효과적인 웹 문서 추천을 위한 동적 사용자 프로파일 생성 기법

윤윤경, 서정연
서강대학교 컴퓨터학과
{runrun@nlprep.sogang.ac.kr, seojy@ccs.sogang.ac.kr}

Dynamic User Profile Creation Method for Effective Recommendation for Documents on the Web

Yoon-Kyung Yoon, Jung-Yun Seo
Dept. of Computer Science, Sogang Univ.

요 약

기하급수적으로 증가하는 인터넷의 정보량에서 최적의 정보를 찾고자 하는 사용자의 요구가 증가함에 따라 개별적 사용자에게 필요한 정보만을 제공하는 것이 필요하다. 이러한 사용자의 요구를 충족시키기 위해 사용자의 행동을 관찰하고 학습하여 사용자 대신 문서를 수집하는 웹 문서 추천 에이전트의 필요성이 대두되었다. 본 논문에서는 웹 문서 추천에 이전트에서 사용되는 프로파일을 효과적으로 생성하고 학습하기 위한 문서 표현 방법, 특징 선택법을 제안한다.

제안된 문서 표현 방법은 슬라이딩 윈도우 방법을 통해 인접한 단어쌍의 문맥 정보를 이용하고, 의존 구조를 이용하며 사용자의 관심 변화에 빨리 적응 할 수 있도록 시간에 대한 가중치를 반영한다.

제안된 방법으로 프로파일을 구성한 웹 문서 추천 에이전트는 사용자의 관심 분야를 효과적으로 반영하고 관심 변화에 빨리 적응하여 사용자에게 알맞은 문서를 추천한다.

1. 서론

최근 수년간 인터넷의 정보량은 기하급수적으로 증가하고 있다. 정보량이 증가하고 정보의 종류가 증가할수록 사용자는 필요한 정보를 찾기 위해 더 많은 시간과 노력을 투자 해야 한다. 이에 따라 방대한 인터넷의 정보 중에서 최적의 정보를 찾고자 하는 사용자의 요구가 증대되고 있으며 이러한 사용자의 요구를 충족시키기 위해서는 개별적 사용자에게 필요한 정보만을 수집하여 제공하는 것이 필요하다. 사용자는 웹 문서(web document)의 정보에 접근하기 위하여 주로 다음 세 가지 방법을 이용한다. 첫 번째 방법은 야후(yahoo)[1]와 같은 검색기를 이용하는 것이다. 검색기를 이용하는 것은 사용자에게 웹 검색을 위해 도움이 되지만 질의 생성의 부담을 주고 사용자 개개인의 특성을 무시하여 특정 사용자에게 알맞은 정보를 제공할 수 없다는 단점이 있다. 두 번째 방법은 자신이 알고 있는 URL 의 하이퍼링크를 따라가면서 찾아가는 것이다. 이 방법은 알고 있는 URL 과 연결되지 않은 웹 문서는 접근할 수 없고 필요한 정보를 찾는데 많은 시간이 필요하다. 세 번째 방법은 메일 푸시 서비스(mail push service)를 통해 접근하는 것이다. 이 방법은 사용자가 푸시받은 정보에 대한 의사를 표현할 수 없기 때문에 사용자는 불필요한 정보를 서비스 받을 가능성이 높다. 그러므로 사용자의 행동을 관찰하고 학습하여 사용자에게 알맞은 웹 페이지를 추천하는 웹 문서 추천 에이전트(agent)가 필요하다. 웹 문서 추천 에이전트는 개인의 프로파일(profile)을 구성하고 이것을 이용하여 사용자의 선호도를 표현한다.

본 논문에서는 웹 문서 추천 에이전트를 구현하기 위해 효과적으로 프로파일을 생성하고 학습하는 방법을 제안한다. 프로파일 생성을 위한 문서의 표현 방법으로는 n-gram 과 함께 의존구조를 이용하며 사용자의 잠재적 관심 변화를 반영하기 위해 시간에 대한 가중치를 부여한다.

[1] http://www.yahoo.co.kr

2. 관련연구

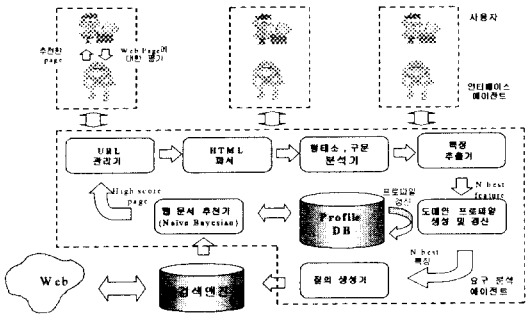
인터넷에 존재하는 방대한 정보에서 사용자의 선호도가 높은 것만을 선택하기 위해서는 여과기 역할을 하는 프로파일이 필요하다. 그러므로 어떤 특징(feature)을 사용하여 어떻게 프로파일을 생성하는가 하는 것이 무엇보다 중요하다. 기존의 시스템들은 프로파일이나 문서 표현을 위해서 uni-gram 만을 사용하거나[1], 인접한 단어를 단순 조합한 n-gram 을 사용한다[2]. 본 논문에서 사용하는 n-gram 의 의미는 특징 추출을 위해 uni-gram 부터 n-gram 까지 모두 사용하는 경우를 뜻한다.

uni-gram 만을 사용한 프로파일 시스템으로는 LIRA[3]등이 있다. LIRA 에서는 문서에 나타난 빈도수가 높은 단어에 대한 가중치를 부여하기 위해서 단어 빈도수(word frequency)에 대한 벡터를 이용하였다. 그러나 한 단어만을 특징으로 사용하였을 경우 인접한 단어 사이의 공기 정보를 사용할 수 없다는 한계가 있다.

n-gram 을 사용할 경우에는 문맥 정보(context information)를 가지고 있기 때문에, 한 단어만을 특징으로 고려하는 uni-gram 보다 효과적이다. 그러나 n-gram 을 이용한 시스템의 경우 인접한 단어쌍에 대한 공기 정보만 사용하기 때문에 문장 내에서 떨어져 있지만 연관도가 있는 의존소와 지배소 사이의 의존 관계에 대한 정보를 반영할 수 없다.

3. 웹 문서 추천 에이전트

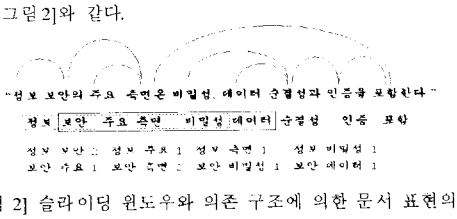
본 논문에서 제안하는 웹 문서 추천 에이전트는 사용자의 행동을 관찰하고 사용자 피드백을 받는 인터페이스 에이전트와 사용자의 피드백과 브라우징(browsing) 결과를 이용하여 사용자의 프로파일을 생성, 이용하여 사용자가 선호할 만한 문서들을 추천하는 요구 분석 에이전트로 구성된다. 제안하는 웹 문서 추천 에이전트는 [그림 1]과 같은 구조를 갖는다.



[그림 1] 웹 문서 추천 에이전트의 구조

기존의 시스템들은 프로파일 구성을 위해서 uni-gram 만 사용하거나, n-gram 을 사용한다. 그러나 한 단어만 사용하였을 경우 단어 사이의 문맥 정보를 반영할 수 없고, 단순한 n-gram 만을 이용하면 떨어져 있지만 연관도가 있는 의존소와 지배소 사이의 문맥 정보를 이용할 수 없다. 이러한 단점을 보안하기 위해서 제안된 프로파일 시스템은 특징으로써 슬라이딩 윈도우(sliding window)와 의존 구조(dependency structure)에 의한 단어쌍을 이용한다[4][5]. 프로파일에 문맥 정보를 반영하기 위해 먼저 슬라이딩 윈도우를 사용하여 인접한 단어의 공기 정보를 추출하고, 의존 구조 분석기를 통해 윈도우의 범위를 벗어나지만 서로 연관이 있는 단어쌍에 대한 공기 정보를 추출한다.

예문 "정보 보안의 중요성은 비밀성, 데이터의 순결성과 인증을 포함한다."에 대해서 슬라이딩 윈도우와 의존 구조를 사용하여 단어쌍을 추출하면 [그림2]와 같다.



[그림 2] 슬라이딩 윈도우와 의존 구조에 의한 문서 표현의 예

[그림2]의 예문에서 '보안 주요' 라는 단어쌍보다 '정보 보안'이라는 단어쌍이 더 많은 문맥 정보를 반영하므로 더 높은 가중치가 부여 되어야 한다. 그러나 기존의 n-gram 만을 사용하였을 경우 두 단어쌍의 빈도수가 1 로써 같은 가중치를 갖게 된다. 슬라이딩 윈도우와 의존구조를 이용할 경우 '정보 보안'이라는 단어쌍이 '보안 주요'라는 단어쌍보다 높은 가중치를 갖게 되므로 사용자가 선호한 문서로부터 올바른 특징을 추출할 수 있다.

제안된 시스템에서는 특징 추출을 위해 단어의 가중치를 구할 때 TF-IDF(Term Frequency, Inverse Document Frequency)[6]방법을 이용한다.

$$word\_weight = (TF(w) + score(D)) \log \frac{P(w|C)}{P(w|\bar{C})} \quad [1]$$

$$P(w|C) = \frac{(DF(w,C)+T)/(\sum_{C'} C')}{(C')/(\sum_{C'} C')} \quad [2]$$

score(D) : D에 대한 사용자 평가점수
$DF(w,C)$: C에서 w가 나타난 문서의 개수
T : $time\_weight$ w : 단어 혹은 단어쌍
C : 사용자가 선호한문서들의 집합
$\bar{C}$: 사용자가 선호하지않은문서들의 집합
$TF(w)$: w의 빈도수

TF-IDF 에서 문서에 대한 사용자 평가를 반영하기 위해 TF-IDF 방법을 식[1]과 같이 수정한다.

사용자의 관심 분야는 시간에 따라 변화하기 때문에 최근 문서는 오래된 문서보다 사용자의 관심을 더 잘 반영할 수 있다. 그러나 기존의 시스템들은 시간에 대한 고려를 하지 않기 때문에 사용자의 관심 분야가 바뀌었을 경우 프로파일 갱신을 할 때 처음보다 더 많은 사용자의 피드백이 필요하다. 그러므로 본 논문에서는 시간에 대한 가중치를 이용하여 불필요한 피드백을 줄이는 특징 선택법을 제안한다. 제안된 시스템은 사용자의 긍정 피드백과 부정 피드백을 반영하기 위해서 사용자가 선호하는 단어 벡터로 구성된 긍정 프로파일과 사용자들이 비선호하는 단어 벡터로 구성된 부정 프로파일을 갖는다. 식[2]에서 시간에 대한 가중치 T는 긍정 프로파일 내의 특징에 대한 가중치를 갱신할 때만 사용된다. 시간에 대한 가중치 T는 [그림3]와 같이 구한다.

```
if (일정 시간동안 브라우징한 문서들 중에서 한번이라도 발생한 특징)
        history = 0;
else    history ++;
if (IsPositiveProfile ){    time_weight = k × history^2  }
else                    {    time_weight = 0;             }
```

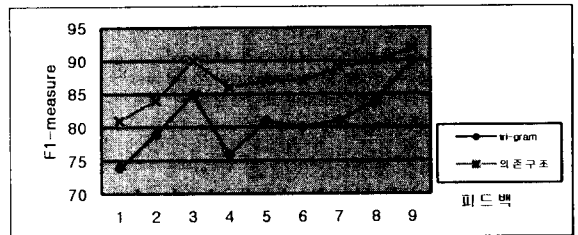
[그림 3] 시간에 대한 가중치 알고리즘

[그림3]에서 history 는 사용자가 접속한 동안 브라우징한 문서들 중에서 긍정 프로파일에 있는 단어쌍이 나타나지 않은 횟수이다. 브라우징한 문서들 중 어느 한 문서에라도 포함 되어 있으면 history 는 0 이 된다. history 가 커지면 커질수록 사용자는 잠재적으로 그 단어쌍에 대하여 점점 관심이 없어진다는 뜻이 되므로 그 단어에 대한 가중치를 줄인다. history 가 큰 단어에 대해서 가중치를 줄이기 위해서 history 가 0 보다 큰 단어쌍을 잠재적 부정 피드백으로 이용한다. 잠재적 부정 피드백은 사용자가 명시적으로 부여한 부정 피드백보다 적은 영향을 미쳐야 하기 때문에 시간에 대한 가중치는 사용자가 브라우징한 문서의 개수와 history 를 어떠한 비율로 반영하는가를 결정하는 k 에 비례한다. k 값이 증가할수록 단어쌍에 대한 가중치는 시간에 대하여 많은 영향을 받고 k 값이 감소할수록 적은 영향을 받기 때문에 적당한 k 값을 정하는 것이 무엇보다 중요하다. 본 논문에서는 실험을 통하여 k 를 0.1 로 설정하였다.

웹 문서 추천 에이전트는 문서 여과 방법으로 내용 기반 여과 방법을 이용하고 검색엔진의 질의 결과에 해당된 문서와 프로파일의 유사도를 구하기 위해 베이지안 모델(Naïve Bayesian classifier)[7]을 사용한다.

4. 실험 및 결과

먼저 본 논문에서 제안한 문서 표현 방법을 이용하는 것이 사용자의 관심문서를 더 효과적으로 기술할 수 있는가를 평가하기 위하여 제안한 문서 표현 방법과 기존의 n-gram 을 특징으로 이용하였을 때의 결과를 비교하였다. 실험에 사용된 데이타는 11 가지의 분야로 나누어진 550 개의 HTML 문서 데이터이다. 11 가지는 모두 취미에 관련된 HTML 문서로 관련 분야는 '바둑', '등산', '낚시', '여행, 관광', '자동차, 자전거, 오토바이', '만화, 애니메이션', '항공', '비행', '게임, 도박', '동물, 애완동물', '곤충' 이다. 실험을 위해 11 개의 분야 중 한 분야는 관심 분야로, 나머지 분야는 비관심 분야로 가정하였다. 시스템은 관심 분야 문서 5 개와 난수 발생시켜서 선택된 비관심 분야 문서 5 개로 이루어진 10 개의 문서 중 사용자의 관심 분야에 적합한 문서인가를 평가하였다. 제안한 문서 표현 방법이 실제로 사용자의 관심 분야를 잘 기술하는가를 평가해야 하지만 프로파일 시스템에 대한 객관적인 평가 방법이 없기 때문에 현재 정보검색 분야에서 많이 사용하고 있는 F1-measure[8]를 이용하여 측정하였다.

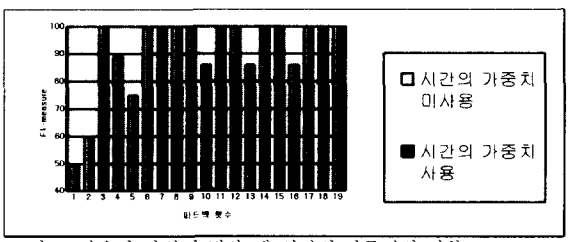


[그림 4] 11 개 분야에 대한 문서 표현 실험결과 (F1-measure 의 총평균)

[그림4]를 살펴보면 사용자의 피드백이 증가할수록 F1-measure 의 값이 증가하므로 프로파일이 사용자의 관심분야에 맞게 학습된다는 것을 알 수 있다. 또한 단순한 tri-gram 을 사용하는 것보다 의존구조를 사용하면 성능이 향상되고 사용자로부터 적은 피드백으로도 사용자의 관심분야와 관계된 문서를 선택할 수 있다는 것을 알 수 있다.

시간에 대한 가중치가 사용자의 관심 변화에 어떤 영향을 미치는가를 알아보기 위해 관심이 고정되어 있을 경우와 고정 되어 있을 않을 경우로 나누어 실험하였다. 먼저 사용자 관심이 고정되어 있는 경우에 문서 표현 실험과 같은 실험 환경에서 F1-measure 값의 변화를 측정한 결과 시간의 가중치를 사용하지 않았을 때 tri-gram 과 의존구조는 F1-measure 값이 각각 81.2%, 87.3%로 나타났고, 사용하였을 경우는 각각 80.9%, 87.4%로 변화가 거의 없었다.

다음 실험으로 사용자의 관심 분야가 변할 때 시간에 대한 가중치가 어떤 영향을 미치는지 알아보았다. 전체적 관심 분야가 스포츠이고 사용자의 관심이 세부적으로 축구에서 야구로 변화하는 가상의 사용자를 가정하고 시간에 대한 가중치를 적용했을 때의 변화를 알아보았다. 실험 데이터는 스포츠와 비 스포츠 관련 HTML 문서가 각각 80 개로 스포츠 관련문서는 축구 관련 HTML 문서 32 개와 야구 관련 HTML 문서 48 개로 구성된다. 가상의 사용자는 스포츠 관련 문서 4 개, 비 스포츠 관련 문서 4 개로 구성된 8 개의 문서 중에서 스포츠 관련 문서는 긍정적 평가를 주고 비 스포츠 관련 문서에 대해서는 부정적 평가를 주었다. 시스템은 4 개의 스포츠 관련 문서를 긍정 피드백으로 4 개의 비 스포츠 관련 문서는 부정 피드백으로 이용한다. 이렇게 학습한 프로파일을 이용하여 다음 8 개의 문서에 대해서 주어진 문서가 스포츠 관련문서인지 아닌지를 예측하였으며 8 개의 문서에 대하여 19 번의 피드백을 주었다. 사용자는 8 번 째 피드백부터 세부 관심 분야가 축구에서 야구로 변화한다.



[그림 5] 사용자 관심이 변할 때 시간의 가중치의 영향

[표 1] 시간의 가중치 사용 여부에 따른 프로파일의 특징 변화

시간의 가중치 미사용	시간의 가중치 사용
축구, 박찬호, 구단, 계약, 팀, 시즌, 월드컵, 프로, 선수, 재계 약, 일본, 연봉, 리그, 이관우, 올 시즌, 입력, 울산, 감독, 메이저리그, 올,드래프트	박찬호, 구단, 축구, 계약, 팀,시즌, 프로, 선수, 재계 약, 일본, 연봉, 리그, 월드컵, 올 시즌,입력, 감독, 메이저리그, 올,이관우,울산,야구

본 실험의 목적은 시간의 가중치를 부여할 경우 관심 분야 문서와 다른 분야문서는 잘 구별하면서 프로파일을 구성한 특징들 중 최근에 브라우징한 문서와 관련된 특징들이 더 높은 가중치를 갖도록 하는가를 확인하는 것이다.

실험 결과[그림5] 사용자 관심이 변할 경우 시간에 대한 가중치를 적용하였을 때 관련분야가 다른 문서들 사이에서 관련 문서를 예측할 때 거의 변화가 없음을 알 수 있었다.

[표1]은 마지막 19 번째 피드백까지 학습시킨 프로파일의 특징 벡터를 높은 가중치 순서로 정렬하였을 때 상위 21 개의 결과이다. [표1]을 살펴보면 시간의 가중치를 사용하지 않았을 경우 축구 관련 특징들인 '축구', '월드컵', '이관우'가 야구 관련 특징인 '박찬호', '메이저리그', '야구'보다 높은 가중치를 갖는 것을 볼 수 있다. 이것은 과거의 관심 분야였던 축구에 관한 특징들이 현재의 관심 분야인 야구에 관한 특징들보다 높은 가중치를 갖는다는 것을 의미한다. 이것은 시간에 대한 가중치를 적용하지 않는 프로파일은 사용자의 현재의 관심을 잘 반영할 수 없다는 것을 의미한다. [표1]에서 시간에 대한 가중치를 적용하였을 경우에는 '축구'보다 '박찬호'가 높은 가중치를 갖고 '이관우' 보다 '메이저리그'가 높은 가중치를 보인다. 이것은 프로파일이 사용자의 현재 관심 분야를 잘 반영한다는 것을 의미한다. 시간에 대한 가중치를 사용할 경우 시간이 지날수록 사용자가 축구에 대한 관심을 보이지 않고 지속적으로 야구에 관심을 보이면 축구 관련 특징에 대한 가중치는 점점 낮아지고 야구에 관련 특징의 가중치는 점점 높아지게 되므로 사용자의 현재의 관심분야를 효과적으로 반영할 수 있다.

5. 결론 및 향후과제

본 논문에서는 기존 프로파일 시스템의 문제점을 해결하고 효과적으로 프로파일을 관리하는 웹 문서 추천 에이전트를 제안하고 구현하였다. 효과적인 프로파일 구성을 위해 슬라이딩 윈도우를 통해 인접한 단어쌍의 문맥 정보를 이용하고, 의존 구조를 통해 의존소와 지배소 사이의 문맥 정보를 이용하는 효과적인 문서 표현 방법을 제안하였다. 또한 사용자의 관심 변화에 대해 효과적으로 적응할 수 있는 시간에 대한 가중치를 이용한 프로파일 특징 추출법을 제안하였다.

제안된 기법을 이용하는 웹 문서 추천 에이전트는 사용자 대신 웹에서 사용자에게 필요한 정보를 수집하고 사용자의 기호에 맞는 정보만을 선별하여 사용자에게 추천한다. 제안된 기법으로 생성된 프로파일을 사용한 웹 문서 추천 에이전트는 사용자의 피드백을 받아 자동적으로 사용자의 관심 분야를 학습하고 사용자의 관심 변화에 효율적으로 대응한다.

본 시스템에서는 시간에 대한 가중치는 비례상수 k 값에 따라 전체 가중치에 영향을 주기 때문에 시간에 대한 가중치를 효과적으로 반영하기 위해 적절한 k 값을 자동으로 구하는 연구가 필요하다. 유사도 측정을 위해 사용하는 단순 베이지안 분류기는 단어 발생의 독립성을 가정하므로 사용자 프로파일을 효과적으로 기술하지 못할 수도 있다. 그러므로 유사도를 구하기 위해 다양한 알고리즘의 적용이 필요하다.

6. 참고 문헌

[1] Michael Pazzani, Jack Muramatsu, Daniel Billsus "Syskill & Webert : Identifying interesting web sites," National Conference on Artificial Intelligence, 1996.

[2] Dunja Mladenic, Marko Grobelnik, "Word sequence as features in text-learning," ERK'98 , IEEE'98, 1998.

[3] Marko Balabanovic, Yoav Shoham, "Learning Information Retrieval Agent : Experiments with Automated Web Browsing" In Proceeding of the AAAI Spring Symposium on Information Gathering from Heterogeneous Distributed Resources, 1995.

[4] 김학수, 고영중, 박수용, 서정연,"문서간 유사도 측정을 통한 효율적인 사용자 요구분석," HCI'99 학술대회 논문집, pp.73-79, 1999.

[5] Harksoo Kim, Youngjoong Ko, Sooyong Park and Jungyun Seo, "Informal Requirements Analysis Supporting System for Human Engineer," In the Proceedings of Conference on IEEE-SMC99, vol. III:1013-1018, Japan, 1999.

[6] lewis, D.,D., Gale, W.,A., "A Sequential Algorithm for Training Text Classifiers," Proceeding of the 7th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Dubline, 1994.

[7] Joachims, T., "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization," Proceedings of the 14th International Conference on Machine Learning ICML97, 1997.

[8] Moulinier, Isabelle, "A Framework for comparing text categorization approaches," In AAAI Spring Symposium : Machine Learning in Information Access, March, 1996.