PDA용 개인화 에이전트 시스템

표석진, 박영택

PDA Personalized Agent System

Suk-Jin Pyo, Young-Tack Park

요 약

무선 인터넷을 이용하는 사용자는 정보의 양의 따른 시간적 통신비용의 증가 문제로 개인화 에이전트가 사용자의 관심에 따라 서비스를 제공하는 기능과 맞춤화된 정보를 제공하는 기능, 지식 기반 방식으로 정보를 예측하는 기능을 가지기를 바라고 있다. 본 논문에서는 이와 같이 무선 인터넷을 사용하는 사용자를 위한 PDA 개인화 에이전트 시스템을 구축하고자 한다. PDA 개인화 에이전트 시스템 구축을 위해 프로파일 기반의 에이전트 엔진과 사용자 프로파일을 이용한 지식기반 방식을 사용한다. 사용자가 웹 페이지에서 행하는 행위들을 모니터링하여 사용자가 관심 가지는 문서를 파악하고 정보 검색을 통해 얻어진 문서를 분석하여 사용자 각각의 관심 문서로 나누어 서비스 하게 된다. 모니터링 되어진 문서를 효과적으로 분석하기 위해 unsupervised clustering 기계학습 방식인 Cobweb을 이용한다. unsupervised 기계학습은 conceptual 방식을 이용하여 검색되어진 정보를 사용자의 관심 분야별로 clustering한다. 클러스터링을 통해 얻어진 결과를 다시 기계학습을 통해 사용자 관심문서에 대한 프로파일을 생성하게 된다. 이렇게 만들어진 프로파일을 룰(Rule)로 만들어 이를 기반으로 사용자에게 서비스하게 된다. 이러한 룰은 사용자의 모니터링 결과로 얻어지기 때문에 주기적으로 업데이트하게 된다. 제안하는 시스템은 인터넷신문이나 웹진 등에서 사용자들에게 뉴스를 전달하기 위한 목적으로 생성하는 뉴스문서를 특정 대상으로 선정하였고 사용자 정보를 이용한 검색을 실시하고 결과로 얻어진 정보를 정보 분류를 통해 PDA나 휴대폰을 통해 사용자에게 제공한다.

Key words : PDA,개인화(Personalized), 에이전트(Agent), 클러스터링(Clustering),모니터링

1. 서 론

오늘날 인터넷을 통해 사용자가 접하는 정보는 무수히 많다. 또 그 정보들은 계속해서 기하급수적으로 늘어나는 추세이다. 이로 인해 사용자가 필요로 하지 않는 정보들을 아무런 여과 없이 접하게 됨으로써 시간과 비용의 낭비를 초래하고 있다. 또한 인터넷상에 존재하는 방대한 정보의 양과 분산의 특징으로 인하여 사용자는 자신이 원하는 정보를 찾는데 어려움과 한계를 가질 수 있다. 이러한 문제점들을 극복하기 위해서 제시되고 연구되어지는 것이 에이전트 시스템들이다. 이러한 에이전트 시스템은 사용자가 찾고자하는 사용자의 관심정보에 적응할 수 있는 시스템들로 발전하고 있으며 이를 위해서 사용자의 기호를 추출함으로써 사용자에게 보다 능동적으로 적응할 수 있는 에이전트 시스템으로 발전해 나갈 것이다.

현재 인터넷을 이용한 웹 서핑은 많이 이용하고 있다. 하지만 사용자 웹 서핑을 하기 위해 자신의 개인 컴퓨터 앞에서만 인터넷을 사용하지는 않을 것이다. 무선 인터넷을 사용할 수 있는 노트북이나, PDA, 휴대폰을 사용하여 장소에 상관없이 인터넷 서핑을 원한다. 현재 보급되어있는 노트북이나 PDA, 휴대폰의 경우 입력 장치, 네트워크 장치 등으로 인한 시간적 비용 문제로 사용자가 원하는 정보를 빠르고 쉽게 얻을 순 있지만 많은 양의 정보를 서비스 받기 위해서는 통신비용의 증가가 필요하게 된다.

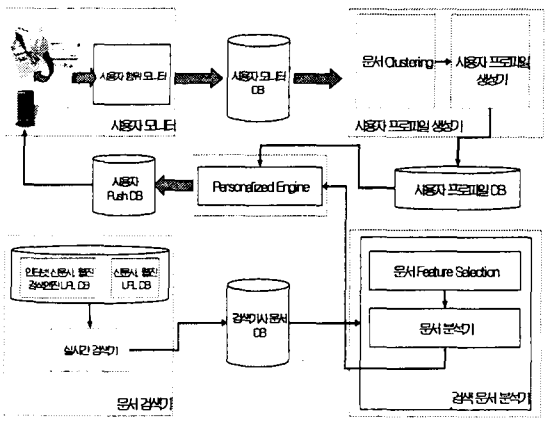
많은 양의 데이터를 제공받기 위해 무선 인터넷을 사용할 경우 생겨나는 근본적인 시간적 통신비용의 절약을 위해 본 논문에서는 사용자가 원하는 정보를 검색하는 에이전트 시스템과 검색 되어진 정보를 사용자의 프로파일을 기반으로 서비스하는 에이전트 시스템을 구현하였다. 본 논문에서의 검색 에이전트 시스템은 인터넷상의 정보들을 사용자가 원하는 정보를 기반으로 검색하여 보다 정확하고 효율적인 정보를 제공한다. 또한 사용자의 행위 정보를 모니터링하여 이를 프로파일화 하고 이렇게 얻어진 정보를 통하여 새롭게 검색되어지는 문서나 정보들을 해당 사용자의 관심 문서임을 자동으로 파악하고 이를 서비스 하게 된다. 이러한 Push 서

* 본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음

비스 시스템은 클러스터링 방법을 이용하여 사용자가 원하는 수많은 정보들에서 정보 분류를 통해 순차적인 나열 보다는 분류되어진 정보를 서비스하여 무선인터넷의 시간적, 비용적 제약을 극복을 통해 통신비용의 절감을 가지게 된다. 제안하는 시스템은 인터넷신문이나 웹진 등에서 사용자들에게 뉴스를 전달하기 위한 목적으로 생성하는 뉴스문서를 특정 대상으로 선정하였고 사용자 정보를 이용한 검색을 실시하고 결과로 얻어진 정보를 정보 분류를 통해 PDA나 휴대폰을 통해 사용자에게 제공한다. 본 논문의 구성은 다음과 같다. 2장에서는 시스템의 계략적인 설명을 하고, 3장에서는 실시간 검색 방식을 설명하고, 4장에서는 사용자 행위 모니터링 방식, 5장에서는 프로파일 생성 방식, 6장에서는 검색되어진 문서를 분석하는 분석기, 7장에서는 사용자에게 서비스하는 개인화 Pusher, 마지막으로 8장에서는 결론 및 향후 연구에 대해 논하겠다.

2. 시스템 구성도

본 논문에서 제안하는 'PDA용 개인화 에이전트 시스템"은 크게 4부분으로 구성된다. 먼저 사용자가 웹 페이지에서 행동하는 행위를 모니터링하는 모니터, 모니터링 되어진 결과를 기반으로 사용자 프로파일을 생성하는 프로파일 생성기, 인터넷 신문과 웹진등에서 문서를 검색하는 실시간 문서 검색기, 검색되어진 문서를 사용자별로 분류하는 분류기로 구성된다. 다음 [그림 1]은 본 시스템을 간략화 시킨 구성도이다.



[그림 1] PDA 개인화 에이전트 시스템 구조도

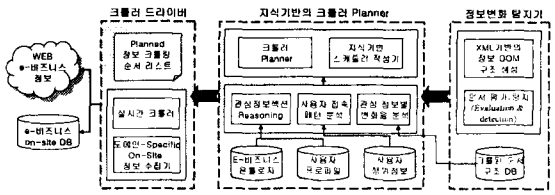
• 사용자 행위 모니터 : 인터넷 웹 페이지에서 사용자의 행위를 모니터하여 모니터 되어진 행위들을 데이터베이스에 저장하게 된다. 모니터링 방법은 웹 페이지에서의 클릭 스트림(Click Stream) 방식을 이용하여 그 여부를 데이터베이스에 저장하는 방식을 사용한다. 모니터링은 후에 사용자 프로파일를 생성하는데 필요한 가장 중요한 자료가 된다.
• 사용자 프로파일 생성기 : 모니터링 되어진 결과가 데이터베이스에 저장되면 주기적으로 사용자의 프로파일을 생성하게 된다. 프로파일 생성을 위해 본 논문에서는 점진적 개념 학습 클러스터링 방식은 Cobweb을 사용하여 모니터링된 결과를 클러스터링하고 이렇게 생성된 클러스터들을 C5.0을 사용하여 Rule 기반인 사용자 프로파일을 만들게 된다. 이렇게 만들어진 사용자 프로파일 룰은 각 사용자별로 데이터베이스에 저장되게 된다.
• 실기간 문서 검색기 : 문서를 검색하는 방법으로 실시간(real-time) 검색기를 사용한다. 실시간 검색기는 시스템에 등록되어 있는 인터넷 신문사와 웹진등의 페이지에 대하여 하이퍼링크(hyper-link)를 기준으로 해당하는 하위 페이지들에 대하여 페이지 수집을 한다. 이때 수집되는 문서의 URL만을 수집한다. 이러한 이유는 문서의 수가 증가함에 따라서 부가되는 문서 저장용량의 부담을 줄이기 위해서 이다.
• 검색 문서 분류기 : 실시간 문서 검색기를 통해 수집된 문서들에 대해서 각 문서의 키워드들을 Feature Selection을 통하여 문서에서의 단어 빈도수를 파악하여 단어 빈도별로 문서를 분류하게 된다. 이렇게 분류되어진 문서를 데이터베이스에 저장한다.
• 뉴스 생성기 : 사용자는 신청한 뉴스 정보를 검색하기 위하여 뉴스 생성기를 이용하게 된다. 뉴스 생성기는 사용자의 뉴스정보를 종류별로 분류하여 보여주게 된다.

3. 실시간 검색 방식

개인화 서비스를 위한 지능형 크롤러은 사용자의 프로파일을 기반으로 사용자가 관심을 가지는 정보를 크롤링하고, 실시간으로 정보의 변화를 탐지하여, 새로운 관심 정보를 찾아 사용자에게 제공하는 연구를 수행한다.

3.1 실시간 검색 엔진

인터넷의 정보의 특징은 방대할 뿐 아니라 매시간 정보가 변화하는 속성을 가진다. 최대한 많은 정보를 보유하려는 기존의 검색서비스에서는 이러한 변화 속도를 따라 갈 수 없다. 그러므로 본 연구에서는 사용자가 관심을 가지는 인터넷 사이트만을 집중적으로 모니터링 함으로 정보가 업데이트되는 즉시 사용자에게 제공하기 위한 엔진인 실시간 웹 크롤러를 구축한다.



[그림 2] 개인화 크롤러

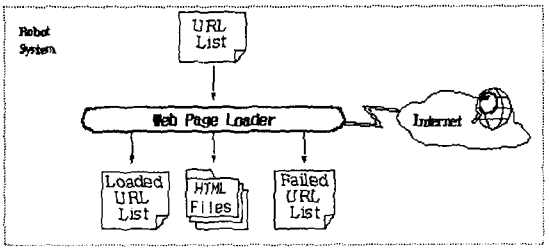
실시간 검색 엔진은 관리자가 미리 정의해 놓은 사이트URL 정보를 참조하여 해당 사이트 범위만을 크롤링 한다. 이러한 크롤링시 가장 주의할 점은 해당되는 사이트의 범위설정 부분이다. 크롤러의 성격상 해당 Seed 정보 페이지를 기반으로 관련 링크들을 추출하고 정리한 후 다음 크롤링 대상 페이지를 찾아가게 되는데 이러한 과정에서 다음과 같은 문제점이 나타난다.

- 사이트 범위를 넘어서는 크롤링 수행
- 게시판과 같은 반복적인 페이지에 대한 리컬시브(Recursive) 크롤링 현상
- 크롤링된 페이지에 대한 반복적인 크롤링으로 인한 시스템 자원낭비

본 연구에서는 실시간 검색 엔진에서 나타나는 위와 같은 문제점들을 해결하기 위하여 다음과 같은 방법을 사용 하였다.

- 특정 사이트 범위는 크롤링되는 페이지의 URL 정보를 참고하여 해당 사이트의 도메인을 규정하고, 크롤러의 수행범위를 제한 시켰다.
- 반복적인 Recursive 크롤링 현상을 막기 위하여 해당 사이트에 대한 크롤링 로그정보를 저장하며, 크롤링 수행시 이를 기반으로 Recursive 크롤링 분제를 해결 하였다.
- 매 크롤링 수행시마다 해당 사이트의 모든 페이지를 수집하는 방식을 탈피하여 수집대상 페이지의 History 정보를 저장함으로써 새로이 업데이트 되거나 추가된 페이지만을 추가적으로 수집하여 기존의 크롤링 DB에 추가하는 방식을 택하였다.

본 연구에서 사용한 위와 같은 방식은 실시간 크롤러의 수행 성능과 속도면에서 많은 이점을 가져왔다.



[그림 3] 실시간 검색기 수행 흐름도

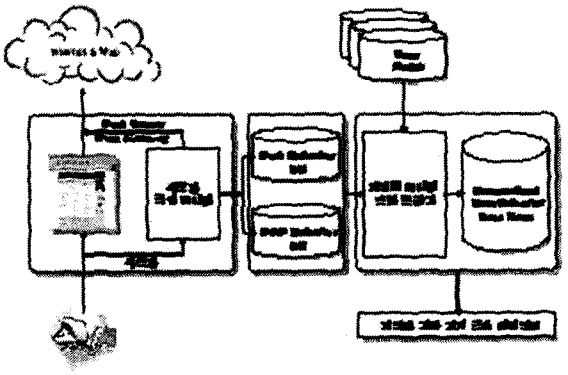
4. 사용자 행위 모니터링 방식

4.1 사용자 행위 모니터링

사용자 행위 모니터링는 학습을 통한 사용자 프로파일 생성을 위한 학습자원을 만들어내는 기능을 수행하는 부분으로 현재 사용자 접속 디바이스를 구별하여 사용자의 브라우저 상에 보여 지는 문서에 대한 URL 및 위치 정보를 유지하며, 사용자의 행위 정보를 추출하여 사용자행위 DB(User Behavior DB)로 저장하는 기능을 수행한다.

4.2 모니터링 구성

개인화 모델 생성을 위한 작업을 수행하는 과정에서 필요한 정보가 개인의 행위 정보와 개인의 Preference 정보이다. 따라서 이 두 가지 정보를 추출하기 위해서는 특정 페이지에 접근한 개인의 모든 정보를 로그화하여 기록할 필요가 있다. 모니터링은 개인의 접근시에 이를 인식하고 접근이 종료될 시점까지의 정보를 추출하는 과정을 말한다. 다음 그림은 이러한 과정을 도식화하였다.



[그림4] 모니터링 에이전트 시스템 구성도

4.3 사용자 행위 데이터베이스 (User Behavior DataBase)

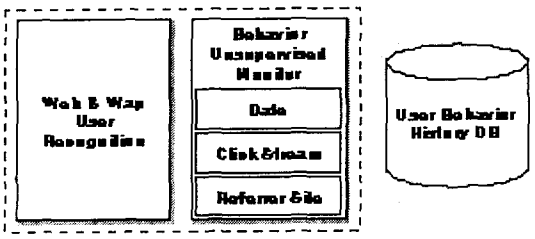
사용자 행위 데이터베이스는 다음과 같은 순서로 생성된다.

- 웹 페이지 모니터 정보 획득 및 분류
- 사용자별 모니터링 히스토리 정보 통합
- 사용자 행위 데이터베이스 입력

웹 페이지에서의 모니터 정보 획득 및 분류 과정이 끝나고 과거의 웹 페이지에서의 모니터링 히스토리 정보 통합 과정을 마친 뒤 각각의 구분된 히스토리 정보들을 하나의 정보군으로 통합하는 과정이 필요하다. 히스토리 정보는 후에 생성될 개인화 모델인 개인 프로파일을 생성하는데 필요한 학습 시스템의 예제정보로 사용된다. 통합된 모니터링 히스토리 정보는 사용자 개인별로 행위 데이터베이스를 구축하는데 이용 된다.

4.4 사용자 행위 정보 추출

사용자 행위 모니터링은 사용자가 서비스 로그인 부터 페이지 내에서의 페이지 이동, 페이지 클릭, 로그 아웃까지의 트랜잭션으로부터 본 연구에서 정의한 사용자 행위 정보들을 사용자 History·데이터베이스에 저장하는 역할을 한다. 이후 이 행위 정보들은 사용자의 관심 정보를 분석하기 위한 기계학습 에이전트의 입력 값으로 사용된다. 아래 그림은 사용자 세션 및 트랙잭션에서의 사용자 행위 정보 추출하는 모듈의 흐름도이다.



[그림 5] 사용자 세션 및 트랜잭션 추출 흐름도

4.4.1 Web user recognition

Web user recognition은 사용자가 유무선 인터넷을 이용하여 페이지내의 Click Stream 정보들이다. 이때 사용자가 Click Stream 정보들은 사용자가 관심을 가지는 정보들이다. 예를 들면 사용자가 뉴스 카테고리 중 정치 카테고리를 선택하고 선택된 정치 뉴스 중 첫 번째 뉴스를 클릭 했을 때의 컨텐츠 정보, 시간 정보, 사용자 연결 IP 정보, 클릭 회수 정보들이다.

4.4.2 Behavior Usupervised Monitor

위에서 언급한 유무선 인터넷을 사용하여 연결하는 사용자의 세션 과 트랜잭션으로부터 본 연구에서 정의한 사용자 행위 정보들을 사용자 행위 히스토리 데이터베이스에 저장하는 모듈이다. 다음은 Unsupervised 모니터 모듈이 추출하는 사용자 행위 정보들이다.

- 클릭한 컨텐츠 정보
- 클릭한 카테고리 정보
- 클릭한 시간
- 클릭한 사용자 아이디

이 정보들은 사용자 디바이스에서 사용자가 컨텐츠를 클릭했을 때 서비스 서버 내에서 동작하는 모니터 에이전트에 의해 정보가 추출되며 User Behavior History 테이블에 저장하게 된다.

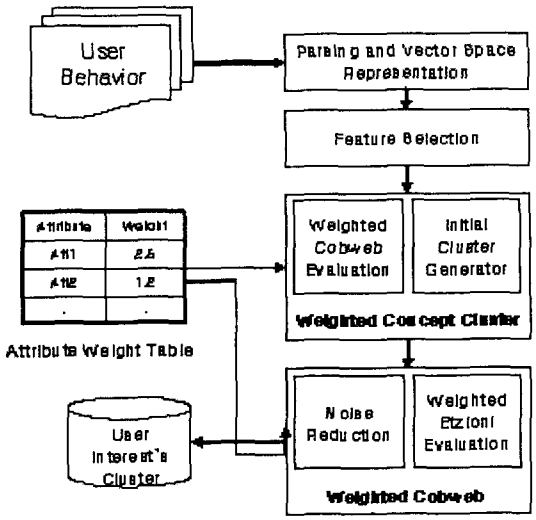
4.4.3 User Behavior History DB

사용자 행위 히스토리 데이터베이스는 사용자가 로그인 한 시간 정보, 사용자가 클릭한 컨텐츠 정보, 어떤 IP 를 사용한지에 대한 정보를 가지고 있다.

User Behavior History DB table schema

- int historyNum : 인덱스 번호
- varchar userID : 사용자 아이디
- varchar servIndex : 서비스 받는 번호
- varchar ipAddress : 사용자 연결 IP 주소
- Date regDate : 등록된 시간
- Int ConnectKinds : 연결 매체 종류(Web & Wap)

5. 사용자 프로파일 생성기



[그림6] 사용자 프로파일 생성

사용자별 모니터링을 통하여 관심 문서의 정보가 추출되면 이를 기반으로 하여 각 사용자의 관심정보 규칙들을 생성할 필요가 있다. 추출된 주요 키워드간의 연관 정보를 생성하고 결정 트리와 규칙 등을 이용하여 사용자의 관심 정보에 대한 구체적인 정보를 파악할 수 있게 된다.

5.1. Conceptual Clustering

Conceptual Clustering이란 실세계 인간의 분류 방식을 모델로 삼아 그 유사한 방식으로 입력 데이터들을 분류하는 방식을 말한다. 이러한 Clustering 처리는 특정 사용자의 관심 데이터들을 각 영역별로 재정의 하여 묶어주는 역할을 수행하며 사용자 Profile 추출에 있어서 중요한 기본 자료로 사용된다.

본 연구에서는 이러한 Conceptual Clustering 방식을 위하여 Top-Down 방식의 Weighted Cobweb 과 Bottom-up 방식의 Weighted Etzioni 방식을 사용한다. 먼저 Weighted Cobweb 방식은 차례대로 입력되는 사용자 관심정보에 대하여 가중치와 노이즈를 제거한 초기 Cluster를 생성하고, Weighted Etzioni 방식을 거치며 정확도가 높은 최종 Cluster를 추출하게 된다. 이러한 방식은 Clustering을 수행하는데 있어서 사용자의 관심 정보에 대한 히스토리를 적용할 수 있으며 특정 관심정보에 대한 가중치를 부여하여 속도와 정확도 측면에서 모두 만족할 만한 결과를 나타내므로 본 연구에 가장 적합한 Clustering 방식이다.

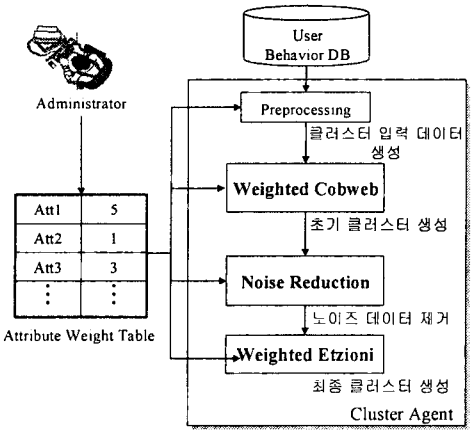
5.2 통합방식 클러스터링 알고리즘

가중치를 적용한 Top-down 방식과 Bottom-up 클러스터링 방식의 장. 단점을 효과적으로 통합하기 위해서는 Cobweb 이 생성하는 문서 분류 트리로부터 효율적인 초기 클러스터 생성이 중요하다.

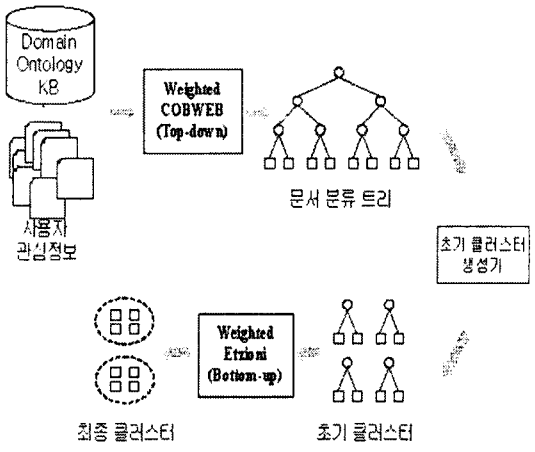
클러스터링 절차는 다음과 같은 4단계로 나뉘어 진다.

[1 단계] 가중치적용 속성 선별 및 가중치 부여
[2 단계] Top-down 클러스터링을 적용하여 문서분류 트리를 생성.
[3 단계] 초기 클러스터 생성 평가함수를 이용하여 초기 클러스터 생성.
[4 단계] Bottom-up 클러스터링을 이용하여 초기 클러스터 병합.

이와 같은 단계를 통하여 최종 클러스터를 생성하게 된다. 이렇게 생성된 최종 클러스터는 이후 학습을 수행하는 단위가 되며 각 클러스터별로 사용자 프로파일의 관심영역을 표현하게 된다. 다음 그림은 통합방식 클러스터링의 과정을 설명하고 있다.



[통합 방식 클러스터링 구조도]



[통합방식 클러스터링 처리 방식]

5.3.1 통합방식 기반 COBWEB, Etzioni 평가함수

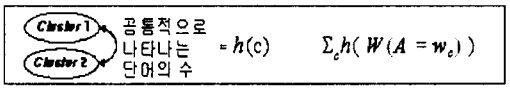
기존 Cobweb의 평가 함수는 입력 데이터의 각 속성들의 중요도를 균등하게 생각하고 평가함수를 계산하였다. 이러한 일률적인 평가함수의 계산 방식은 입력 데이터의 순서에 따른 클러스터링을 수행할 때 다음과 같은 문제점을 내포하고 있다. 먼저 클러스터링을 수행하는데 있어서 고려되는 모든 데이터의 속성들에 대한 중요도를 똑같이 인식한다. 그럼으로써 서로 별다른 연관성이 없는 데이터들에 대해서도 작은 유사도에 의한 클러스터 생성을 수행하고 있다. 결과적으로 이러한 작은 유사도는 노이즈 데이터를 포함하게되는 원인이 되며 초기 클러스터에 대한 정확도를 떨어뜨리는 원인이 된다.

이러한 단점을 극복 하고자 각 속성에 대한 가중치를 설정하여 각 속성간의 중요도를 부여하고 이를 평가함수에 반영한 가중치 기반 Cobweb 평가함수를 제안한다. 다음은 속성에 따른 가중치 부여를 위하여 변형된 Cobweb의 평가함수이다.

$$\frac{\sum_{k=1}^{K} P(C_k)\sum_{i=1}^{I}\sum_{j=1}^{J}[P(A_i=V_{ij}\mid C_k)W(A_i)]^2-\sum_{i=1}^{I}\sum_{j=1}^{J}P(A_i=V_{ij})^2}{k}$$

기본적인 Cobweb 평가함수에 대해서는 앞에서 설명하였으므로 해당하는 속성에 대한 가중치 부여 부분에 대해서 설명하겠다. 위 식에서 $P(A_i = V_{ij} \mid C_k)$ 다음에 부여된 $W(A_i)$는 해당하는 속성값에 부여된 가중치 값이다. 여기서 $P(A_i = V_{ij} \mid C_k)$ 는 주어진 분류에 대하여 개체의 속성이 특정값을 가지는 확률값을 나타내기 때문에 해당 속성의 가중치 값을 나타내는 $W(A_i)$의 값을 곱하여 줌으로써 전체 평가함수의 계산결과는 증가 또는 감소할 것이다.

다음은 Etzioni의 Bottom-up 클러스터링 평가 함수(GQF)의 가중치를 부여한 평가함수를 제안한다. GQF 평가함수에서는 클러스터 c 의 응집도를 클러스터에 속하는 모든 문서들에 공통적으로 나타나는 속성의 수로 정의하고 $h(c)$로 표시했었다. 따라서 이 속성을 이용하여 가중치를 부여한다.



$W(A)$: Attribute A 에 대한 가중치 값 , w : h(c) 에 포함된 단어

본 논문에서는 사용자 모니터링 행위 정보를 전처리 과정으로 클러스터링 방법을 사용하여 C4.5 기계학습에 입력 값으로 생성한다.

5.3. 사용자 관심 정보 기반 규칙 생성

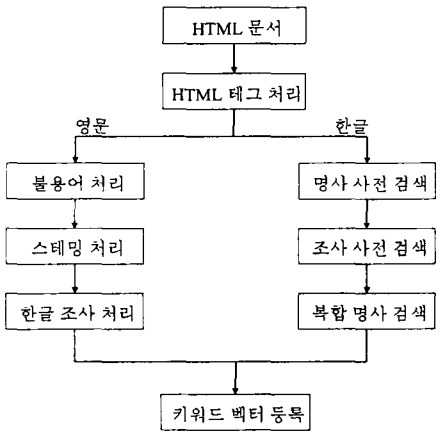
5.3.1. C4.5 기계학습 시스템

본 연구에서 학습 작업에 사용하고 있는 C4.5는 ID3계열의 귀납적 기계 학습 시스템이다. C4.5는 어느 특정 종류별로 모아놓은 레코드들을 대상으로 각각의 특성에 따라 밝혀진 패턴을 발견하고 분석

한다. 분석된 정보는 각각의 유사한 분야별로 나누어지는 분류모델(classification model)을 생성한다. 분류된 분야별 모델은 각각의 속성값에 따라 트리 형태에 의하여 생성된다. C4.5는 이러한 결정 트리를 이용하여 일정한 룰을 생성한다. 생성된 룰은 키워드에 따른 카테고리의 분류를 가능하게 하며 마찬가지로 카테고리에 따른 키워드도 추출할 수 있게 한다. 본 연구에서는 사용자의 행위 정보를 가지고 룰을 생성하게 된다.

C4.5 학습 시스템은 먼저 트리의 형태로 학습결과가 표시되는데 생성되는 룰은 이를 간략화된 형태의 트리를 이용하여 표시된다. 이 정보는 사용자가 특정 카테고리에서 자신의 기호에 맞게 선택한 웹 화면들의 키워드 정보들 간의 전체적인 규칙을 도출해낸 것이다. 그러므로 이는 사용자가 수많은 정보들 중에 자신이 지정한 카테고리의 범주내에 속한다고 판단하는, 즉 사용자가 관심문서를 선택하는 데 있어서 기준이 되는 키워드들인 것이다. 이렇게 결정트리와 규칙 등을 구성하고 있는 키워드들이 본 시스템의 프로파일 생성기를 통해 사용자의 관심 정보에 대한 구체적인 정보를 파악할 수 있게 된다.

6. 문서 분석기



[그림 7] 관심문서 전처리 과정 흐름도

실시간 검색기로부터 검색 되어진 문서는 사용자의 관심 문서 여부 파악을 위해서 전처리를 필요로 한다. 이 전처리 과정에는 입력 문서들로부터 HTML 테그와 불용어 제거 작업, 스테밍 작업등을 수행하여 키워드들을 추출한다. 이렇게 생성된 키워드들 중에서 사용자의 관심 영역을 잘 표현하는 중요 키워드을 추출하기 위해서 특성 추출(Feature Selection) 기법을 적용하여 각 카테고리를 대표하는 키워드 벡터 리스트를 생성한다.

관심문서의 전처리는 위에서 설명한 바와 같이 사용자가 관심을 표시한 문서를 입력으로 하여 키워드를 추출하는 과정을 말한다. 여기서 추출된 키워드는 사용자가 관심 있어 할 가능성이 있는 키워드로서 문서 전처리의 다음 단계인 귀납적 기계학습의 입력 값으로 사용된다. 본 연구의 가장 큰 목적은 사용자의 관심사항을 알아내는데 있다. 이를 위해서는 문서 전처리 단계에서 사용자가 관심 있어 할 만한 키워드를 추출하는 것이 매우 중요하다. 관심문서의 전처리 과정은 위의 그림에서 기술한 순서로 진행되어 중요한 키워드들을 추출해낸다. 먼저 입력으로 사용자가 관심을 표시한 HTML 문서를 읽어들여 HTML 테그를 제거한다. 그리고 입력 토큰이 영문인지 한글인지에 따라 처리 순서가 달라진다. 영문일 경우에는 불용어 처리, 스테밍 처리, 한글 조사 처리의 단계를 거쳐 키워드로 저장이 된다. 한글일 경우에는 명사 사전 검색, 조사 사전 검색, 복합 명사 처리의 단계를 거쳐서 키워드로 저장이 된다. 각 단계에서 사용되는 자료로는 영문에서는 불용어 리스트, 한글 조사 사전이 사용되고, 한글에서는 한글 명사 사전, 한글 조사 사전, 한글 불용어 리스트가 사용된다.

6.1 Feature selection을 이용한 문서 분석기

많은 양의 학습예제 중에서 특정 영역에 속하는 예제들의 특징을 추출해내는 특성 추출기법이 Feature Selection이다. 사용자의 Behavior DB를 구성하는 모든 속성들 중에서 실질적으로 학습 작업에 중요한 비중을 갖는 양질의 속성만을 선정하는 작업이다. 이를 위해 영역을 구성하는 속성의 특성을 대표적으로 표현하는 속성 값들만을 추출해내기 위해 특성 추출 기법을 적용한다.

Feature를 추출하는데 있어서 가장 중요한 Feature Selection Algorithm은 학습예제로 구성된 예제 집합에 기입된 각각의 Feature들의 특징을 정의하는 텍스트 학습 기법들이 주로 사용되는데 본 연구에서는 이들 중에서 Exp기법을 사용한다. Exp의 수식은 다음과 같다.

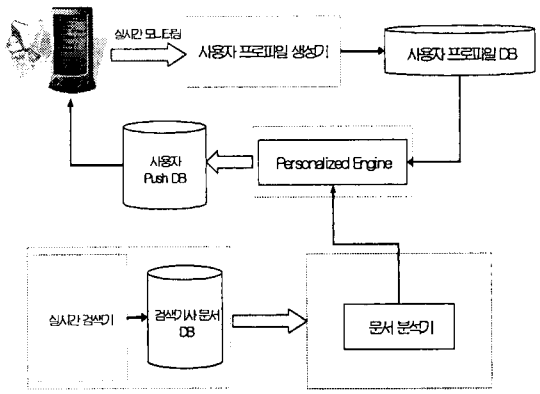
$$ExpP(A) = \{ e\}^{P(W|C_1)-P(W|C_2)} \}$$

위에서 $P(W|C_1)$는 특정 클래스 C_1에서 특정 Feature W가 나온 확률 값으로서 위의 계산 값을 모든 Feature에 적용하여 웨이트(Weight)를 주는 방법을 사용하였다. 이처럼 각 Feature마다 가중치를 설정해주는 작업은 추출될 최종 중요 Feature에 커다란 긍정적 영향을 끼치게 된다. 이렇게 만들어진 문서의 feature 들을 이용하여 사용자 프로파일과 매칭을 통해 관심 문서 여부를 파악하게 된다.

7. Personalization PDA pusher

사용자 프로파일을 통해 만들어진 사용자 규칙 룰을 이용하여 개개인의 규칙에 맞게 서비스할 필요가 있다. 사용자의 모니터링을 통해 얻어진 여러 정보를 통합하고 이를 클러스터링을 통해 분류하고 C4.5 기계학습을 이용하여 최종적으로 사용자 관심

정보를 추출하게 된다. 이렇게 만들어진 사용자 관심 정보를 기반으로 새롭게 검색되어져 저장되는 문서들을 HTML 태그 제거기, 불용어 제거, 조사 제거를 통하여 문서의 중요 Contents를 추출한 후 Feature Selection을 이용하여 문서 분석을 하게 된다. 이러한 문서 분석을 통하여 사용자의 관심정보와 매칭이 되는 문서라고 판단 될 경우 새로운 정보가 검색 되었다는 메시지와 함께 제공되게 된다.



[그림 8] PDA Pusher 구조도

본 연구에서 개인화 PDA 에이전트 모듈을 통해 사용자와의 통신을 하게 된다. PDA안에 내장되어지는 모듈은 어플리케이션의 형태로 서버와의 통신을 할 수 있는 클라이언트 모듈과 서버에서 PDA로부터 전달되어지는 정보를 받아 처리하는 서버 모듈로 구분된다. 클라이언트 모듈은 사용자의 사용 여부와 로그인을 통해 접속했을 때 새로운 정보의 업데이트 여부를 알려주며 서버로부터 사용자 관심정보 기반으로 만들어진 서비스 내역을 제공한다. 또한 이렇게 만들어진 서비스를 제공하고 동시에 모니터링을 함으로써 그 정보를 사용자 프로파일을 업데이트 하는데 주기적으로 이용하게 된다. 사용자의 모니터링 정보는 PDA에서 자동으로 정장되며 이는 후에 접속을 했을 때 서버로 전달되어 사용자의 프로파일을 업데이트 하게 되는데 사용되어진다.

8. 결론 및 향후 연구

현대 사회는 '정보의 홍수'라는 단어로 표현된다. 하지만 1990년대부터 급속히 확산되기 시작한 인터넷 환경에 힘입어 2000년대는 '정보 빅뱅' 이라는 표현이 생겨나게 되었다. 이렇게 하루에도 수천만 페이지에 달하는 엄청난 정보가 새로이 생성되는 시점에서 인터넷 사용자들은 자신이 원하는 정보를 찾기가 더욱 어려워지고 정보검색을 위해 투자하는 시간 또한 증가하고 있다. 이와 더불어 기존의 유선 인터넷환경을 뛰어 넘는 무선인터넷 환경으로의 변화가 생겨났다. 무선인터넷은 개인 휴대폰과 PDA, 무선 노트북, 핸드헬드 PC와 같은 여러 디바이스를 매개로 언제 어디서든지 원하는 정보를 사용자에게 전달할 수 있는 길을 열어 주었다. 이러한 급속한 인터넷 환경 변화와 사용자 접속 디바이스의 다양화의 중심에는 '사용자에게 원하는 정보를 언제 어느 때든지 전달' 하겠다는 기본 개념이 있다고 하겠다.

본 'PDA용 개인화 에이전트 시스템'에 대한 연구는 이러한 사용자들의 정보습득요구를 더욱 지능화, 자동화하기 위한 연구이다. 사용자가 원하는 정보를 직접 찾아다니는 것이 아니라 최소한의 사용자 프로파일을 기반으로 자동화된 정보검색 에이전트의 정보 수집과 검색, 학습에이전트를 통한 사용자 맞춤형 정보 생성하는 시스템을 연구하였다.

9. 참고 문헌

[1] 데이터 마이닝 이론과 실제 (저자 최국렬 외)

[2] "A Survey of Data Mining Software Tools" by Michael Goebel and Le Gruenwald, ACM SIGKDD Exploration, June 1999, Volume 1, Issue 1

[3] "Predictive Data Mining: A Practical Guide" by Sholom M. Weiss and Nitin Indurkhya , Morgan Kaufmann Publishers, Inc. 1998.

[4] http://www.itbiz.co.kr 기사 참고.

[5] Fisher, D. H., & Langley, P., "Methods of conceptual clustering and their relation to numerical taxonomy," In W. Gale (Ed.), Artificial intelligence and statistics, Reading MA: Addison Wesley, 1986.

[6] Gennari, J. H., Langley, P., & Fisher, D. H., "Models of incremental concept formation," Artificial Intelligence, 40, pp. 11-61, 1989.

[7] Oren Zamir, Oren Etzioni, Omid Madani and Richard M. Karp, "Fast and Intuitive Clustering of Web Documents," KDD'97

[8] Mark Devaney, Ashwin Ram, "Efficient Feature Selection in Conceptual Clustring", Machine Learning: Proceeding of the Fourteenth International Conference, Nashville, 1997

[9] Doug Fisher, "Interative Optimization and Simplification of Hierarchical Clusterings," AI Access foundation and Morgan Kaufmann Publishers, 1996.

[10] Gennari, J. H., Langley, P., & Fisher, D. H., "Models of incremental concept formation," Artificial Intelligence, 40, pp. 11-61, 1989.

[11] Gluck, M., & Corter, J., "Information, uncertainty and the utility of categories," Proceedings of the Seventh Annual Conference of the Cognitive Science Society, pp. 283-287, Irvine, CA: Lawrence Erlbaum, 1985.

[12] Jorma Laaksonen, Erkki Oja, "Classification with Learning k-Nearest Neighbors", Proceedings of ICNN'96, pp. 1480-1483, Washington, DC, 1996

[13] T.M. Mitchell, "Machine Learning", McGraw Hill, 1997.
[14] Kathleen Mckusick, Kevin Thompson, "COBWEB/3 : A Portable Implementation", NASA Ames Research Center, Technical Report FIA-90-6-18-2, 1990
[15] Wettschereck, D., Aha, D. W., & Mohri, T. "A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms", Artificial Intelligence Review, pp. 273-314, 11, 1997
[16] 양찬범, "웹 에이전트를 위한 통합방식 문서 클러스터링" 숭실대학교 석사학위논문, 2000
[17] 소영준, "사용자 관심도 추출을 위한 모니터 에이전트 시스템" 숭실대학교 석사학위논문, 2000
[18] 이성열, "점진적 클러스터링에서의 노이즈 제거" 숭실대학교 석사학위논문, 2001
[19] Orkut Buyukkokten, Hector Garcia-Molina, and Andres Paepcke. "Accordion Summarization for End-Game Browsing on PDAs and Cellular Phones." In Proceedings of the Conference on Human Factor in Computing Systems, 2001.
[20] Orkut Buyukkokten, Hector Garcia-Molina, and Andres Paepcke. and Terry Winograd. "Power Browser: Efficient Web Browsing for PDAs." In Proceedings of the Conference on Human Factors in Computing System, 2000.