

사용자 활동 점수에 기반한 태그 검색 개선 (Improving Tag Search Based on User Activeness Scores)

임 영 석 [†] 이 강 표 [†] 김 현 우 ^{**} 안 재 민 [†] 김 형 주 ^{***}
(Young-Seok Lim) (Kang-Pyo Lee) (Hyun-Woo Kim) (Jae-Min Ahn) (Hyoung-Joo Kim)

요 약 웹2.0 기술의 출현과 함께 소셜 태깅 서비스들이 큰 호응을 얻고 있다. 태깅에서 사용되는 태그들은 웹상의 자원들을 기술하는 단순한 키워드들의 나열이다. 구조는 단순하지만 많은 사용자들에 의해 협력적으로 생산되는 특성으로 인해 문서를 잘 추상화하므로, 정보 검색 분야에 있어 유용한 메타데이터로서 여겨진다. 특히 검색의 성능을 향상시키기 위해 태그를 사용하는 연구들이 최근 들어 활발하게 진행되어 왔다. 태그는 그 태그를 부여하는 사용자의 참여 정도에 따라 메타데이터로서의 품질이 확연하게 달라지며, 보다 활동적인 사용자도 존재하고 덜 활동적인 사용자도 존재한다. 이로 인해 사용자들의 활동성은 태그를 이용한 검색 결과에도 영향을 미치게 된다. 본 논문에서는 사용자들의 참여 정도를 고려하여 검색 성능을 개선하기 위해 소셜 태깅 서비스인 del.icio.us의 사용자들을 활동성 관점에서 관찰하였다. 그리고 이를 바탕으로 개개인의 다양한 사용자 활동을 모델링하는 방법을 제안하였다. 또한 이를 통해 사용자들의 활동성을 점수화하여 검색에 활용함으로써 검색을 개선하는 방법을 제안한다. 끝으로 제안하는 알고리즘이 실제 검색의 만족도를 향상시킬 수 있음을 실험을 통해 확인하였다.

키워드 : 태그, 검색, 소셜 태깅, 사용자 활동, 딜리셔스

Abstract With the advent of Web 2.0 technologies, social tagging service has gained a great popularity. The tags used in tagging systems are a simple list of keywords describing the resources on the web. Even though their structure is simple, tags abstract documents well and are regarded as useful metadata in the field of information retrieval, since they are produced cooperatively by many users. Specially, to improve the performance of search, a number of approaches using tags have been active to date. Depending on the participation of the users to assign the tags, the quality of tags as metadata varies remarkably. Furthermore more active and less active users exist. Because of this, the activeness of each user assigning the tags will also affect the search results. In this paper, we consider the degree of user participation in order to improve the search performance. We observe the del.icio.us social tagging service, from the perspective of the users activeness. And we propose a modeling for a variety of individual user activities. In addition, we propose an algorithm improving tag search by scoring the user activeness. Finally, we evaluate the proposed method to verify improving the satisfaction of the actual search through experiments.

Key words : Tag, Search, Collaborative Tagging, User Activity, Delicious

· 본 연구는 BK-21 정보기술 사업단의 연구결과로 수행되었음

[†] 비 회 원 : 서울대학교 컴퓨터공학부
yslim@idb.snu.ac.kr
kplee@idb.snu.ac.kr
jmahn@idb.snu.ac.kr

^{**} 학생회원 : 서울대학교 컴퓨터공학부
hwkim@idb.snu.ac.kr

^{***} 종신회원 : 서울대학교 컴퓨터공학부 교수
hjk@snu.ac.kr

논문접수 : 2010년 7월 13일
심사완료 : 2011년 1월 24일

Copyright©2011 한국정보과학회 : 개인 목적이거나 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. 이외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.
정보과학회논문지 : 컴퓨팅의 실제 및 레터 제17권 제3호(2011.3)

1. 서 론

많은 사람들은 웹을 통해 조금이라도 빨리 필요한 정보에 대한 좋은 검색 결과를 얻기를 원한다. 이러한 욕구로 인해 웹 상에 존재하는 방대한 양의 자원에 대한 정보 검색의 품질을 높이기 위한 시도는 매우 활발하게 진행되고 있다. 일반적인 웹 검색의 과정은 사용자가 원하는 정보를 얻기 위해 검색엔진(search engine)의 입력 창에 해당 질의(query)를 입력하면 검색엔진은 이 질의를 분석하고 각각 고유한 알고리즘으로 데이터베이스에 수집된 문서 집합들 중 가장 적합하다고 생각되는 문서를 순위화해서 사용자에게 보여주게 된다.

웹 검색 결과의 품질을 높이기 위해 크게 세가지 접근 방식이 이루어지고 있다. 첫 번째는 사용자가 입력한 질의 자체에 대한 분석이다. 사용자가 입력하는 질의는 검색을 위해 충분하지 않을 수도 있고 혹은 더 좋은 결과를 위해 추가적인 단어가 더해질 수도 있다. 이와 같은 상황을 고려하여 질의를 조작하는 질의 확장(query expansion) 기술이 사용된다. 두 번째는 검색 알고리즘에 사용되는 문서 속성에 대한 접근 방법이다. 웹 문서는 작성자가 입력한 단어나 어구뿐 아니라 타 문서로의 링크관계, 최신성, 사용자 클릭 수 등 다양한 속성을 가지고 있는데 이에 대한 분석을 통해 검색 결과를 향상시키고자 하는 접근이다. 마지막으로 실제 수식화를 위한 방법에 대한 접근이다. 전통적으로 정보 검색 분야에서 많이 사용되어온 TF-IDF[1], BM25[2]와 같은 문서와 단어의 가중치(weight)를 이용한 방법, 질의와 문서의 벡터 비교 방법 등이 있으며 최근에는 각 속성의 가중치를 학습을 통해 결정하는 기계학습(machine learning) 방법이 널리 사용되고 있다.

여러 가지 접근 방법 중 검색 알고리즘에 사용되는 문서 속성에 대한 접근으로써 태그를 사용하는 방법이 각광을 받고 있다. 최근 몇 동안 웹 2.0 기술의 등장으로 del.icio.us[3], CiteULike[4], Flickr[5], YouTube[6]와 같은 소셜 태깅 서비스들이 호응을 얻고 있다. 전 세계의 다양한 사람들이 수많은 주제에 대해 북마크, 웹문서, 사진, 동영상 자료 등을 게시하면 여기에 사용자들은 그와 관련된 태그(Tag)라는 키워드를 작성함으로써 검색이나 분류 시에 자원의 공유가 가능해진다. 태그는 그 자체로도 문서를 잘 추상화하면서도 간결하기 때문에 사용자가 이해하기에 더욱 쉽다. 이러한 태그가 가장 많이 사용되는 곳이 바로 소셜 태깅 서비스이다. 소셜 태깅 서비스에서는 지금도 믿을 수 없는 속도로 많은 양의 태그가 생산되고 있다.

태그는 이와 같이 양적인 측면과 질적인 측면에서 풍부한 메타 데이터이기 때문에 실제로 대중분류법(folksonomy), 시각화(visualization), 시맨틱웹(Semantic Web) 등 많은 분야에서 활용하고 있지만 그 중에서도 특히 웹 검색에서 유용하게 이용될 수 있다. 웹 검색에서 태그를 활용하기 위한 대부분의 연구에서는 del.icio.us[3]와 같은 소셜 북마킹 서비스의 데이터를 이용하고 있다. 해당 북마킹 서비스에 등록된 문서에 부여된 태그들과 실제 웹 문서의 내용을 종합적으로 고려해서 검색에 대한 결과를 얻게 된다.

Google[7]의 PageRank[8] 알고리즘은 최근까지 가장 큰 각광을 받아온 대표적인 웹 검색 알고리즘으로서 문서들 사이의 링크관계를 고려하는 평판(reputation) 기반의 알고리즘이다. 그러나 여기서의 평판은 웹 문서들

작성한 사용자들 간의 평판일 뿐 실제 정보를 검색하는 사용자의 평판이 반영되었다고 보기는 힘들다. 반면 소셜 태깅 서비스에서의 태그는 실제 사용자가 제안하는 키워드이므로 이를 검색에 이용하게 됨으로써 사용자들의 의견을 구체적으로 반영할 수 있게 된다. 즉, 태그는 다양한 사람들로부터 도출된 집단지성(the collective intelligence)으로서, 웹 검색에 있어 메타데이터로서의 긍정적인 역할을 할 수 있는 것이다. 하지만 이전의 연구에서는 문서에서의 태그의 수와 같은 문서-태그 이진 관계(binary relationship)에 대해서만 고려하거나 사용자를 고려하더라도 개개인의 태그 활동을 고려하지는 못하였다.

이에 따라, 소셜 태깅 시스템에서 각 사용자들의 태그 활동을 관찰하고 기존의 방법들을 개선하여 사용자-태그-문서 사이의 삼각관계(ternary relationship)를 활용하여 사용자 활동 점수를 얻을 수 있는 모델에 대해 연구하였다. 그리고 이 모델을 기반으로 실제 검색 결과를 얻는 방법을 제안하였다.

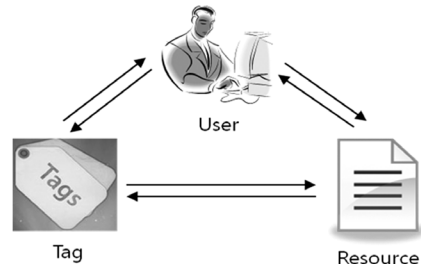


그림 1 소셜 태깅 시스템에서의 삼각관계

본 논문의 구성은 다음과 같다. 2장에서는 태그를 이용한 검색 개선에 대한 관련 연구들에 대해서 살펴본다. 3장에서는 사용자 활동을 고려한 소셜 태깅 시스템 검색에 대하여 제안한다. 즉, 소셜 태깅 시스템에서의 사용자에게 대한 관찰을 통해 사용자 점수화를 위한 알고리즘을 제안하고 이를 이용해서 문서를 순위화하고 질의에 대한 결과를 얻어내는 방법을 제안한다. 4장에서는 3장에서 제안한 사용자 활동 점수를 이용한 태그 검색의 개선 시스템에 대한 실험과 평가에 대해 설명한다. 실험에 사용된 데이터, 평가방법, 실험결과, 결과에 대한 분석에 대해 설명한다. 마지막으로 5장에서는 제안한 개선 방법에 대한 결론에 대해 언급하기로 한다.

2. 관련연구

태그 검색은 일반적인 웹 검색의 특별한 형태로서 웹 검색에서 질의와 문서의 관련성을 판단하기 위해 기본적으로 사용하는 문서의 단어나 어구 대신 소셜 태깅

시스템에서의 각 자원에 부여된 태그를 검색에 활용한다. 태그 검색과 웹 검색은 매칭에 사용되는 용어만 다르기 때문에 기본적으로 웹 검색에서의 다양한 선행 연구가 태그 검색에서도 상당 부분 연관되어 있다.

검색의 기본적인 기술은 8,90년대에 주로 연구되었던 질의와 문서 내 단어의 가중치의 관련성을 계산하는 TF-IDF[1], BM25[2]를 기본으로 사용하는 경우가 많다. 문서 내에서 해당 질의어가 많이 나타날수록 그리고 해당 문서의 길이가 짧을수록 더욱더 중요한 문서임을 주장한다. 이와 관련된 확장 연구들이 꾸준히 진행되고 있다.

90년대 후반부터 Google[7]의 검색엔진에 사용되는 알고리즘으로 널리 알려진 Brin과 Page의 PageRank[8]와 같은 문서의 전역 순위를 계산하는 방법들이 연구되고 있다. PageRank[8]는 웹 문서들의 링크관계를 사용해서 문서들의 순서를 정렬하게 된다. 평판에 따른 문서들의 전역 순서를 얻을 수 있지만 웹 문서 작성자들의 평판만 반영된다. 또한 웹 문서의 내용을 고려하지 않은 순서화이므로 질의와 문서 내용의 관련성을 고려한 추가적인 알고리즘이 필요하다. 하지만 평판의 전달을 통한 순서화 방법은 이후 많은 연구에서 사용되고 있다.

YuskeYanbe[9]의 논문에서는 소셜 태깅 시스템에서의 태그 검색의 가능성을 제시하고 PageRank[8]와의 결합을 보여주었다. 소셜 태깅 시스템에서의 한 문서에 대한 사용자들의 총 태깅 횟수를 SBRank로 정의하고 이를 PageRank와 선형 결합하는 방법을 제안하였다. 소셜 태깅 시스템을 기존 웹 검색에 접목함으로써 태그 검색, 감정 질의를 지원할 수 있음을 보여주었지만 소셜 태깅 시스템의 사용자 개개인의 특성은 활용하지 못하였다. 또한 질의와 문서를 벡터로 놓고 비교하거나 최근의 기계학습을 이용한 속성들의 가중치를 자동으로 계산하는 방법이 아닌 단순 선형 결합만 사용하였다. 반면 E.Agichtein[10]의 논문에서는 RankNet이라는 기계학습 방법을 이용하여 사용자 클릭수나 직접적인 사용자 피드백 등의 속성을 결합한 웹 검색의 개선에 대해 잘 보여주고 있다.

소셜 태깅 시스템의 정보만 사용한 태그의 순위와 관련된 연구도 진행되고 있다. Golder[11]는 소셜 북마크 서비스의 문서들의 태그의 상태 비율이 일정한 시간이 지나면 상대적으로 일정해짐을 보였다. 이는 모방을 통한 대중의 지식으로써 태그가 안정적이라는 사실을 보여준다. 또한 문서가 아닌 동영상, 사진과 같은 내용에 대한 정보가 없는 자원에 대한 검색에 대해 태그를 사용할 수 있음을 보여주는 연구들이 있다. Liu[12]의 연구에서는 태그와 사진의 이미지 정보간의 확률 관계를 통해 해당 사진에서의 태그에 대한 순위를 얻는다. 이를 통해

태그 추천, 태그 검색 등에 활용하는 예를 보여준다.

Bao[13]의 연구는 본 연구의 동기가 되는 논문으로써 역시 웹 검색 결과의 향상을 위해 태그를 이용하는 문제를 연구하고 있다. 기존의 SimRank[14]를 활용하여 질의와 유사 태그간의 매칭을 시도하는 SocialSimRank와 소셜 태깅 시스템에서의 문서-사용자-태그 삼각관계를 상호강화관계로 보고 문서의 전역 순위를 구하는 SocialPageRank를 제안하였다. 특히 본 연구에서 주목한 부분은 SocialPageRank이다. 웹 검색에서의 전역 문서 순위를 정하기 위해 소셜 태깅 시스템의 삼각관계를 고려하지만 전역적인 평판만 반영하고 있다. 즉, 사용자, 태그, 문서의 기본 점수를 설정하고 높은 점수의 사용자가 태깅한 태그는 높은 점수를 얻고 높은 점수의 태그가 많이 달린 문서는 높은 점수를 얻으며 높은 점수의 문서에 태깅한 사용자는 높은 점수를 얻게 되는 상호연결 과정을 매트릭스 연산을 통해 계산하는 방법이다. 하지만 전역적인 태그의 점수가 높다 할지라도 특정 문서에서 해당 태그의 의미가 항상 중요시 될 수는 없다. 또한 전역적인 점수가 높은 문서에 태깅을 많이 한 사용자가 해당 문서에 태깅을 함으로서 더 많은 기여를 했다고 볼 수는 없다.

이에 본 연구에서는 위 연구와 달리 웹 검색이 아닌 태그 검색에서 해당 문서에서의 특정 태그의 지역적인 중요도를 고려하여 사용자의 태깅 활동을 모델링하고 이를 이용하여 소셜 태깅 시스템에서의 태그 검색을 개선하고자 한다.

3. 사용자활동을 고려한 소셜 태깅 시스템 검색

이 장에서는 사용자활동을 고려한 소셜 태깅 시스템의 검색 모델에 대해서 소개한다. 3.1절에서는 소셜 태깅 시스템의 사용자 활동에 대한 관찰에 대해 서술한다. 3.2절에서는 3.1절의 관찰을 바탕으로 사용자 활동을 모델링할 수 있는 방법을 제안하고 3.3절에서는 사용자 점수를 바탕으로 문서의 순위를 얻는 것에 대해 설명한다. 마지막으로 3.4절에서는 문서의 순위를 속성으로 활용하여 기존의 방법과 결합하는 방법에 대해 설명한다.

3.1 소셜 태깅 시스템에서의 사용자 활동 관찰

소셜 태깅 시스템에서의 태그는 특정 문서에서의 지역적인 중요도를 가진다. 전체 문서 집합에서 많이 사용되는 태그라 할지라도 특정 문서에서는 의미 없는 태그일지도 모른다. 또한 평판이 높은 문서에 태그를 부여하는 사용자가 다같이 높은 순위를 가지는 사용자일 수는 없다. SocialPageRank[13]에서의 위와 같은 문제를 해결하기 위해 태그의 전역 순위가 아닌 지역적인 중요도를 이용해서 사용자 활동을 모델링하고자 한다.

소셜 태깅 시스템은 세계의 다양한 배경을 가진 사용

자들이 자신의 편의 혹은 생각에 따라 다양한 태그를 부여하는 시스템이다. 따라서 사용자마다 다양한 태깅 활동을 보인다. 사용자 활동을 모델링하기 위해 크게 세 가지 측면에서 접근할 수 있다. 사용자의 태그의 양, 태그의 품질, 태깅 활동의 일관성이 바로 그 세 가지 측면이다.

사용자의 태깅 활동을 살펴보면 사용자마다 한 문서에 부여하는 태그의 양이 다르다는 것을 알 수 있다. 1,2개의 태그를 부여하는 덜 활동적인 사용자가 있는 반면 5개 이상 혹은 10개가 넘는 태그를 부여하는 보다 활동적인 사용자가 있음을 볼 수 있다. 문서의 내용을 고려하지 않는 태그 검색의 특성 때문에 태그가 문서의 내용을 잘 추상화할수록 검색의 품질이 향상된다. 문서의 내용을 잘 추상화하는 태그들로 구성된 태그 집합이 필요하다는 점에서 볼 때 여러 개의 태그를 통해 문서를 표현하는 사용자들의 태깅 활동이 검색의 품질을 높인다고 볼 수 있다. 따라서 사용자의 활동을 모델링하는 경우에 많은 양의 태그를 부여하는 사용자와 적은 양의 태그를 부여하는 사용자의 활동을 차별화해야 한다.

태그의 질적인 측면에서 사용자의 태깅 활동을 살펴볼 수 있다. 태그의 양적인 측면을 만족시키는 사용자라 할지라도 그 태그가 실제로 해당 문서와 연관되어 있는지 살펴볼 필요가 있다. 반대로 태그의 양은 부족할지라도 해당 문서를 잘 표현하는 태그를 부여하는 사용자도 있을 수 있다. 예를 들어, apple.com[15]의 문서와 연관성이 높은 “apple”, “mac”, “ipod” 등의 태그를 부여하는 사용자가 있는 반면 “test”, “mar:))”와 같은 상관없거나 개인적인 태그를 부여하는 사용자가 있다. 따라서 태그 검색 개선을 위해 사용자 태깅 활동을 모델링하기 위해서는 사용자의 태깅 활동을 질적인 측면에서도 고려해야 한다.

마지막으로 소셜 태깅 시스템 사용자들의 태깅 활동에 있어서의 일관성에 대하여 생각해 볼 필요가 있다. 사용자들이 태깅 활동을 함에 있어서 특정 문서에는 많은 양의 태그를 부여하고 또 다른 문서들에서는 적은 양의 태그를 부여하거나 문서마다 부여하는 태그의 질적인 정도가 달라지는 등 일관성이 없다면 사용자 활동을 고려한 검색 개선의 의미가 약해질 수 있다.

그림 2의 그래프는 실험을 위해 del.icio.us 서비스에서 실제로 수집한 약 16만 명의 문서당 평균 태그 수에 따른 사용자 수를 보여준다. 사용자당 평균 3.1개의 태그를 1개의 문서에 부여하는 것으로 나타났으며 최대 34개의 태그를 한 문서에 부여하는 사용자도 존재하였다. 즉, 그래프의 평균 태그 수의 스펙트럼에서 볼 수 있듯이 일관되게 태그를 많이 부여하는 유저가 있는 반면 적게 부여하는 유저가 존재한다. 태그의 양적인 측면

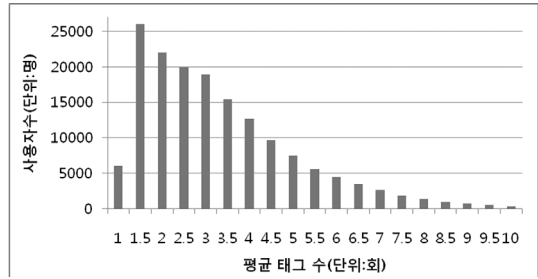


그림 2 평균 태그 수에 따른 사용자 변화

뿐만 아니라 이후에 제안하는 태그의 품질을 고려하는 방법을 통해 실험적인 결과를 살펴보면 질적인 측면에 있어서도 일관성을 발견할 수 있다.

이상의 세 가지 관찰을 통해 소셜 태깅 시스템의 사용자 활동을 모델링하여 태그 검색을 개선할 수 있는 가능성이 있음을 알 수 있고 실제로 이를 모델링하는 방법을 3.2절에서 제안하였다.

3.2 사용자 활동 모델링

이 절에서는 3.1 절에서 설명한 소셜 태깅 시스템에서의 사용자 관찰의 세 가지 측면을 고려하여 실제 검색에서 중요한 속성으로 사용될 수 있는 사용자 활동을 모델링 할 수 있는 알고리즘을 제안한다. 즉, 사용자 활동을 점수화하여 소셜 태깅 시스템 사용자들을 점수화한다. 이상의 관찰을 통해 알 수 있듯이 많은 양의 품질이 좋은 태그를 항상 일관되게 부여하는 사용자는 높은 점수를 얻게 하고자 하는 것이 이 모델링의 핵심이라 할 수 있다.

먼저 태그의 품질은 앞서서도 설명했듯이 전체 태깅 시스템에서의 전역적인 평균보다는 특정 문서에서의 역할이 중요하므로 태그 t_i 의 문서 d_j 에서의 관련성을 다음과 같이 정의하였다.

$$Rel(t_i, d_j) = \frac{TagCount(t_i, d_j)}{TotCount(d_j)}$$

여기서 $TotCount(d_j)$ 는 문서 d_j 에 부여된 총 태깅 횟수를 말하며 $TagCount(t_i, d_j)$ 는 태그 t_i 가 문서 d_j 에 부여된 횟수를 나타낸다.

그러면 사용자 u_k 가 문서 d_j 에 부여한 태그의 집합을 $T_{jk} = \{t_{jk1}, t_{jk2}, \dots, t_{jkm}\}$ 라 할 때 사용자 u_k 가 문서 d_j 에 부여한 태그의 품질은 $\sum_{i=1}^m Rel(t_{jki}, d_j)$ 로 표현할 수 있다. 이를 통해 한 문서에서 한 사용자의 태깅 활동의 양 또한 고려할 수 있게 된다. 따라서 사용자 u_k 가 태깅을 한 문서 집합을 $D_k = \{d_{k1}, d_{k2}, \dots, d_{kn}\}$ 라 할 때 사용자 u_k 의 태깅 활동의 평균적인 품질은 다음과 같다.

$$AvgTagQuality(u_k) = \frac{\sum_{j=1}^n \sum_{i=1}^m Rel(t_{jki}, d_{kj})}{n}$$

사용자의 평균 태그 품질을 구함으로써 사용자의 한 문서 당 태그의 양과 품질을 표현할 수 있다. 여기서 또 하나 고려해야 될 점이 있다. 사용자의 평균 태그 품질이 높다는 것은 그만큼 관련성이 높은 태그를 문서마다 많이 달았다는 것을 의미한다. 하지만 평균적인 수치이므로 사용자의 태깅한 문서 집합 D_k 의 크기가 작으면 작을수록 그 수치는 신뢰도가 떨어진다. 극단적인 예로 태깅 시스템내의 유일하게 문서 하나에만 관련성이 높은 태그를 부여하게 되면 이 사용자의 평균 태그 품질은 상당히 높게 나타난다. 하지만 이 사용자가 더 활동적인 사용자인지 덜 활동적인 사용자인지는 아직 판단할 수 없다. 따라서 이에 대한 페널티를 주기 위해 사용자가 태깅한 문서 집합의 크기도 사용자 태깅 활동을 모델링하는데 포함시켰다. 즉 많은 문서에 태깅을 한 사용자일수록 평균적인 태그의 품질 수치를 더 많이 반영할 수 있도록 다음과 같이 표현하였다.

$$UserScore(u_k) = AvgTagQuality(u_k) \left(- \frac{1}{\mu \times ResourceCount(u_k) + 1} + 1 \right)$$

여기서 $ResourceCount(u_k)$ 는 사용자 u_k 가 태깅한 총 문서집합의 크기를 나타낸다. 또한 μ 값은 신뢰상수로써 실제 실험결과 사용자들의 평균적인 총 문서집합의 크기가 460개였으므로 계산상의 편의를 위해 실험적인 값으로 0.008로 설정하였다. 이 경우 500개의 문서에 태깅을 한 사용자는 평균 태그 품질 값의 80%가 반영된 사용자 점수를 얻게 된다.

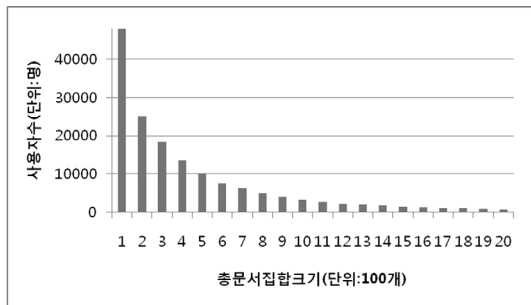


그림 3 문서집합크기에 대한 사용자수

실험을 위해 수집한 실제 del.icio.us 서비스의 16만명의 데이터에 대한 사용자 태깅 문서집합의 크기에 대한 결과를 그림 3의 그래프에서 살펴볼 수 있다. 4000개가 넘는 문서에 태깅을 한 사용자도 12명이 존재하였다. 2008년에 발표된 del.icio.us 사이트의 사용자수가 530만명이라는 것을 감안하면 실제 평균적인 문서집합의 크기는 더욱더 증가할 것이다.

또한 그림 4의 그래프는 앞에서 설명한 사용자 활동

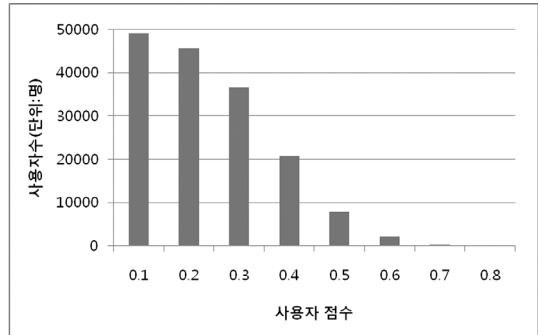


그림 4 사용자 점수에 따른 사용자수

모델링을 통한 실제 사용자 점수를 보여주고 있다. 사용자 점수의 최대 값은 1이고 0.7-0.8 사이의 사용자도 30명이 존재한다. 많은 사용자가 0-0.3이내의 범위에 속하지만 0-0.8까지 다양한 태깅 활동 결과를 보인다.

지금까지 소셜 태깅 시스템에서의 사용자 활동을 모델링하기 위해 태그의 품질과 양, 그리고 사용자 패턴의 세가지 측면을 고려하였으며 이를 바탕으로 사용자 점수화 알고리즘을 제안하였다. 3.3절에서는 제안한 사용자 점수를 바탕으로 태그와 문서의 관련성에 대해 고려하고 이를 순위화하는 방법을 제안한다.

3.3 사용자 활동 점수를 통한 문서 순위화

태그 검색에서 최종적으로 얻는 결과는 특정 질의에 대한 문서의 순위이다. 따라서 앞에서 제안한 사용자 활동 모델링을 바탕으로 특정 질의에 대한 사용자들의 점수를 관련 문서들과 연결시킴으로써 사용자 활동 속성을 태그 검색에 활용할 수 있도록 하였다. 이 논문에서는 질의 확장을 하지 않으므로 기본적으로 특정 질의를 태그로 포함하고 있는 문서에 대한 순위화만이 필요하다. 문서의 순위화는 바로 태그 검색의 결과가 될 수 있지만 보다 나은 검색 결과를 위해 3.4절에서 설명하는 최종 결과를 위한 제정렬의 한가지 속성으로 사용된다.

문서 순위화를 위해 먼저 태그 검색 시스템에 입력되는 질의 집합을 $Q = \{q_1, q_2, \dots, q_m\}$ 로 정의한다. 그러면 순위화를 해야 되는 문서 집합은 $D = \{D_{q_1}, D_{q_2}, \dots, D_{q_m}\}$ 으로 표현할 수 있다. 여기서 D_{q_i} 은 질의 q_i 을 태그로 포함하는 문서집합을 뜻한다. 또한 문서집합 D_{q_i} 들의 문서 원소는 서로 중복될 있다. 즉, 질의 집합 Q 의 원소를 여러 개 포함하는 문서가 존재할 수 있다. 이제 문서의 순위화를 위해 질의 집합 Q 에 대한 특정 문서 d_i 의 문서 점수를 다음과 같이 정의한다.

$$ResourceScore(d_i, Q) = \sum_Q \sum_{U_{ji}} UserScore(u_k)$$

사용자 집합 U_{ji} 는 문서 d_i 에 질의 집합 내 원소 태그 q_i 를 태깅한 사용자들의 집합이다. 질의 집합 Q 에 대한 문서 점수는 질의 집합의 각 원소에 해당하는 태

그를 부여한 사용자들의 사용자 점수의 합이 높을수록 높아진다.

최종적으로 얻고자 하는, 질의 집합 Q에 대한 문서 순위화의 결과는 문서 집합 D 내의 모든 원소들에 대한 문서 점수를 정렬함으로써 얻을 수 있게 된다. 이 질의 처음에 언급했던 바와 같이 문서 점수를 통한 순위화 결과를 바로 태그 검색의 결과로 사용할 수도 있다. 하지만 이 경우 질의 집합과 관련된 태그의 양이나 품질의 중요도가 떨어지게 되어 결과가 나빠질 수 있다. 이를 방지하기 위한 재정렬 방법을 3.4절에서 다룬다.

3.4 문서 점수 속성을 이용한 재정렬

이 절에서는 앞에서 제안한 사용자 활동 점수 속성의 결과인 문서 점수를 검색에 사용되는 하나의 속성으로 활용해서 질의 집합과 관련된 태그의 양이나 질적인 측면도 고려하고자 한다. 즉 질의 집합의 원소와 문서 내 태그와의 가중치 관계를 고려하는 기본 속성을 생각해 본다. 태깅 시스템내의 문서에서 태그는 그 문서를 추상화하는 키워드이기 때문에 일반 웹 문서에서의 내용상의 단어로 볼 수 있다. 따라서 질의 집합과 문서 내 단어 사이의 가중치를 계산하는 기존의 방법을 태깅 시스템에 도입하고자 한다.

일반적인 웹 검색에서 질의 집합과 문서 내 단어 사이의 가중치를 계산하는 방법으로 TF-IDF[1], BM25[2] 기법 등이 있다. 그 중 가장 좋은 결과를 보이는 것으로 알려진 BM25를 태깅 시스템에서의 태그 검색을 위한 기본 속성으로 사용하였다. 질의 집합 $Q = \{q_1, q_2, \dots, q_m\}$ 에 대한 문서 d의 BM25 값을 계산하기 위해 다음의 식을 사용하였다.

$$Score(d, Q) = \sum_{i=1}^m IDF(q_i) \cdot \frac{f(q_i, d) \cdot (k+1)}{f(q_i, d) + k \cdot \left(1 - b + b \cdot \frac{|d|}{avgdl}\right)}$$

여기서 $f(q_i, d)$ 는 문서 d에서 q_i 의 나타난 빈도 즉, 태깅 횟수를 나타내고 $|d|$ 는 문서 d의 길이 즉, 총 태깅 횟수를 나타낸다. 또한 avgdl은 전체 문서 집합에 있는 문서들의 평균 길이로 태깅 시스템 내의 전체 문서들의 평균 총 태깅 횟수로 표현할 수 있다. 실험에 사용한 데이터의 avgdl 값은 31.7449였으며 k와 b는 상수로써 $k=2$, $b=0.75$ 의 값을 사용하였다. 마지막으로 $IDF(q_i)$ 는 일반적인 웹 검색에서 사용되는 역문서빈도(Inverse Document frequency)로써 다음의 식을 사용하였다.

$$IDF(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

여기서 N은 전체 문서집합의 크기로서 실험에 사용한 값은 37508367이다. $n(q_i)$ 는 q_i 질의를 태그로 포함하는 문서의 수이다.

태깅 시스템에 도입된 문서의 BM25 값은 태그 검색에서 사용될 수 있는 기본적인 속성이다. 이는 일반적인 웹 검색에서의 BM25 기법의 역할과 유사하다. 따라서 일반적인 웹 검색의 기존 연구들과서와 마찬가지로 태그 검색에서도 앞에서 제안한 사용자 활동 점수를 통한 문서 점수 속성을 BM25 값과의 결합을 통해 재정렬 한다.

서로 다른 속성들의 가중치를 계산하기 위한 많은 방법들이 존재한다. 2장에서 살펴본 바와 같이 주로 PageRank와 결합하는 방법이 많고 선형 결합하거나 Microsoft의 연구[10]처럼 휴리스틱 결합을 사용하기도 한다. 최근에는 기계학습을 이용해 특성들 간의 가중치를 결합하는 방법도 많이 사용되고 있다. 기계학습을 통한 방법은 학습을 통해 각 속성들의 가중치를 자동적으로 정해준다는 장점이 있다. 이에 이 논문에서도 SVM을 바탕으로 만들어져 공개된 SVMRank[16]를 사용하여 사용자 활동 점수를 통한 문서 점수 속성과 BM25를 이용한 기본 속성을 결합하였다.

4. 성능평가

이 장에서는 소셜 태깅 시스템에서의 사용자 태깅 활동 모델링을 통하여 태그 검색이 개선될 수 있음을 보이는 실험을 위한 실험 환경을 설명하고 실험 결과를 분석한다. 앞에서 설명한 기존의 방법들과의 비교 실험을 통해 제안한 방법의 태그 검색 성능을 평가한다.

4.1 실험 데이터

실험을 위해 사용한 데이터셋(dataset)은 소셜 북마킹 서비스인 del.icio.us에서 2010년 4월 19일부터 2010년 4월 21일까지 수집된 162,373명의 사용자 데이터로서 여기서 사용자는 del.icio.us가 제공하는 인기 있는(popular)태그로 검색되는 문서를 북마킹한 사용자를 수집하였다. 제안한 방법의 특성상 사용자의 전체 태깅 활동에 대한 실험이 요구되므로 각 사용자의 전체 태깅 활동을 수집하였다. 수집된 사용자들의 전체 데이터는 4,620,404개의 태그, 37,508,367개의 문서, 331,009,284개의 태깅 활동으로 구성되어 있다. 여기서 태깅 활동이란 한 사용자가 하나의 문서에 하나의 태그를 부여하는 활동을 말한다. 즉 하나의 문서에 여러 개의 태그를 부여하는 경우 각각의 태그 별로 태깅 활동이 저장된다. 이 사용자들 중 북마킹은 하였지만 태그를 부여하지 않았거나 다른 사람이 태그를 부여하지 않은 문서에만 태그를 부여한 사용자가 1,247명이었다. 이 사용자들에 대해서는 사용자 점수를 0점 처리하였으며 그 외 사용자들의 태깅 활동에 있어서도 다른 사람이 태그를 달지 않은 문서에 태그를 부여한 경우는 스팸머(spammer)와 같은 비정상적인 사용자를 방지하기 위해 계산에서 제외된다.

표 1 실험에서 사용된 질의 집합

분야	질의
Computers	{apple, iphone}
Computers	{multimedia, software}
Computers	{programming, history}
Computers	{database, oracle}
Arts	{television, guide}
Arts	{movie, database}
Shopping	{skin, care}
Home	{cooking, pizza}
Reference	{google, maps}
Games	{video, game, music}

또한 제안한 방법의 기본 속성인 BM25와의 SVMRank의 기계학습을 통한 결합을 위해 ODP(Open Directory Project)[17]에서 질의를 선정하였다. 2-3개의 단어로 된 질의 10개를 선정하여 9개의 질의에 대한 값들을 학습하고 1개의 질의를 테스트하는 방식으로 10개의 질의에 대한 결과 값을 산출 하였다.

표 1은 실험에 사용하기 위해 선정된 질의를 보여준다. ODP의 6개 분야에서 각각 1~4개의 질의를 선정하였다.

각 질의마다 BM25 속성을 기준으로 상위 100개의 문서를 추출하였고 이 문서들의 값을 실험에 사용하였다. 제안한 방법과의 비교를 위한 다른 방법들도 같은 질의에 대해서 학습하였다.

4.2 성능 평가방법

제안한 방법의 성능을 평가하기 위해 4개의 방법에 대하여 비교하였으며 웹 검색 엔진 알고리즘의 효율성의 척도인 NDCG[18]를 성능 평가의 척도로 사용하였다. NDCG는 다음과 같은 방법으로 계산하였다.

$$N_q = M_q \sum_{j=1}^k (2^{r(j)} - 1) / \log(1 + j)$$

질의 q에 대해 M_q 는 NDCG 값을 1로 만드는 완벽 정렬을 위한 정규화 상수이다. NDCG의 평가를 위해 JSP를 사용한 평가 UI를 구현하였으며 12명의 실험 참가자들을 대상으로 실험을 진행하였다. 1인당 2개의 질의에 대하여 평가하도록 하였으며 각 질의는 BM25 방법을 기준으로 상위 100개의 문서에 대하여 평가가 이루어져 개인당 총 200개의 문서에 대하여 평가 결과를 얻었다. 문서당 관련도 점수 $r(j)$ 는 0-3점으로 설정하였다. 표 1에서 사용된 질의집합을 살펴보면 단일 질의로 구성된 질의 집합이 아닌 2개 내지 3개의 단어로 구성된 복합 질의임을 알 수 있다. 이는 단일 질의로 구성해서 실험 참가자들로부터 평가를 얻을 경우 동점인 문서가 많거나 실험참가자에 따른 편차가 클 수 있으므로 질의의 의미를 명확하게 해서 이러한 문제를 줄이기 위

표 2 비교를 위한 방법

비교 방법
BM25 (Baseline)
BM25+SocialPageRank (SPR)
UserActivityScore (UAS)
BM25+UserActivityScore (Our Approach)

함이다. 실험을 통해 비교한 방법 4가지는 표 2와 같다.

웹 검색의 기본적인 방법인 BM25를 기본선으로 설정하였으며 BM25와 Bao의 SocialPageRank의 결합, 사용자 활동 점수, BM25와 사용자 활동 점수의 결합에 대하여 서로 비교하였다. 여기서의 결합은 모두 SVMRank를 사용하여 학습한 결과를 이용하였다.

일반적인 웹 검색과 마찬가지로 상위 k개에 대한 검색 효율을 보이기 위해 기본선인 BM25의 상위 100개 문서에 대하여 각 방법들로 재정렬한 순위에 대하여 상위 k개에 대한 NDCG 값을 계산하였으며 정렬의 효율성을 보이기 위해 BM25의 상위 10개의 문서에 대하여 각각의 방법으로 재정렬한 상위 k개에 대한 NDCG 값도 계산하였다.

4.3 실험 결과

4.2절에서 설명한 기본 방법들과의 비교 실험 결과는 다음과 같다. 각 질의의 NDCG값의 평균값에 대한 결과이며 표준편차는 0.012~0.05정도의 수치를 가진다. 이와 같은 수치는 사용자가 직접 평가해야 하는 NDCG의 단점으로 인해 적은 양의 질의 집합을 통한 실험으로 인한 결과이다.

먼저 그림 5는 BM25를 이용해서 뽑은 상위 100개의 문서에 대한 상위 k의 NDCG 값을 보여준다.

그림에서 보는 바와 같이 BM25를 제외한 나머지 방법이 기본선인 BM25 보다는 상당히 좋은 결과를 보인다. 이를 통해 문서의 내용만 고려하기 보다 부가적인

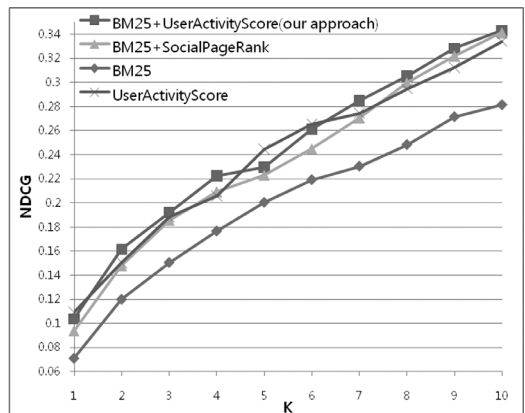


그림 5 BM25기준 상위 100개의 문서에 대한 top-k NDCG

속성이 검색에 활용될 경우 검색의 결과가 더욱 향상됨을 알 수 있다. NDCG 값으로 살펴보면 본 연구뿐 아니라 기존의 연구들도 BM25에 비해 0.03~0.05 정도의 성능 향상을 보인다. 또한 제안한 방법인 BM25와 사용자 활동 점수를 결합한 방법은 Bao가 제안한 SocialPageRank 보다는 모든 k에 대해서 좋은 결과를 보이고 BM25와 결합하지 않은 사용자 활동 점수만 사용했을 때와 비교해서는 k가 5일 때를 제외하고는 항상 좋은 결과를 보임을 알 수 있다. 이는 제안한 방법이 실제 소셜 태깅 시스템에서의 태그 검색의 품질을 향상시킨다는 것을 나타낸다. 반면, k가 5일 때 사용자 활동 점수만 가지고 평가한 방법이 결과가 더 좋게 나오지만 사용자 활동 점수만 가지고 검색한 경우는 k가 변화함에 따라 결과 변동이 심한 것을 볼 수 있다. 즉, 질의와 문서의 태그 집합과의 관계를 고려하지 않은 해당 질의를 태그로 부여한 사용자의 점수만 고려할 경우 일정한 결과를 낼 수 없음을 알 수 있다. 이 결과는 기존의 질의와 문서 내용의 단어 가중치 관계를 이용한 방법[19]과 다른 부가적인 속성과의 기계학습을 통한 순위 결합을 통해 보다 안정적이고 향상된 검색 결과를 얻을 수 있음을 나타낸다.

그림 6은 BM25를 이용해서 뽑은 상위 10개의 문서에 대해 각각의 방법을 사용해서 재정렬했을 때의 상위 k의 NDCG 값을 보여준다.

그림 6에서 볼 수 있듯이 모든 k에 대해서 제안한 방법인 BM25와 사용자 활동 점수와의 결합이 가장 좋은 결과를 보인다. 그림 5에서 BM25에 비해 전체적인 NDCG 값의 향상이 0.03~0.05였던 반면 그림 6에서는 제안한 방법이 BM25와 SocialPageRank의 결합 방법에 비해서 0.05 이상의 성능 향상을 보여준다. 즉 제안한 방법은 그림 5의 실험 결과에서 볼 수 있듯이 전체적인 검색의 품질을 향상시킬 뿐만 아니라 그림 6의 실험 결과에서처럼 상위 문서들의 순위 정렬에서도 더 나은 결과를 제공함을 알 수 있다. 그림 5의 실험 결과인 상위 100개의 문서에 대한 실험에서 사용자 활동 점수만 고려한 방법이 변화가 심하지만 BM25에 비해 좋은 결과를 낸 것에 반해 상위 10개의 문서에 대한 재정렬 실험에서는 BM25보다도 낮은 실험 결과를 보인다. 이는 앞서서도 살펴본 바와 같이 질의와 문서의 내용, 즉 태그 집합과의 관계가 검색의 바탕이 되고 사용자 활동 점수는 더 나은 검색 결과를 위한 속성으로 사용되어야 함을 보여준다.

또한 Bao의 SocialPageRank 역시 BM25에 비해 나은 결과를 보여주지 못하는 것은 전체적인 검색 성능은 향상시키지만 소규모 문서의 정렬에 있어서는 좋지 못함을 나타낸다.

이상의 결과에서 볼 수 있듯이 제안한 방법인 BM25와 사용자 활동 점수의 결합을 이용한 태그 검색이 전

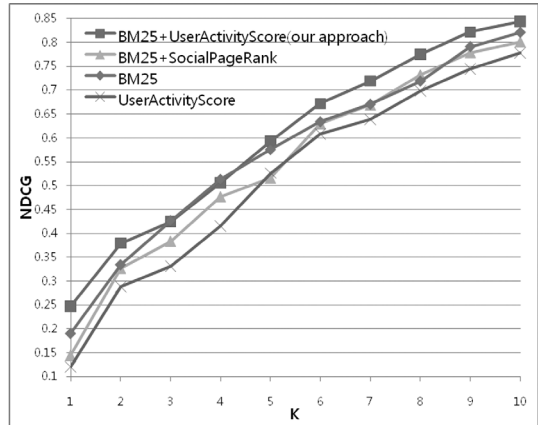


그림 6 BM25기준 상위 10개의 문서에 대한 top-k NDCG

제적인 검색 성능뿐 아니라 특정 문서 집합의 정렬에서도 좋은 결과를 나타낸다.

5. 결론 및 향후연구

일반적인 웹 검색에 있어서 웹 2.0의 등장과 함께 나타난 사용자의 참여로 이루어진 소셜 태깅 시스템을 활용하여 성능을 향상시키고자 하는 많은 연구들이 있었다. 그러나 소셜 태깅 시스템의 사용자 개개인의 참여 정도에 대해서는 고려하지 않았다.

본 논문에서는 소셜 태깅 시스템의 사용자 개개인의 활동을 모델링 하여 이를 사용자 점수로 환산하였고 질의와 문서 내 단어의 가중치 관계를 이용하는 BM25와 결합하여 태그 검색에 활용하였다.

4장의 실험 결과에서 보여주는 바와 같이 제안한 방법인 사용자 활동 점수에 기반한 태그 검색은 기존의 검색 방법들보다 전체적인 검색 결과 뿐 아니라 특정 문서 집합의 순위 정렬에 있어서도 더 나은 결과를 보여준다. 이는 의도한 바와 같이 무의미하게 태그를 다는 사용자의 기여도를 줄이고 유용하고 사용자들이 공감하는 태그를 많이 다는 사용자들의 태깅활동이 실제로 검색을 하는 사용자들에게 더 유익한 정보를 제공함을 말한다. 그러므로 소셜 태깅 시스템의 개개인의 사용자 활동 점수를 통해 태그 검색을 개선하는 연구는 충분히 의미를 가진다.

또한 제안한 방법은 지역적인 태그의 중요도를 고려하므로 웹 활동에서 항상 문제시 되는 스팸어의 활동이 검색 결과를 저해하는 부분을 상당 부분 해소할 수 있다.

하지만 본 연구에서 제안한 사용자 활동 점수에 기반한 태그 검색 개선 방법은 사용자의 태깅 활동이 새로 추가 될 때마다 문서에서의 태그의 가중치가 달라지므로 사용자 점수가 재계산 되어야 한다. 이는 빠르게 증가하는 태그의 양에 대처하는데 어려움이 따른다. 또한

제한한 방법을 태그 검색이 아닌 다른 검색 시스템에서 이용하고자 할 경우 태그가 아닌 다른 속성을 이용하여 모델링 해야 하는 어려움이 따른다. 따라서 이를 해결하기 위한 연구가 필요하다.

결론적으로 실제 검색 시스템을 사용하고 평가하는 사용자 개개인의 참여도가 다르다는 사실을 관찰하고 이를 고려한 검색 방법의 필요성을 인지하였으며 실제로 이를 활용한 본 논문의 제한한 방법을 통해 향상된 검색 결과를 얻을 수 있었다는 점에서 본 연구가 충분히 가치 있는 연구임을 알 수 있다.

참 고 문 헌

[1] Salton, G. and C. Buckley, Term-weighting approaches in automatic text retrieval, *Information Processing & Management*, vol.24, no.5, pp.513-523, 1988.

[2] Robertson, S., et al. Okapi at TREC-3, Text REtrieval Conference, 1992.

[3] Delicious, <http://delicious.com/>.

[4] CiteULike, <http://www.citeulike.org/>.

[5] Flickr, <http://www.flickr.com/>.

[6] YouTube, <http://www.youtube.com/>.

[7] Google, <http://www.google.com>.

[8] Page, L., et al., The PageRank Citation Ranking: Bringing Order to the Web, Technical report, Stanford Digital Library Technologies Project, 1998.

[9] Yanbe, Y., et al., Can social bookmarking enhance search in the web? *ACM/IEEE-CS joint conference on Digital libraries*, vol.7, pp.107-116, 2007.

[10] Agichtein, E., E. Brill, and S. Dumais, Improving web search ranking by incorporating user behavior information, *International ACM SIGIR conference on Research and development in information retrieval*, vol.29, pp.19-26, 2006.

[11] Golder, S. and B. Huberman, Usage patterns of collaborative tagging systems, *Journal of Information Science*, vol.32, no.2, pp.198-208, 2006.

[12] Liu, D., et al. Tag ranking. *International conference on World wide web*, vol.18, 2009.

[13] Bao, S., et al. Optimizing web search using social annotations, *International conference on World Wide Web*, vol.16, 2007.

[14] Jeh, G. and J. Widom. SimRank: a measure of structural-context similarity, *ACM SIGKDD international conference on Knowledge discovery and data mining*, vol.8, 2002.

[15] Apple, <http://www.apple.com>.

[16] Joachims, T. Optimizing search engines using clickthrough data, *ACM SIGKDD international conference on Knowledge discovery and data mining*, vol.8, 2002.

[17] Open Directory Project, <http://www.dmoz.org/>.

[18] Järvelin, K. and J. Kekäläinen, Cumulated gain-

based evaluation of IR techniques. *ACM Transactions on Information Systems*, vol.20, no.4, pp.422-446, 2002.

[19] Choi,Sung-Hwan and Chung,Young-Mee. The Impact of Combining Term Weights on Retrieval Effectiveness, *Korean Institute of Information Scientists and Engineers*, vol.29, no.1, pp.481-483, 2002.



임 영 석

2008년 KAIST 전산학과(학사). 2008년~2010년 서울대학교 컴퓨터공학부(석사). 관심분야는 SNS, 데이터베이스, 웹 2.0



이 강 표

2004년 연세대학교 컴퓨터학과(학사) 2006년 서울대학교 컴퓨터공학부(석사) 2007년~현재 서울대학교 컴퓨터공학부 박사과정 재학중. 관심분야는 데이터베이스, 웹2.0, 정보검색, 시맨틱웹



김 현 우

2007년 KAIST 전산학과(학사). 2007년~현재 서울대학교 컴퓨터공학부 석박통합과정 재학중. 관심분야는 웹2.0, 태그, 데이터베이스



안 재 민

2008년 연세대학교 컴퓨터학과(학사) 2008년~2010년 서울대학교 컴퓨터공학부(석사). 관심분야는 웹2.0, 스캠 필터링, 데이터베이스



김 형 주

1978년~1982년 서울대학교 컴퓨터공학부(학사). 1982년~1985년 University of Texas at Austin Computer Science(석사). 1985년~1988년 University of Texas at Austin Computer Science(박사). 1991년~현재 서울대학교 공과대학 컴퓨터공학부 교수. 관심분야는 데이터베이스, XML, 시맨틱 웹, 전자상거래, IT 정책