적합성 피드백에 의한 웹 에이전트의 프로파일 학습

한정기, 김준태
동국대학교 컴퓨터공학과

Profile Learning of Web Agent by Relevance Feedback

Junggee Han, Juntae Kim
Department of Computer Engineering, Dongguk University

요 약

웹 에이전트는 사용자가 좀 더 손쉽게 웹 상의 정보를 얻을 수 있게 하는 것을 목표로 하는 인터넷 정보 검색 도구이다. 본 논문에서는 개인용 웹 에이전트 시스템에서 적합성 피드백(Relevance Feedback)에 의해 사용자의 취향을 학습하는 방법을 제시하고, 실험을 통하여 제시된 적합성 피드백에 의한 학습 방법이 사용자의 취향을 성공적으로 학습함을 보였다. 적합성 피드백의 두 가지 형태인 양성 피드백과 음성 피드백이 학습에 미치는 영향에 대하여 양쪽 피드백을 혼용할 때와 각각 한가지만 사용할 때로 나누어 실험하였으며, 피드백이 진행되면서 검색 결과의 정확도가 변화하는 정도를 관찰하였다.

1. 서론

웹 검색 엔진과 같은 성공적인 인터넷 정보 검색 도구에 힘입어 오늘날 인터넷이 가지고 있는 정보의 양과 사용자의 수는 빠르게 증가하고 있다 웹 에이전트란 사용자가 좀 더 손쉽게 웹 상의 정보를 얻을 수 있게 하는 것을 목표로 한 인터넷 정보 검색 도구이다[6] 이러한 웹 에이전트는 사용자를 대신하여 사용자의 취향에 맞는 웹 페이지들을 찾아주며 사용자의 취향을 점진적으로 학습함으로써 보다 효율적으로 개인이 원하는 정보를 얻을 수 있도록 하여 준다

본 논문에서는 적합성 피드백에 의한 학습 방법으로 개인용 웹 에이전트를 구현하고, 실험을 통해 학습 효과를 검증한다 또한 실험 시 양성(positive) 피드백만 사용한 경우, 음성(negative) 피드백만 사용한 경우, 두 피드백을 혼용하여 사용할 경우의 3가지 경우로 나누어 실험하여 양성 및 음성 피드백이 학습에 미치는 영향에 대하여 알아본다 마지막으로 피드백 회수, 즉 학습 회수에 따른 학습 결과도 살펴보기로 한다. 2장에서는 일반적인 웹 에이전트와 본 논문에서 사용한 에이전트 시스템에 대해 설명하였고, 3장에 학습 방법을, 4장에 실험 결과를 보였으며, 결론 및 향후 과제를 5장에 기술하였다.

2. 개인용 웹 정보 수집 에이전트

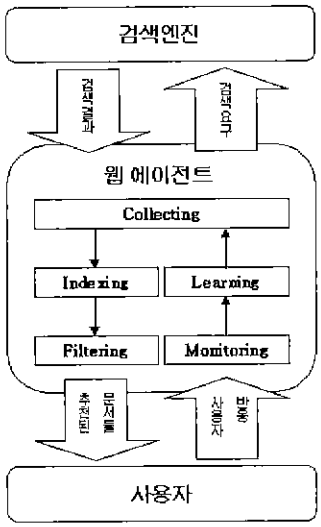
개인용 웹 에이전트는 웹과 개인 사용자 중간에 위치하면서 사용자의 명시적 혹은 묵시적 반응으로부터 사용자의 취향을 학습하여 점진적으로 좀더 사용자의 취향에 근접한 정보를 제공해 주는 것을 그 목표로 한다. 웹 에이전트 시스템에서의 학습 예로는 베이지안 확률(Bayesian Probability)을 이용한 학습을 하는 Personal Webwatcher[1]와 Syskill&Webert[2], 그리고 결정트리(Decision Tree)를 이용한 학습을 하는 Infofinder[3] 등이 있다.

이러한 에이전트들은 사용자의 취향을 프로파일로 표현하고 사용자로부터의 피드백에 따라 프로파일을 수정함으로써 학습을 진행한다 프로파일의 표현 형식은 사용되는 응용분야나 프로파일이 나타내야 하는 정보에 따라 여러 형태가 가능하겠지만 비교적 많이 사용하는 대표적인 형태의 프로파일은 단어와 그 단어에 대한 가중치의 쌍들로 구성된 프로파일이다 높은 가중치를 가진 단어일수록 사용자의 취향에 관련된 문서를 구별하는데 더 중요하게 사용된다.

웹 에이전트의 프로파일은 위와 같은 단어와 단어 가중치 쌍들 이외에 사용자 자신에 대한 정보를 추가적으로 프로파일의 일부로 나타낼 수도 있다 예를 들어 개인의 교육 수준, 성별, 나이, 사는 지역, 관련 분야에 대한 친밀도, 외국어 실력, 구독하는 잡지, 독서 습관 등과 같은 수많은 개인정보가 프로파일에 포함된다면 좀 더 개인에게 적합한 문서를 추천해 줄 수 있을 것이다.[4]

본 논문에서 사용된 개인용 웹 에이전트는 [그림1] 과 같이 다섯 단계로 구성되어 있다. 사용자 관심사를 나타내는 프로파일을 기초로 하여 수집기가 웹 문서를 모아 오고(Collecting), 수집된 문서는 색인기에 의해 분석된다(Indexing). 분석된 문서 정보와 프로파일을 기반으로 하여 여과기는 중요 문서를 걸러내서 사용자에게 추천하고(Filtering), 추천된 문서에 대한 사용자 반응을 감시기가 감시한다(Monitoring). 감시된 사용자 반응 정보를 기초로 하여 학습기가 프로파일을 갱

신함으로써 학습이 진행된다(Learning). 본 논문에서는 일반적으로 많이 사용하는 단어와 단어 가중치의 쌍들로 구성된 프로파일을 사용하였다.



[그림 1] 본 논문에 사용된 개인용 웹 에이전트 시스템 구조도

3. 적합성 피드백에 의한 프로파일 학습

웹 에이전트에서의 학습이란 명시적, 혹은 묵시적으로 획득된 사용자의 피드백(User Feedback)으로부터 사용자의 관심사를 나타내는 프로파일을 갱신하는 것을 의미하고, 사용자의 취향에 좀더 근접한 프로파일을 작성하는 것을 그 목적으로 한다.

적합성 피드백(Relevance Feedback)에 의한 프로파일 학습 방법은 적합성 피드백에 의한 질문 수정 방식을 프로파일 학습에 응용한 것이다. 대화식 온라인 검색시스템의 사용자는 보다 낳은 검색 결과를 얻기 위해 탐색과정을 반복할 수 있다. 사용자가 탐색된 결과를 보고 질문을 수정하는 것을 사용자 피드백(User Feedback)에 의한 질문 수정이라고 하며, 특히 검색문헌의 내용을 나타내는 정보와 검색문헌의 적합성을 판단한 정보에 의해 질문을 수정하는 것을 적합성 피드백(Relevance Feedback)에 의한 질문 수정이라고 한다 이때 사용자가 해당 문헌의 적합성을 긍정적으로 판단한 문헌을 적합문헌, 그 피드백을 양성 피드백(Positive Feedback)이라고 하며, 사용자가 해당문헌의 적합성을 부정적으로 판단한 문헌을 부적합문헌, 그 피드백을 음성 피드백(Negative Feedback)이라고 한다.[5]

본 논문에서는 적합성 피드백으로 질문을 수정하는 대신 프로파일을 수정함으로써 프로파일에 대한 학습을 진행시킨다. 프로파일 수정은 사용자의 새로운 적합성 피드백이 발생하면 지금까지 발생한 모든 피드백들과 새 피드백을 합하여 평균을 내는 형식으로 진행하게 하였다 프로파일에 갱신에 적용시킨 적합성 피드백을 이용한 프로파일 수정 방법은 다음과 같은 식으로 표현된다

$$P' = \frac{1}{N+1} * \{(\alpha * N * P) + (\beta * A) - (\gamma * B)\}$$

α, β, γ : 상수, 각 항의 비중을 나타냄
P' : 갱신된 프로파일의 벡터 표현
P : 기존의 프로파일의 벡터 표현
A : 적합한 웹 문서의 벡터 표현
B : 부적합 웹 문서의 벡터 표현
N : 프로파일 갱신 횟수

α, β, γ는 기존 프로파일, 적합한 웹 문서, 부적합한 웹 문서에 대한 각각의 가중치로 이 값들의 조정함으로써 새로운 피드백에 대하여 학습하는 성향을 조정 할 수 있다. N은 지금까지의 프로파일 갱신 횟수, 즉 지금까지의 적합성 피드백의 개수를 의미한다

프로파일과 웹 문서는 단어와 단어 가중치 쌍들의 벡터로 표현하였다 웹 문서의 벡터표현을 위한 단어가중치 방법은 웹 문서 내 단어빈도(Term Frequency, TF)를 사용하였다. 일반적으로 문서 표현 시 단어 가중치 방법으로 많이 사용되는 역문헌빈도를 고려한 단어빈도(Term Frequency Inverted Document Frequency, TFIDF)를 웹 에이전트에 적용하기에는 전체 문서집단을 규정하기 어렵다는 문제점을 안고 있다 만일 웹 에이전트가 모아온 문서들을 전체 문서집단으로 하여 TFIDF를 적용한다면 중요한 단어의 가중치를 떨어뜨리는 역효과를 가져올 수 있다. 왜냐하면 웹 에이전트는 프로파일을 참고로 하여 문서들을 모아오기 때문이다. 즉, 프로파일에서 높은 가중치를 갖는 단어일수록 수집해 온 문서들 사이에 빈번히 나타날 가능성이 높아서 웹 문서 표현 시 TFIDF를 적용하면 프로파일 내에서 가중치가 높은 단어가 문서 표현에서 상대적으로 낮은 가중치를 부여받게 되는 것이다.

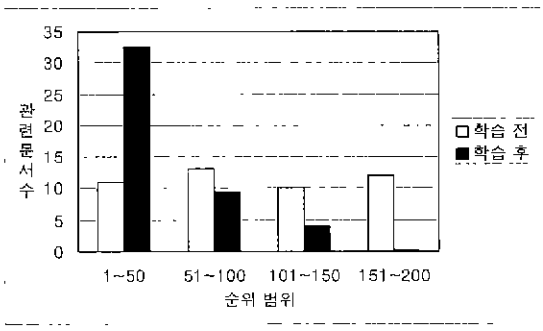
4. 실험 및 결과

본 논문에서는 기존의 검색엔진이 두 가지 이상의 뜻을 가진 단어에 관한 문서 중 한가지 뜻에 대한 문서만을 선별하지 못하는데 착안하여 다음과 같은 실험 방법을 구상하여 보았다. 먼저 '한글'이라는 키워드로 AltaVista Korea에 질의를 던져 상위 200개의 문서를 모았다. 모아온 문서를 워드프로세서 소프트웨어들 중 하나인 '아래아 한글'과 관련된 문서와 관련되지 않은 문서 두 종류로 구분하였다. 사용자가 원하는 정보는 '아래아 한글'에 관한 문서로 가정한다. 총 문서 200개 중 '아래아 한글'과 관련된 문서는 모두 46개 문서였다.

사용자가 '아래아 한글'에 관련된 문서나 관련되지 않은 문서를 피드백으로 주면 이를 바탕으로 주어진 학습 방법을 이용하여 프로파일을 갱신하고, 갱신된 프로파일을 기준으로 총 문서에 대한 순위를 재조정함으로써 사용자가 관심 있어 하는 '아래아 한글'에 관한 문서가 상위의 순위에 오는 정도로 학습의 진행여부 및 학습 효과를 평가할 수 있다.

[그림 2]는 검색엔진에서의 관련 문서 분포(학습 전)와 20개 문서로 사용자 피드백이 있은 후 관련문서 분포(학습 후)를 보여 준다. 20개 문서의 사용자 피드백 중 양성 피드백과 음성 피드백은 각각 10개씩으로 하였으며 각 피드백으로 선정되는 문서는 랜덤하게 선택하였다. [그림 2]는 10회 반복한 실

험 결과의 평균치이다. 실험결과를 보면 학습이 진행된 후 관련 문서가 보다 많이 상위로 올라옴으로써 적합성 피드백을 이용한 학습이 효과가 있음을 알 수 있다.



[그림 2] 학습 전후의 관련문서 분포 비교

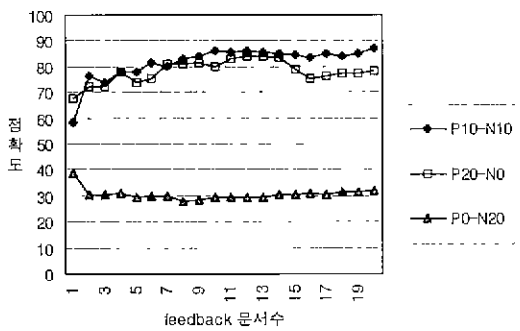
[표 1]은 양성 피드백과 음성 피드백의 유무 및 혼용이 학습 결과에 미치는 영향에 대하여 비교한 것을 보여준다. P10-N10은 20개의 피드백 중 양성 피드백(Positive feedback)과 음성 피드백(Negative feedback)이 각각 10개씩 있었음을 의미한다. 각 순위 범위 아래의 숫자는 순위 범위 안에 포함된 관련문서 수를 나타낸다. 위에서와 같이 이 실험도 10회 반복한 실험 결과의 평균치이다. 관련 문서가 어느 정도 분류되는지를 나타내기 위하여 총 관련문서(46개의 문서) 중 몇 개의 문서가 50위권 순위 안에 포함되는지를 '상위50내 진입율'로 나타내었다. 실험 결과 양성 피드백과 음성 피드백을 혼용한 경우가 가장 우수한 학습 효과를 보였다.

[표 1] 양성피드백과 음성피드백의 유무 및 혼용이 학습결과에 미치는 영향 비교

	1~10 순위	11~20 순위	21~30 순위	31~40 순위	41~50 순위	상위50내 진입율
검색엔진	2	3	1	3	2	23.9%
P10-N10	9.5	7.9	5.9	5.5	3.7	70.7%
P20-N0	9.6	6.1	4.8	3.3	2.9	58.0%
P0-N20	4.1	2.3	1.8	2.9	3.9	32.6%

[그림 3]은 피드백이 1개부터 20개까지 주어질 때마다 상위 20위권 내에 몇 퍼센트의 문서가 관련 문서인가를 나타낸다. 보통 상용되는 검색 엔진에서 검색결과로 한번에 보여 줄 수 있는 문서의 수가 10개 정도인 것을 감안하여 사용자의 입장에서 웹 에이전트를 사용할 때 몇 번의 피드백으로 한 두 화면 내에 원하는 정보를 얻을 수 있는 가를 나타내는 것이 이 실험의 목적이다. 웹 에이전트가 사용자의 편의를 도모하는 소프트웨어임을 고려하여 볼 때 학습성과에 대한 이와 같은 평가 방법은 타당하다고 하겠다.

검색엔진은 상위 20개의 문서 중 5개의 관련 문서, 즉 25%의 정확도를 나타내었다. 실험 결과, 적합성 피드백에 의한 학습 방법 중 양성 피드백과 음성 피드백을 혼용한 경우와 양성 피드백만을 사용한 경우에는 약 2~3개의 피드백 후에는 70% 정도 이상의 정확도로 관련 문서를 추천하는 것으로 나타났다 하지만 음성 피드백만 사용한 경우에는 정확도의 향상이 거의 이루어지지 않았다



[그림 3] 피드백 회수에 따른 상위 20위권 내 관련 문서 비율

5. 결론 및 향후과제

본 논문에서는 개인용 웹 에이전트의 프로파일 학습에 적합성 피드백(Relevance Feedback)에 의한 방법을 적용해 보고 실험을 통해 적합성 피드백에 의한 학습 방법이 사용자의 취향을 올바로 학습하는지 알아보았다 실험 결과 적합성 피드백에 의한 프로파일 수정이 사용자의 취향을 학습하는데 효과가 있음을 보였고, 양성 피드백과 음성 피드백을 혼용한 경우에 가장 학습 효과가 높음을 보였다. 본 실험의 경우 양성 피드백과 음성 피드백을 혼용했을 때에는 2개 정도의 문서를 피드백하여 단순한 검색 엔진의 결과보다 매우 높은 70% 이상의 정확도로 웹 페이지를 추천할 수 있었다.

문서를 표현할 때 TF만을 사용하면 "mail"과 같은 많이 사용되나 중요하지 않은 단어가 높은 가중치를 나타내는 문제 및 문서 길이를 고려해야 하는 문제가 발생할 수 있다 웹 에이전트의 특성에 맞는 불용어 사전을 작성하고 예외적인 TF 값에 영향을 받지 않도록 TF의 범위를 제한한다면 이러한 문제를 해결할 수 있을 것이다.

참고문헌

[1] Dunja Mladenic, "Personal WebWatcher: design and implementation" , Carnegie Mellon University, 1996.

[2] Michael Pazzani, Jack Muramatsu & Daniel Billsus. "Syskill & Webert: Identifying interesting web sites" 1996 AAAI Proceedings Volume One page 54, August

[3] Bruce Krulwich, Chad Burkey. "The InfoFinder Agent: Learning User Interests through Heuris" IEEE Expert/Intelligent Systems & Their Applications, Vol. 12, No 5, September/October 1997

[4] Rober R. Korfhage. "Information Storage and Retrieval : p145 Chapter 6. User Profiles and Their Use"

[5] 정영미, "정보검색론" 구미무역 출판부, 1993.

[6] 최중민, "에이전트의 개요와 연구방향" 한국정보과학회 정보과학학회지, "특집 에이전트 시스템" 제15권 제3호 3월 1997년