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

한기준 ○, 장진철, 이문용
한국과학기술원 지식서비스공학과
{keejun.han, dreaming, munyi}@kaist.ac.kr

Folksonomy based Personalized Search Utilizing Query Proximity and Graph Profiling Method

Keejun Han○, Jinchul Jang, Mun Y. Yi
Department of Knowledge Service Engineering, KAIST

요 약

본 논문은 Folksonomy 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.

1. 서 론

최근 사용자 태그 기능을 활발하게 제공하는 서비스들이 증가함에 따라, Folksonomy라고 불리는 사용자 태그 정보를 활용하여 개인화 검색의 성능을 향상시키려는 시도가 많은 주목을 받고 있다 [1]. 일반적인 Folksonomy 데이터를 활용한 개인화 검색에서, 사용자와 문서의 프로파일은 사용자가 직접 쓴 태그와 각 태그의 가중치들의 집합으로 표현된다. 이때 각 태그의 가중치들은 TF (term frequency) [2], TF-IDF (term frequency-inverse document frequency) [2], bm25 [3] 등의 방법으로 구해진다. 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 주어진 질의어들 간의 숨겨진 관계를 분석하여 정보 검색 요구에 반영하면 사용자의 검색 만족도를 향상시킬 수 있음이 알려져 있다 [4].

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 있게 되어 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.

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 문서내 프로파일에서 해당 질의어들이 얼마나 긴밀한 관계를 맺고 있는지 측정한 근접성 값과 융합하여 최종적으로 개인화된 검색 결과를 사용자에게 제공한다.

본 연구에서 제안한 그래프 기반의 개인화 검색 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 Folksonomy 데이터셋 상에서 검증하였다.

* 이 논문은 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2011-0029185).

2. 기존 기술과 제안하는 방법간의 차이점

2.1 사용자 및 리소스 프로파일링 구축

상기에서 언급한 바와 같이 기존 Folksonomy 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 중 하나인 'A'는 미래를 그린 공상과학 영화이지만 가족간의 사랑의 의미도 일정 부분 담겨있는 영화이다. 그러나 해당 영화를 시청한 사용자들이 남긴 태그들에 기반하여 기존 검색 방식 중에서 tf 기반의 프로파일 방식으로 표현된 영화 A의 프로파일은 다음과 같다.

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

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 [5] 등에도 적용되며, bm25역시 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. 사용자 및 리소스에서 그래프 기반의 프로파일을 추출하는 방법 및 근접성 정보를 고려한 재정렬 과정은 3장에서 자세히 설명한다.

2.2 사용자 프로파일과 리소스 프로파일의 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 Cosine Similarity [2], Scalar similarity [3] 등의 방법을 통하여 측정되었다. 이외에도 Folksonomy 데이터의 특성을 반영한 여러

측정법들이 제안되었으나 [5], 이는 모두 프로파일의 형식을 벡터 기반으로 표현하는 것을 가정하여 제안되어 그래프 기반의 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 리소스간의 관련 정도를 측정하고자 한다.

3. 제안하는 방법

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.

3.1 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. 기존의 여러 검색 방법 중에서도 bm25는 [4]에서 근접성을 고려하는 모델의 베이스라인으로 사용하였기 때문에, 관련 연구와의 비슷한 실험 조건을 형성하기 위하여 bm25 기법을 베이스라인으로 선정하였다. 따라서 리소스 r 이 주어진 질의어 Q에 적합한 정도를 나타내는 점수 NS (Non-personalized Score)를 구하는 공식은 다음과 같다. bm25가 잘 알려진 모델이기에 본 논문에서는 지면 한계상 bm25 모델 내 변수들에 관한 자세한 설명은 생략한다.

$$NS(Q,r) = bm25(Q,r)$$

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

최종적으로 전체 리소스들 중 질의어 Q에 관한 NS점수가 높은 리소스부터 초기 검색 결과 상위에 반환되고, 이를 바탕으로 초기 검색 결과 재정렬을 위한 프로세스가 시작된다. 본 연구에는 상위에 위치한 50개의 리소스를 대상으로 재정렬을 실시하였다.

3.2 사용자 및 리소스 프로파일링 구축 방법 제안

사용자 u의 그래프 기반 사용자 프로파일 $G_u = (V, E)$에서, V는 사용자 u가 남긴 태그들을 의미하는 노드(node)들의 집합이고 E는 V에 속하는 점들을 연결하는 선(edge)으로, V에 속하는 태그 tag_i과 tag_j의 근접성 정보 $w_{tag_i tag_j}$ 는 다음과 같이 정의된다.

$$w_{tag_i tag_j} = \text{co} - \text{occur}_u(tag_i, tag_j) \quad (2)$$

이때 $\text{co} - \text{occur}_u$는 태그 tag_i와 tag_j가 사용자 u에 의해 함께 사용된 횟수로, 그 횟수가 클수록 사용자 u는 두 태그의 관련성을 높게 정의함을 의미한다. 또한, 리소스 r의 그래프 기반 프로파일 $G_r = (V, E)$ 에서 V는 리소스 r에 사용자들이 남긴 태그들을 의미하는 노드(node)들의 집합을 의미하며, E는 V에 속하는 점들을 연결하는 선으로, V에 속하는 태그 tag_i과 tag_j의 근접성 정보 $w_{tag_i - tag_j}$는 다음과 같이 정의된다.

$$w_{tag_i tag_j} = \text{co} - \text{occur}_r(tag_i, tag_j) \quad (3)$$

이때의 $\text{co} - \text{occur}_r$은 태그 tag_i와 tag_j를 리소스 r에 동시에 사용한 사용자의 총 수로 그 수가 클수록 두 태그의 관련성이 높은 것을 의미한다.

3.3 사용자 및 리소스 프로파일 간 유사도 계산 방법 제안

본 논문에서는 사용자 프로파일 G_u와 리소스 프로파일 G_r간의 유사도 값인 D(Distance)를 구하기 위해 Closeness [6], Maximum Common Subgraph (MCS) [6], edit distance [7] 기법들을 활용한다. 처음 두 알고리즘은 두 그래프간의 노드간 유사도만을 고려하며 마지막 알고리즘은 노드 및 노드를 연결하는 선들의 정보 모두 고려한다. 각 알고리즘의 대한 상세설명은 분량 제한상 본 논문에서는 생략한다. 각 공식에 따른 실험 결과 변화의 추이분석은 4.3장 및 표2에 언급되어 있다.

$$D(G_u, G_r) \begin{cases} D_{CLOSENESS}(G_u, G_r) = \frac{1}{n}\sum_{v \in V} \frac{|G_u \cap G_r|}{|G_u \cup G_r|} \\ D_{MCS}(G_u, G_r) = 1 - \frac{|MCS(G_u, G_r)|}{\max(|G_u|, |G_r|)} \\ D_{EDIT\_DISTANCE}(G_u, G_r) = \min(C(\varepsilon)) \end{cases} \quad (4)$$

3.4 리소스 프로파일 내 질의어 근접성 측정 방법 제안

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.

$$P(Q, G_r) = \exp(-DS(Q, G_r) \times \alpha) \quad (5)$$

$$DS(Q, G_r) = avg\left(\frac{Distance(q_1, q_2)}{maxDistance(G_r)}\right) \times n \quad (6)$$

이때, $Distance(q_1, q_2)$는 질의어 q_1과 q_2간의 최단거리이며, n은 질의어의 개수, $maxDistance(G_r)$ 는 리소스 프로파일 G_r 내에서의 가장 먼 거리로 정규화(Normalization)를 위한 값이다. 구해진 $DS(Q, G_r)$이 거리를 의미함으로 이를 근접성 값으로 변환하기 위해 음수로 변환하고, 실제 단어간의 거리의 특성을 잘 반영하기 위해 공식 (5)와 같이 Convex 함수로 변환한다. 이때 볼록도의 정도를 조절하는 α는 실험을 통하여 0.6으로 고정되었다 (4.3장 및 그림 2의 (a) 참조).

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

$$PS(G_u, G_r, Q) = D(G_u, G_r) + \beta * P(Q, G_r) \quad (7)$$

이때 근접성 정보의 반영도를 조절하는 β는 0과 1사이의 값으로 실험을 통해 0.4로 고정되었다 (4.3장 및 그림 2 의 (b) 참조).

4. 실험

4.1 실험 설계

우리는 여러 Folksonomy 데이터셋들 중 데이터의 밀도가 그래프 기반의 알고리즘에 가장 큰 영향을 미침을 고려하여 [8], 데이터 밀집도가 상대적으로 낮은 CiteULike 북마크 데이터셋(리소스당 평균 태그 수 2.87)과 상대적으로 높은 편인 MovieLens 영화 데이터셋(리소스당 평균 태그 수 9.08)을 실험 데이터셋으로 선정하였다. 다음 표 1은 선정된 두 데이터셋들의 속성을 소개하는 표이다.

표 1. 실험 데이터 셋 소개

	CiteULike	MovieLens
사용된 사용자 수	109	71,567
사용된 리소스 수	44,038	10,681
사용된 태그 수	160,374	95,580
사용자 프로파일 속 평균 태그 개수	64.82	10.6221
리소스 프로파일 속 평균 태그 개수	2.87	9.08

또한 우리는 [2,3]에서 제안되고 검증된 실험설계에 기반하여, 검색된 리소스들 가운데 사용자의 평가 데이터에 존재하는 리소스들이 사용자에게 적합한 문서로 판단하였다. 이러한 적합 문서를 찾기 위하여, 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.

평가 척도로는 MRR과 Precision이 사용되었다. MRR은 검색 결과에서 적합 문서가 등장한 위치 r의 역수, 즉 1/r값을 갖고, Precision은 적합한 문서가 상위 N개의 결과 내에 포함되면 1의 값을 갖는다 (p@5,p@10,p@20). 두 척도 모두 적합한 문서가 높은 순위에 있을수록 높은 수치를 가진다.

4.2 비교 대상

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.

4.3 실험 결과

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. 기존의 개인화 검색 방법인 bm25와 비교했을 때, Edit distance기반 유사도 계산법은 사용된 데이터셋에 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. 이는, 그래프의 구조적 특성 및 각 edge들의 가중치까지 모두 고려하는 Edit distance에 비하여, Closeness는 두 그래프를 구성하고 있는 노드들의 집합이 동일한 경우에만 잘 작동하며, MCS역시 그래프의 구조는 고려되나 각 edge들의 가중치들을 고려하지 않기 때문에 실제적인 그래프 기반의 개인화 검색 성능 향상의 제한을 지니고 있는 것으로 보인다. 이에 우리는 각 데이터셋에서 균일한 향상 정도를 보이는 Edit distance를 본 그래프 기반의 프로파일의 유사도를 계산하는 방법으로 최종 채택하였다.

표 2. 두 프로파일간의 유사도 계산법에 따른 성능 차이

	비개인화	개인화	그래프 기반 개인화		
데이터셋	Baseline	bm25	Closeness	mcs	Edit
CiteULike	0.3557	0.4559	0.4221	0.3999	0.4778
MovieLens	0.2887	0.3644	0.3701	0.3778	0.3912

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

- 두 데이터셋에서 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 Folksonomy 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 보다 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 데이터셋에 관계없이 두 개의 평가 척도 모두에서 비교된 대상들에 비해 더 뛰어난 성능을 가진 것을 보여주고 있다. 즉, 우리의 제안하는 방법은 다른 기법들과 비교하여 MRR기준으로 비교 대상에 비하여 평균적으로 약 7~43%의 유의한 향상 (Wilcoxon test, $p < 0.05$)을 보였고, Precision 기준으로는 평균적으로 약 10~58%의 유의한 향상 (Wilcoxon test, $p < 0.05$)을 보였다.

최종적으로, 그림 2는 우리가 제안하는 방법에서 사용되는 파라미터들의 값에 따라 변하는 MRR을 기록한 것이다. β 값을 0으로 고정하고 α 에 대한 민감도 실험을 수행했을 때, 두 데이터셋 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 최적값에 수렴함을 확인할 수 있었다.

5. 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, 향후 연구에서는 본 모델의 실제 시스템 구현을 위한 확장성 및 효용성 검증이 필요하다.

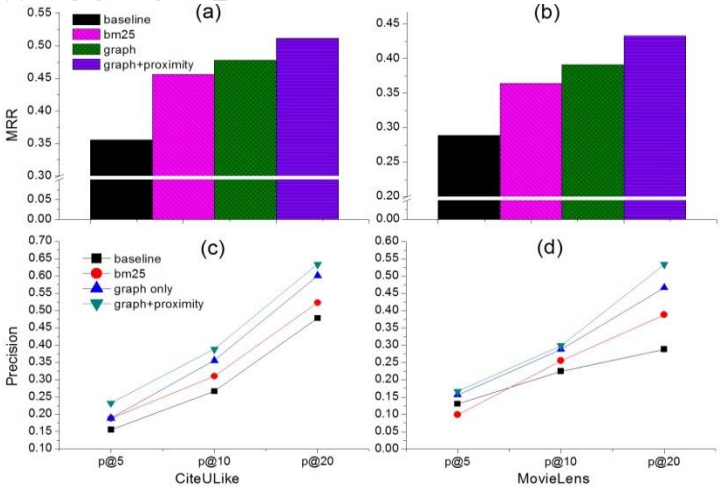
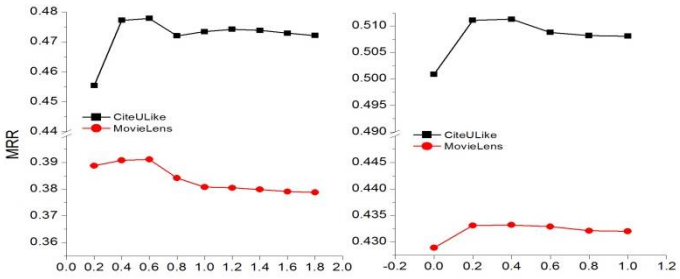


그림1. 개인화 모델에 따른 MRR 및 Precision 결과 요약



(a) α값 변화 (b) β값 변화

그림2. 파라미터 α, β의 변화에 따른 MRR 척도의 변화

참고 문헌

[1] Golder, A. G., Huberman, B. A.: Usage patterns of collaborative tagging systems. Journal of information science. 32 (2), pp. 198-208 (2006)
[2] Xu, S., Bao, S., Fei, B., Su, Z., Yu, Y.: Exploring folksonomy for personalized search. In Proc. of SIGIR 2008. pp. 155-162, ACM Press, New York (2008)
[3] Vallet, D., Cantador, Ivan., Jose, J. M.: Personalizing web search with folksonomy-based user and document profiles. In Proc. of ECIR . pp. 420-431 (2010)
[4] Tao, T., Zhai, C.: An exploration of proximity measures in information retrieval. In Proc. of SIGIR, pp. 295-302 (2007)
[5] Xie, H., Li, Q., Cai, Y.: Community-aware profiling for personalized search in folksonomy. Journal of computer science and technology. 27(3), pp.599-610 (2012).
[6] Levi, G.: A note on the derivation of maximal common subgraphs of two directed or undirected graphs. Calcolo 9(4), 341-354 (1973)
[7] Daoud, M., Tamine, L., Boughanem, M.: A personalized graph-based document ranking model using a semantic user profile. In Proc. of UMAP (2010)
[8] Shepitsen, A., Gemmell, J., Mobasher, B., Burke, R.: Personalized recommendation in social tagging systems using hierarchical clustering. In Proc. of RecSys, pp. 259-266 (2008)