하이퍼링크 정보를 이용한 정보여과 시스템에서의 사용자 프로파일 학습

박민규, 김준태
동국대학교 컴퓨터공학과

Learning User Profile in Information Filtering System by Using Hyperlink Information

Minkyu Park and Juntae Kim
Department of Computer Engineering, Dongguk University

요 약

본 논문에서는 정보여과 시스템에서 웹 페이지를 수집하고 여과하는 과정과 사용자 프로파일을 학습하는 과정에 하이퍼링크 정보를 이용하는 방법을 제안한다. 사용자가 원하는 웹 페이지를 추천하기 위해 사용자 프로파일을 하이퍼링크 정보를 이용해 만들고 사용자의 반응(feedback)에 따라 사용자 프로파일을 조정한다. 가중치 조정에 있어서 학습 효과를 높이기 위해 사용자가 반응을 보인 웹 페이지에서 출발해 링크로 연결된 모든 페이지에 대해 깊이에 따라 가중치를 조정하는 가중치 전파 알고리즘(Weight Propagation Algorithm)을 제안한다. 적은 사용자의 반응으로도 프로파일 내의 많은 페이지에 영향을 줄 수 있어 높은 학습 효과를 기대할 수 있다.

1. 서론

인터넷 보편화에 따른 온라인 문서의 폭발적인 증가로 일반 사용자들은 웹으로부터 다양한 정보를 접하게 되었다. 하지만 정보의 양이 증가됨에 따라 불필요한 정보도 함께 증가하는 상황에서 자신이 원하는 정보를 정확히 찾아낸다는 것은 매우 어려운 일이 되었다. 일반 사용자들은 방대한 웹에서부터 원하는 정보를 얻기 위해 야후(Yahoo)나 알타비스타(AltaVista) 등의 검색 엔진을 사용하고 있지만, 사용자의 요구를 충족하기에는 많은 시간과 노력을 들어야만 한다.

이러한 문제점을 해결하기 위해 다양한 연구가 수행되어 왔다. 이러한 연구들로 사용자의 검색 질의에 대해 다양한 검색 결과를 종합적으로 분석하여 정보를 제공하는 메타서치엔진, 사용자를 대신하여 사용자의 취향을 분석하고 웹 페이지를 수집하고 여과하여 추천하는 웹 에이전트에 관한 연구 등이 수행되고 있다. 이러한 시스템들은 대부분 웹 페이지를 여과하기 위해 내용기반 방법을 사용하고 있다. 하지만 웹 페이지는 HTML 태그와 같은 비문자 정보가 많은 의미를 가진다. 기존의 내용기반 방법으로 웹 페이지를 여과하는 경우, 인덱스 페이지(index page)나 이미지 맵(image map)과 같이 문자 정보가 거의 포함되지 않은 페이지인 경우 한계를 드러내고 있다. 또한 단순히 주어진 단어가 한 두 개 더 많은 페이지가 반드시 더 중요한 페이지라고 할 수는 없다.

본 논문에서는 내용기반 방법이 가지는 한계를 극복하기 위해 정보여과 시스템에서 웹 페이지를 수집하고 여과하는 과정과 사용자의 취향을 학습하는 과정에 하이퍼링크 정보를 이용하는 방법을 제안한다. 학습을 위한 사용자 프로파일의 가중치 조정에 있어서는 학습 효과를 높이기 위한 가중치 전파 알고리즘을 제안한다. 가중치 전파 알고리즘의 효과를 보이기 위해 사용자의 반응에 따라 학습이 진행되는 과정에서 가중치 전파 알고리즘을 사용한 경우와 사용하지 않은 경우를 비교하는 실험을 수행한다.

2. 관련 연구

하이퍼링크에 관한 연구는 문서들과 그 문서가 가지는 인용문들의 구조를 연구하는 Bibliometrics[4]와 많은 관련성을 가진다. Bibliometrics에서 문서의 유사도를 분석하는 기본적인 방법으로 bibliographic coupling과 cocitation을 사용한다. bibliographic coupling은 두 문서 모두를 가리키는 문서들의 수를 의미하며, cocitation은 두 문서 모두가 가리키는 문서들의 수를 의미한다[1].

하이퍼링크에 관한 연구들은 대부분 그래프 이론에서 출발하였는데, Page[3]는 방향을 갖는 링크 그래프에서 PageRank를 계산하는 방법으로 순위를 계산하였고, Rao[5]는 페이지들을 그룹화하고 분류하는데 링크 기술과 내용 기반 기술을 동시에 이용하여 유사도를 계산하는 방법이 시도하였다. Kleinberg[1]는 권위(authority) 페이지와 허브(hub) 페이지를 정의하여 사용자가 원하는 페이지를 찾는데 사용하였다. 권위 페이지는 주제에 관해 많은 정보를 포함한 페이지이고 허브 페이지는 주제에 관해 정보를 가지는 페이지의 링크를 많이 가지는 페이지를 의미한다. Kleinberg는 좋은 허브 페이지와 권위 있는 페이지를 찾는 알고리즘으로 HITS(Hypertext-Induced Topic Search) 알고리즘을 제안하였다.

본 논문에서 제안하는 하이퍼링크 이용 방법은 Kleinberg

의 방법과 유사하나 학습과정에서 하이퍼링크 정보를 이용하는 가중치 전파를 함으로써 학습효과를 높이도록 하였다.

3. 하이퍼링크 정보를 이용한 프로파일 학습

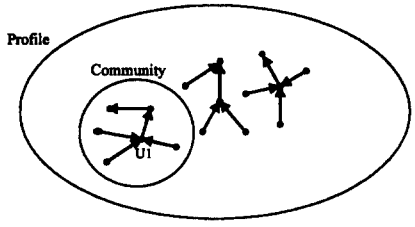
3.1 하이퍼링크 구조

하이퍼링크 정보를 이용하기 위해서 HTML 문서로부터 <a href = *url*> 태그를 수집하여 *url*을 추출한다. 추출한 *url*들이 상호 연결되어 있을 때 페이지 P1에서 P2로 하이퍼링크가 있고 페이지 P2에서 P3으로 하이퍼링크가 있다면 페이지 P2를 기준으로 P1을 '인페이지' (Inpage)로 정의하고, P3을 '아웃페이지' (Outpage)로 정의한다. 예를 들면, P1 페이지 내의 <a href = "P2">가 존재하면 P1이 P2를 가리킨다고 말하고 P2를 기준으로 P1은 '인페이지'가 되는 것이다. 하이퍼링크로 연결된 *url*들이 같은 호스트에 존재하는 경우에는 링크 정보가 가치 없는 경우가 많으므로 다른 호스트 이름을 갖는 하이퍼링크 정보만을 이용한다.

3.2 사용자 프로파일

사용자 프로파일은 사용자의 관심(interest)을 문자, 숫자, 기호로 표현한 것이다. 사용자의 취향을 반영하여 사용자 프로파일을 구성하고 사용자를 모델링하여 변화하는 사용자의 취향에 접근한다.

본 논문에서는 사용자 프로파일을 방향을 갖는 가중치 그래프 G로 표현한다. 그래프 G에서 정점(V)은 웹 페이지들의 주소(*url*)를 나타내고, 방향을 갖는 간선($E(U1,U2)$)은 웹 페이지 U1이 웹 페이지 U2를 가리키는 하이퍼링크를 나타낸다. 각각의 웹 페이지들은 가중치(weight)를 가지고 W로 표시한다. 웹 페이지들을 이용해 프로파일을 그래프로 구성하면 상호 연결된 군(community)이 생성되는데 이러한 군은 C로 표시한다.



[그림1] 프로파일 구조

[그림1]은 하이퍼링크를 이용한 프로파일의 한 예이다. 이 프로파일은 3개의 군과 26개의 웹 페이지의 주소로 구성되고, 각각의 웹 페이지들은 가중치와 '인페이지'들과 '아웃페이지'들을 가진다. 예를 들어, 웹 페이지 U1은 3개의 '인페이지'와 1개의 '아웃페이지'를 가진다.

3.3 웹 페이지 수집과 여과

본 논문에서는 학습을 수행하기 위해서 프로파일에 웹 페이지 주소를 추가, 삭제하는 작업과 가중치 조정하는 작업을 수행한다.

우선 초기 질의를 구성하여 검색엔진에 질의 후 검색 결과 리스트에서 관련 웹 페이지를 수집한다. 질의는 검색엔진 주소와 검색 엔진에서 사용되는 연산자(&,|)들과 기호들로 구성한다. 검색결과로 웹 페이지가 수집되고 나면 하이퍼링크 정보를 이용하여 수집된 웹 페이지를 가리키는 페이지와 수집된 웹 페이지가 가리키는 페이지를 수집하여 관련 페이지의 집합을 확장한다. 확장된 웹 페이지 집합을 방향을 갖는 그래프로 표현하고 통합한다. 같은 호스트 이름을 갖는 페이지들은 홈페이지를 찾아 하나로 통합한다. 통합 과정을 거쳐 그래프를 재구성한 후 모든 웹 페이지의 가중치를 계산하고 기준값 이상이 되는 웹 페이지들을 추출하여 사용자에게 제공하고 프로파일을 구성한다.

웹 페이지의 가중치를 계산하는 식은 다음과 같다.

$$W_U = \alpha\left(\frac{\sum W_{IU}}{\sum W}\right)+(1-\alpha)\left(\frac{\sum W_{OU}}{\sum W}\right)$$

W_U : page *url* U의 가중치
$\sum W$: 모든 page의 가중치 합
$\sum W_{IU}$: U의 Inpage의 가중치 합
$\sum W_{OU}$: U의 Outpage의 가중치 합
α : W_{IU} 와 W_{OU} 상대적인 중요도 ($0 \le \alpha \le 1$)

이 식에서 상수 α는 '인페이지'와 '아웃페이지'의 상대적인 중요도를 나타낸다. 이 식의 의미는 자신을 가리키는 페이지(Inpage)와 자신이 가리키는 페이지(Outpage)가 많은 페이지에 높은 가중치를 주는 것이다. 페이지 U의 '인페이지'와 '아웃페이지'를 전체 페이지의 가중치의 합으로 나누어 계산하므로 W_U는 항상 0과 1사이의 값을 갖는다.

3.4 프로파일 학습

사용자의 관심을 학습하기 위하여 추출된 웹 페이지를 사용자에게 추천하고 사용자의 반응을 받는다. 본 논문에서는 명시적 피드백을 사용한 인터페이스를 사용자에게 제공하고 사용자의 반응에 따라 가중치 전파 알고리즘을 수행하여 프로파일을 변경한다.

가중치 전파 알고리즘은 사용자가 반응을 보인 웹 페이지에서 출발해 링크로 연결된 모든 페이지에 대하여 깊이에 따라 가중치를 변경하는 알고리즘이다. 본 논문의 가중치 전파 알고리즘은 적은 사용자의 반응으로도 프로파일 내의 많은 페이지에 영향을 줄 수 있다. 또, 급격하게 사용자의 취향이 변할 때 프로파일 내의 군들의 전체적인 가중치도 함께 급격하게 변화하게 되어 사용자의 취향에 적응할 수 있다.

가중치 전파 알고리즘에서 각 페이지에 가중치를 조정하는 과정은 Breadth First Search와 비슷하다. [그림2]에 알고리즘을 나타내었다.

U의 깊이(depth)는 1로 초기화되고, 사용자의 반응을 받은 초기 페이지(initial page)를 큐(queue)에 넣는다. 큐가 비었는지를 체크하여 큐가 비었으면 끝나고 원소가 있으면 계속 수행한다. 큐의 가장 왼쪽의 원소를 U로 정의하고 U의 가중치가 조정되었는지를 체크하여 조정되지 않았으면 가중치를 조정한 뒤 U의 '아웃페이지'들을 큐에 넣는다. 다시 U

의 가장 왼쪽의 '아웃페이지'가 U가 되어 가중치가 깊이에 따라 수정되며 계속해서 다른 '아웃페이지'들도 깊이에 따라 가중치가 수정된다.

```
d of initial page = 1
put initial page on queue[]
while queue[] > 0 do
    begin
        remove leftmost page from queue[], call it U
        if (U is not adjusted) adjust weight of U
        if (U is not extended and depth of U < 5)
            begin
                put Outpages of U on queue[]
                d of Outpage = d of U + 1
            end
    end
```

[그림2] 하이퍼링크를 이용한 가중치 전파 알고리즘

가중치 전파 알고리즘에서 가중치를 조정하는 식은 다음과 같다. (±는 초기페이지가 관련 문서인 경우 +, 비관련 문서인 경우 -로 계산한다.)

$$i)\ W'_U = \beta W_U \pm \frac{(1-\beta)}{d}\left\{\alpha\left(\frac{\sum W_{IU}}{\sum W}\right)+(1-\alpha)\left(\frac{\sum W_{OU}}{\sum W}\right)\right\}$$

$$ii)\ W'_U = \beta W_U \pm \frac{(1-\beta)}{d}\left(\frac{\alpha\gamma}{\sum W}\right),\ \text{if}\ \sum W_{IU} = \sum W_{OU} = 0$$

β: 학습 비율($0 \le \beta \le 1$), γ: 1/10000, d: depth

이 식에서 상수 β는 기존의 가중치와 사용자의 반응의 상대적인 중요도를 나타내는 것으로, β가 0에 가까울수록 사용자 반응에 따른 가중치 변화가 커지게 된다. 상수 γ는 '인페이지'와 '아웃페이지'가 존재하지 않는 경우 단 하나의 링크가 존재하는 경우보다 적은 값을 가지기 위해 1/10000을 사용하였다. 이 식을 적용하면 사용자가 관련 문서라고 반응한 페이지로부터 링크로 연결된 페이지들은 가중치가 증가하게 된다.

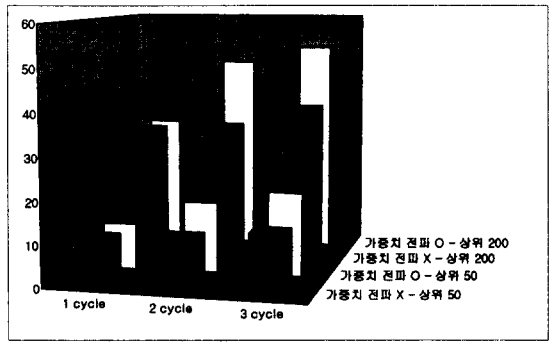
프로파일의 가중치가 모두 조정되고 나면 프로파일 내의 가장 높은 평균 가중치를 갖는 군에 존재하는 상위 5개의 페이지 내에서 단어를 추출해 단어의 빈도수를 계산하고 이러한 과정 속에서 가장 높은 빈도수를 갖는 단어를 추출하여 다시 검색 엔진에 질의하여 프로파일을 확장한다.

4. 실험 및 결과

가중치 전파 알고리즘의 효과를 알아보기 위해 가중치 전파 알고리즘을 사용하여 학습하는 경우와 가중치 전파 알고리즘을 사용하지 않고 학습하는 경우를 비교하는 실험을 수행하였다. 본 실험에서 사용한 가중치 조정 식의 상수 α값과 β값은 실험을 통해 얻은 값으로 1/4, 1/2를 사용하였다.

평가방법은 사용자가 'Artificial Intelligence'에 관심있다고 가정하고, 본 논문에서 제시한 하이퍼링크를 이용한 정보여과 시스템을 이용하여 수집한 웹 페이지를 높은 가중치 순으로 정렬한 뒤, 야후의 Artificial Intelligence 카테고리와 그 서브 카테고리들에서 추출한 *url*들과 비교하여 일치하는 수를 상위 5개, 10개, 50개, 100개, 200개에 대하여 비교하였다. 각 학습단계에서는 상위 10개의 페이지에 대해서 사용자가 페이지에 대한 반응(피드백)을 하도록 하였다.

[그림3]은 가중치 전파 알고리즘 효과 실험 결과로 상위 50개와 200개중에 일치하는 *url*의 수를 학습 단계별로 나타낸 것이다. 가중치 전파에 관계없이 모두 학습이 진행될수록 관련 페이지의 수가 증가하는 것을 알 수 있다. 하지만 가중치 전파 알고리즘을 사용한 경우가 학습이 진행될수록 일치하는 페이지의 수가 빠르게 증가하는 것을 볼 수 있었다. 또, 가중치 전파 알고리즘을 사용한 경우 상위 200개에서 상위 50개에 비해 일치하는 수가 많이 증가하는 것도 볼 수 있었다. 즉, 가중치 전파 알고리즘을 사용하면 전체적으로 프로파일의 변화를 기대할 수 있고 높은 학습효과를 기대할 수 있다.



[그림3] 상위 50개, 200개 중 일치하는 *url* 수의 변화

5. 결론

본 논문에서는 정보 여과 시스템에서 웹 페이지를 수집하고 여과하여 사용자 프로파일을 학습하는 과정에 하이퍼링크 정보를 이용하여 수행하는 방법을 제안하였다. 학습과정에서는 가중치 전파 알고리즘을 사용하여 학습의 효율을 향상시켰으며 실험을 통하여 본 논문에서 제안한 방법이 효과적임을 보였다.

참고문헌

[1] Kleinberg. J, Authoritative sources in a hyperlinked environment, *Proceeding of 9th ACMSIAM symposium on Discrete Algorithms.*

[2] Krishna B., Monika R.H., Improved Algorithms for Topic Distillation in a Hyperlinked Environment , *Proceeding of ACM SIGIR. 97*, 1997

[3] L. Page, "PageRank: Bringing order to the Web", stanford Digital Libraries working paper 1997-0072

[4] H.D. White and K.W. McCain, "Bibliometrics", *in: Ann. Rev. Info. Sci. and Technology*, Elsevier, 1989, pp. 119-186.

[5] P.Pirolli, J.Pitkow, R.Rao, Silk from a Sow's Ear: Extracting Usable Structures from the Web, *Proceeding of ACM SIGCHI Conference on Human factors in Computing* , 1996