질의어 단위 프로파일링을 이용한 북마크 기반 개인화 검색

(Bookmark-Based Personalized Search through Query-Level User Profiling)

김 현 지 † 배 동 환 †
(Hyun Ji Kim) (Dong Hwan Bae)
고 민 삼 † 이 문 용 ††
(Minsam Ko) (Mun Yong Yi)

요 약 본 논문에서는 개인화 검색 시 사용자의 단일 프로파일이 개인의 다양한 정보 요구를 만족시키지 못하는 문제를 개선하고자, 질의어에 따라 프로파일을 조정하는 방법을 제안한다. 특히, 제안하는 방법은 질의어에 관해 사용자가 중요하게 생각하는 단어들을 북마크 데이터로부터 추출하여 프로파일 조정에 활용한다. 유명 북마크 서비스 제공자인 CiteULike의 데이터를 활용한 실험에서, 본 연구에서 제안하는 방법이 단일 프로파일에 기반한 기존의 방법보다 더 뛰어난 개인화 검색 결과를 제공함을 확인할 수 있었다.

키워드: 사용자 프로파일, 개인화 검색, 북마크

Abstract The approach of a single user profiling for personalized search does not satisfy a person's diverse information needs. This paper presents a new approach, query-level user profiling, for personalized search in order to improve the situation. Our proposed method extracts query-related words from each user's bookmarking data and makes use of them in the process of adjusting the user's profile. We tested our approach using the field data obtained from CiteULike, a popular bookmarking service provider on the Web. Our experiment results show that the proposed approach through query-level user profiling outperforms the existing personalized search method based on single user profiling.

Keywords: user profile, personalized search, bookmark

1. 서 론

최근 개인의 정보 요구에 맞춤화 된 검색 결과를 제공하는 개인화 검색이 주목을 받고 있다[1]. 개인화 검색은 같은 질의어라 해도 개인의 관심사에 따라 다른 검색 결과를 반환한다. 일반적으로 개인화 검색은 (1) 사용자의 관심사를 표현하는 단어들로 구성된 사용자 프로파일을 만들고, (2) 해당 프로파일에서 입력된 질의어와 관련된 문서들의 가중치를 높여주는 재정렬 순서로 진행된다.

대부분의 개인화 검색 방법들은 사용자 별로 하나의 프로파일만을 만들어 사용한다. 그러나 여러 영역에 걸친 사용자의 다양한 관심사는 단일 프로파일에 모두 표현하기에 한계가 있다. 예를 들어, 오랜 기간 데이터 마이닝에 관심 있던 사람이 새롭게 HCI에 관심을 가지게 되었다고 하자. 데이터 마이닝에 관한 사용자의 행동 기록이 많기 때문에, 프로파일은 데이터 마이닝을 중심으로 구축되어 있을 것이다. 그 결과, HCI 관련 단어들은 상대적으로 프로파일에서 소외되고, 결국 HCI 관련 검색에 대한 개인화 효과는 떨어지게 된다.

본 논문에서 우리는 단일 프로파일에 기반한 방법의 문제를 개선하기 위해 질의어에 따라 프로파일을 조정하여 개인화 검색에 사용하는 방법을 제안한다. 사용자가 질의어를 입력하면, 제안하는 방법은 입력된 질의어와 연관하여 사용자가 중요하게 생각하는 단어들이 개인화 과정에서 더 중요하게 고려되도록 프로파일을 조정한다. 따라서 원래의 프로파일에서는 상대적으로 소외되었던 단어들도 특정 질의어에 대해서는 더 중요하게 고려될 수 있게 된다. (본 논문에서 질의어와 관련하여 사용자가 중요하게 생각하는 단어는 "질의어 관련 단어"라 표기되고, 조정된 프로파일은 "질의어 단위 프로파일"이라 표기된다.)

특히, 제안하는 방법은 질의어 관련 단어를 추출하기

· 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2011-0029185)
· 이 논문은 2012 한국컴퓨터종합학술대회에서 '질의어 단위 사용자 프로파일을 이용한 북마크 기반 개인화 검색 방법'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : KAIST 지식서비스공학과
jacehyun@gmail.com
bdhwan@gmail.com
minsam.ko@gmail.com
†† 정 회 원 : KAIST 지식서비스공학과 교수
munyi@kaist.ac.kr
(Corresponding author임)

논문접수 : 2012년 8월 20일
심사완료 : 2012년 11월 19일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제19권 제3호(2013.3)

위해 북마크 데이터를 사용한다. 북마크 데이터에서 사용자, 태그, 그리고 문서를 각각 사용자, 질의어, 그리고 사용자에게 적합한 문서로 가정한다. 이 가정에 기반하여 어떤 사용자가 단어 “A”로 태그를 단 문서들로부터 단어들을 추출하고, 추출된 단어들을 단어 “A”에 대한 이 사용자의 질의어 관련 단어들로 사용한다. 북마크는 사용자의 관심사를 직접적으로 표현하기 때문에 기존의 검색 방법들[2,3]에서 사용된 클릭 데이터보다 정확하게 사용자의 의도를 알 수 있다.

우리는 유명한 북마크 서비스인 CiteULike의 데이터를 사용하여 제안하는 방법의 성능을 기존 방법과 비교해 보았다. 실험결과는 제안하는 방법이 단일 프로파일에 기반한 기존의 개인화 검색 방법보다 더 뛰어난 성능을 가진 것을 보여준다.

2. 관련 연구

2.1 사용자 프로파일

개인화 검색에서 프로파일은 질의어에 대한 사용자의 정보 요구를 알아내는데 사용된다. 질의어들은 길이가 짧고[4] 그 뜻이 모호한 경우[5]가 많다. 따라서 사용자의 정보 요구에 맞는 개인화 검색을 위해서는 정확한 프로파일을 구성하는 것이 필수이다.

그 동안 프로파일 구축을 위해 클릭[2,3,6], 북마크[7], 사회적 관계[8] 등과 같은 다양한 데이터가 사용되었다. 본 논문이 제안하는 방법은 북마크 데이터를 활용한다. 북마크 데이터는 사용자가 선호하는 문서에 단 태그들을 포함하며, 기존 연구들에서는 문서의 특징과 사용자의 관심사를 알아내는데 주로 사용되어 왔다[7]. 본 논문에서는 북마크 데이터의 태그를 사용자가 입력할 법한 질의어로 간주하여, 개인화 검색을 위한 질의어 단위 프로파일을 만드는데 활용한다.

2.2 질의어 단위 프로파일

많은 개인화 검색 방법이 단일 프로파일에 기반하고 있는 반면, 클릭데이터로 질의어 단위 프로파일을 구축한 개인화 검색 연구들도 있었다[2,3,6]. 이들은 검색 결과 중 사용자가 클릭 한 문서들을 질의어에 적합한 문서라 가정하고, 이 문서들에서 추출된 단어들을 질의어 관련 단어들로 사용하였다.

하지만 클릭한 문서 전부가 사용자에게 적합한 문서는 아닐 수 있다[2]. 단순히 문서의 검색 순위가 높아서 클릭 했을 수도 있고, 제목과 스니펫(Snippet)을 보고 클릭했으나 본문은 만족스럽지 않았을 수도 있다. 본 연구는 사용자가 본문을 직접 확인하고 저장한 북마크를 사용하기 때문에, 보다 정확한 질의어 단위 프로파일을 만들 수 있다.

2.3 검색 결과의 재정렬

프로파일에 기반한 검색 결과 재정렬 방법에는 크게 콘텐트 기반 방법과 협업 기반 방법이 있다.

먼저 콘텐트 기반 방법은 개인의 프로파일과 문서 사이간 유사도에 따라 검색된 문서의 순위를 결정한다. 프로파일과 문서를 비교하기 위해 Chirita[9]는 문서에 대한 Open Directory Project(ODP) 데이터를 사용했고, Tan[3]은 통계적 언어 모델을 사용했다.

반면 협업 기반 방법은 사용자들의 프로파일 간 유사도에 근거하여, 나와 비슷한 관심을 가지는 사람들이 만족한 정도에 따라 문서들의 검색 순위를 결정한다. 대표적으로는 추천 시스템에서 사용하는 협업 필터링을 통해 검색 순위를 결정하는 방법[10]과 기존 협업 필터링에 질의어를 더 고려한 방법인 CubeSVD[6]가 있다.

본 논문이 제안하는 질의어 단위 프로파일은 두 방법 모두에 쉽게 적용될 수 있으나, 본 논문에서는 계산 비용이 상대적으로 적은 콘텐트 기반 방법에 먼저 적용해 보았다. 앞으로 질의어 단위 프로파일을 협업 기반 방법에도 적용해 볼 예정이다.

3. 질의어 단위 사용자 프로파일을 이용한 북마크 기반 개인화 검색 방법

3.1 전체 수행 과정

그림 1은 제안하는 방법이 수행되는 전체 과정을 보여준다.

그림 1 제안하는 방법의 전체 흐름도

제안하는 방법은 “문서 모델링,” “사용자 프로파일 구축,” “질의어 단위 프로파일 구축,” 그리고 “검색결과 재정렬”의 네 단계로 진행된다. 다음 장에서는 각 과정에 대한 자세한 설명이 이어진다.

3.2 문서 모델링

먼저 문서들을 연산 가능한 형태로 변환하기 위해 문서 모델링을 수행한다. 이 과정을 통해 문서에서 중요한 단어들이 뽑혀지고 각 문서가 하나의 단어 벡터 형태로 변환된다.

단어 벡터는 문서를 표현하는 단어들과 그 단어들이 해당 문서에서 중요한 정도를 나타내는 가중치들로 구성된다. 문서를 표현하는 단어들은 문서의 본문이나 서지 정보를 이용해도 되지만, 본 논문에서는 문서에 대해

사람들이 단 모든 태그들을 사용한다. 단어들의 가중치로는 tf-idf (Term Frequency - Inverse Document Frequency)에 의해 계산된 값을 사용한다. 식 (1)은 문서 d에 대한 단어 벡터를 나타낸다. t_n은 문서 d에 포함된 단어들이며, n은 문서 d가 포함하고 있는 단어들의 총 수를 의미한다. $tfidf_{t_n,d}$는 d에서 t_n이 가지는 tf-idf 값을 나타낸다.

$$d = \{(t_1, tfidf_{t_1,d}), ..., (t_n, tfidf_{t_n,d})\} \quad (1)$$

이렇게 모델링 된 문서들은 사용자 프로파일 구축 및 검색 결과 재정렬 과정의 기본 재료가 된다.

3.2 사용자 프로파일 구성

이어서 사용자의 전반적인 관심사 및 성향을 표현하는 사용자 프로파일을 구축한다. 전체 사용자 프로파일 구축 과정 없이 질의어가 입력되었을 때 질의어 단위 프로파일을 동적으로 구축하는 것도 가능하다. 하지만 미리 전체 사용자 프로파일을 만들어 놓으면 질의어 입력 때마다 프로파일을 다시 구축할 필요 없이 미리 구축해 놓은 프로파일에서 질의어와 관련된 단어들의 가중치들만 조정하면 되기 때문에 연산 효율이 향상되는 장점이 있다.

사용자 프로파일은 사용자가 태그를 단 문서들에 등장한 단어들의 가중치들이 합쳐져 하나의 단어 벡터로 표현된다. 결과적으로 사용자가 태그를 단 문서들에서 자주 발견되는 단어일수록 더 높은 가중치를 가지게 된다. 식 (2)는 사용자 u의 프로파일 P_u를 나타낸다.

$$P_u = \{(t_1, w_{t_1,u}), ..., (t_m, w_{t_m,u})\} \quad (2)$$

그리고 다음 식 (3)을 통해 사용자 u의 프로파일에서 단어 t의 가중치 $w_{t,u}$를 계산한다. 식 (3)에서 D_u는 사용자 u의 북마크 목록에 있는 모든 문서들의 집합이다.

$$w_{t,u} = \sum_{d \in D_u} tfidf_{t,d} \quad (3)$$

예를 들어 Machine Learning과 HCI분야에 관심이 있는 사용자 u_1의 북마크 목록은 표 1과 같을 수 있다. 북마크 목록에 기반하여 u_1에 대한 사용자 프로파일 P_{u_1}은 {(classification, 5.5), (SVM, 1.1), (Bayes, 1.2), (interaction, 6.3), (usability, 1.8), (interface, 1.4)}과 같은 형태로 구축될 것이다. 그 결과 프로파일 내에서 Machine Learning과 HCI 관련 단어들이 비슷한 가중치를 가지게 된다.

표 1 사용자 북마크 목록 예시

문서	단어 벡터	태그
d_1	{(classification,3.2), …, (SVM,1.1)}	Machine Learning
d_2	{(classification,2.3), …, (Bayes,1.2)}	Machine Learning
d_3	{(interaction,2.5), …, (usability 1.8)}	HCI
d_4	{(interaction,3.8), …, (interface,1.4)}	HCI

3.3 질의어 단위 사용자 프로파일 구성

사용자가 검색을 위해 질의어를 입력하면, 사용자의 프로파일은 질의어 단위 프로파일로 변환된다. 이 과정을 통해 프로파일 내 질의어 관련 단어들의 가중치가 높아지고, 검색 순위 결정 시 이 단어들이 더 중요하게 고려된다. 질의어 관련 단어는 사용자의 북마크 데이터에서 추출된다. 사용자가 특정 태그를 단 문서들을, 그 태그가 질의어로 사용되었을 때 이 사용자에게 적합한 문서로 여긴다. 그리고 질의어에 적합한 문서들에 등장한 단어들을 질의어 관련 단어로 사용한다. 식 (4)는 사용자 u의 질의어 q에 대한 질의어 단위 프로파일 $P^*_{u,q}$을 나타낸다.

$$P^*_{u,q} = \{(t_1, w^*_{t_1,u,q}), ..., (t_m, w^*_{t_m,u,q})\} \quad (4)$$

$w^*_{t,u,q}$는 질의어 단위 프로파일 내 단어들의 조정된 가중치를 의미하며 식 (5)를 통해 계산된다.

$$w^*_{t,u,q} = (df(t, D_{u,q}) + \beta) * w_{t,u} \quad (5)$$

식 (5)에서 $d_f(t, D_{u,q})$는 사용자 u가 질의어 q를 태그로 단 문서 집합 $D_{u,q}$에서 단어 t가 등장한 문서의 수를 반환하는 함수다. β는 $D_{u,q}$에서 한번도 등장하지 않는 단어들(즉, $df(t, D_{u,q} = 0)$)이 질의어 단위 프로파일에서 가중치가 0이 되어 완전히 제외되는 것을 방지하기 위한 변수이다. 질의어 관련 단어가 아닌 단어들도 전체적으로 봤을 때 사용자에게 의미가 있을 수 있다. 따라서 단어들을 완전히 제외하는 것보다는 가중치를 상대적으로 낮추지만 질의어 관련 단어들과 함께 사용하여 사용자의 전반적인 관심사가 같이 고려되도록 하였다. β값이 커질수록 프로파일 내에서 질의어 관련 단어들의 영향력이 높아지며, 반대로 이 값이 0에 가까울수록 프로파일이 질의어에 맞게 조정되는 효과가 미비해진다. 본 논문에서는 이 값을 0.5로 사용하였다. 그리고 만약 입력된 질의어가 사용자의 태그 목록에 존재하지 않을 경우에는 질의어 단위 프로파일을 구축하는 것이 불가능하기 때문에 조정 과정 없이 처음에 구성한 사용자 프로파일을 다음 단계에서 그대로 사용한다.

위의 예제(표 1)의 사용자 u1이 질의어로 HCI를 사용했다면, 질의어 단위 프로파일에서는 HCI로 태그 된 문서들에서 나온 단어들인 interaction, usability, interface 등에 더 높은 가중치가 부여된다. 원래 전체 프로파일에서는 machine learning에 대한 단어들과 HCI에 대한 단어들이 비슷한 가중치로 프로파일이 구성되었지만 질의어 HCI에 대한 프로파일은 HCI에 구체화 되어 있어 사용자 u1의 HCI에 대한 관심사를 더 잘 표현하게 된다.

3.4 검색 결과 재조정

마지막으로 일반 검색 방법을 통해 반환된, 입력된 질의어에 적합한 문서들을, 질의어 단위 프로파일 $P^*_{u,q}$

과 비교하여 보다 유사한 문서를 상위 검색 순위에 올려준다.

각 문서들과 질의어 단위 프로파일간의 코사인 유사도(Cosine similarity)를 계산하여 유사도 값이 높은 문서를 상위 순위에 놓는다. 식 (6)은 문서 d와 사용자 u의 질의어 q에 대한 질의어 단위 프로파일 $P^*_{u,q}$ 간의 유사도를 구하는 식이다.

$$Sim\left(d, P^*_{u,q}\right) = \frac{\sum_{t \in d \cap P^*_{u,q}} tfidf_{t,d^*} w^*_{t,u,q}}{\sqrt{\sum_{t \in d} tfidf_{t,d}^2} \sqrt{\sum_{t \in P^*_{u,q}} w^*_{t,u,q}^2}} \quad (6)$$

4. 실험 결과 및 분석

4.1 실험 데이터

우리는 2011년에 생성된 CiteULike의 북마크 데이터를 실험에 사용하였다. 전체 데이터의 양이 방대 하기 때문에 실험 세팅 및 결과 해석이 용이 하도록 컴퓨터 공학 관련 태그를 많이 사용한 사람들을 평가 대상으로 샘플링 했다. 이 샘플링을 위해 컴퓨터 공학의 주제어들(“social network,” “semantic,” “machine learning,” “mobile”)을 선택하였고, 주제어 별로 주제어를 태그로 많이 사용한 상위 30명씩을 선택했다. 중복 선택된 경우를 제외하여 총 109명의 사용자들이 선택되었다. 선택된 사용자들은 44,038개 문서들에 160,374개 태그들을 달았다. 각 사용자의 북마크들은 북마크 된 시간에 따라 반반씩 나누어져 학습 데이터와 평가 데이터로 사용되었다.

4.2 평가 절차

평가는 다음의 순서로 진행되었다. 먼저 사용자 별로 학습 데이터를 사용해 프로파일을 구성하였다. 그리고 사용자 별로 각 평가 질의어에 대해 CiteULike의 일반 검색이 반환한 문서들 중 상위 50개 문서들을 개인화 방법에 따라 재정렬하였다. 평가 질의어로는 10명 이상의 사용자들이 사용한 태그들(총 216 단어)을 사용하였다. 마지막으로 평가 척도를 사용해 재정렬 된 검색 결과를 점수화 하여, 사용자에게 적합한 문서가 얼마나 상위 순위에 있는지를 보았다.

평가 척도로는 MRR, NDCG, 그리고 Precision이 사용되었다. 세 척도 모두 적합한 문서가 높은 순위에 있을수록 높은 수치를 가진다. NDCG와 Precision은 검색 결과 중 상위 5개 문서에 대해 측정하였고, 각각 NDCG@5와 P@5로 표기한다.

우리는 문서가 사용자에게 적합한지 여부를 판단하기 위하여 “회상”과 “발견”이라는 두 기준을 정의하였다. 회상은 사용자가 전에 본 적합한 문서를 다시 찾은 것을 의미한다. 회상의 관점에서는 검색된 문서들 가운데 사용자의 학습 데이터에 존재하는 문서들이 사용자에게 적합한 문서로 판단된다. 그리고 발견은 사용자가 새롭게 적합한 문서를 찾은 것을 의미한다. 발견의 관점에서는 검색된 문서들 가운데 사용자의 평가 데이터에 존재하는 문서들이 사용자에게 적합한 문서로 판단된다.

4.3 비교 방법

우리는 두 가지 비교 방법들을 준비했다. 첫 번째 비교 방법은 CiteULike에서 제공하는 일반 검색 방법이다(베이스라인). 그리고 다른 하나는 단일 프로파일 기반 개인화 검색 방법이다(단일 프로파일 방법). 단일 프로파일 방법은 질의어를 고려하지 않고 사용자 당 하나의 프로파일을 사용한다. 질의어에 따라 프로파일을 조정하는 단계를 제외하고는 제안하는 방법과 같은 과정을 따른다.

4.4 실험 결과

실험 결과는 제안하는 방법이 회상과 발견 모두에 대해서 비교하는 기존의 방법들에 비해 더 뛰어난 성능을 가진 것을 보여준다(그림 2, 그림 3, 그리고 그림 4 참고). 특히 회상에 대한 평가에서 큰 격차를 보였다. 제안하는 방법은 “회상” 면에서 베이스라인의 성능을 81%~191% 향상시켰고, 단일 프로파일 방법의 성능을 16%~26% 향상시켰다. 단일 프로파일 방법도 베이스라

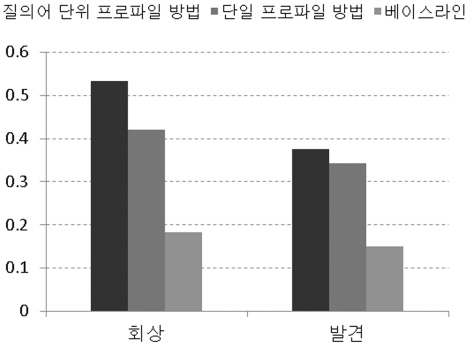

그림 2 Mean Reciprocal Rank (MRR)

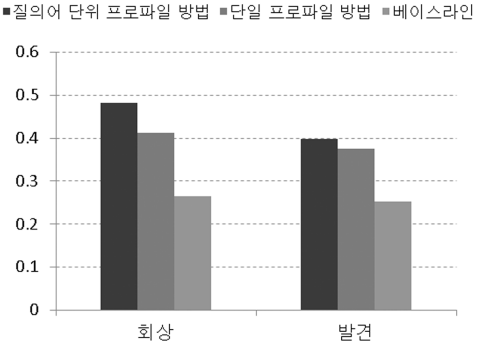

그림 3 NDCG@5

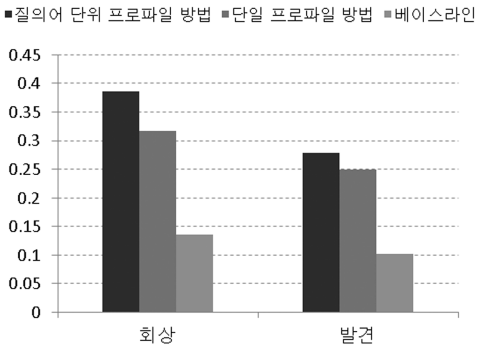

그림 4 Precision@5

인에 비해 향상된 결과를 보여 줬지만, 제안하는 방법에 비해서는 성능이 떨어지는 것을 볼 수 있었다.

발견에 대한 평가에서도 제안하는 방법은 가장 좋은 성능을 보였다. 하지만 회상에 비해 성능 향상이 작았다. 본 논문에서 사용자 프로파일은 한 사용자의 과거 기록만을 참고하여 만들어졌기 때문에 개인화 검색 시 그 사람의 과거 기록에 편향된 결과를 보여주기 쉽다. 우리는 이러한 점을 보완하기 위하여 다른 사람들과의 관계성을 고려하여 질의어 단위 프로파일을 확장하는 방법에 대해 연구하고 있다.

5. 결론 및 향후 연구

본 논문은 북마크 데이터를 이용해 형성한 질의어 단위 프로파일링을 개인화 검색에 적용하는 방법을 제안하였다. CiteULike 데이터를 사용한 성능 평가 실험을 통해 단일 프로파일에 비하여 개인화 검색의 성능을 향상시킨 것을 확인하였다.

우리 사회가 통섭의 시대로 접어들면서 다양한 영역에 관심을 가지는 사람들의 수는 날로 늘어나고 있다. 개인의 다양한 정보 요구들을 잘 반영하는 질의어 단위 프로파일은 개인화 검색은 물론 추천 시스템이나 협업 시스템 등에 매우 유용할 것이다.

참 고 문 헌

[1] N. J. Belkin, "Some(what) Grand Challenges for Information Retrieval," *ACM SIGIR Forum*, vol.42, Issue1, 2008.

[2] Z. Dou, et al., "Evaluating the Effectiveness of Personalized Web Search," *IEEE Transactions on Knowledge and Data Engineering*, vol.21, Issue8, 2009.

[3] B. Tan, X. Shen, C. Zhai, "Mining Long-Term Search History to Improve Search Accuracy," in *proc. of KDD*, 2006.

[4] C. Silverstein, et al., "Analysis of a Very Large Web Search Engine Query Log," *ACM SIGIR Forum*, vol.33, Issue1, 1999.

[5] R. Krovetz, W. B. Croft, "Lexical Ambiguity and Information Retrieval," *Information Systems*, vol.10, Issue2, 1992.

[6] J. Sun, et al., "CUBESVD: a Novel Approach to Personalized Web Search," in *proc. of WWW*, 2005.

[7] D. Vallet, et. al., "Exploiting Social Tagging Profiles to Personalize Web Search," in *proc. of FQAS*, 2009.

[8] D. Carmel, et al., "Personalized Social Search Based on the User's Socizal Network," in *proc. of CIKM*, 2009.

[9] P. A. Chirita, et al., "Using ODP Metadata to Personalize Search," in *proc. of SIGIR*, 2005.

[10] K. Sugiyama, et al., "User-Oriented Adaptive Web Information Retrieval Based on Implicit Observations," in *proc. of APWeb*, 2004.