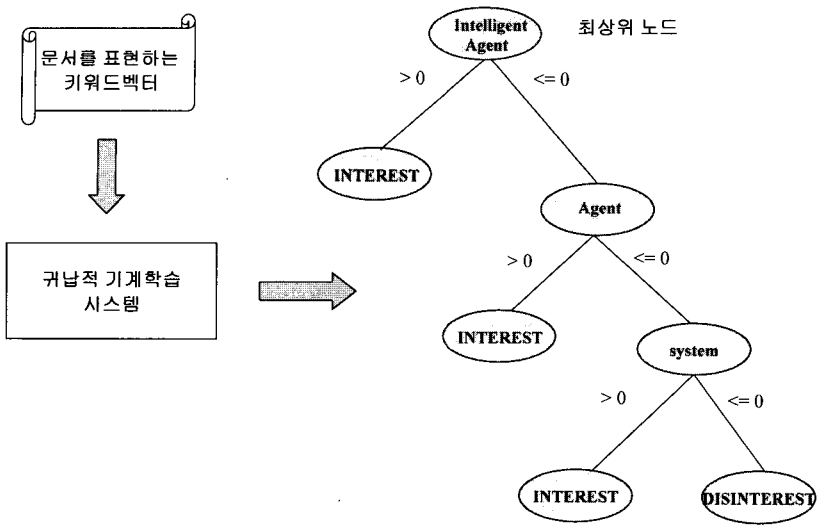

그림 3 학습을 통한 특성 추출 키워드들의 결정 트리

키워드를 추출하기 위한 목적으로 이전 키워드를 삭제하고 다시 학습을 수행하는 재귀적 학습 방식을 적용하였다.

5.1 사용자 프로파일 표현 방식

문서의 관심도를 구분 지을 수 있는 중요 키워드의 추출된 형태는 각 키워드간의 상관관계 없이 단순한 영역을 대표하는 키워드 군으로 표현된다. 이러한 다량의 키워드들은 기계학습 작업을 통한 결정 트리를 구성하여 서로 밀접한 규칙을 갖는 키워드군 형태로 학습되어 사용자 프로파일을 구축하게 된다. 이 과정에서 관심 영역을 결정짓는 가장 영향력 있는 키워드는 결정 트리의 첫 번째 노드에 위치한 키워드로 해당 노드에 연결된 다른 중요 키워드들과 함께 관심 영역에 대한 키워드 규칙을 생성한다.

[그림3]은 단순히 나열된 키워드들이 학습을 통해 결정 트리를 구성하는 과정을 나타낸 것이다. 관심영역에서 특성 추출 기법을 통해 좌측의 다량의 키워드 집합이 선정된다. 이러한 단순한 키워드 집합이 귀납적 기계학습 작업을 통하여 우측과 같은 결정 트리를 생성 키워드간의 규칙을 생성한다.

5.2 귀납적 기계학습 기반 학습

본 논문에서는 사용자의 관심 모델을 표현하기 위해 엔트로피 개념을 활용한 귀납적 기계 학습 방법을 적용한다. 적용된 학습 시스템은 C4.5학습 시스템으로 영역별로 모아놓은 문서들을 대상으로 각각의 특성을 발견하고 분석한다[10]. 분석된 정보는 영역을 분류하는 모델을 구성하고 이 모델은 각각의 속성 값에 따라 트리 형태에 의하여 생성된다. C4.5는 이러한 결정 트리를 이용하여 일정한 규칙을 생성하고 이 규칙들은 키워드에 따른 카테고리의 분류를 가능하게 할 뿐만 아니라 카테고리에 따른 키워드도 추출할 수 있게 한다. 이로써 특성 추출기법을 사용하여 추출된 키워드들은 위의 그림3과 같이 C4.5를 통해 결정 트리를 생성하여 일정한 키워드 구조를 갖게 된다.

이와 같은 C4.5시스템의 학습예제 입력 값은 사용자의 관심 문서와 비관심 문서의 키워드들의 출현 빈번도 정보를 키워드 벡터 형식으로 표현한 것이다. 이는 사용자가 브라우징한 문서들을 해당 관심 영역의 관심 문서와 비관심 문서를 추출된 키워드의 특성 값으로 표현한 벡터이다. 본 논문에서는 입력되는 키워드의 속성을 선정하는데 있어서 본래의 C4.5시스템이 모든 단어에 자체적으로 가중치를 설정하는 단계가 없는 이유로 N-그램 단위의 중요도가 설정되도록 C4.5시스템을 수정하여 적용하였다.

5.3 재귀적 학습 모듈

학습 작업을 통하여 관심 영역을 구체적으로 표현하는 키워드 구조가 결정되었다. 이러한 키워드 구조는 앞에서도 언급했듯이 결정 트리를 구성하는 첫 번째 노드에 위치한 키워드의 중요도가 매우 높기 때문에 루트 노드에 따라 하위 노드를 구성하는 키워드가 달라지게 된다. 그러므로 만약 루트 노드의 키워드가 관심 영역에

서 차지하는 비중이 매우 높을 경우 이보다 낮은 비중을 갖고 이에 따른 다른 구조를 갖는 키워드가 추출되지 않을 수 있다. 예를 들어 다음 [그림4]와 같이 특정 영역내의 관심 문서들에서 키워드$_1$의 "Graphic"이 커다란 비중을 갖고 있을 경우 영역을 대표할 수 있는 키워드$_3$의 "MPEG4"가 추출되지 않을 수 있다.

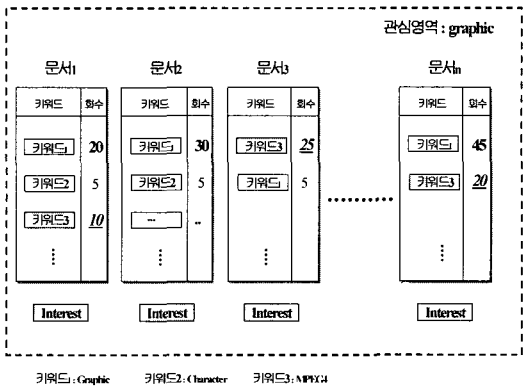

그림 4 다양한 트리 구조 추출을 제약하는 문서표현의 예

이러한 경우에 두 번째로 비중이 큰 키워드를 추출해 내기 위해서는 루트 노드의 키워드가 영역을 구성하는 키워드 구조에 미치는 영향을 배제해야 할 것이다. 이를 위해 본 논문에서는 첫 번째 학습을 수행하여 가장 비중이 큰 키워드와 이와 관련한 키워드 구조를 추출하고 난 후에 다시 학습을 수행하는데 이때 주어지는 학습의 입력 자료에 변화를 준다. 학습에 입력되는 키워드 벡터 정보에 저장되어 있는 루트 노드의 키워드를 삭제함으로써 전체적으로 편중된 키워드 구조를 배제한다. 변화된 키워드 벡터는 다시 학습 시스템의 입력 값으로 설정되어 두 번째 큰 비중을 갖는 키워드를 루트 노드로 하는 새로운 키워드 결정 트리를 생성하게 된다.

이러한 작업은 위의 [그림 4]의 키워드 구조를 갖는 관심영역 클래스를 비관심 영역 클래스의 2-클래스 학습 작업을 통해 학습결과로 생성되는 트리 구조의 루트 노드가 생성되지 않을 때까지 여러 번 C4.5학습 시스템을 수행시킨다.

본 논문에서는 구축된 사용자 프로파일을 검색에이전트와 푸쉬 에이전트에 적용하였다. 검색에이전트는 사용자가 입력한 질의어가 사용자 프로파일의 구조에 속해 있는 경우 다른 노드의 관련 키워드들을 확장하여 사용자에게 검색 질의어로 추천해주는 기능을 한다. 푸쉬 에이전트는 프로파일을 구성하는 키워드 구조를 검색어로 설정 관련 문서를 수집하여 이를 사용자 관심문서로 인식 사용자에게 푸쉬 해주는 기능을 한다.

6. 사용자 프로파일 재학습

사용자 프로파일의 정확도는 관심영역으로 설정된 문서들의 정확도와 밀접한 관련이 있다. 설정된 문서가 사용자 관심 영역에 해당되지 않거나 거리가 있는 경우 이로 인해 사용자 프로파일의 정확도가 낮아질 수 있다. 이 경우 푸쉬 에이전트가 제공한 사용자 관심문서가 이를 받아보는 실제 사용자의 관심도와 정확히 일치하지 않게 된다. 본 논문에서는 이러한 경우 사용자 관심영역을 보다 정확하게 표현할 수 있도록 사용자 프로파일을 수정하기 위해 피드백 기능을 제공한다. 사용자는 자신의 관심도와 관련 없는 푸쉬 정보를 사용자의 비관심 영역에 전송함으로써 추가된 문서를 포함한 학습 작업을 통해 사용자 프로파일이 수정되도록 한다. 이 같은 여러 번의 수정을 통해 사용자 프로파일의 정확도는 더욱 향상될 수 있다. 이러한 검색된 정보에 대한 사용자의 피드백기능을 제공함으로써 사용자에게 향상된 정보를 제공하는데 이용된다. 이와 같은 사용자 피드백에 의한 사용자 프로파일 재구축 작업은 웹 에이전트가 수행하는 학습 작업에 사용자의 더욱 정확한 관심도를 입력받기 위한 것일 뿐만 아니라 지속적으로 사용자 프로파일을 수정, 보완하는데 중요성이 있다.

7. 실험

본 논문에서 구현한 사용자 적응형 웹 에이전트는 클라이언트/서버 기반 시스템으로 C4.5학습 시스템은 C언어로 구현되었으며 이외의 모든 모듈은 java언어로 구현하였다. 클라이언트 부분은 윈도우 환경의 브라우저에 적용되는 Java 프락시를 구현한 모니터 에이전트를 구축하였다. 서버 부분에서는 C언어로 구현되어 있는 C4.5 학습 시스템을 이용하였고 이를 제외한 모든 학습 모듈, 적응형 에이전트 부분 등의 모든 모듈을 Java언어로 구현하였다. 실험은 1명의 사용자가 생성한 9개의 특정 관심 영역에 대해 각각 30개의 문서를 대상으로 이에 대한 관심도를 학습하였다.

7.1 사용자 프로파일 구축

위의 [그림5]는 학습된 결과 생성된 사용자 프로파일을 나타낸 것이다. 아래의 관심 영역 이름인 'JDBC', 'Visual C++', 'Y2k'등은 각각 "자바를 이용한 데이타베이스 연결 시스템", "마이크로 소프트 사의 프로그래밍 툴", "2000년 표기 문제"등을 의미하는 것으로 사용

자가 직접 생성한 것이다. 각 카테고리의 관심문서가 포함하고 있는 키워드들은 학습 작업을 거쳐 결정 트리를 구성하는 그림5와 같은 키워드들로 추출되었다.

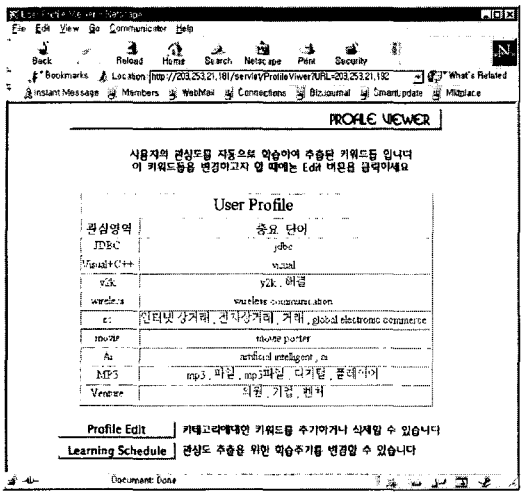

그림 5 사용자 프로파일의 시스템 출력 화면

7.2 적응형 웹 검색 에이전트

[그림 6]은 적응형 웹 검색 에이전트를 이용하여 사용자가 찾고자 하는 대상에 대해 질의어 "상거래 "를 입력하였을 경우 에이전트가 사용자 프로파일의 키워드를 추천한 예이다.

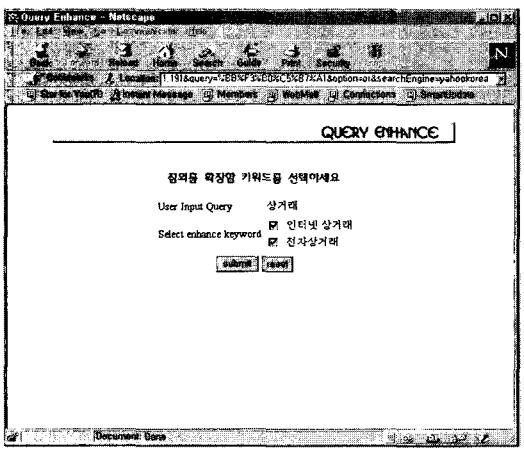

그림 6 사용자에게 질의 단어를 추천하는 화면

이 경우 웹 에이전트 시스템은 사용자 프로파일에 기록되어 있는 "인터넷 상거래"와 "전자상거래"의 두 키워드를 추천한다.

7.3 사용자 피드백에 의한 프로파일 재학습

구축된 사용자 프로파일을 기반으로 제공되는 적응형 웹 푸쉬 에이전트의 결과 중 사용자가 설정한 비관심 정보를 피드백으로 학습 서버에 전송하여 프로파일이 재학습되는 과정이다. [그림 7]은 사용자 신문의 해당 비관심문서를 전송하였을 경우 재학습된 프로파일을 기반으로 더욱 향상된 푸쉬서비스가 제공된 화면이다.

8. 결론 및 향후연구

본 논문은 지능형 웹 에이전트를 구현하기 위해, 웹 문서 내에서 중요한 키워드를 추출하기 위해 다양한 문서처리방법과 특성 추출 기법과 사용자 요구정보를 분석하는 사용자 프로파일 구축 기술, 이러한 프로파일을 적용하여 웹을 검색하는 기술을 제시하였다[9]. 먼저, 웹 에이전트는 사용자가 자신의 관심영역으로 생성한 임의의 영역을 관심문서로 구성한다. 수집된 문서들에서는 특성 추출 기법을 통해 관심영역을 표현하는 다량의 중요 키워드가 추출되고 이들은 귀납적 기계학습을 통해 관심영역을 결정짓는 키워드 트리 구조를 생성하였다. 그리고, 이렇게 구축된 결정 트리 구조는 키워드간의 구체적인 규칙을 생성하여 사용자 프로파일로 작성되었다.

구축된 사용자 프로파일은 질의어를 확장하고 사용자 관심 문서를 추측하여 푸쉬하는데 적용하였다. 특히 본 논문에서는 사용자 프로파일을 작성하는데 있어서 특성 추출기법과 N-그램을 적용한 C4.5학습을 수행함으로써 결정 트리의 구조를 보다 정확히 하였고 다양한 키워드 구조를 추출하기 위하여 재학습을 수행하였다. 또한 푸쉬 정보에 대한 사용자의 부정적 피드백을 받아 프로파일의 정확도를 향상하였다.

본 논문에서는 사용자 프로파일을 구축하는데 감독 학습 방식을 적용하였다. 이는 학습 결과가 사용자의 관심 영역을 보다 정확히 표현할 수 있는 장점이 있는 반면 사용자에게 자신의 관심도를 항상 표시하게 하는 불편을 수반한다. 향후에는 이러한 사용자의 관심도를 별도의 입력 없이 추출하는 비감독 학습 방식을 적용하여 사용자에게 보다 편리한 에이전트 시스템으로의 확장이 가능할 것이다.

또한, 본 논문에서 적용한 학습 시스템인 C4.5 시스템은 *incremental learning* 기법을 지원하지 않는다. 하지만 사용자가 생성한 각 관심영역에 저장되는 데이타의 양이 많지 않고 또한 사용자의 새로운 관심영역을 사용자가 직접 생성하는 작업으로 해결할 수 있을 것이

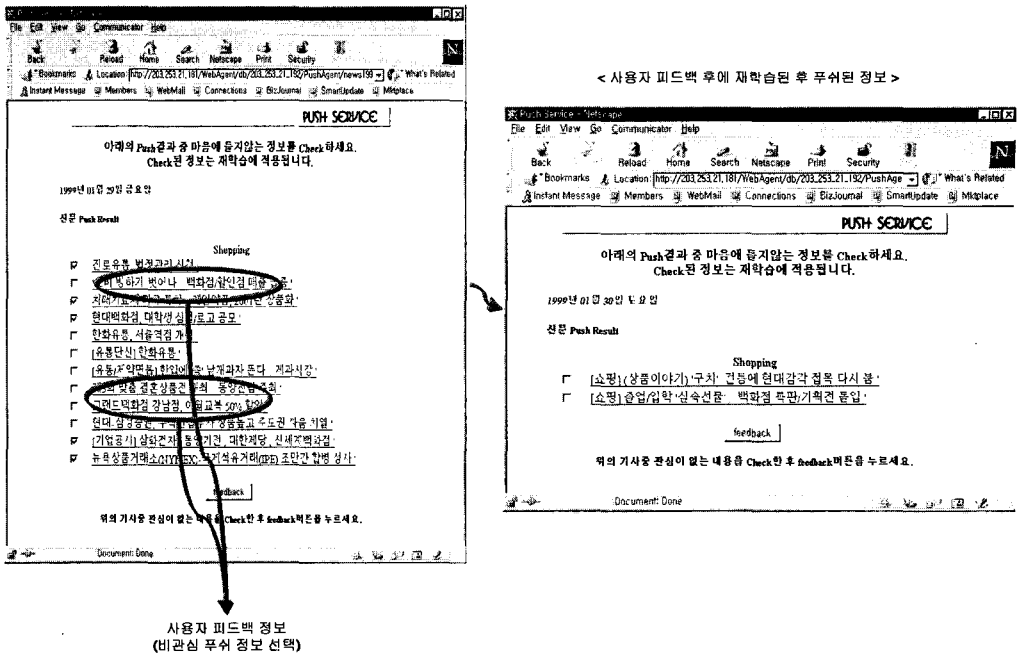

그림 7 사용자 피드백에 의한 프로파일 재구축

다. 하지만 보다 정확한 사용자의 관심도를 반영하기 위해서 향후 본 시스템에 적용되고 있는 학습 시스템을 incremental learning 기법을 지원하는 시스템으로 구축해야 할 것이다.

참 고 문 헌

[1] *Marko Balabanovic and Yoav Shoham*, "Learning Information Retrieval Agents: Experiments with Automated Web Browsing," AAAI Spring Symposium on Information Gathering, Stanford, CA, March 1995.

[2] *L. Dent, J. Boticario, J. McDermott, T. Mitchell, D. Zabowski*, "A Personal Learning apprentice," In Proceedings of the 11th International Conference on Machine Learning, July 1994.

[3] *O. Etzioni, D. Weld*, "A Softbot-Based Interface to the internet," Communication of ACM, July '94

[4] *O. Etzioni, S. Hanks, T. jiang, R. M. Karp, O. Madani, O. Waarts, Efficient* "Information Gathering on the Internet," ARPA F30602-95-1-0024.

[5] *Yezdi Lashkari, Max Metral, Pattie Maes*, " Collaborative Interface Agents," Conference of the American Association for Artificial Intelligence, Seattle, August 1994.

[6] *Dunja Mladenic*, Personal WebWatcher: Implementation and Design, Technical Report IJS-DP-7472, October, 1996.

[7] *Salton, G., and Buckey, C.* "Term weighting approaches in automatic text retreival," Technical Report 87-881, Cornell University, Department of Computer Science 1987.

[8] *Thorsten Joachims*, " A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization," March 1996.

[9] Dunja, Mladenic "Feature subset selection in text-learning," Department for Intelligent Systems, J.Stefan Institute, 1997.

[10] J *R. Quilan*, "C4.5 Programes for Machine Learning," San Mateo, CA:Morgan, Kaufaman,1993.

[11] *Bruce Krulwich*, "Learning document category description through the extraction of semantically significant phrases," Center for Strategic Technology Research Andersen Consulting LLP 100 South Wacker Drive, Chicago, IL 60606, 1995.

[12] *Bruce Krulwich, Chad Burkey* "The InfoFinder Agent: Learning User Interest through Heuristic Phrase Extraction," AgentSoft Ltd, Andersen Consulting LLP, 1995.

소 영 준

1998년 숭실대학교 전자계산학과 (학사). 2000년 숭실대학교 대학원 컴퓨터학과 (석사). 2000년 ~ 현재 숭실대학교 대학원 컴퓨터학과 박사과정 재학중. 관심분야는 인공지능, 에이전트.

박 영 택

1978년 서울대학교 전자공학과 (학사). 1980년 한국과학기술원 전산학과 (석사). 1981년 ~ 현재 숭실대학교 정보과학대학 컴퓨터학부 교수. 1992년 University of Illinois at Urbana-Champaign 전산학과 (박사). 관심분야는 인공지능, 에이전트.