사용자의 동적인 관심변화를 학습하는 개인화된 뉴스 에이전트

고경희[O] 오경환
서강대학교 컴퓨터학과
garahan@ailab.sogang.ac.kr, kwoh@ccs.sogang.ac.kr

Learning Dynamic Changes of User Interests in Personalized News Agent

Ko Kyoung-Hee[O] Oh Kyoung Whan
Department of Computer Science, Sogang University

요 약

정보여과 시스템은 사용자의 관심사를 정확하게 알아내야 하고(specialization), 시간에 따른 변화에 적응할 수 있어야 하며(adaptation), 사용자의 잠재적인 관심사를 발견하기 위해 새로운 도메인을 탐험할 수 있어야 한다(exploration). 본 논문에서는 온라인 뉴스 기사를 여과하여 사용자와 관련이 있는 뉴스 기사를 추천하는 뉴스 에이전트를 설계, 구현하고자 한다. Specialization, adaptation의 두 가지 요구사항을 충족시키기 위해 사용자의 관심사를 도메인별로 분리하고 각 도메인은 long-term과 short-term으로 나눈다. Exploration의 요구사항을 충족시키기 위해서는 카테고리 교차(crossover) 연산을 사용한다. 실험 결과, 사용자에 대한 사전 정보가 전혀 없는 상태에도 불구하고 빠른 적응능력을 보였다. long-term과 short-term의 분리는 사용자의 관심사에 급격한 변화가 일어난 후에도 시스템이 빠르게 적응할 수 있음을 보여주었다. 또한 카테고리 교차 연산을 통해 사용자의 새로운 관심사 탐험을 수행해 낼 수 있음을 보여주었다.

* 본 논문은 BK21 연구지원 사업에 의해 이루어졌음

1. 서론

현대의 컴퓨터 기술의 발달과 웹의 인기로 신문사와 잡지의 온라인 사이트, 유즈넷 뉴스그룹 등의 수많은 정보 소스로부터 시시각각 변화하는 뉴스 기사를 손쉽게 접할 수 있게 되었다. 그러나 너무 많은 정보속에서 정보의 취사선택이 문제가 되었고 사용자와 관련이 있는 아이템만을 선택하여 적절한 시기에 추천하는 정보여과 시스템의 중요성도 증가하게 되었다.

여과 시스템이 다양한 사용자 개개인에 맞는 서비스를 수행하기 위해서는 다음과 같은 세 가지 요구사항을 만족해야 한다[1]. 여과 시스템은 사용자의 특정한 관심사에 부합하는 서비스를 제공해야 한다(specialization). 여과 시스템은 장기간에 걸쳐 사용자와 상호작용한다. 사용자의 관심사가 변경될 경우 여과 시스템은 변화가 발생되었음을 알아야 하고 여기에 적응해 가야 한다(adaptation). 여과 시스템은 사용자의 잠재적인 관심사를 발견하기 위해 새로운 도메인을 탐험할 수 있어야 한다. 이러한 탐험능력은 실제로는 해당되나 현재로서는 알려지지 않은 사용자의 관심사를 찾을 수 있다. 또한 시스템의 적응 능력을 향상시킬 수 있다(exploration).

본 논문에서는 이 세 가지 요구사항을 만족시키는 뉴스 에이전트를 설계, 구현하고자 한다.

2. 연구배경

정보여과를 위해 사용자 프로파일을 학습하는 기존의 많은 연구들은 대부분 문서와 사용자 프로파일을 벡터공간 모델을 이용하여 모델링하였다[2].

이정수의 논문에서 사용자 프로파일은 선택된 키워드들을 요소로 하는 벡터이다[3]. 그는 뉴스 기사를 조회, 저장, 인쇄, 갈무리 하는 등의 사용자 행동에서 묵시적인 적합성 피드백을 받아서 사용자 프로파일을 학습하였다. 이를 이 후의 다른 연구와 구별하기 위해 1-Descriptor 형태의 사용자 프로파일이라고 하겠다. Webmate에서는 여러 도메인에 걸친 사용자의 관심사를 다중 벡터를 이용하여 모델링하였고, NewT는 프로파일을 에이전트의 population으로 모델링하였다[1]. 사용자의 피드백을 프로파일의 적합도를 증감시키는데 이용하였고 복제, 교차와 돌연변이의 유전자 알고리즘 연산자를 통해 관심사의 변화에 적응함과 동시에 사용자의 새로운 관심 영역을 탐험하였다.

상대적으로 꾸준한 사용자의 요구 형태도 시간이 흐름에 따라 천천히 변화하기도 하고, 때로는 급작스럽게 바뀔 수 있다. 그래서 사용자의 정보요구 형태를 short-term(단기)와 long-term(장기)으로 분리하여 모델링하는 연구가 시작되었다. NewsDude는 사용자의 short-term과 long-term 관심사를 각각 모델링하고 학습하는데 기계학습 방법을 이용하였다[4]. Alipse는 여러 도메인에 걸친 사용자의 관심사를 다중 벡터를 이용하여 모델링하였고 각 카테고리는 long-term, positive short-term, negative short-term의 세 가지 descriptor로 표현하였다[2]. 이를 3-Descriptor 형태의 사용자 프로파일이라고 하였다.

3. 사용자의 동적인 관심변화를 학습하는 개인화된 뉴스 에이전트

3.1 사용자 프로파일의 표현

본 논문에서는 사용자 프로파일을 n개의 사용자 관심사 도메인별로 분리한다. 각 카테고리는 long-term과 short-term 관심사의 LP(Long-term Profile)와 SP(Short-term Profile)로 나뉜다(그림 1). 이를 2-Descriptor 형태의 사용자 프로파일이라고 하겠다. 각 카테고리는 cValue라는 카테고리의 적합도를 나타내는

값을 가지고 있다. 이 cValue는 해당 카테고리가 받은 긍정적 피드백 값의 평균으로 구한다. LP와 SP는 피드백을 받은 문서로부터 학습된 키워드들로 이루어진다. 또한 LP와 SP는 각각 사용자의 관심도 가중치(interest weight) w_{lp}, w_{sp}를 가지고 있다. 이것은 관심도의 강도 레벨(interest intensity level)이다. LP의 경우 학습된 문서의 수를 dCount에 저장한다.

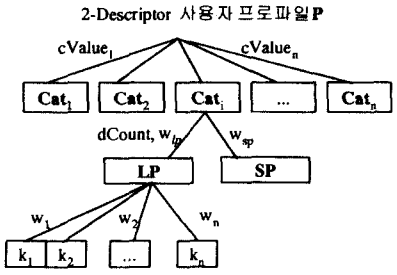

그림 1 사용자 프로파일의 2-Descriptor 표현

3.2 문서의 전처리와 키워드의 가중치 계산

먼저 프로토콜의 헤더와 HTML 태그를 제거한 다음 단어별로 분리, 추출한다. 530개의 리스트를 이용하여 불용어를 제거하고 마침표와 쉼표 등의 구두점과 숫자, 특수문자를 제거한다. 다음으로 Porter의 알고리즘을 이용하여 어간처리를 하였다. 이렇게 전처리 과정을 거친 키워드의 가중치는 TF-IDF 방법을 이용하여 계산하였다[3]. 계산된 가중치가 높은 순으로 100개의 키워드만을 사용했고 정규화 과정을 거쳐 단위벡터가 되도록 하였다.

3.3 문서의 순위부여 (Ranking)

먼저 문서 D가 어느 관심사 도메인에 속하는가, 즉 어느 카테고리에 속하는가를 결정해야 한다. 이를 위해서 각 카테고리의 SP와 LP와의 코사인 유사도를 계산하고, 이 중 큰 값이 이 카테고리 Cat_i와 문서 D와의 relevance 값 $Rel(Cat_i, D)$이 된다(식 1). 가장 큰 relevance 값을 가진 카테고리가 결정되면 사용자 프로파일 P와 문서 D와의 ranking score를 계산한다(식 2). Ranking score는 SP와 D의 코사인 유사도에 사용자의 관심도 가중치인 w_{sp}를 곱한 결과와, LP와 D의 코사인 유사도에 w_{lp}를 곱한 결과 중 큰 값으로 선택한다. 이 ranking score가 큰 값 순으로 문서를 정렬하여 사용자에게 보여준다.

$$Rel(Cat_i, D) = \max\{sim(SP_i, D), sim(LP_i, D)\}$$
$$\text{where } sim(P, D) = \cos\theta = \frac{P \cdot D}{|P| \times |D|} = P \cdot D \qquad \text{(식 1)}$$

$$Score(P, D) = \max\{w_{lp} \times sim(LP_c, D), w_{sp} \times sim(SP_c, D)\}$$
$$\text{where } c = \arg\max_i\{Rel(Cat_i, D)\} \qquad \text{(식 2)}$$

3.4 사용자 프로파일의 학습

여과 시스템이 추천한 문서에 대한 사용자의 피드백이 주어지면 사용자 프로파일의 학습이 이루어진다[3]. 사용자의 피드백은 7단계 (-1.0, -0.7, -0.3, 0, 0.3, 0.7, 1.0)로 나뉘며 사용자로부터 명시적으로 획득하였다.

3.4.1 기존 카테고리의 갱신

사용자의 피드백을 받은 문서 D가 속한 카테고리 Cat_i는 SP와 LP로 나뉘었기 때문에 학습방법도 각각 달리한다[2]. 먼저 SP의 경우, (식 3)에서처럼 가중치를 변경한다. 예를 들어 사용자의 피드백이 1.0인 경우, 이 문서는 현재 사용자의 관심사와 매우 잘 부합된다고 볼 수 있다. 따라서 Cat_i의 SP는 이 문서의 키워드들로 완전히 대체된다. 피드백이 0.0이면 피드백을 받은 문서는 프로파일에 아무런 영향을 주지 않는다. 피드백이 -1.0일 때는 SP는 -D로 완전히 대체된다. short-term 관심사의 경우, 즉각적으로 반응하려는 경향이 있고 그 결과 사용자 피드백이 즉시 반영되어야 하기 때문에 SP는 사용자 피드백을 학습율 α로 사용한 것이다.

$$Cat_i \leftarrow Cat_i \times (1 - |\alpha|) + \alpha \times D$$
$$\text{where } \alpha = \text{user's feedback for SP}, \alpha = \frac{1}{dCount + 1} + 0.05 \text{ for LP} \qquad \text{(식 3)}$$

LP의 경우도 마찬가지로 (식 3)을 이용한다. 단, long-term 관심사는 점진적으로 변화하기 때문에 프로파일의 갱신도 충분히 작아야 한다. 따라서 SP처럼 사용자의 피드백 값을 사용하는 것이 아니라 현재까지 피드백을 받은 문서의 수 dCount로 학습률 α를 구한다. 더 많은 피드백이 주어질수록 나중에 학습된 문서의 공헌도는 점점 작아진다. 따라서 앞서 학습된 사용자의 흥미가 SP와 같이 심한 변화를 겪지 않고 유지될 수 있다.

사용자 관심도 가중치의 학습은 (식 4)에서처럼 이루어진다. w_{sp}는 [-1,1] 사이의 값을 가진다. w_{lp}는 long-term 관심사의 점진적인 변화를 모델링할 수 있는 bipolar sigmoid 함수 $f(x) = 2/(1+\exp^{-x}) - 1$ 를 이용한다.

$$w_{sp} \leftarrow w_{sp} + (1 - |w_{sp}|) \times \alpha, w_{lp} \leftarrow f(f^{-1}(w_{lp}) \pm \alpha)$$
$$\text{where } f(x) = \frac{2}{1 + \exp^{-x}} - 1 \qquad \text{(식 4)}$$

3.4.2 새로운 카테고리의 생성

SP와 LP의 키워드는 피드백을 받은 문서의 키워드들로 초기화된다. w_{sp}는 사용자의 피드백으로, w_{lp}는 $f(\alpha)$로 초기화된다.

3.5 카테고리 교차를 통한 새로운 관심사의 탐험

일정 기간이 지나면 기존의 카테고리를 이용하여 새로운 카테고리를 생성한다. 이 때 어느 카테고리를 선택할지의 기준은 cValue 값이다. cValue의 값이 가장 높은 두 개의 카테고리가 선택되고 SP, LP 각각 교차를 수행한다. 두 카테고리내 키워드를 가중치 순으로 정렬하였을 때 1/3과 2/3 되는 지점을 기준으로 뒤바꾼다. 새로운 카테고리의 관심도 가중치는 두 개의 기존 카테고리가 가지고 있는 관심도 가중치에 각각의 cValue 값을 곱한 평균값으로 구하였다.

4. 실험 및 분석

실험은 유즈넷 뉴스 그룹 중 clari.web.* 뉴스 기사를 대상으로 하였다. clari.web.* 뉴스 기사는 8개 카테고리,

190여개의 세부카테고리로 분류되며 매일 평균 4000여개의 기사가 posting된다.

4.1 실험 1

실험 1은 1-Descriptor와 2-Descriptor와의 비교실험이다. 이 비교실험을 통해서 본 논문에서 구현한 2-Descriptor의 타당성을 검증하고자 한다.

실험 1은 2050개의 뉴스 기사를 사용한 모의실험이다. 매 cycle에서 100개의 문서를 임의로 선택한 후, 실험에서 가정한 타겟 프로파일 P_t와 시스템이 학습해 나가야 할 시스템 프로파일 P_s를 기준으로 각각 ranking score를 계산하여 순위를 정한다. 순위가 정해진 문서의 집합을 각각 D_{target}, D_{system}이라고 하자. 실험의 평가기준은 정확도로서 D_{target}과 D_{system}의 상위 n개의 문서 중 얼마나 일치하는지를 백분율로 표시한 것이다. 본 실험에서는 상위 10개의 문서를 비교하였다. 20번째 cycle에서 사용자의 관심사가 급격하게 변하는 것을 모의실험하기 위해서 P_t를 새로운 내용으로 교체하는데 이를 타켓 프로파일의 inversion이라고 한다.

그림 2 비교실험 결과

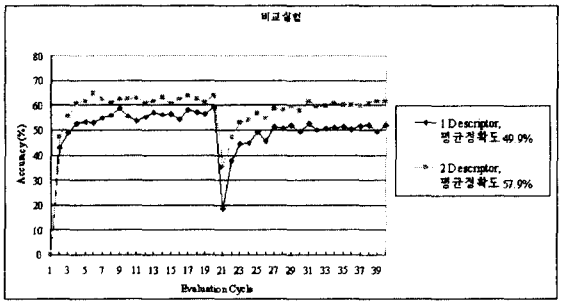

그림 2는 비교실험의 결과이다. 1번째 cycle에서는 시스템 프로파일 P_s가 빈 파일이기 때문에 정확도는 당연히 0이 된다. 처음 100개 문서의 피드백으로 1-Descriptor와 2-Descriptor 프로파일 모두, 2번째 cycle에서 정확도가 급격하게 향상되었다. 전반적으로 2-Descriptor의 정확도가 높게 나타나는데 이것은 사용자의 관심사 도메인별로 프로파일을 구성하기 때문이다. 타겟 프로파일 inversion이 일어난 21번째 cycle에서 정확도가 급격하게 떨어졌으나 22번째 cycle에서 다시 정확도가 향상되었다. 전체적으로 2-Descriptor의 정확도가 높았을 뿐만 아니라 타겟 프로파일 inversion 이후에는 2-Descriptor와 1-Descriptor의 정확도 차이가 심화되었다. 이는 2-Descriptor의 경우 long-term 관심사를 LP에서 유지하기 때문이다.

4.2 실험 2

실험 2는 카테고리의 교차로 생성된 새로운 카테고리가 시스템의 새로운 관심사 탐험 능력에 어느 정도 영향을 미치는가에 대한 관찰이다. 이를 serendipity(횡재) 실험이라고 할 수 있다. 실험 2는 3명의 실제 사용자가 2주간 수행했다.

전체 4000 여개의 문서 중 평균 ranking score 이상의 뉴스 기사만을 선택하여 사용자에게 보여주었다. 이중 평균 30% 정도의 문서가 카테고리의 교차 결과로 새로이 사용자에게 추천된 것이다. 사용자들에게 이런 기사 중 기대하지 않은 카테고리로부터 선택되었는데, 읽어보니 뜻밖에 흥미로운 기사였다고 판단되면 선택하게 하여 따로 저장하였다. 그 결과 하루 평균 5개씩의 기사가 serendipity 능력을 보인다고 사용자들로부터 피드백을 받을 수 있었다. 그림 3의 초기 프로파일 칼럼은 실험 시작 후 첫날 생성된 사용자들의 프로파일중 일부이다. 중간 및 최종 프로파일은 실험 과정 중과 실험이 종료된 후 최종적인 프로파일을 정리한 것 중 일부이다. 이들 중에는 자연스럽게 변화된 관심사도 있지만, 카테고리의 교차로 생긴 것도 있다. 💡 아이콘이 있는 것은 사용자들이 serendipity 능력이라고 판단한 기사들의 카테고리들이다. 주로 평소에 관심을 갖지 않았던 도메인으로부터 선택된 기사들이었고, 결과적으로 카테고리의 교차로 새로운 관심사를 탐험하려는 본 논문의 취지가 잘 드러났다고 볼 수 있다.

그림 3 카테고리 교차 실험 결과

초기 프로파일	중간 및 최종 프로파일
	biz.earnings.releases
	biz.industry.agriculture
	biz.industry.automotive
💡	biz.industry.banking.releases
	biz.industry.health.care
	biz.industry.media.entertainment.releases
	biz.stocks.report.asia
💡	biz.stocks.report.usa.misc
	biz.top
	biz.world_trade
living.entertainment.misc	living.entertainment.misc
living.movies	living.movies
	living.consumer
	living.history
	living.music
	living.royalty
	living.tv
	local.california.northern
💡	local.california.sfbay.trouble
💡	local.louisiana
	local.ohio
💡	local.washington

5. 결론

본 논문에서는 온라인 뉴스 기사를 여과하여 사용자와 관련이 있는 뉴스 기사를 추천하는 뉴스 에이전트를 설계, 구현하였다. 향후 문서 추천시 협동적 여과를 이용하면 serendipity 효과를 더 높일 수 있을 것이라고 본다.

참고문헌

[1] Beerud Dilip Sheth, "A Learning Approach to Personalized Information Filtering", MIT 전자공학과 & 컴퓨터학과 대학원 석사논문, 8-19쪽, 1994

[2] Dwi H. Widyantoro, "Dynamic Modeling and Learning User Profile in Personalized News Agent", Texas A&M 컴퓨터학과 대학원 석사논문, 10-57쪽, 1999

[3] 이정수, "적합성 피드백을 이용한 정보여과 에이전트", 서강대학교 대학원 석사논문, 5-30쪽, 1997

[4] Billsus D., and Pazzani M. J. , "A Hybrid User Model for News Classification", Proceedings of the 7th International Conference on User Modeling, pp. 103-105, 1999