ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 41, No. 12, pp. 1117-1125, 2014. 12
http://dx.doi.org/10.5626/JOK.2014.41.12.1117

질의어의 근접성 정보 및 그래프 프로파일링 기법을 이용한 태그 기반 개인화 검색

(Exploiting Query Proximity and Graph Profiling Method for Tag-based Personalized Search in Folksonomy)

한 기 준 † 장 진 철 †† 이 문 용 †††
(Keejun Han) (Jincheul Jang) (Mun Yong Yi)

요 약 최근 폭소노미라고 불리는 데이터들이 사용자의 의도 파악 및 흥미를 분석하는 데에 매우 유용하게 쓰이고 있다. 본 논문은 폭소노미 데이터를 이용한 개인화 검색에서, 기존의 벡터 기반 프로파일링 및 유사도 계산 모델의 한계점을 지적하고, 이러한 한계를 극복하기 위한 방법으로 그래프 기반의 프로파일링 및 유사도 계산법을 제안한다. 최종적으로 그래프 기반의 개인화 검색 모델에 추가적으로 질의어간의 근접성까지 고려한 보다 발전된 개인화 검색 기법을 제안하였다. 본 연구에서는 복수의 데이터셋을 사용한 객관적인 성능 평가 실험을 통해 제안한 모델이 기존의 벡터 스페이스 모델에 기반한 프로파일링 기법 및 프로파일 간의 유사도 계산 기법보다 더 뛰어난 개인화 검색 결과를 제공함을 확인하였다. 또한 추가적인 파라미터 실험을 통하여, 제안하는 모델은 어떠한 형태의 데이터셋에도 쉽게 적용가능함을 보였다.

키워드: 협업적 태깅 시스템, 개인화 검색, 사용자 프로파일, 문서 프로파일

Abstract Folksonomy data, which is derived from social tagging systems, is a useful source for understanding a user's intention and interest. Using the folksonomy data, it is possible to create an accurate user profile which can be utilized to build a personalized search system. However there are limitations in some of the traditional methods such as Vector Space Model(VSM) for user profiling and similarity computation. This paper suggests a novel method with graph-based user and document profile which uses the proximity information of query terms to improve personalized search. We demonstrate the performance of the suggested method by comparing its performance with several state-of-the-art VSM based personalization models in two different folksonomy datasets. The results show that the proposed model constantly outperforms the other state-of-the-art personalization models. Furthermore, the parameter sensitivity results show that the proposed model is parameter-free in that it is not affected by the idiosyncratic nature of datasets.

Keywords: collaborative tagging systems, personalized search, user profile, document profile

· 본 연구는 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2011-0024560)
· 이 논문은 2014 한국컴퓨터종합학술대회에서 '질의어의 근접성 정보 및 그래프 프로파일링 기법을 이용한 태그 기반 개인화 검색'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 한국과학기술원 지식서비스공학과
keejun.han@kaist.ac.kr
†† 비 회 원 : 한국과학기술원 지식서비스공학과
jcjang@kaist.ac.kr
††† 종신회원 : 한국과학기술원 지식서비스공학과 교수(KAIST)
munyi@kaist.ac.kr
(Corresponding author임)

논문접수 : 2014년 7월 15일
(Received 15 July 2014)
논문수정 : 2014년 9월 15일
(Revised 15 September 2014)
심사완료 : 2014년 9월 16일
(Accepted 16 September 2014)

Copyright©2014 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

정보과학회논문지 제41권 제12호(2014. 12)

1. 서 론

개인화 검색은 동일한 질의어에 대해 사용자의 선호도에 따라 서로 다른 분석 결과를 제공하는 검색 기법으로 다양한 개인의 정보 요구에 맞춤화된 검색 결과를 제공하여 사용자의 검색 만족도를 높이는 기법이다[1]. 개인화 검색의 일반적인 수행 과정은 다음과 같다. (1) 사용자의 관심사를 표현하는 단어들로 구성된 사용자 프로파일을 구축하고, (2) 문서에 대해 설명하는 단어들로 구성된 문서 프로파일을 구축한 뒤, (3) 이를 바탕으로 사용자 및 문서 프로파일의 유사도를 측정하여 그 유사도가 높은 문서들의 검색 순위를 높여주는 재정렬 순서로 진행된다. 따라서, 효과적인 개인화 검색을 위해서는 각 프로파일을 이루는 단어들의 품질이 중요하며, 프로파일간의 유사도 측정 기법이 정교할수록 그 개인화 검색의 성능은 증가하게 된다.

본 논문에서는 사용자 및 문서 프로파일을 구성하는 단어 셋을 폭소노미에 의해 형성된 사용자 태그 정보를 활용하여 구축하고자 한다. 폭소노미는 다중의 사용자가 특정 콘텐츠에 협업적으로 태그를 달아 그 콘텐츠의 성격을 규명하고 분류를 돕는 행위 또는 시스템을 일컫는 말로 협업적 태깅이라고도 불린다. 이러한 폭소노미 데이터를 축적하여 활용하는 서비스에는 여러 종류가 있다. 대표적으로 사용자가 흥미 있는 주제를 몇 개의 단어로 표현하여 나타내는 Del.icio.us, 사용자가 좋아하는 노래를 태그하는 Last.fm, 사용자가 좋아하는 사진을 업로드하거나 태그로 표현할 수 있는 Flicker, 해당 영화에 대해 사용자들이 영화를 표현하는 태그를 입력할 수 있게 하는 IMDB 등의 서비스에서 이러한 데이터를 활발히 이용하고 있다. 해외는 물론 국내에서도 이러한 태그 기능을 활용하는 서비스들이 증가하고 있다. 이렇게 서비스를 이용하여 사용자가 직접 남긴 태그 데이터, 즉 폭소노미 데이터는 사용자의 의도를 명시적으로 파악할 수 있으며 문서에 달린 태그를 분석하면 해당 문서를 보다 효과적으로 설명하는 데에 사용될 수 있음이 알려져 있다[2]. 또한 이러한 폭소노미 데이터를 활용하면 기존 검색 성능의 향상뿐만이 아니라[3], 개인화 검색의 성능 향상에도 기여할 수 있다[4].

일반적인 폭소노미 데이터를 활용한 개인화 검색에서, 사용자와 문서의 프로파일은 사용자가 직접 쓴 태그와 각 태그의 가중치들의 집합으로 구성된 벡터로 표현된다. 이 때, 각 태그의 가중치들은 TF(term frequency)[5], TF-IDF(term frequency-inverse document frequency)[4], BM25[6] 등의 방법으로 계산된다. 하지만, 이러한 기술들은 프로파일을 구성하는 각 태그 사이의 근접성(proximity) 정보를 고려하지 않아 질의어에 대한 사용자의 정보 요구를 정확하게 파악하는 것에 한계를 지닌다. 대부분의 사용자 질의어들은 그 길이가 짧고, 뜻이 모호하여, 질의어 자체만이 아니라 주어진 질의어 간의 숨겨진 관계를 분석하여 정보 검색 요구에 반영하면 사용자의 검색 만족도를 향상시킬 수 있다[7].

따라서 본 논문에서 우리는 사용자와 문서의 프로파일을 기존의 벡터형태가 아닌 그래프 형태로 표현하고 이를 개인화 검색에 이용하는 방법을 제안한다. 그래프 형태의 프로파일링 기법은 기존의 기법과는 달리 단어간의 시맨틱 정보를 표현할 수 있어 각 단어간의 거리 정보를 측정할 수 있게 된다. 이는 주어진 질의어들이 해당 프로파일에서 어떠한 관계를 지니고 있는지 파악하는 단초를 제공하여 최종적으로 근접성 정보를 검색에 반영할 수 있도록 해준다.

하지만 이러한 그래프 기반의 접근법은 기존 관련 연구에서 제안된 프로파일 간의 유사도를 계산하는 방식에도 변화가 필요함을 암시한다. 기존의 관련 연구에서는 벡터 형태로 이루어진 두 프로파일 간의 유사도를 계산하기 위해 두 벡터 사이의 유사도를 측정하는 코사인 유사도[4], 스칼라 유사도[6] 등의 방법을 사용했으나, 본 연구에서 제안하는 프로파일은 벡터기반이 아닌 그래프 기반이므로, 두 그래프 간의 유사도를 측정하는 새로운 유사도 측정법이 필요하다.

따라서 우리가 제안하는 방법은 (1) 질의어와 관련된 문서들이 반환된 초기 검색 결과에서 (2) 반환된 문서의 프로파일과 해당 사용자의 프로파일이 얼마나 유사한지 그래프 기반 유사도 기법으로 계산하고, (3) 이를 최종적으로 문서 내 프로파일에서 해당 질의어들이 얼마나 긴밀한 관계를 맺고 있는지 측정한 근접성 값과 융합하여 최종적으로 개인화된 검색 결과를 사용자에게 제공한다.

본 연구는 사용자 및 문서 프로파일을 모두 그래프로 나타내며, 특히 문서 내에 위치한 질의어의 근접성 정보를 검색 결과의 성능 향상에 활용하는 것으로, 폭소노미 데이터를 활용하여 태그 기반의 개인화 검색에서 처음으로 제안하는 그래프 기반의 개인화 검색 방법으로 기존 연구와 비교하여 독창성이 있다. 본 연구에서 제안한 기법은 기존의 개인화 방법들과 비교하였을 때 월등한 성능 향상을 보였으며, 또한 질의어간의 근접성 정보를 활용하여 제안하는 방법에 결합하면 그 개인화 검색 성능이 일관되게 진보하는 것을 복수개의 폭소노미 데이터 셋 상에서 검증하였다.

2. 관련 연구

본 장에서는, 본 연구와 관련된 여러 관련 문헌들을 1) 개인화 검색 2) 폭소노미를 이용한 개인화 검색으로 분류하여 소개한다. 이어지는 본 장의 세부 부분은 해당

분류에 따른 관련 문헌들을 소개한다.

2.1 개인화 검색

현재 개인화 검색은 크게 두 개의 연구 방향으로 나누어져 있는데, 첫 번째는 질의어 확장(Query Expansion) 방법이며 두 번째는 검색결과 재정렬(Result Reranking)이다[1]. 본 논문에서는 검색결과 재정렬 개인화 방식을 활용하여 연구를 진행하였다.

질의어 확장은 사용자에 의해 입력된 초기 질의어를 수정하거나 검색 결과 향상을 위해 새로운 질의어를 시스템에서 추가하는 방법이다. 또한, 사용자의 특성을 추가적으로 반영하기 위해 초기 질의어의 가중치를 조정하는 방식도 크게 질의어 확장에 해당한다고 볼 수 있다[8].

그에 반해 검색결과 재정렬은 질의어와 문서와의 적합성(Relevance)만이 반영된 초기의 결과가 주어졌을 때, 각각의 사용자 선호도를 반영하기 위해 초기 검색 결과를 재정렬하는 방법이다. 대부분의 개인화 재정렬 방법들은 다양한 사용자 데이터로부터 개인의 선호도를 추출하여 이를 바탕으로 사용자 프로파일을 구축하고, 해당 프로파일에 적합한 문서들을 상위에 올리고, 해당 프로파일과 관련이 없는 문서를 초기 검색 결과에서 제외하는 과정으로 개인화 검색을 진행한다. 관련문헌 [9]에서는 ODP(Open Directory Project) 구조를 이용하여 사용자와 문서를 벡터로 표현하여 두 벡터간의 거리를 측정하여 개인화 검색을 수행하는 방식을 제안하였고, 관련문헌 [10]에서는 통계적 언어 모델을 사용하기도 하였다.

또한, 관련문헌 [11]에서는 해당 사용자의 과거 데이터만이 아니라 사용자와 유사한 행태를 가지는 사용자들의 데이터까지 추가적으로 고려하기 위해 추천 시스템에서 주로 사용되는 협업 필터링(Collaborative Filtering)을 활용하여 사용자 프로파일을 구축하는 방식이 제안되기도 하였으며, 관련문헌 [12]에서는 이차 행렬을 특정한 구조로 분해하는 기존의 특이값 분해 방법(Singular Vector Decomposition 혹은 SVD)을 사용자, 질의어, 문서의 세가지 특성을 반영할 수 있는 삼차 행렬에서 사용할 수 있게 확장한 CubeSVD 방식을 도입하여 개인화 검색의 성능을 향상시켰다. 관련문헌 [13]에서는 이러한 다양한 검색 기법들을 실제 사용자들이 축적한 대용량 로그 데이터를 활용하여 검증함으로써, 이러한 검색 기법들이 실제 서비스에서도 사용자의 검색 만족도 향상에 기여할 수 있음을 보였다.

2.2 폭소노미를 활용한 개인화 검색

태그 기능을 제공하는 서비스들이 최근 증가함에 따라, 이러한 폭소노미 데이터를 개인화 검색에 활용하는 방식이 최근 증가하였다. 관련문헌 [5]는 사용자와 문서의 태그들 중에서 가장 많이 사용된 태그 정보(TF)를 활용하여 사용자와 문서의 벡터 프로파일을 구축하고, 이를 바탕으로 초기의 검색 결과를 재정렬 하는 방식을 제안하였다.

이에 반하여, 관련문헌 [4]에서는 사용자 프로파일을 구성하는 각 태그의 가중치를 TF-IDF에 기반하여 제안된 term frequency-inverse user frequency (TF-IUF) 방법을 통하여 다음과 같이 계산하였다.

$$w = \mathrm{tf} * \mathrm{idf} = \mathrm{tf} * \log\frac{N}{n_t} \tag{1}$$

식 (1)에서 tf는 사용자가 해당 태그를 사용한 횟수이며, N은 총 사용자의 수, n_t는 해당 태그를 사용한 사용자의 수이다. 이와 유사하게 문서 프로파일도 TF-IDF를 활용하여 term frequency-inverse resource frequency (TF-IRF)로 나타낼 수 있다.

또한, 관련문헌 [6]에서는 기존의 BM25 랭킹 모델을 개인화 검색에 활용하기 위해 사용자 프로파일 내 각 태그의 가중치를 다음과 같이 계산하였다.

$$\begin{aligned} w &= bm25(t_x) \\ &= \mathrm{iuf}(t_x) \cdot \frac{tf \cdot (k_1 + 1)}{tf + k_1 \cdot (1 - b + b \cdot \frac{|u_l|}{avgu_l})} \end{aligned} \tag{2}$$

식 (2)에서 k_1과 b는 2와 0.75로 정의된 파라미터 값이며, $|u_l|$은 사용자 프로파일 벡터의 길이, $avgu_l$은 전체 사용자 프로파일 벡터 길이의 평균을 의미한다. 이와 유사하게 문서 프로파일도 BM25를 이용하여 표현할 수 있다.

관련문헌 [6]은 벡터 스페이스 기반의 TF-IDF와 통계적 기법인 BM25를 동시에 활용하여 개인화 검색의 성능을 보다 더 향상시키기 위한 하이브리드 기법을 제안하였다.

최근에는 기존의 제안된 TF, TF-IDF, BM25보다 더 발전된 모델인 Normalized Term Frequency (NTF) 모델이 관련문헌 [14]에서 제안되었다. NTF 모델에서 사용자 프로파일 내 태그 t_x의 가중치는 다음과 같이 계산된다.

$$w(t_x) = \frac{N_{c,x}}{N_c} \tag{3}$$

식 (3)에서 N_c는 사용자 c에 의하여 태그된 문서들의 개수이고, $N_{c,x}$는 사용자 c가 태그 t_x를 활용한 횟수를 의미한다. 이와 유사하게 문서의 프로파일도 NTF를 활용하여 계산된다.

이와 같은 기법으로 계산된 사용자와 문서의 벡터 프로파일은 코사인 유사도[4]나, 코사인 유사도에 기반을 둔 여러 벡터 기반 유사도 계산 기법을 활용하여 [6,14], 두 프로파일이 어느 정도 유사한지 계산하게 된다. 만약 두 프로파일간의 유사도가 높다면 해당 문서는 사용자가 주로 관심 있는 주제들과 유사한 문서일 확률이 높

을 것이며, 반대로 두 프로파일 간 유사도가 낮다면 해당 문서는 사용자의 질의어와 관련이 높은 문서라고 할지라도, 사용자가 관심 있는 주제와는 관련성이 상대적으로 떨어질 수 있음을 의미한다.

관련연구 [15]은 그래프기반의 유사도 측정방식을 개인화 검색에 활용한 사례이다. 본 연구에서는 여러 사용자에 의해 협업적으로 구축된 폭소노미 데이터와 질의어의 근접성 정보를 추가로 활용하여 그래프 기반 개인화 검색의 성능을 향상시키고자 하였다.

3. 연구 동기

본 장에서는 기존 폭소노미 데이터를 활용한 개인화 검색 기법의 한계점을 분석함과 동시에 이를 바탕으로 본 연구에서 제안하는 기법과 기존 기법과의 차이점을 기술한다.

3.1 사용자 및 문서 프로파일 구축

상기에서 언급한 바와 같이, 기존 폭소노미 기반의 개인화 검색에서 자주 사용되는 기법인 TF, TF-IDF, BM25등은 사용자가 남긴 태그와 그 태그의 가중치만을 고려하여 사용자 프로파일을 구축한다. 또한, 여러 사용자가 리소스에 남긴 태그들과, 그 태그들의 가중치를 고려하여 리소스 프로파일을 구축한다. 그러나 이러한 방법은 실제 사용자의 정보 검색 요구와 맞지 않는 경우가 존재한다. 예를 들어, 영화 검색 및 태깅 서비스에서 제공되는 영화 중에는 미래를 그린 공상과학 영화이지만 가족간의 사랑의 의미도 일정 부분 담겨있는 영화인 'A'가 있다고 가정해볼 수 있다. 해당 영화를 시청한 사용자들이 남긴 태그들에 기반하여 기존 검색 방식 중에서 TF 기반의 프로파일 방식으로 표현된 영화 A의 프로파일은 다음과 같다.

$$\overrightarrow{R_A} = \{(\text{미래}, 56), (\text{과학}, 36), \cdots, (\text{가족}, 18), (\text{사랑}, 3)\}$$

기존의 방식은 해당 영화에 대한 상대적인 주제를 비교적 잘 설명하고 있기는 하지만, 각 주제들간의 관계에 대해서는 명확하게 설명하고 있지 않다. 이는 A를 검색하기 위해 '미래, 사랑'을 질의어로 입력할 경우 해당 영화 프로파일 중 사랑의 가중치가 크지 않기 때문에 최종 검색 결과에서 A가 상위에 검색될 확률을 감소시키는 요인이 된다. 이는 유사한 기법인 TF-IDF, NTF 등에도 적용되며, BM25 역시 TF와 IDF값을 변수로 포함하기에 동일한 한계를 지닌다.

이에 반하여, 본 연구에서 제안하는 방법은 기존 기술과는 달리 사용자 및 리소스를 태그 기반의 그래프로 표현한다. 이를 통하여 표현된 그래프 기반의 영화 A의 프로파일을 이용하면, '사랑'의 가중치가 적더라도, 또 다른 질의어인 '미래'와 어느 정도의 근접성을 가지고 있는지 파악할 수 있어, 이 정보를 통하여 해당 영화를 검색 결과에서 상위로 올리는 재정렬 과정이 가능해진다. 실제로 이러한 질의어의 근접성을 이용하여 검색 프로세스에 반영하게 되면 전체적인 검색 성능 향상에 도움을 주는 것으로 관련 문헌[7,16]에 나타나 있다. 사용자 및 리소스에서 그래프 기반의 프로파일을 추출하는 방법 및 근접성 정보를 고려한 재정렬 과정은 다음 장에서 자세히 설명한다.

3.2 사용자 및 문서 프로파일간 유사도 계산

기존의 개인화 검색방법에서, 개인화는 사용자 프로파일과 리소스 프로파일 간의 유사도를 측정하여 이루어진다. 두 개의 프로파일이 벡터 형식으로 이루어져 있기 때문에, 이들간의 유사도는 주로 코사인 유사도[4], 스칼라 유사도[6] 등의 방법을 통하여 측정되었다. 이외에도 폭소노미 데이터의 특성을 반영한 여러 측정법들이 제안되었으나[14], 이는 모두 프로파일의 형식을 벡터 기반으로 표현하는 것을 가정하여 제안되어 그래프 기반 프로파일의 유사도를 비교하는데 사용하는 것은 한계를 지닌다.

따라서, 개인화를 위해 두 프로파일간의 유사도를 계산할 때 상기 설명된 기존의 유사도 기법들을 사용하는 것은 그래프 상에서 주어진 단어간의 시맨틱 정보 및 그래프의 구조 활용이 불가능하고, 결국 태그의 가중치 정보만을 이용하므로, 그래프 기반의 프로파일을 활용하는 방식에는 적절하지 않다. 이에 우리는 그래프 기반의 프로파일에 특성화된 유사도 기법을 활용하여 사용자와 문서간의 관련 정도를 측정하고자 한다.

4. 제안하는 방법

본 장에서 제안하는 방법은 크게 다음과 같은 단계로 이루어진다. 1) 사용자의 질의어와 연관된 리소스들의 초기 검색 결과를 반환한다. 2) 사용자가 남긴 태그 정보를 바탕으로 사용자의 프로파일과 초기 검색 결과 내 리소스의 프로파일을 그래프 기반으로 구축한다. 3) 초기 검색 결과 내 리소스들의 프로파일과 사용자의 프로파일간의 유사도를 계산하고 리소스 내 질의어간의 근접성 정보와 융합하여 최종 개인화 검색 결과를 반환한다. 이어지는 본 장의 세부 부분에서는 각 과정에 대하여 자세히 설명한다.

4.1 초기 검색 결과 획득

개인화 검색을 위해선 먼저, 주어진 질의어 $q_1, q_2, \cdots, q_n$의 집합 Q와 어느 정도 연관이 있는 리소스들을 찾아 초기 검색 결과를 얻어내야 한다. 만약 이러한 과정을 거치지 않고 바로 사용자와 문서 간의 선호도만을 고려하여 최종 검색 결과를 제공한다면 사용자의 질의어와 관련이 없는 문서도 단순히 사용자가 선호할 만한 주제라는 이유로 검색 결과에 상위에 반영되는 문제가 발생하게 된다.

본 논문에서는, 기존의 여러 검색 방법 중에서도 BM25 모델을 질의어와 문서간의 관련성을 측정하여 초기 결과를 반영하는 방법으로 사용하였다. 이는 관련 문헌 [7]에서 BM25 모델이 근접성을 고려하는 모델의 베이스라인으로 사용하였기 때문에, 관련 연구와의 비슷한 실험 조건을 형성하기 위하여 해당 기법을 선정하였다.

주어진 문서 r과 주어진 질의어 집합 Q에 관련 정도를 나타내는 점수 NS(Non-personalized Score)를 구하는 식은 다음과 같다.

$$NS(Q,r) = bm25(Q,r) = \sum idf(q_i) * \frac{f(q_i,r) * (k_1+1)}{f(q_i,r) + k_1 * (1 - b + b \cdot \frac{|r|}{avgdl})} \quad (4)$$

식 (4)에서 k_1과 b는 2와 0.75로 정의된 파라미터 값이며 $|r|$은 문서 프로파일 벡터의 길이, $avgdl$은 전체 문서 프로파일 벡터 길이의 평균을 의미한다.

최종적으로 전체 문서들 중 질의어 집합 Q에 관한 NS점수가 높은 문서부터 내림차순으로 초기 검색 결과 상위에 반환되게 된다. 이때, 반환된 초기의 검색 결과는 사용자의 질의어 집합 Q문서들간의 적합도만을 고려하며, 사용자의 선호도는 해당 초기 문서 랭킹에 반영되어 있지 않다. 따라서, 이 초기 검색 결과를 재정렬하여 최종 개인화가 반영된 검색 결과를 사용자에게 제공하게 된다. 본 연구에는 상위에 위치한 50개의 문서를 대상으로 재정렬을 실시하였다.

4.2 사용자 및 문서 프로파일의 구축

본 연구에서 구축하려고 하는 사용자 및 문서 프로파일은 그래프 형태이다. 예를 들어, 어떤 사용자 u의 프로파일은 해당 사용자가 남긴 태그를 그래프의 노드 집합(Vertex)으로 표현하고, 각 태그 사이의 관계는 링크 집합(Edge)으로 표현하여 $G_u = (V, E)$로 나타낸다. 이 그래프 G_u는 방향성이 없고 각 링크는 해당 링크에 연결된 두 노드 tag_i와 tag_j가 서로 얼마나 가까운지를 나타내는 가중치 값 $w_{i,j}$를 갖는 그래프이다. 가중치 값 $w_{i,j}$는 다음과 같이 정의된다.

$$w_{i,j} = \text{co} - \text{occur}_u(tag_i, tag_j) \quad (5)$$

식 (5)에서 이때 co-occur$_u$는 태그 tag_i와 tag_j가 사용자 u에 의해 함께 사용된 횟수로, 그 횟수가 클수록 사용자 u는 두 태그의 관련성을 높게 정의함을 의미한다.

또한 문서 r의 그래프 기반 프로파일 $G_r = (V, E)$에서 V는 리소스 r에 사용자들이 남긴 태그들을 의미하는 노드들의 집합을 의미하며, E는 V에 속하는 점들을 연결하는 링크로, V에 속하는 태그 tag_x와 tag_y의 가중치 값 $w_{x,y}$는 다음과 같이 정의된다.

$$w_{x,y} = \text{co} - \text{occur}_r(tag_x, tag_y) \quad (6)$$

이때의 co-occur$_r$은 태그 tag_i와 tag_j를 리소스 r에 동시에 사용한 사용자의 총 수로 그 수가 클수록 두 태그의 관련성이 높은 것을 의미한다. 구축된 그래프 G_u와 G_r은 계산 편의를 위하여 $|V| \times |V|$ 차원을 갖는 정사각형행렬 형태로 변환되어 저장된다.

4.3 사용자 및 문서 프로파일간 유사도 계산

본 논문에서는 사용자 프로파일 G_u와 리소스 프로파일 G_r간의 유사도 값인 D(Distance)를 구하기 위해 Closeness[17], Maximum Common Subgraph(MCS) [18], Edit Distance[15] 기법들을 활용한다.

Closeness는 두 그래프 간의 형태적 유사성을 집합론적인 관점에서 측정하는 방법으로, 두 그래프간에 공통인 노드가 많을수록 두 그래프간의 유사도는 1에 가까운 값을 갖게 되며, 반대의 경우는 0에 가까운 값을 갖게 된다. 두 그래프 G_u와 G_r의 유사도 값 D를 Closeness로 구하는 식은 다음과 같다.

$$D(G_u, G_r) = D_{Closeness}(G_u, G_r) = \frac{1}{n}\sum_{v \in V} \frac{|G_u \cap G_r|}{|G_u \cup G_r|} \quad (7)$$

식 (7)에서, $|G_u \cup G_r|$은 두 그래프 G_u, G_r에 공통되게 속한 노드 v와 연결된 이웃 노드들 (neighborhood)의 합집합이며, $|G_u \cap G_r|$은 이웃 노드들의 교집합, n은 $|G_u \cup G_r|$의 크기를 의미한다.

MCS는 주어진 두 그래프에 공통인 부분 그래프 중 가장 많은 노드의 개수를 지닌 부분 그래프를 의미한다. 두 그래프 G_u와 G_r의 유사도 값 D를 MCS로 구하는 식은 다음과 같다.

$$D(G_u, G_r) = D_{MCS}(G_u, G_r) = 1 - \frac{|MCS(G_u, G_r)|}{\max(|G_u|, |G_r|)} \quad (8)$$

식 (8)에서, $|G_u|$와 $|G_r|$은 그래프 G_u와 G_r의 노드의 개수이고, $|\text{MCS}(G_u, G_r)|$는 G_u와 G_r의 부분 그래프 중 가장 큰 그래프의 노드의 개수, $\max(|G_u|, |G_r|)$는 두 그래프 G_u와 G_r 중 노드의 개수가 더 많은 그래프를 의미한다.

그러나 상기의 두 알고리즘은 두 그래프간의 노드 간 유사도만을 고려하며 링크 정보를 고려하지 못한다는 한계를 지니고 있다. 이를 고려하여 제안하는 마지막 알고리즘은 노드 및 노드를 연결하는 링크들의 연결 정보까지 모두 고려하는 Edit Distance이다. Edit Distance는 한 그래프가 다른 그래프로 변형될 때까지 필요한 노드 및 링크의 삭제, 추가 및 교환 등의 작업의 최소 개수를 의미한다. 두 그래프 G_u와 G_r의 유사도 값 D를 Edit distance로 구하는 식은 다음과 같다.

$$D(G_u, G_r) = D_{Edit_distance}(G_u, G_r) = \min(C(\varepsilon)) \quad (9)$$

식 (9)에서, $C(\varepsilon)$는 그래프 G_u가 G_r로 (혹은 G_r에서 G_u로) 변형되는데 필요한 작업의 개수를 의미한다.

두 그래프 G_u와 G_r의 유사도 값 D는 식 (7)에서 (9)

중 선택된 그래프 유사도 계산법에 따라 각각 다른 값을 갖게 되며 이는 본 논문에서 제안하는 방법의 최종 성능에 크게 영향을 미친다. 각 식에 따른 실험 결과 변화 분석은 다음 장에 언급되어 있다.

4.4 근접성 정보를 반영한 최종 개인화 검색 결과

본 논문에서 제안하는 알고리즘은 어떤 문서 r이 사용자 u의 관심과 얼마나 일치하는지를 4.3장에 소개한 방법으로 측정함에 동시에, 사용자 u의 질의어 $q_1, q_2, \cdots, q_n$의 집합 Q가 문서 r 속에서 얼마나 긴밀한 관계를 갖는지를 측정하여 개인화 검색 결과에 반영한다. 질의어 집합 Q에 속하는 질의어 $q_1, q_2, \cdots, q_n$들의 문서 프로파일 G_r 내에서의 근접성 값 $P(Q, G_r)$은 다음과 같다. 질의어 $q_1, q_2, \cdots, q_n$의 집합 Q들의 리소스 프로파일 G_r 속에서의 근접성 값 $P(Q, G_r)$은 다음과 같다.

$$P(Q, G_r) = \exp\left(-DS(Q, G_r) \times \alpha\right) \quad (10)$$

$$DS(Q, G_r) = avg\left(\frac{Distance(q_1, q_2)}{maxDistance(G_r)}\right) \times n \quad (11)$$

식 (11)에서, $Distance(q_1, q_2)$는 질의어 q_1과 q_2 간의 최단거리이며, n은 질의어의 개수, $maxDistance(G_r)$는 리소스 프로파일 G_r 내에서의 가장 먼 거리로 정규화(Normalization)를 위한 값이다.

관련문헌 [7]에 따르면, 어떤 두 노드 간의 근접성 정보는 다음과 같은 두 개의 특성을 지닌다. 첫째, 그래프 상에서 노드 간의 거리가 작을수록 해당 노드 간의 근접성 정보는 급격히 커진다. 둘째는 단어들 간의 거리가 멀더라도 여전히 어느 정도 느슨한 관계를 갖게 되므로, 근접성 점수가 급격하게 낮아지지 않는다. 이러한 두 가지 조건을 만족하는 두 노드 간의 거리 정보와 근접성 정보의 관계는 볼록 곡선(Convex curve)의 형태로 표현할 수 있다.

따라서, 이러한 볼록 곡선의 형태를 갖는 근접성 점수를 얻기 위해 $DS(Q, G_r)$점수를 음수로 변환하고, 실제 단어간의 거리의 특성을 잘 반영하기 위해 식 (10)과 같이 Convex 함수로 변환한다. 이때 볼록도의 정도를 조절하는 α는 실험을 통하여 0.6으로 고정되었다(5.3장 참조).

최종적으로, 우리는 사용자와 문서 간의 유사도와 각 문서에서 주어진 질의어들이 갖는 근접성 정보를 다음과 같이 융합하여 최종 개인화 점수 PS(Personalized Score)를 제안하고, 이 점수가 높은 리소스부터 최종 개인화 검색 결과 상위에 반환된다.

$$PS(G_u, G_r, Q) = D(G_u, G_r) + \beta * P(Q, G_r) \quad (12)$$

식 (12)에서, 근접성 정보의 반영도를 조절하는 β는 0과 1 사이의 값으로 β값이 0인 경우 질의어의 근접성 정보는 최종 개인화 검색 점수에 반영되지 않으며, 반대로 1에 가까워질수록 최종 개인화 검색에 질의어의 근접성 점수가 영향을 미치는 정도가 증가한다. 본 논문에서는 β값이 0.4인 경우 가장 효과적인 성능을 보임을 실험을 통해 알아내었다 (5.3장 및 그림 2 의 (b) 참조).

5. 실 험

본 장에서는 제안하는 방법의 성능 평가를 위한 성능 평가 프레임워크를 제안하고, 실험을 위해 사용한 데이터 및 실험 환경을 소개한 후 최종적으로 본 실험의 결과를 설명한다.

5.1 성능 평가 프레임워크

개인화 검색의 성능을 평가하는 것은 (1) 실제 시스템을 이용할 사용자를 고용하여 평가하는 방식과 (2) 검색 시스템의 로그 데이터를 분석하는 방법, 크게 2가지로 분류할 수 있다. 첫 번째 방법의 경우 객관적인 성능 평가를 위해 많은 사용자를 고용하고, 오랫동안 시스템을 사용하게 하는 비용의 문제가 크게 발생하여, 본 논문에서는 실제 이용 가능한 데이터들을 분석하여 실험에 이용하는 방법을 채택하였다.

그러나, 온라인 상에서 접근가능한 폭소노미 데이터를 개인화 검색의 성능 평가를 위해 바로 사용하는 것은 여러 가지 어려움이 따른다. 그 중에서도 가장 큰 어려움은, 그러한 폭소노미 데이터 셋의 경우 질의어에 대해 어떠한 문서의 적합성을 표현하는 데이터를 획득하기 어렵다는 것에 있다. 로그 데이터를 분석하여 검색 성능을 평가하는 경우, 일반적으로 연구자들은 주어진 질의어에 의해 반환된 검색 결과에서 클릭된 문서들은 그 질의어의 관련 있는 정답 셋이라고 가정한다. 이와 비슷하게, 폭소노미 데이터 셋에서 사용자가 태그를 단 문서는 해당 사용자가 찾고자 하는 문서라고 볼 수 있다. 예를 들어, 사용자가 어떠한 문서에 '개인화'라는 태그를 달았다면, 우리는 그 문서를 해당 사용자가 '개인화'를 질의어로 입력했을 때 찾기 원하는 문서로 가정할 수 있다. 따라서, 만약 해당 문서의 순위가 개인화되지 않은 검색 결과보다 개인화된 검색 결과에서 더 높다면 우리는 제안하는 방법의 검색 성능이 잘 작동한다고 정의할 수 있다. 실제로 이러한 성능 평가 프레임워크는 객관적인 개인화 성능 평가를 위해 여러 폭소노미 기반의 개인화 검색 관련 문헌 [4,6,14]에서 사용되었다. 본 논문에서는 해당 사용자가 리소스에 입력한 태그 중 상위 3개를 질의어로 선정하여 이를 초기 검색 결과를 얻는데 사용하였다. 이를 위하여, 90%의 태그정보가 사용자 및 리소스 프로파일을 구축하는데 사용되었고, 10%의 태그정보가 실험용으로 사용되었다.

5.2 실험 환경

우리는 여러 폭소노미 데이터 셋들 중 데이터의 밀도

표 1 실험 데이터 셋 소개
Table 1 Description of Datasets Used in Experiment

Data Set	# of Users	# of Resources	# of tags	Avg. # of tags in user profiles	Avg. # of tags in resource profiles
CiteULike	109	44.04	160.37	64.82	2.87
MovieLens	71.567	10.68	95.58	10.62	9.06

가 그래프 기반의 알고리즘에 가장 큰 영향을 미침을 고려하여[19], 데이터 밀집도가 상대적으로 낮은 CiteULike 북마크 데이터 셋(리소스당 평균 태그 수 2.87)과 상대적으로 높은 편인 MovieLens 영화 데이터 셋(리소스당 평균 태그 수 9.06)을 실험 데이터 셋으로 선정하였다. CiteULike은 사용자들이 관심있는 논문에 대해 직접 태그를 달아 활용할 수 있게 한 서비스이며 MovieLens는 영화에 대해 사용자들이 태그를 직접 작성할 수 있게 한 서비스로 해당 서비스에서 배포된 데이터셋들은 관련 연구 [14,20]에서 실험 데이터로 사용되었다. 다음 표 1은 선정된 두 데이터 셋의 속성을 소개하는 표이다.

기존 기법과의 평가를 위하여 우리는 비교 대상으로 1) BM25로 얻은 초기 검색 결과(baseline) 및 2) BM25 및 코사인 유사도로 개인화된 검색 결과 (BM25), 3) 근접성 정보가 고려되지 않은 그래프 기반의 검색 기법(공식 (6)에서 β=0 인 경우, 이하 graph only)들을 우리가 제안하는 4) 그래프 기반의 개인화 검색 및 근접성 정보를 모두 고려한 알고리즘(graph+proximity)과 비교하였다. 2, 3번 및 최종 우리가 제안하는 방법의 경우, 1번 베이스라인 방법을 통해 질의어에 적합한 리소스 최대 50개를 초기 검색 결과로 반환한 결과를 바탕으로 각각의 알고리즘을 적용한 재정렬 과정을 통해 최종 검색 결과를 반환하였다.

개인화 검색의 성능 평가 척도로는 주로 MRR, MAP, NDCG, Precision @N 등이 사용된다. 하지만 본 연구에서 사용하는 프레임워크의 경우 주어진 질의어 셋에 적합한 문서는 전체 문서 중 하나이므로, 복수개의 적합 문서가 있는 실험 환경에 주로 사용되는 NDCG와 MAP는 사용할 수 없다. 따라서, 우리는 첫 번째 평가 척도로 MRR을 사용한다. MRR은 제안하는 방법이 얼마나 빠르게 사용자가 주어진 질의에 적합한 문서에 도달할 수 있는지를 측정하는 척도이다. MRR은 다음과 같이 계산된다.

$$MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{rank(R_{q_i})} \tag{12}$$

식 (12)에서, n은 평가에 사용된 질의어 집합의 개수이며, q_i 는 쿼리, $rank(R_{q_i})$ 는 해당 질의어에 적합한 문서의 최종 검색 결과 내 순위이다.

또한 제안하는 기법의 정확성을 측정하기 위한 척도로 우리는 P@N을 사용하였다.

$$P@N(q_i) = \begin{cases} 1, if\ rank_i \le N \\ 0, if\ rank_i > N \end{cases} \tag{13}$$

$$P@N = \sum_{I=1}^{n}\frac{P@N(q_i)}{n} \tag{14}$$

식 (13)에서, N은 최종 검색 결과 내 상위 N개의 문서이고, 이를 바탕으로, 평균 $P@N$ 은 식 (14)와 같이 얻어진다. 본 논문에서는 상위 5, 10, 20 개의 문서에서의 $P@N$ 을 측정하였다(즉, N=5, 10, 20).

5.3 실험 결과 및 분석

먼저, 3.3에서 제안한 프로파일간의 유사도를 계산하는 여러 방법 중에 제안하는 알고리즘에 가장 적합한 방법을 찾아내기 위한 실험을 실시하여 표 2와 같은 MRR 척도 결과를 얻었다. 이때 식 (10)과 식 (12)의 α 및 β 값은 파라미터 민감도 실험을 통해 각각 0.6과 0.4로 고정되었다(그림 2 참조). 기존의 개인화 검색 방법인 BM25와 비교했을 때, Edit Distance (표 2의 ED) 기반 유사도 계산법은 사용된 데이터 셋에 관계없이 통계적으로 유의한 우수한 성능을 보였다(Wilcoxon test, $p < 0.05$). 그러나 그 외에 Closeness와 MCS 그래프 유사도 계산법은 오히려 CiteULike에서는 기존의 방법보다 더 낮은 성능을 보였고, MovieLens에서는 성능 향상의 정도가 통계적으로 유의하지 않은 것으로 나타났다.

이와 같은 결과를 이해하기 위해선, Closeness와 MCS, 그리고 Edit Distance의 그래프 유사도를 측정하는 차이를 이해해야 한다. Closeness와 MCS의 경우, 그 세부적인 알고리즘은 차이가 있으나 본질적으로 두 개의 그래프가 유사한지를 판단할 때에는 노드 간의 구조적 유사도 만을 주로 고려한다. 이는 두 개의 그래프가 동일한 노드를 많이 가지고 있다면 그 정확도가 향상된다는 장점이 있으나, 본 연구에서 사용하는 두 개의 그래프는 각각 사용자 데이터에서 추출된 사용자 그래프와,

표 2 두 프로파일 간 유사도 계산법에 따른 성능 차이
Table 2 Performance of Different Similarity Schemes to Measure Two Profiles

Data Set	Non-graph based		Graph-based		
	Baseline	BM25	Closeness	MCS	**ED**
CiteULike	0.355	0.456	0.422	0.400	**0.478**
MovieLens	0.288	0.364	0.370	0.378	**0.392**

문서에 달린 태그로부터 추출된 문서 그래프로 동일한 노드 집합을 갖게 될 확률이 상대적으로 적어, 노드 간의 유사성만을 고려하는 기법은 개인화 검색 성능의 향상을 크게 기대할 수 없는 것으로 보인다. 게다가, 두 그래프의 공통된 부분인 서브 그래프의 유사도를 비교하는 MCS와 달리, Closeness의 경우 개념적으로 두 개의 그래프가 서로 동일한 노드를 지니고 있을 때에만 확실한 유사도 비교가 가능하여, MCS보다 데이터 특성에 훨씬 의존적인 경향을 보인다.

이에 반하여, Edit Distance는 그 알고리즘의 간편함에 비하여 그래프의 구조적 특성 및 각 링크들의 가중치까지 모두 고려할 수 있게 되어 데이터 셋에 관계없이 실제적인 그래프 기반의 개인화 검색 성능 향상을 보이는 것으로 나타난다. 이에 우리는 각 데이터 셋에서 균일한 향상 정도를 보이는 Edit Distance를 본 그래프 기반의 프로파일의 유사도를 계산하는 방법으로 최종 채택하였다.

이를 바탕으로, 그림 1의 (a)와 (b)는 각각 CiteULike와 MovieLens 데이터 셋에 대한 최종 개인화 성능을 나타내는 MRR 척도의 결과이고, (c)와 (d)는 정확성 척도의 결과이다. 이를 토대로 개인화 검색에 관한 다음과 같은 분석 결과를 도출할 수 있다.

- 두 데이터 셋에서 모두 개인화 검색을 적용하는 것은 일반적인 검색 결과와 비교했을 때, 큰 성능의 향상을 보였다. 이는 태그 정보를 활용하여 개인화 검색에 적용하는 것을 통해 사용자의 검색 만족도를 향상시킬 수 있음을 의미한다.
- 기존에 폭소노미 기반의 개인화 검색 기법에서 주로 사용된 벡터 기반의 프로파일링 모델 및 유사도 계산 기법과 본 논문에서 제안하는 그래프 기반의 프로파일링 모델 및 유사도 계산 기법을 비교하였을 때, 우리가 제안한 그래프 기반의 개인화 모델이 기존의 모델보다 더 좋은 성능을 보여주었다. 또한 비교적 데이터가 균일하지 않아 프로파일 구축이 쉽지 않은 경우에도, 제안하는 모델이 더 좋은 성능을 보였다.
- 마지막으로, 그래프 기반의 개인화 모델과 더불어 질의어의 근접성 정보를 모두 고려하는 우리의 방식이 데이터 셋에 관계없이 두 개의 평가 척도 모두에서 비교된 대상들에 비해 더 뛰어난 성능을 가진 것을 보여주고 있다. 즉, 우리가 제안하는 방법은 다른 기법들과 비교하여 MRR기준으로 비교 대상에 비하여 평균적으로 약 7~43%의 유의한 향상(Wilcoxon test, $p < 0.05$)을 보였고, 정확성 기준으로는 평균적으로 약 10~58%의 유의한 향상(Wilcoxon test, $p < 0.05$)을 보였다.

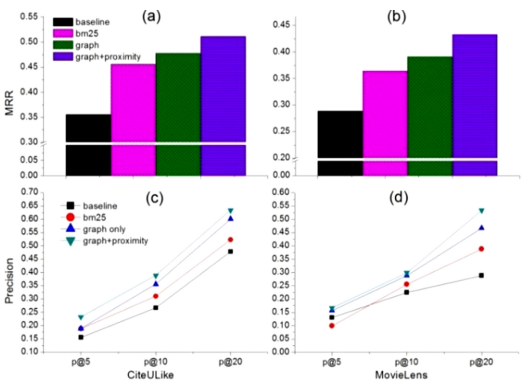
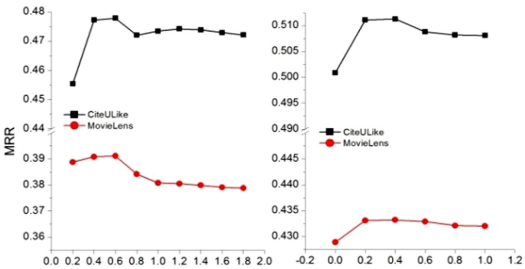


그림 1 개인화 모델에 따른 MRR 및 정확성 변화
Fig. 1 MRR and Precision Performance of Different Personalization Models



(a) parameter $\alpha(\beta = 0)$ (b) parameter $\beta(\alpha = .6)$
그림 2 파라미터 변화에 따른 MRR 척도의 변화
Fig. 2 MRR Performance by Varying Parameters

최종적으로, 그림 2는 우리가 제안하는 방법에서 사용되는 파라미터들의 값에 따라 변하는 MRR을 기록한 것이다. β값을 0으로 고정하고 α에 대한 민감도 실험을 수행했을 때, 두 데이터 셋 모두에서 α가 0.6정도일 때 가장 최적의 결과를 얻을 수 있었으며, 이를 토대로 α를 0.6으로 고정한 후 수행한 β에 관한 민감도 실험에서는 0.4정도에서 최적 값에 수렴함을 확인할 수 있었다. 이러한 결과는 제안하는 모델이 데이터 셋에 관계없이 항상 일정한 값에서 최적의 성능 향상을 보일 수 있다는 것을 의미하여, 이를 토대로 제안하는 모델은 데이터 셋에 독립적이며 파라미터에도 독립적인 알고리즘으로 향후 실제 서비스에 쉽게 적용 가능함을 의미한다.

6. 결론 및 향후연구

본 논문은 단어간의 근접성을 측정할 수 있는 그래프 기반의 프로파일링 기법 및 유사도 측정법을 기반으로 질의어간의 근접성까지 고려한 보다 발전된 태그 기반의 개인화 검색 기법을 제안하였다. 객관화된 개인화 성능 평가 실험을 통해 우리가 제안한 모델이 기존의 벡터 기반의 모델에 비해 개인화 성능을 향상시킨 것을 확인하였다.

또한, 그래프 기반 프로파일링 기법은 데이터의 크기가 커질 경우 복잡도가 증가할 수 있으나, 향후 연구에서 제안하는 모델을 실제 서비스 환경에 적용하고 이를 통하여 제안하는 모델의 확장성 및 효용성을 검증하는 것이 필요하다. 한편, 본 연구에서는 여러 그래프 기반의 유사도 기법 중에서 Closeness, MCS, Edit distance만을 사용했으나, 향후 연구에서는 보다 다양한 그래프 기반의 유사도 기법을 비교하며 제안하는 모델의 성능을 보다 향상시킬 수 있는 유사도 기법을 찾으려는 노력 역시 지속되어야 할 것이다.

References

[1] J. Pitcow, H. Schutze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel, "Personalized Search," *Communications of the ACM*, Vol. 45, No. 9, pp. 50-55, 2002.

[2] C. M. A. Yeung, N. Gibbins, and N. Shadbolt, "A Study of User Profile Generation from Folksonomies," *Proc. of the Social Web and Knowledge Management Workshop, WWW Conf.*, 2008.

[3] S. Bao, G. Xue, X. Wu, Y, Yu, B. Fei, and Z. Su, "Optimizing Web Search Using Social Annotations," *Proc. of WWW*, pp. 501-510, 2007.

[4] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu, "Exploring Folksonomy for Personalized Search," *Proc. of SIGIR*, pp. 155-162, 2008.

[5] M. G. Noll, and C. Meinel, "Web Search Personalization Via Social Bookmarking and Tagging," *Proc. of ISWC*, pp. 367-380, 2007.

[6] D. Vallet, I. Cantador, and J. M. Jose, "Personalizing Web Search with Folksonomy-based User and Document Profiles," *Proc. of ECIR*, pp. 420-431, 2010.

[7] T. Tao, and C. Zhai, "An Exploration of Proximity Measures in Information Retrieval," *Proc. of SIGIR*, pp. 295-302, 2007.

[8] P. A. Chirita, C. S. Firan, and W. Nejdl, "Personalized Query Expansion for the Web," *Proc. of SIGIR*, pp. 7-14, 2007.

[9] P. A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschutter, "Using ODP Metadata to Personalized Search," *Proc. of SIGIR*, pp. 178-185, 2005.

[10] B. Tan, X. Shen, and C. Zhai, "Mining Long-term Search History to Improve Search Accuracy," *Proc. of KDD*, pp. 718-723, 2006.

[11] K. Sugiyama, K. Hatano, M. Yoshikawa, and S. Uemura, "User-Oriented Adaptive Web Information Retrieval Based on Implicit Observations," *Proc. of ApWeb*, pp. 636-643, 2004.

[12] J. T. Sun, H. J. Zeng, H. Liu, Y. Lu, and Z. Chen, "CubeSVD: A Novel Approach to Personalized Web Search," *Proc. of WWW*, pp. 382-390, 2005.

[13] Z. Dou, R. Song, J. R. Wen, "A Large-scale Evaluation and Analysis of Personalized Search Strategies," *Proc. of WWW*, pp. 581-590, 2007.

[14] Y. Cai, Q. Li, "Personalized Search by Tagbased User Profile and Resource Profile in Collaborative Tagging Systems," *Proc. of CIKM*, pp. 969-978, 2010.

[15] M. Daoud, L. Tamine, M. Bouhanem, "A Personalized Graph-based Document Ranking Model Using a Semantic User Profile," *Proc. of UMAP*, pp. 171-182, 2010.

[16] J. Zhao, Y. Yun, "A Proximity Language Model for Information Retrieval," *Proc. of SIGIR*, pp. 291-298, 2009.

[17] G. Levi, "A Note on the Derivation of Maximal Common Sub-graphs of Two Directed or Undirected Graphs," *Calcolo*, Vol. 9, No. 4, pp. 341-354, 1973.

[18] H. Bunke, X. Jiang, and A. Kandel, "On the Minimum Common Supergraph of Two Graph," *Computing*, Vol. 65, No. 1, pp. 13-25, 2000.

[19] A. Shepitsen, J. Gemmell, B. Mobasher, R. Burke, "Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering," *Proc. of RecSys*, pp. 259-266, 2008.

[20] S. Lee, M. Ko, K. Han, J. Lee, "On Finding Fine-Granularity User Communities by Profile Decomposition," *Proc. of ASONAM*, pp. 631-639, 2012.

한 기 준

2010년 호주국립대학교 정보통신공학과(학사). 2013년 한국과학기술원 지식서비스공학과(석사). 2013년~현재 한국과학기술원 지식서비스공학과 박사과정 재학 관심분야는 정보검색 및 개인화 서비스

장 진 철

2010년 연세대학교 정보산업공학과(학사) 2012년 한국과학기술원 지식서비스공학과(석사). 2013년~현재 한국과학기술원 지식서비스공학과 박사과정 재학. 관심분야는 개인화를 위한 사용자경험 연구

이 문 용

1998년 University of Maryland 정보시스템(박사). 1998년~2004년 University of South Carolina 조교수. 2005년~2009년 University of South Carolina 부교수(tenured). 2009년~2013년 한국과학기술원 지식서비스공학과 부교수. 2013년~현재 한국과학기술원 지식서비스공학과 교수(tenured) 관심분야는 Knowledge Engineering, Semantic Web, Personalization, User Experience