사용자 그룹을 이용한 효과적인 정보 여과 및 학습 방법에 관한 연구*

송미란(mirans@cs.sookmyung.ac.kr), 김교정(kiochkim@sookmyung.ac.kr)
숙명여자대학교 전산학과

A Study on Efficient Information Filtering and Learning Using User Group Filtering

Song Mi Ran, Kim Kio Chung
Dept. of Computer Science, Sookmyung Women's University

요약

인터넷의 발달은 정보의 폭발적인 증가를 가져오게 되었고 더불어 일반인은 어디서나 쉽게 정보를 습득할 수 있게 되었지만 늘어나는 정보의 양이 원하는 정보의 습득을 방해하게 되었다. 이러한 정보 과잉현상을 해결하기 위해 사용자가 원하는 정보만을 여과해 주는 정보 여과 시스템이 연구되고 있다. 정보 여과 시스템은 사용자의 관심도를 파악하기 위해 사용자 프로파일을 구축하고 이를 학습을 통해 갱신한다.

하지만 기존의 개인 프로파일을 이용한 정보 여과 시스템은 개인의 관심도를 분석하기 위해 에이전트가 학습하는 시간이 너무 오래 걸린다는 단점과 사용자의 능력에 따라 적합한 문서를 검색하기 위한 정보가 너무 한쪽으로만 치우치는 우려가 있다. 따라서 본 논문은 효과적인 프로파일 학습을 위해 비슷한 관심도를 갖는 다른 사용자로부터 학습을 받는 방법을 제안한다. 이를 위해 그룹 프로파일을 구축하는 방법과 그룹 프로파일을 이용한 효과적인 정보 여과 방법, 그리고 그룹 프로파일 학습방법에 대해 기술한다.

1. 서론

인터넷의 발달과 더불어 일반인이 습득할 수 있는 정보의 양이 크게 증가하고 있다. 그렇지만 정보량의 증가가 정보 습득의 증가를 의미하지는 않는다. 범람하는 정보로 인하여 원하는 정보를 사용자가 얻지 못하는 경우도 많이 발생하게 된다. 이러한 정보 과잉 현상을 완화시켜 주기 위한 효율적인 방법들이 연구되고 있는데, 그 중에 하나가 에이전트를 이용한 방법이다[1, 2, 3, 7]. 인터넷에서 수많은 검색 엔진들이 사용자의 효율적인 정보 수집을 도와주지만 수집된 정보 모두가 사용자가 원하는 정보일 수는 없다. 따라서 사용자의 요구에 적응할 수 있는 적응형 개인 웹 에이전트(Adaptive Personal Web Agent)들이 사용자의 정보 습득을 도와주기 위해 연구되고 있다. 이러한 개인 웹 에이전트들은 검색된 정보 중 사용자가 원하는 정보만을 골라내어 제공하여 주는 정보 여과 시스템에서 사용된다.

정보 여과 시스템은 사용자 개인의 관심도를 표현하는 사용자 프로파일을 구축하여 정보를 여과한다. 그렇지만 사용자가 자신의 관심도를 정확히 지적해 내는 것은 쉽지 않고, 사용자의 관심을 학습을 통해 습득하는 것 또한 오랜 시간이 걸린다[7]. 그래서 본 논문에서는 이를 개선하기 위해 사용자와 비슷한 다른 사용자의 프로파일로부터 학습할 수 있는 그룹 프로파일을 이용한 그룹 정보 여과를 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 사용자의 정보 습득 방법인 정보 검색과 정보 여과에 대해 그 차이점과 정보 여과 시스템의 과정을 알아보고, 3장에서는 그룹 프로파일을 이용한 정보 여과 시스템의 구성과 각 단계를 기술하고, 끝으로 4장에서는 향후 과제와 결론을 맺도록 한다.

2. 정보 검색과 정보 여과

정보 범람에 대처하기 위해서 일반적으로 정보 검색과 정보 여과 방법을 사용한다. 이들은 관심 있는 주제와 부합하는 문서를 찾아내고 특정 패턴과 일치하는 문서를 찾아낸다는 것에 있어서 비슷한 역할을 한다[5]. 일반적으로 정보 여과와 정보 검색은 비슷한 개념으로 이해되거나 별다른 구분 없이 혼용되어 쓰이는 경우가 많다.

Belkin과 Corft는 정보 여과와 정보 검색의 차이점에 대해 기술한 바 있다. 정보 여과는 시간에 따라 변화하는 정보 원천(dynamic information source)을 대상으로 하는 반면, 정보 검색은 일반적으로 정적인 특성을 갖는 대규모의 정보 원천(static information source)을 대상으로 한다는데 그 차이점이 있다. 또한, 정보 여과에서는 장기적인 사용자의 관심을 처리하기 위해 사용자 프로파일을 구축한다[4].

정보 여과 시스템의 사용자들은 장기적으로 변동이 적은 관심과 취향을 보이고 있다고 가정하고 있으며, 이러한 가정은 정보 여과와 정보 검색의 또 다른 차이점으로 인식되고 있다. 즉, 정보 여과에서와 달리 정보 검색을 이용하는 사용자들의 정보 요구 특성은 간헐적이며, 단기적인 면을 보인다.[8]

정보 여과 시스템은 다음과 같은 과정으로 이루어진다. 사용자의 관심도를 파악하여 사용자 프로파일을 구축하는 사용자 모델링 단계와 검색된 문서를 표현하기 위한 문서 특징 추출 단계가 우선 수행되고, 다음으로 검색된 문서가 사용자의 관심문서인지를 판단하기 위한 사용자 프로파일과 문서의 유사성이 측정되며, 피드백을 통한 학습으로 사용자의 특성에 프로파일을 일치시키고 변화하는 사용자의 관심에 적응해 나간다.

* 본 연구는 여자 대학 연구 기반 확충 사업의 지원으로 수행되었습니다.

3. 사용자 그룹을 이용한 정보 여과 시스템

기존의 개인 프로파일을 이용한 정보 여과 시스템은 개인의 관심도를 분석하기 위해 에이전트가 학습하는 시간이 너무 오래 걸린다는 단점과 사용자의 능력에 따라 적합한 문서를 검색하기 위한 정보가 너무 한쪽으로만 치우친다는 우려가 있다. 따라서 본 논문에서는 이미 학습되어 있는 다른 사용자의 프로파일로부터 에이전트가 학습을 받도록 하여 학습 시간의 단축과 학습의 효율을 높이는 것을 목적으로 한다.

3.1 사용자 모델링

사용자 모델링은 사용자의 관심분야와 습관의 프로파일을 생성해 내는 과정으로 정의 할 수 있다[2]. 정보 여과 시스템에서 개인 웹 에이전트는 사용자의 행위를 관찰하면서 사용자의 관심도를 추출하고 이를 이용하여 사용자에게 보다 편리한 정보 습득 환경을 제공하는 것을 목적으로 한다.

본 연구에서는 다른 사용자 프로파일로부터 학습을 받도록 하므로 비슷한 관심을 갖는 사용자들을 찾아야 한다. 이는 사용자들 간의 친밀도를 측정함으로서 수행할 수 있다. 한 명의 사용자에 의해 학습을 받기보다는 사용자들간의 친밀도에 근거하여 비슷한 관심도를 보이는 사용자들의 그룹을 형성하여 학습을 받도록 하는 것이 더 효율적일 것이므로 사용자 프로파일 뿐 아니라 그룹 프로파일을 구축한다. 사용자 프로파일을 사용자 그룹 프로파일과 비교하여 개인 프로파일이라고 본 논문에서는 이를 불인다.

본 논문에서 제안하는 개인 프로파일과 그룹 프로파일의 구조는 각각 <식 1>과 <식 2>와 같다. 개인 프로파일은 키워드에 대한 개인의 관심도로 이루어지고, 그룹 프로파일은 그룹의 키워드 관심도와 그룹에 대한 사용자의 친밀도로 이루어진다.

$$\vec{P}_i = (w^{p}_{i1}, w^{p}_{i2}, \cdots, w^{p}_{ik}) \qquad \text{<식 1>}$$

$\vec{P}_i$: i번째 사용자 개인의 관심도 벡터

w^{p}_{ik} : i번째 사용자의 k번째 키워드에 대한 관심도

$$\vec{G} = (w^{g}_{1}, w^{g}_{2}, \cdots, w^{g}_{k})$$

$$\vec{M} = (w^{m}_{1}, w^{m}_{2}, \cdots, w^{m}_{j}) \qquad \text{<식 2>}$$

$\vec{G}$: 그룹의 그룹 관심도 벡터

w^{g}_{k} : 그룹의 k번째 키워드에 대한 관심도

$\vec{M}$: 사용자의 그룹 친밀도 벡터

w^{m}_{j} : 사용자 j의 그룹에 대한 친밀도

그룹 프로파일은 그룹의 키워드 선호도 벡터($\vec{G}$)와 사용자가 그룹에 속하는 그룹 친밀도 벡터($\vec{M}$)로 이루어지는데, 그룹 관심도와 사용자 관심도간의 친밀도 계산은 <식 6>의 벡터 코사인 유사도에 의해 측정한다. 측정된 값에 따라 일정 임계치 이상인 사용자만 그룹에 포함시키고 임계치 이하의 사용자는 친밀도를 0으로 한다. 사용자가 그룹에 적은 친밀도를 가진다할지라도 그러한 사용자가 많으면 그룹의 관심도에 영향을 미칠 수 있기 때문이다. 그래서 일정 수준 이상의 친밀도를 갖는 사용자만 그룹 친밀도 벡터에 포함시킨다.

그리고 그룹 프로파일에서 그룹의 키워드 선호도는 그룹에 속해있는 사용자들의 키워드 선호도에 의해 표현된다. 사용자들의 키워드 관심도와 사용자가 그룹에 속하는 친밀도로 그룹의 키워드 선호도를 계산하면 <식 3>과 같다.

$$w^{g}_{ik} = \frac{\sum_{i=0}^{j} w^{m}_{ij} w^{p}_{ik}}{j} \qquad \text{<식 3>}$$

j : 사용자 수

이렇게 구축된 사용자 프로파일은 학습과정을 통하여 점차적으로 사용자의 실제 특성에 근접해 나갈 수 있어야 하며 변화하는 사용자의 특성에 적응해 나갈 수 있어야 한다. 이를 위해 다양한 학습 방법이 이용되어 왔는데 본 연구에서 사용한 학습 방법은 3.4절에서 설명하도록 한다.

3.2 문서 표현

사용자 프로파일과 문서의 유사도 측정을 통해 일정 임계치 이상의 문서만이 사용자에게 제공되는데, 문서는 유사도 측정이 가능하도록 적절한 방법으로 표현되어야 한다. 본 연구에서는 문서의 표현을 각 문서로부터 제한된 수의 키워드를 추출해 내는 것으로 정의하고 이에 대한 방법을 제시한다. 사용자의 관심은 순수하게 키워드와 그와 연관된 가중치에 의해서만 표현된다고 가정한다. 비록 이것은 명백한 한계점이 있기는 하지만 일반적으로 많은 정보 검색 시스템에서 이 방법을 사용하고 있다[3].

가장 일반적으로 사용되는 문서 표현 방법은 벡터 공간 표현 방법이다. 문서를 문서 벡터 $\vec{D}$로 표현하면 벡터 $\vec{D}_i$는 i번째 문서 벡터를 의미한다. 벡터 $\vec{D}_i$는 i번째 문서 속에 있는 키워드들의 가중치로 이루어진다. 가중치는 TF-IDF[1]방식을 사용하여 계산된다. 그런데 이 방식은 문서의 길이에 따라 가중치의 차이가 나기 때문에 정규화 시켜야 한다. 이 벡터의 구조와 가중치 계산이 각각 <식4>과 <식5>에 나타난다[3].

$$\vec{D}_i = [w^{d}_{i1}, w^{d}_{i2}, \cdots, w^{d}_{ik}] \qquad \text{<식 4>}$$

w^{d}_{ik} : i번째 문서에서 k번째 키워드의 가중치

$$w^{d}_{k} = \frac{\left(0.5 + 0.5\frac{tf(k)}{tf_{max}}\right)\left(\log\frac{n}{df(k)}\right)}{\sqrt{\sum_{d_j \in D}\left(0.5 + 0.5\frac{tf(j)}{tf_{max}}\right)^2\left(\log\frac{n}{df(j)}\right)^2}} \qquad \text{<식 5>}$$

$tf(k)$: 문서 D에 나타난 k번째 키워드의 횟수

$df(k)$: k번째 키워드를 포함하는 문서의 개수

n : 전체 문서의 수

tf_{max} : 문서 D안의 모든 단어 중 tf의 최대 값

3.3 유사도 측정

사용자 프로파일과 문서 벡터 사이의 유사도 측정에는 유클리드 거리 측정법과 코사인 유사도 측정법 두 가지가 있다[9]. 유클리드 거리 측정법은 문서의 양이 늘어날수록 계산량이 많아진다는 단점 때문에 코사인 유사도 측정법이 주로 사용되는데, 코사인 측정법은 비교대상인 벡터가 희소 행렬로 표현될 경우 계산량을 줄일 수 있는 장점이 있기 때문이다. 이 유사도 측정식이 <식 6>에 나타난다[8]

$$S^{p}_{i} = \frac{\vec{D}_i \cdot \vec{P}}{|\vec{D}_i||\vec{P}|} \qquad \text{<식 6>}$$

S^{p}_{i} : i번째 문서와 개인 프로파일과의 유사도

D_i : i번째 문서 벡터

P : 사용자 관심도 벡터(사용자 프로파일)

또한 본 논문에서 제안하는 그룹 프로파일에 의한 유사도 측정은 <식 7>과 같다.

$$S^{gl}_{i} = \frac{\vec{D}_i \cdot \vec{G}_l}{|\vec{D}_i||\vec{G}_l|} \qquad \text{<식 7>}$$

S^{gl}_{i} : i번째 문서와 l번째 그룹프로파일과의 유사도

3.1절에서 구축된 프로파일들과 3.2절에서 표현된 문서벡터를 이용하여 문서들이 <식 6>과 <식 7>의 벡터 코사인 방법으로 여과된다. 개인 프로파일에 의해 여과된 문서와 그룹 프로파일로 여과된 문서의 합

1) TF-IDF : TF × IDF

TF : Term Frequency, 키워드 빈도수. 문서에서 해당 키워드가 나타나는 횟수

IDF : Inverse Document Frequency, 역문서 빈도수. 전체 문서에서 해당 키워드가 나타나는 문서의 수

집합이 유사도와 함께 사용자에게 제공된다. 그러나 하나의 문서에 대해 개인 프로파일에 의해 여과된 유사도와 그룹 프로파일에 의해 여과된 유사도는 달라지므로 문서의 유사도를 재계산할 필요가 있다. i번째 문서에 대한 개인 프로파일에 의한 유사도를 S_i^p 라하고, 그룹 프로파일에 의한 유사도를 S_i^g 라 할 때, 사용자 k에게 제공되는 그 문서의 유사도 S_i 는 <식 8>에 의해 재계산된다. 여기에서 개인 프로파일에 의해 여과된 문서 유사도의 가중치는 1이고 그룹 프로파일에 의해 여과된 문서 유사도의 가중치는 사용자 k의 그룹 친밀도임을 알 수 있다.

$$S_i = \frac{S_i^p + S_i^g w_k^m}{j+1} \qquad <식\ 8>$$

예를 들어 10개의 문서 벡터 $\vec{D_1}$부터 $\vec{D_{10}}$까지가 있다고 가정하고 유사도 S_i가 0.5이상인 문서들만 여과한다고 가정하자. 개인 프로파일로 여과된 문서는 문서 1, 문서 5, 문서 8이고 그룹 프로파일을 이용하여 여과된 문서는 문서 1, 문서 3, 문서 8일 때, 사용자에게 제공되는 문서는 문서 1, 문서 3, 문서 5, 문서 8이다. 여기서 문서 1에 대한 유사도 S_1과 문서 8에 대한 유사도 S_8은 사용자의 그룹 친밀도에 의해 다시 계산되어 사용자에게 제공된다. 즉, 문서 1에 대해서 개인 프로파일에 의한 유사도 S_1^p가 0.5이고 그룹 i의 프로파일에 의한 유사도 S_1^g 는 0.6이라 할 때, 사용자 l이 그룹 i에 속하는 친밀도 w_k^m 을 0.9라 하면 문서 1의 유사도 S_1은 <식 8>에 의하여 0.52가 된다. 문서 8도 이와 같은 방법으로 유사도가 재계산되어 사용자에게 제공된다.

3.4 프로파일 학습

학습과 적응은 거의 같은 개념으로 사용된다. 시스템은 학습을 통해 사용자의 관심에 적응해 나간다

연관 피드백에 의한 학습으로 사용자 프로파일을 사용자의 관심과 습관에 적응시켜 나간다. 문서 벡터 $\vec{D_i}$에 대해 사용자가 피드백 e_i를 주었다면 사용자 프로파일 벡터 P는 <식 9>에 의해 갱신된다[3].

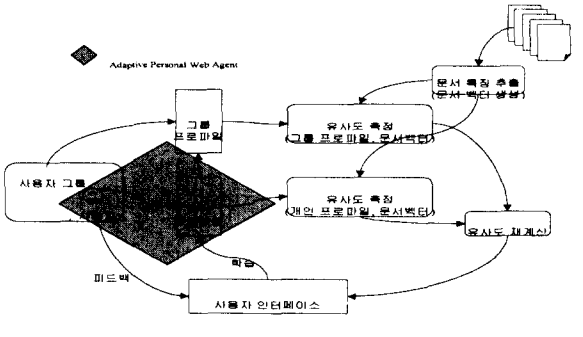
$$\vec{P} = \vec{P} + e_i \vec{D_i} \qquad <식\ 9>$$

n : 정보 여과시스템을 통해 여과되어 사용자로부터 피드백을 받은 문서의 수

사용자는 3.3의 유사도 측정에 의해 제공된 문서에 대해 적합성 평가를 하게 되고 이 평가를 통해 프로파일이 학습되어 사용자의 관심도에 적응하게 된다. 또한 그룹 내에 속한 다른 사용자들의 프로파일도 그룹 친밀도를 이용한 <식 10>에 의해 갱신되어 비슷한 관심도를 보이는 사용자에 의해 학습이 이루어진다.

$$\vec{P} = \vec{P} + e_i \vec{D_i} w_k^m \qquad <식\ 10>$$

본 논문에서 제안하는 그룹프로파일을 이용한 정보여과시스템의 구성도는 <그림 1>과 같고 본 연구의 구현 화면은 <그림 2>와 같다.



<그림 1> 그룹프로파일을 이용한 정보여과 시스템 구성도

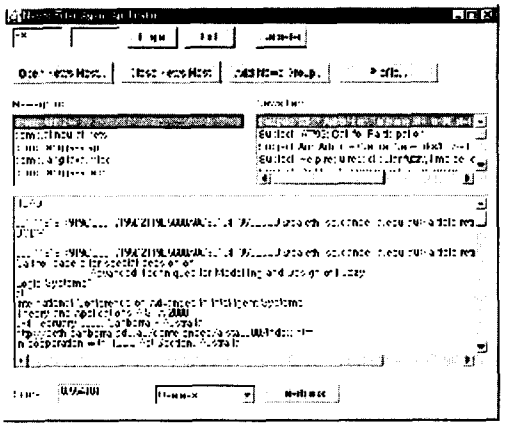


그림 2 정보 여과 시스템 구현 화면

4. 결론 및 향후 과제

본 논문은 범람하는 정보를 효과적으로 습득하기 위한 정보 여과 시스템에서의 사용자 관심도 학습을 위해 사용자 프로파일을 다른 사용자 프로파일로부터 학습하는 방법을 제안한다. 이는 사용자 프로파일을 학습하는데 있어 생길 수 있는 많은 시행착오와 학습 시간을 줄일 수 있다. 다른 사용자로부터 학습을 하기 위해 비슷한 사용자의 그룹을 형성하고 그룹의 프로파일을 생성하는 방법과 그룹 프로파일을 이용하여 효과적으로 정보를 여과하는 방법, 그리고 프로파일을 학습하는 방법을 연구하였다.

본 연구의 구현은 뉴스 그룹을 여과하는 것으로 자바를 사용하여 구현하였으며 실험을 통한 성능검증을 향후과제로 한다. 한편 인터넷에서 웹브라우저를 기반으로 정보 여과를 할 수 있도록 하고자 한다. 또한 그룹 정보 여과를 위한 그룹 형성과 그룹 프로파일 생성에 많은 연구가 필요하다.

참고문헌

[1] Moukas A., "Amalthaea: Information Discovery and Filtering using a Multiagent Evolving Ecosystem", Proceedings of the Conference on Practical Application of Intelligent Agents & Multi-Agent Technology, London, 1996

[2] Robert Armstrong, "WebWatcher: A Learning Apprentice for the World Wide Web", AAAI, 1999

[3] Marko Balabanovie, "An Adaptive Agent for Automated Web Browsing", Stanford University Digital Library Project Working Paper, 1995

[4] Douglas W. Oard, "User Modeling for Information Filtering", http://www.clis.umd.edu/dlrg/filter/papers/umir.html

[5] Badrul M. Sarwar, "Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System", Proceedings of CSCW '98, Seattle, Washington, ACM, 1998

[6] 백혜정, 박영택, "웹 에이전트를 위한 사용자 관심도 학습", 정보과학회 가을 학술발표논문집, 1997

[7] 우선미, 유춘식, 김용성, 김순기, "사용자 프로파일과 그룹 프로파일을 이용한 문서 순위결정", 정보과학회 봄 학술발표논문집, 1999

[8] 이정수, 윤소정, 오경환, "행동 분석을 이용한 적응형 정보여과 에이전트", 한국 인지학회 논문지 제 9권 제 2호, 1999