개인 웹 에이전트를 위한 사용자 프로파일 구축

이 상 섭, 소 영 준, 박 영 택
숭실대학교 정보과학대학 컴퓨터학부

Construction of User Profile for Personal Web Agents

Sang-Sub Lee, Young-Jun So, Young-Tack Park {{sslee, so, park}@multi.soongsil.ac.kr}
Dept. of Computer Science, Soongsil Univ.

요약

본 논문에서 구현하고자 하는 웹기반 사용자별 에이전트는 웹을 이용해 정보를 검색하는 사용자들에 대한 사용자 관심도를 사용자의 웹검색 행위를 감시하는 모니터 에이전트에 사용자가 직접 기술하게 하고 이를 별도의 학습서버를 두어 사용자별 프로파일을 만들어 이를 사용자가 확인 및 편집할 수 있게 하였다. 서버에서의 학습 과정은 웹 브라우저를 통하여 수집된 정보를 바탕으로 사용자가 관심을 가지는 웹 문서의 일반적인 내용에 대한 관심 정확도를 높이는 일련의 단어 점제 과정을 통하여 최적의 관심 키워드를 추출하는 작업으로 이루어지며 이는 표현 모델인 사용자 프로파일을 구축하여, 관심 문서를 검색하는데 적절한 정보를 제시하는 것을 목적으로 한다. 이 시스템에서 적용되는 학습 방식은 사용자의 웹 문서 관심도에 의존하므로 웹 문서에 나타나는 텍스트들을 대상으로 C4.5 학습 시스템을 적용한다

1 서론

개인 웹 에이전트는 사용자가 웹을 이용하여 문서를 탐색하는 데 도움을 주는 기능을 필요로 한다 이와 같은 기능을 수행하기 위해서, 개인 웹 에이전트는 각 사용자가 선호하는 문서에 대한 정보를 표현하는 사용자 프로파일을 활용하게 된다 기존의 웹 에이전트는 이와 같은 사용자 프로파일을 각 사용자가 수동으로 입력하는 방식을 취하고 있다 또한, 많은 웹 에이전트가 활용하는 사용자 프로파일은 사용자가 이해할 수 없는 형태인 키워드 백터로 표현되고 있다. 이와 같은 문제점들은 에이전트가 사용자의 일상적인 타스크를 위임받아 수행하고, 사용자와 대화하는 가능을 현저히 떨어뜨리게 된다

이와 같은 문제점을 극복하기 위해서 본 논문에서는 각 사용자가 관심을 가지는 문서에 대한 정보를 스스로 학습하고, 학습된 결과를 사용자가 이해할 수 있는 형태로 표현하는 개인 웹 에이전트에 대해서 서술하고자한다[Mladenic 96]. 이와 같은 개인 웹 에이전트는 사용자가 관심을 가지는 문서를 같은 성질을 가지는 문서들로 분류하고, 분류된 문서들에 귀납적 기계학습을 재귀적으로 적용하여 n-그램 키워드로 구성된 규칙을 생성하게 된다 각 규칙은 사용자가 해당 분류 문서들에 대한 관심을 대표할 수 있는 키워드를 포함하고, 경우에 따라서는 관심없는 문서들도 표현하게 된다.

본 연구에서는 각각의 사용자가 에이전트의 도움을 받을 수 있고, 동시에 사용자가 에이전트를 제어할 수 있는 기본적인 기능을 실현할 수 있는 개인 웹 에이전트 시스템을 구현하였다. 현재, 개인 웹 에이전트는 모니터 에이전트와 학습 에이전트로 구성되어 있다 모니터 에이전트는 각 사용자의 로컬 머신에 위치하고 있으면서 사용자의 브라우징 행위를 모니터 하여 그 결과를 서버에 위치한 학습 에이전트에 전달한다. 학습 에이전트는 여러 사용자의 요구를 처리하기 위해서 각 사용자별로 데이터베이스를 구축하고 귀납적 기계학습 프로그램인 C4.5를 제귀적으로 활용하여 사용자 프로파일을 생성한다

2.웹 에이전트 시스템 구성

본 논문에서 구현한 웹 기반 에이전트 시스템의 전체적인 시스템 구성은 크게 두 부분으로 구성되어 있다. 첫째, 로컬 머신에 위치하는 모니터 에이전트로 사용자 프로파일 생성을 위한 학습 자원을 만들어내는 모듈이다 이 모니터 에이전트는 현재 사용자의 브라우저상에 보여지는 문서에 대한 URL를 저장하고, 본 문서 내에 에이전트 메뉴를 삽입 시용자가 직접 자신의 관심도를 표현할 수 있게 하며, 기술된 사용자 관심도를 학습서버에 보내는 기능을 수행한다[Thorsten 95]. 또한 학습서버를 통해서 만들어진 사용자 프로파일에 대한 내용을 서버에 연결하여 사용자에게 보이고 이를 사용자 스스로 수정, 편집할 수 있게 한다 둘째, 모니터 에이전트에 의해서 만들어진 사용자 관심 정보를 분석하고 관심분야별 키워드를 전처리과정을 거쳐 추출하여 이를 C4.5 학습 모듈을 이용해 각 관심분야별 중요 키워드로 구성된 사용자 프로파일을 생성하는 학습 에이전트 서버로 구성된다

2.1 모니터 에이전트

모니터 에이전트는 학습을 통한 사용자 프로파일 생성을 위한 학습 자원을 만들어내는 기능을 수행하는 부분으로 그 기능과 구조는 다음과 같다 첫째, 현재 사용자의 브라우저상에 보여지는 문서에 대한 URL 정보를 유지하고, 본 문서에 대한 사용자의 관심표현시 이 정보를 학습서버로 보내는 사용자 브라우저 감시 기능과 둘째, 사용자의 관심도 표현을 위한 에이전트 메뉴를 생성하는 것으로, 사용자는 이 메뉴

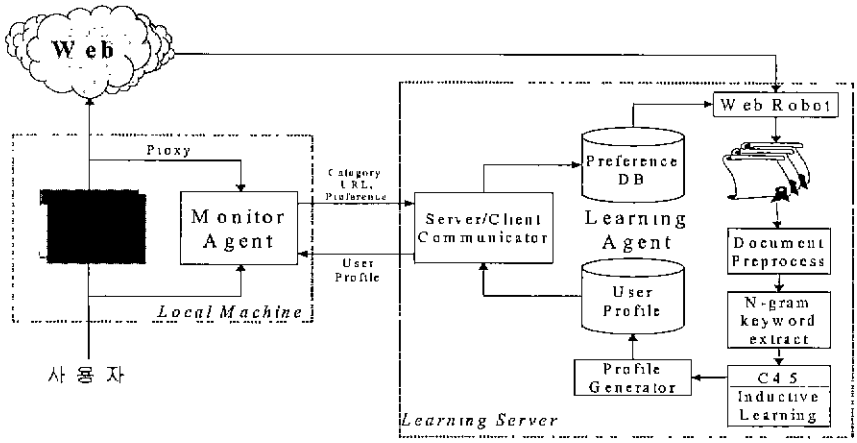


그림 1. 개인 웹 에이전트 구조

를 이용하여 자신의 관심도를 표현하게 되는데, 이는 에이전트 메뉴의 주요 기능으로 관심분야별 학습을 위한 사용자 정의 관심분야생성 및 편집 기능과 본 문서에 대한 관심, 비관심도를 사용자가 표현 할 수 있게 해 준다. 즉 본 문서에 대한 관심분야를 선택하고 이에 대한 해당 관심분야가 존재하지 않을 때, 새로운 관심분야를 생성할 수 있게 한다. 또한 사용자에 의해서 정의된 관심분야내에서 본 문서에 대한 관심도를 Hot', 'Cold'의 형태로 정의, 이에 대한 사용자 관심도를 저장, 특정 관심분야 중심으로한 URL정보와 관심도를 학습을 위한 자원으로 학습서버에 보내는 기능을 수행한다. 마지막으로 학습서버를 통해서 만들어진 사용자 프로파일을 사용자가 직접 수정, 편집할 수 있게 하여 학습된 결과를 사용자가 직접 확인할 수 있게 하는 기능을 수행한다

2.2 학습 에이전트 서버

학습서버는 로컬머신에 위치한 모니터 에이전트가 보내오는 사용자 관심 정보에 대해서 사용자가 직접 정의한 관심분야별 중요 키워드에

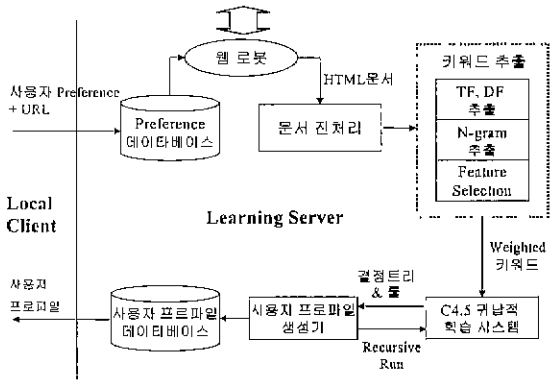


그림 2. 학습 에이전트 구조

대한 학습작업을 통해서 사용자 프로파일을 생성한다

모니터 에이전트로부터 보내지는 사용자 관심정보 즉 웹 브라우징를 통하여 사용자가 미리 열람한 문서에 대해 수집된 정보가 학습서버내의 사용자 DB에 저장되면 학습서버는 이를 일정한 시간간격으로 학습을 하게 되는데 그 과정을 요약하면 다음과 같다 첫째, 학습 서버내의 학습스케줄모듈을 통해서 사용자 관심 DB를 감시 일정한 양의 정보가 쌓이면 이를 학습서버의 학습모듈에 알리고 학습모듈은 특정 사용자의 프로파일 구축을 위한 학습작업을 수행한다. 둘째, 학습모듈은 각 관심분야에 대한 관심문서 URL정보를 이용, 웹 로봇을 통해서 문서를 가져오게 한다 모든 문서가 로컬 디스크에 저장되면 각 문서 내에 나타나는 키워드를 추출하고 관심 키워드와 비 관심키워드를 분류해내는 전처리 과정을 수행하여 더욱 신뢰성 있는 관심키워드군으로 변형되는데 이는 C45학습 시스템을 이용한 기계학습 과정을 통해서 일정한 형태의 룰을 생성하도록 한다. 셋째, 학습 작업을 거쳐 얻어지게 되는 룰 들에서 추출된 키워드들과 관심분야에 관한 일정한 규칙을 배경으로 하여 사용자 프로파일을 생성하는 기능을 수행한다.

2.3 문서 전처리 및 키워드 추출 과정

문서 전처리 과정은 사용자가 관심을 표현한 웹문서내의 모든 단어들 중에 내재되어 있는 수많은 비중요 단어들을 제거하고 학습 과정에 쓰일만한 중요한 단어들만을 추출하여 학습 결과의 정확도를 향상시키는데 결정적인 기능을 수행하게 된다. 이 과정은 전체적으로 4개의 집린 작업으로 이루어져 있는데 이는 다음과 같다 첫째, 웹 문서내의 정보들이 HTML언어와 자연언어의 형태로 이루어져 있으므로 이를 일정한 의미를 갖는 품사로만 된 단어들로 변형시켜 주는 불용어제거 과정이 필수적인데 이 과정에서는 불용어 사전과 한글의 경우에 있어서는 명사와 조사 사전을 이용한 비교 추출/제거 방식과 스테밍 알고리즘을 이용하여 동의어를 동일 형태로 수집해 주는 단어 형태 변형 추출 방식의 두 가지 방식을 이용하였다 둘째, 이러한 과정을 거쳐 추출된 일련의 단어들은 이 후에 하나의 단어(1-그램)만이 아닌 서로 이어진 단어들 즉, 2-그램, 3-그램 등의 n-그램 추출 방식을 적용하게 되는데 이 작업으로 인하여 더욱 상세하고 중요한 단어군들을 추출할 수 있게 된다. 셋째, 이로써 제 생성된 키워드 벡터들은 각 문서 내에서 서로간의 상관관계와 문서들간의 각 단어들과의 관계를 도출해 내기 위한 작업을 수행하게 된다. 이는 단어들이 자신이 속한 문서 내에서 추출된 횟수인 TF(Term frequency)와 일정 단어를 모든 문서를 통틀어 포함하고 있는 문서의 개수인 DF(Document frequency)를 기술하는 키워드 벡터 생성에 그 목적이 있다 여기까지 웹 문서에서 추출된 단어의 정제 과정과 단어 발생 빈번도에 대한 측정과정이 완료되었다 네 번째

작업으로 앞의 과정들에 의해 생성된 키워드들에 대해 일정한 웨이트(weight)를 설정하는 작업이 수행된다. 이는 앞의 세 번째 작업에서 보여진 TF, DF추출 과정에서는 밝혀지지 않은 사용자 관심 관심분야와 각 단어와의 관계를 알아내 이를 토대로 각 단어마다의 TF, DF의 정보와 더불어 관심분야에 의한 집합들을 추출하여 데이터에 대한 관심도 추출의 정확도를 측정하는데 이용된다[Mladenic 97]. 본 논문에서는 트레이닝 다큐먼트에 기입된 각각의 단어들의 특징을 정의하는 텍스트 학습기법중에서 Exp기법을 사용하였다. Exp의 수식은 다음과 같다

$$\mathrm{Exp}\,P(A) = e^{P(W|C1) - P(W|C2)}$$

위에서 P(W|C1)은 특정 클래스 C1에서 특정 단어 W가 나온 확률 값으로서 위의 계산 값을 모든 단어에 적용하여 웨이트를 주는 방법을 사용하였다. 이처럼 각 단어마다 웨이트를 설정해주는 작업은 생성될 최종 중요 단어에 커다란 긍정적 영향을 끼치게 된다

2.4 C4.5를 이용한 학습 시스템

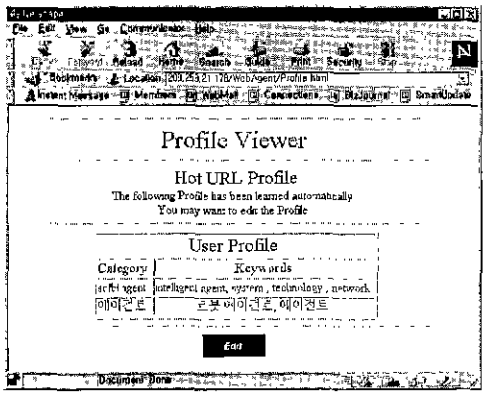
본 논문에서는 관심분야별 중요 단어를 추출해 내는 기계 학습 시스템으로 결정트리를 이용하여 분류모델을 형성하는 C4.5를 적용하였다[Quinlan92]. C4.5시스템의 학습예제 집합을 생성하기 위하여 위의 전처리 과정을 거친 키워드 단위의 속성이 입력되도록 하였다. 이 시스템에는 2가지 형태의 파일들이 입력되는데 첫 번째, 분류된 관심분야와 키워드벡터들로 기록되는 파일과 두 번째, 관심분야별로 각 문서에서 발견된 키워드 횟수와 전처리과정에서 결정된 웨이트들이 키워드 순서대로 기입된 파일의 형태로 구성된다. 이 입력되는 키워드단위의 속성을 선택하는데 있어서 본래의 C4.5시스템이 모든 단어에 자체적으로 웨이트를 설정하는 단계가 없는 이유로 1-그램의 단어보다는 2-그램, 3-그램 등의 순으로 중요도의 우선 순위가 설정되도록 C4.5시스템을 수정하였다. 최종 간편한 형태의 결정트리를 생성하는 것을 목적으로 하는 이 C4.5시스템의 클래스들은 첫 번째 클래스로 관심 있는 특정 분야를 그리고 두 번째 클래스로 사용자가 관심 없다고 표시한 모든 문서들이 포함되어 있는 분야로 설정하여 학습결과의 형태가 두 클래스 중 첫 번째 즉 관심 있다고 표기한 분야의 중요 키워드가 추출되는 형태가 되도록 설정하였다

2.5 사용자 프로파일 구축

위의 C4.5 학습 시스템은 먼저 트리의 형태로 학습결과가 표시되는데 생성되는 룰은 이를 간략화된 형태의 트리를 이용하여 표시된다. 이 데이터는 프로파일 생성기에 의해 룰에 표시된 키워드들을 역시 n-그램의 단위가 높은 순으로 프로파일에 기입된다. 되도록 다수의 중요 키워드가 추출되도록 하기 위하여 생성된 룰에서 우선 순위가 가장 높은 형태로 추출된 키워드를 제외하고 C4.5 입력 파일을 수정한 다음 다시 학습작업을 수행하게 된다. 가능한 한 1-그램 이상의 키워드가 추출될 수 있도록 이러 차례 C4.5입력 양식을 수정하여 다시 반복 수행된다. 이 기능으로 인하여 추출된 키워드들의 정확도는 더욱 높아지게 된다. 이렇게 제작이 저장된 프로파일은 검색추출과 동시에 다시 클라이언트로 보내지게 되며 이로써 사용자는 자신의 관심도를 구체적으로 인식할 수 있을 뿐 아니라 이 프로파일에 기준 하여 인터넷상의 정보들 중에서 자신이 관심 있어 하는 것들을 걸러 내을 수 있는 작업도 가능하게 된다. 또한 프로파일에 기입된 관심분야와 각 중요 키워드들은 사용자가 임의로 편집할 수 있도록 했으며 수정된 정보는 바로 서버에 전송되어 항상 서버와 사용자의 프로파일의 형태가 동일하도록 설정되었다. 사용자는 이들을 중심으로 인터넷 검색 측면에서 더욱 다양하게 관심 문서의 필터링을 용이하게 해 줄 것이다

3. 실험

실험에서는 위의 설명한 단계들을 통해 키워드 벡터를 추출한 후 C4.5학습 시스템을 이용하여 결정트리를 구한 후 생성된 룰을 이용하여 사용자 프로파일을 생성하는 순으로 진행되었다. 실험은 두종류로 나뉘어 이루어졌으며 대상 문서는 첫 번째의 경우 영문으로 구성된 소프트 에이전트를 관심분야로 두 번째의 경우에는 한글로 구성된 에이전트를 관심분야로 하여 실험하였다. 다음은 두 관심분야별 모두 20개씩의 문서들을 대상으로 한 실험결과를 나타낸 그림이다.



4. 향후 연구과제 및 결론

본 논문에서 사용자는 모니터 에이전트를 통하여 웹 문서에 대한 자신의 관심도를 표시하고 학습 시스템에 의하여 관심분야에 관한 구체적인 키워드 정보를 갖게 된다. 구축된 프로파일은 사용자가 웹의 특정 정보를 찾기 위해서 검색엔진을 사용하는 행위에 자문을 줄 수 있는 기능으로 확장될 것이며, 이는 사용자의 검색질의를 분석, 구축된 프로파일에서 이외 연관성 있는 키워드를 찾아내어 검색질의를 확장하는 기능을 추가할 것이다. 이 기능은 구축된 프로파일을 효율적으로 응용하여 사용자에게 더욱 정확한 양질의 검색 결과를 나타내주는 등의 형태로 이용될 것이다.

5. 참고문헌

[Mladenic 96] *Dunja Mladenic*, Personal WebWatcher: Implementation and Design, Technical Report *IJS-DP-7472*, October, 1996

[Mladenic 97] *Dunja Mladnic*, Feature subset selection, Department of Intelligent Systems, J.Stefen Institute, Jamova 39, 1100 Ljubljana, Slovenia, 1997

[Quinlan 93] J *R Quilan*, "C4.5 Programes for Machine Learning", San Mateo, CA Morgan, Kaufaman,1993

[Thorsten 95] *Thorsten Joachims, Tom Mitchell, Dayne Freitag, and Robert Armstrong*. "WebWatcher' Machine Learning and Hypertext", Fachgruppentreffen Maschinelles Lernen, Dortmund, Germany, August 1995