태그 네트워크를 이용한 개인화 북마크 추천시스템

Personalized Bookmark Recommendation System Using Tag Network

엄태영(Tae Young Eom)*, 김우주(Wooju Kim)**, 박상언(Sangun Park)***

초 록

웹 2.0을 이끌어가는 원동력이라고 할 수 있는 일반 개인 사용자의 참여와 공유는 블로그, 소셜 네트워크(Social Network), 집단지성, 소셜 북마크(Social Bookmark), 태깅(Tagging) 등의 다양한 형태로 나타나고 있다. 이 중에서 소셜 북마크는 개인이 사용하는 북마크를 웹에 추가하여 공유함으로써, 다수의 사람들이 유용하다고 생각하는 북마크에 대한 정보를 기반으로 한 다양한 서비스를 제공하는 개념이다. 딜리셔스(Delicious.com)는 소셜 북마크 서비스의 대표적인 사례라고 할 수 있으며, 북마크에 사용자들이 붙인 태그를 이용하여 검색 서비스를 제공한다. 본 논문은 북마크 검색에 대해 개인화된 검색결과를 추천하기 위하여 사용자 태그를 기반으로 하여 딜리셔스가 제공하는 북마크들의 순위를 재순위화 하는 방법론을 제안하였다. 또한 태그유사도를 기반으로 한 태그 네트워크를 이용하여 사용자의 검색어에 의미적으로 유사한 다른 태그들도 순위에 반영될 수 있도록 하였다. 그리고 실험을 통하여 딜리셔스가 제시하는 순위에 비해 본 논문에서 제안하는 시스템의 재순위화 결과가 사용자들에게 더 만족스러우며 정확성도 높음을 확인하였다.

ABSTRACT

The participation and share between personal users are the driving force of Web 2.0, and easily found in blog, social network, collective intelligence, social bookmarking and tagging. Among those applications, the social bookmarking lets Internet users to store bookmarks online and share them, and provides various services based on shared bookmarks which people think important.Delicious.com is the representative site of social bookmarking services, and provides a bookmark search service by using tags which users attach to the bookmarks. Our paper suggests a method re-ranking the ranks from Delicious.com based on user tags in order to provide personalized bookmark recommendations. Moreover, a method to consider bookmarks which have tags not directly related to the user query keywords is suggested by using tag network based on Jaccard similarity coefficient. The performance of suggested system is verified with experiments that compare the ranks by Delicious.com with new ranks of our system.

키워드 : 웹 2.0, 태그, 북마크, 태그 네트워크, 소셜 북마킹
Web 2.0, Tagging, Tag Network, Bookmark, Social Bookmarking

이 논문은 2010년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(한국연구재단-2010-327-B00182).

* 파인디지털 컨버전스사업팀
** 연세대학교 공과대학 산업정보시스템공학과
*** 교신저자, 경기대학교 경상대학 경영정보학과

2010년 09월 20일 접수, 2010년 10월 16일 심사완료 후 2010년 11월 05일 게재확정.

1. 서 론

웹 2.0은 개방과 참여, 공유로 대표되는 인터넷 환경으로 정의되고 있으며[4], 그 중에서도 전문가가 아닌 일반 개인 사용자의 참여와 공유가 웹 2.0을 이끌어가는 원동력이라고 할 수 있을 것이다. 블로그, 소셜 네트워크(Social Network), 집단지성, 소셜 북마크(Social Bookmark), 태깅(Tagging) 등은 이와 같이 개인의 참여와 공유에 의해 활발하게 제공되고 있는 서비스로서, 웹 2.0의 대표적인 분야라고 할 수 있다. 그 중에서도 소셜 북마크는 많은 사람들이 사용하고 있는 북마크를 개인의 컴퓨터에만 저장하는 대신 웹에 업로드하여 공유함으로써 보다 다양한 서비스를 제공하는 서비스이다. 북마크의 공유를 통해 많은 사람들이 유용하게 생각하는 북마크에 대한 정보를 얻을 수 있다는 장점이 있다. 북마크 정보를 공유함으로써 협동적인 웹 브라우징을 지원하고자 하는 연구는 국내에서도 이미 오래 전에 제시된 바 있다[16]. 최근에는 웹 2.0의 유행과 더불어 북마크와 태그를 활용한 검색에 대한 연구가 매우 폭발적으로 진행되고 있다. 태그를 검색대상에 추가하거나 태그 유사도를 측정하여 블로그를 추천하고자 한 연구[12, 14]가 있으며, 클러스터링 기법을 활용하여 태그들을 군집화[19]하고 그 결과를 추천에 활용한 연구[6]도 있다. 그 외에 북마크와 태그를 동시에 연구 대상으로 이용한 연구[13, 18]도 있으며, 태그 자체를 추천하거나[17] 태그 연관 그래프를 만들어 검색에 활용하고자 한 연구[11]도 있다.

딜리셔스(Delicious.com)는 소셜 북마크 서비스의 대표적인 사례라고 할 수 있다. 딜리셔스에서는 북마크에 사용자들이 붙인 태그를 이용하여 검색 서비스를 제공한다. 이는 소셜 태깅(Social Tagging)에 의한 포크소노미(Folksonomy)방식으로[3], 사용자들이 각 북마크에 붙인 태그들을 인덱스로 활용하여 사용자의 검색어와 가장 연관성이 높은 북마크를 제공하는 서비스이다. 일반 개인 사용자들이 붙인 태그를 활용함으로써 검색의 정확도를 높인다는 점에서 웹 2.0의 개념을 잘 활용한다고 볼 수 있을 것이다. 태그를 기반으로 한 분류체계는 미리 정의한 분류체계의 택소노미에서 벗어나 유연하고 역동적인 정보를 분류할 수 있는 구조를 제공한다[19]. 따라서 전통적인 정보검색 환경과 달리 태그는 정보에 대한 직관적인 이해를 기반으로 정보의 검색이 가능하도록 돕고, 이용자들이 검색 과정에서 우연한 발견을 통해 정보를 획득할 수도 있도록 한다[19].

그러나 이와 같은 장점에도 불구하고 현재 딜리셔스의 검색 서비스는 첫째, 모든 이용자에게 동일한 검색결과를 제공함으로써 개인화가 되고 있지 못하며, 둘째, 태그를 직접적으로 비교함으로써 의미적으로 관계가 있는 다른 태그가 달린 북마크에 대해 고려가 되지 못하고 있다는 단점이 있다. 둘째로 제시된 문제점을 보완하기 위하여, 딜리셔스에서는 검색결과를 보여주는 화면에서, 사용자가 이용한 검색어와 연관성을 가진 태그들을 필터로 사용할 수 있도록 하고 있다. 그러나, 처음부터 이러한 연관관계가 검색순위에 영향을 끼치지는 않고 있다. 본 연구는 태그 네트워크를 사용함으로써 사용자가 수동으로 태그 필터를 사용하여 검색순위를 조절하는 수고

를 덜 수 있으며, 개인화된 결과를 제공함으로써 사용자는 자신의 취향에 가까운 북마크를 상위 랭크로 추천 받을 수 있게 지원할 수 있다. 즉 본 연구에서는 사용자가 평소에 주로 사용한 태그들을 검색에 반영함으로써 개인화된 검색을 제공하고, 태그 간의 연관 관계로 구성된 태그 네트워크를 구축 및 사용함으로써 사용자의 검색 의도에 보다 적합한 검색 결과를 제공하고자 한다.

본 논문에서 제안하는 시스템은 딜리셔스에서 제공하는 순위를, 개인화 관점과 태그 간의 연관성 관점을 반영하여 재순위하는 방식으로 구축되었다. 개인화된 순위를 제공하기 위하여 사용자가 지금까지 딜리셔스에 등록한 태그 정보를 태그 벡터로 변환하여, 태그 벡터로 변환된 각 북마크의 태그 집합과 비교함으로써, 각 북마크와 사용자 태그와의 유사도를 계산하였다. 사용자가 검색에 사용한 검색어들도 태그 벡터로 변환하여 마찬가지로 각 북마크에 대한 태그 벡터와의 유사도를 계산하고 이 두 유사도를 합하여 북마크들을 재순위화 하였다. 또한 태그 간의 연관 관계를 반영하기 위하여 태그 쌍에 대한 북마크에서의 동시출현빈도를 이용하여 태그 네트워크를 구축하였으며, 구축된 태그 네트워크를 이용하여 북마크의 태그 벡터를 변환함으로써 북마크에 직접적으로 붙어 있는 태그 외에 다른 연관 태그들도 순위화 과정에서 고려될 수 있도록 하였다.

본 논문의 구조는 다음과 같다. 먼저 관련 연구를 제 2장에서 정리하고, 제 3장에서는 본 논문에서 제안하고자 하는 태그 네트워크와 개인화 검색에 대하여 정리하였다. 제 4장에서는 구축된 시스템에 대한 성능평가를 기술하였으며, 마지막으로 제 5장에서는 본 연구의 결론 및 향후 연구 방향에 대하여 정리하였다.

2. 관련 연구

2.1 태그와 북마크 활용에 대한 관련 연구

최근 웹 2.0의 유행과 더불어 북마크와 태그를 활용하고자 하는 연구가 폭발적으로 진행되고 있다. 태그 정보를 활용하고자 한 최근의 연구로 태그를 확장하여 태그 유사도를 측정함으로써 이미 구독중인 블로그와 유사한 블로그를 추천하고자 한 연구가 있다[14]. 이 연구에서는 각 블로그에 달려 있는 태그들에 대해 유사도를 계산하는 식을 제안함으로써 블로그의 유사성을 평가하고 추천하였다. 구독 중인 블로그가 아닌 새로운 블로그에 대한 검색을 지원하지 않는다는 점, 간단한 유사도 계산식에 의해 블로그 유사도가 계산된다는 점에서 본 연구와는 많은 차이가 있다. 블로그 검색에 태그를 활용한 다른 연구로는 검색대상에 블로그 본문만을 허용한 경우와 본문 외에 태그를 추가하여 검색대상을 사용한 경우를 비교하여 후자가 더 나은 성능을 보이고 있음을 보인 연구가 있다[12]. 성능의 검증 측면에서 시사점이 있으나, 방법론 측면에서는 단순히 태그를 검색범위 안에 포함한 것이라 할 수 있다.

태그 간의 직접적인 유사성 대신에 태그들이 함께 출현하는 빈도를 이용하고자 한 연

구로는, 연관 태그를 군집화 하는데 클러스터링 기법을 적용하는데 있어, 동시출현빈도에 기반한 다양한 태그간 유사도 함수와 클러스터링 알고리즘을 적용함으로써 여러 클러스터링 기법을 비교하고자 한 연구가 있다[19]. 그러나 이 연구에서는 딜리셔스가 제공하는 태그의 동시출현빈도 대신 직접 추출한 샘플링된 문서들을 기반으로 직접 동시출현빈도를 구한다는 점에서, 딜리셔스의 검색결과를 향상시키는데 사용하기에는 정확성의 한계가 있다. 이 외에 태그 유사성 대신 태그 출현에 대한 조건부확률을 이용하여 웹 페이지 간의 유사성을 측정하는 최근의 연구가 있다[10]. 이 연구에서는 웹 페이지들을 태그에 따라 K개의 추상클래스로 나누고, 각 추상클래스에서 태그들이 나타날 확률을 구한 다음 이를 기반으로 하여 두 웹 페이지에 있는 모든 태그들이 동일한 추상 클래스에서 나타날 확률을 구함으로써 웹 페이지 간의 유사성을 측정하였다. 이는 두 웹 페이지에 있는 태그들 간의 직접적인 유사도를 이용하여 웹 페이지 간의 유사성을 측정하는 방법에 비해 더 나은 성과를 보이는 것으로 제시되고 있다. 그러나 앞선 연구와 마찬가지로 특정 검색어를 기반으로 검색을 수행하는 시스템에서의 활용방안은 제시되지 않고 있다.

북마크와 태그를 동시에 고려한 연구 중에서 흥미로운 주제로, 소셜 북마킹을 이용하는 사용자가 고의적으로 시스템을 악용하는 스패머인지를 그 사용자가 사용한 태그를 분석함으로써 판별하고자 한 연구가 있다[13]. 또한 딜리셔스와 같은 협력적 북마킹 서비스를 대상으로 하여 사용자의 태깅 행태를 분석함으로써 웹 문서에 다는 태그의 수는 사용자 특성에 더 많이 기인하며, 따라서 사용자에 대한 보상과 인센티브를 부여해야 한다고 제시한 논문[18]도 있다.

본 논문이 이미 존재하는 북마크에 대한 태그 정보를 이용하여 웹 페이지를 추천하고자 하는 것에 비해 북마크에 달 태그 자체를 추천하는 연구로, 로지스틱 회귀분석을 이용하여 소셜 북마크 시스템에서 등록하고자 하는 북마크에 대해 태그를 추천하고자 한 연구가 있다[17]. 이 연구에서는 후보 키워드를 소스로부터 추출하고 키워드 정확도에 기반하여 가중치를 부여한 후에, 그 결과에 대해 로지스틱 회귀분석 기법을 적용하여 사용자 자료, 동일한 리소스 전체 태그집합, 사용자 전체 태그집합에서 고루 가중치가 높은 키워드를 태그로 추천하였다.

2.2 태그 정보를 기반으로 한 개인화 검색

태그 연관 그래프를 그래픽 기반 태그 연관 검색에 활용하고자 한 연구로 김운용[11] 등의 연구가 있다. 이 연구에서는 태그 연관 그래프를 관리하기 위한 알고리즘을 제시하고 그래픽 인터페이스 환경에서 이 연관성을 따라 검색을 진행하는 방안을 제시하고 있다. 그러나, 가장 중요한 태그 연관성을 얻을 수 있는 방안이 제시되어 있지 않고, 검색 방식도 사용자가 직접 태그 연관 그래프를 탐색하는 형태로 제안하고 있어 일반적인 검색도구로는 실용성이 낮다고 할 수 있다.

태그의 활용과 관련하여 국내에서 진행된 가장 최근의 연구 중 블로그와 인터넷 카페에서 많이 사용되는 태그를 개인화 검색에

활용하는 연구[15]가 있다. 이 연구는 태그 카테고리가 있다는 가정 하에 이 태그 카테고리를 이용하여 사용자의 관심 프로파일을 생성하고 기존의 검색엔진이 제시한 웹 페이지의 순위를 이 관심 프로파일을 이용하여 수정하는 형태로 개인화 검색방안을 제시하고 있다. 먼저 사용자의 관심 프로파일은 이미 생성되어 있는 태그 카테고리에 개인의 관심 순위를 할당하여 작성하는데, 이는 사용자가 모든 태그 카테고리에 자신의 관심도를 차별화하여 점수로 할당해야 한다는, 현실적으로 쉽지 않은 과정을 이용하고 있다. 이렇게 해서 사용자의 관심 카테고리에 속한 태그의 점수를 기존 검색엔진이 제시한 순위의 조정에 이용함으로써 개인화된 순위를 제시하고 있다. 본 연구는 딜리셔스가 제시하는 검색순위를 재조정한다는 점에서 이 연구와 유사점이 있으나, 사용자의 태그에 대한 관심도를 자동으로 얻어오기 위한 방안을 제시하고 있으며, 단순히 사용자의 관심도만을 반영하는 것이 아니라, 태그 네트워크를 이용하여 사용자가 사용한 검색어와의 연관성을 검색과정에 이용한다는 점에서 차별성이 있다.

Shepitsen 등[6]은 쿼리 태그를 바탕으로 추천 웹 페이지들을 수집하고, 태그 정보로 구성된 사용자의 프로파일과 태그 클러스터를 이용하여 개인화 추천 웹 페이지들을 제시하였다. 사용자 프로파일의 태그와 웹 페이지들의 태그들을 대상으로 클러스터링 작업을 수행하고, 그 결과와 선택된 태그로부터 추천된 웹 페이지들과의 유사성을 비교해 추천하는 방식을 사용하였다. 사용자 프로파일상의 태그와 웹 페이지들이 가지고 있는 태그들을 클러스터링하기 때문에 선택된 군집 외에 다른 군집에 속한 태그들의 정보는 활용되지 않는다는 단점이 있다. 이에 비해 본 연구에서는 전체 태그 네트워크의 정보를 이용함으로써 군집이 다른 경우라도 네트워크상에 연결이 되어 있으면 관련 태그들의 정보를 보다 구체적으로 반영할 수 있다는 장점이 있다. 이 외에 딜리셔스 사이트에서 각 웹 페이지에 대해 등록한 사용자들의 태그 일치도를 통해서 사용자에게 웹 페이지를 추천한 연구가 있으며[2], 페이지 랭크를 이용하여 초기 검색결과를 생성한 후 각 검색 결과와 사용자 프로파일을 비교하여 검색결과를 재순위화 함으로써 개인화 검색결과를 제공하고자 한 연구가 있다[5]. 마지막으로 사용자의 관심과 페이지 주제에 대한 코사인 유사도를 사용자 쿼리와 웹 페이지의 코사인 유사도와 합하여 검색 결과를 제공하고자 한 연구가 있다[9]. 이 연구는 기존의 연구들을 이용하고 있다는 점에서 아이디어 측면의 참신성은 떨어지나 다양한 실험을 통한 분석을 수행하였다.

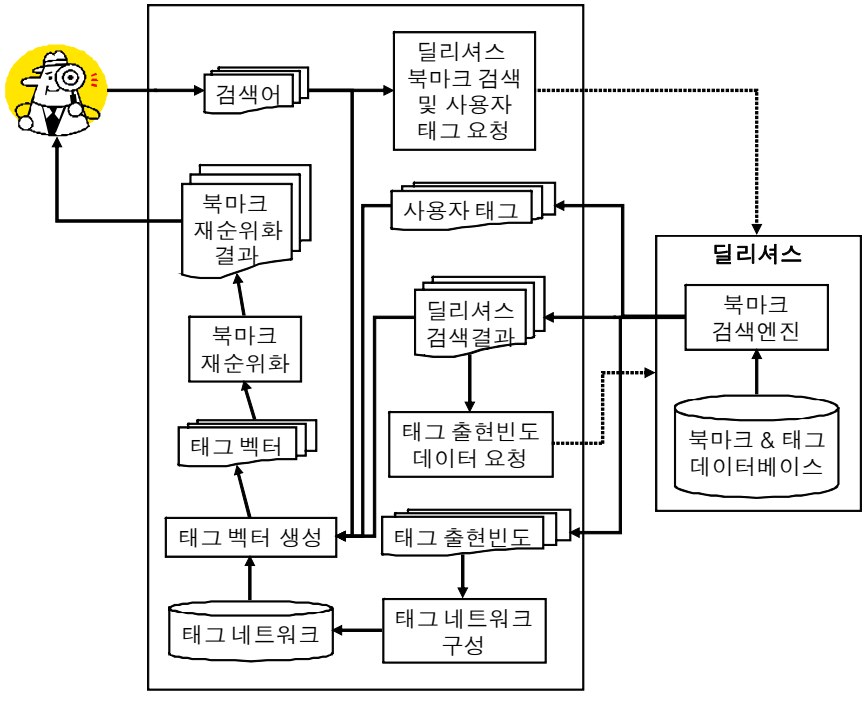
3. 태그 네트워크를 이용한 개인화 북마크 추천시스템

3.1 개인화 북마크 검색 시스템 방안과 시스템 구성

본 논문에서 제안하고자 하는 시스템의 목적은, 첫째 사용자가 평소에 사용한 태그를 검색 결과에 반영하여 사용자의 취향에 따라 개인화된 검색결과를 제공하고, 둘째 태그 간

의 연관성을 나타내는 태그 네트워크를 구성하여 이용함으로써 사용자가 검색에 이용한 검색어가 직접 태그로 등록되어 있지 않더라도 태그 네트워크 상에서 의미적으로 가까운 태그가 사용된 북마크의 검색순위를 높이는 것이다. 따라서 시스템은 크게 태그 네트워크를 구성하고 이를 이용하는 부분과 개인화된 북마크 검색 결과를 제공하는 부분으로 나뉘게 된다.

<그림 1>은 본 논문에서 제안하고자 하는 북마크 검색 시스템의 구조를 보여준다. 본 논문에서는 딜리셔스가 제공하는 북마크 검색결과의 순위를 조정함으로써 제시한 두 목표를 달성하고자 한다. 따라서 사용자가 제시한 검색어를 이용하여 먼저 딜리셔스로부터 일정 순위까지의 북마크 검색결과를 가져온다. 다음 단계에서는 검색결과에 있는 모든 북마크에 대해 태그를 추출하여 각 태그들의 출현빈도와 가능한 태그 쌍의 출현빈도를 딜리셔스로부터 가져와 이 결과를 바탕으로 태그 네트워크를 구축한다. 태그 네트워크의 상세한 구축 방안에 대해서는 제 3.2절에서 설명하고자 한다. 태그 네트워크가 완성되면 이 정보를 기반으로 하여 사용자가 제시한 검색어, 사용자가 지금까지 사용한 태그, 그리고 각 북마크에 대하여 태그 벡터를 생성하고 이 태그 벡터를 비교함으로써 딜리셔스가 제공한 북마크 검색결과를 재순위화하고, 그 결과를 사용자에게 제시한다. 태그 벡터의 생성 방법에 대해서는 제 3.3절에서 상세히 설명하



〈그림 1〉 개인화 북마크 검색 시스템 구조도

고자 한다.

3.2 태그 네트워크의 구축과 태그간 유사도 계산

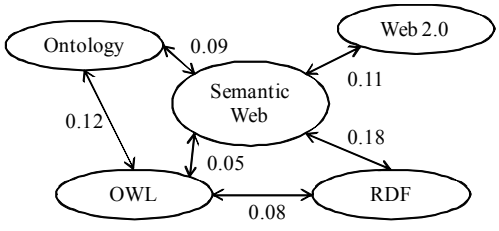
태그 네트워크의 목적은 딜리셔스가 제공한 북마크들에 달려 있는 태그들 간의 연관성을 표현하는데 있다. 본 논문에서는 Begelman 등[2]이 제시한, 태그 수를 바탕으로 한 태그 관계도 구성에 관한 연구를 참조하여, Jaccard Similarity Coefficient[8]를 사용함으로써 각 태그 쌍의 유사도를 계산하였다. Jaccard Similarity Coefficient의 계산식은 다음과 같다.

$$J(A,\ B) = \frac{|A \cap B|}{|A \cup B|}$$

딜리셔스는 특정 태그집합에 대해 태그집합이 사용된 북마크의 수를 제공한다. 이를 이용하여 모든 태그들의 출현빈도와 태그 쌍의 출현빈도를 얻는 것이 가능하다. 예를 들어 OWL과 RDF에 대해 Jaccard Similarity Coefficient를 계산하고자 한다면, OWL과 RDF 그리고 OWL, RDF 쌍에 대해 출현빈도를 얻어온 후 다음 식과 같은 방법으로 둘 간의 유사도를 계산하는 것이 가능하다.

$$J(S\_OWL,\ S\_RDF) = \frac{|S\_OWL \cap S\_RDF|}{|S\_OWL| + |S\_REF| - |S\_OWL \cap S\_RDF|}$$

위 식에서 S_OWL은 OWL 태그가 사용된 북마크들의 집합, S_RDF는 RDF 태그가 사용된 북마크들의 집합을 나타낸다. <그림 2>는 Ontology, OWL, RDF, Semantic Web, Web 2.0이라는 5개의 태그에 대해 구축한 태그 네크워크의 예를 보여 준다.



〈그림 2〉 Jaccard Similarity Coefficient를 이용한 태그 네트워크

위의 태그 네트워크에서 직접 연결되지 않는 태그 쌍에 대해서는 유사도의 곱을 이용하였다. 두 태그 간에 가능한 경로가 하나 이상인 경우에는 모든 경로에 대해 유사도를 계산하고 그 중에서 최대값을 선택하여 해당 태그 쌍의 유사도로 정의하였다. 예를 들어 위 태그 네트워크에서 OWL과 Web 2.0사이에는 OWL → Ontology → Semantic → Web → Web 2.0, OWL → Semantic Web → Web 2.0, OWL → RDF → Semantic → Web → Web 2.0의 세 경로가 존재한다. OWL → Ontology → Semantic → Web → Web 2.0 경로에서 OWL과 Web 2.0의 유사도는 0.12*0.09*0.11 = 0.0012이고 OWL → RDF → Semantic Web → Web 2.0 경로에서의 유사도는 0.08*0.18*0.11 = 0.0016인 반면, OWL → Semantic Web → Web 2.0 경로의 유사도는 0.05*0.11 = 0.0055이다. 따라서 OWL과 Web 2.0의 최종 유사도는 0.0055가 된다.

3.3 개인화 북마크 추천과정

개인화 북마크 추천과정에서는 딜리셔스에서 검색결과로 받은 북마크들을 대상으로 첫째, 사용자 검색어와의 유사성을 계산하고, 둘째, 사용자가 평소에 사용한 태그와의 유사성을 계산하여 이 둘을 합산함으로써 각 북마크들의 순위를 결정한다. 따라서 이 과정의 입력은 사용자 검색어, 사용자 태그 그리고 딜리셔스의 검색결과이고, 태그 간 유사성을 계산하는 과정에서 태그 네트워크를 이용하게 된다.

3.3.1 개인화 북마크 추천을 위한 태그 벡터의 생성

여기서 설명되는 내용은 <그림 1>의 "태그 벡터 생성" 과정에 해당된다. <그림 1>과 앞에서 설명된 바와 같이 개인화 북마크 추천을 위해서는 사용자 검색어, 사용자 태그 그리고 딜리셔스 검색결과가 요구된다. 검색결과로 받은 북마크와 사용자 태그, 사용자 검색어는 각각 다음과 같이 표현한다.

$B = \{b_1, b_2, b_3, \cdots, b_n\}$: 딜리셔스 검색결과 북마크들의 집합

$BT_j = \{tb_1, tb_2, tb_3, \cdots, tb_k\}$: 각 북마크의 태그 집합

$UT = \{tu_1, tu_2, tu_3, \cdots, tu_l\}$: 사용자 태그 집합

$QT = \{tq_1, tq_2, tq_3, \cdots, tq_p\}$: 사용자 검색어 태그 집합

개인화 북마트 추천에 대한 이해를 돕기 위해 다음과 같이 간단한 예제를 설정하였다. 먼저 사용자 검색어는Semantic Web이고, 사용자가 지금까지 사용한 태그는 OWL이라고 하자. 두 개의 북마크를 딜리셔스의 검색서비스로부터 제공받았는데 첫째 북마크에 붙은 태그는 Ontology, RDF, Semantic Web이고 둘째 북마크에 붙은 태그는 RDF, Web 2.0, Semantic Web라고 하자. 이 예제를 위에서 제시한 표현으로 나타내면 다음과 같다.

$B = \{b_1, b_2\}$,
$BT_1 = \{\text{Ontology, RDF, Semantic Web}\}$,
$BT_2 = \{\text{RDF, Web 2.0, Semantic Web}\}$,
$UT = \{\text{Semantic Web}\}$,
$QT = \{\text{OWL}\}$

개인화 검색을 위한 준비단계의 첫 단계로 위의 북마크, 사용자 태그 집합을 모두 합하여 다음과 같이 전체 태그집합을 생성한다.

$$TS = LT_1 \cup LT_2 \cup \cdots \cup LT_n \cup UV \cup QT$$
$$= \{t_1, t_2, t_3, \cdots\cdots t_m\} : \text{전체 태그 집합}$$

예제에서 TS는 {Ontology, RDF, OWL, Web 2.0, Semantic Web}이 된다. 이 전체 태그 집합을 이용하여 다음과 같이 사용자 태그, 사용자 검색어에 대하여 최초의 태그 벡터를 생성한다.

V_{user}(사용자 태그에 대한 태그 벡터) = $[u_1, \cdots, u_i, \cdots, u_m]$, ui = 1 : t_i가 UT에 있음, u_i = 0 : t_i가 UT에 없음.

V_{query}(사용자 검색어에 대한 태그 벡터) = $[q_1, \cdots, q_i, \cdots, q_m]$, q_i = 1 : t_i가 QT에 있음, q_i = 0 : t_i가 QT에 없음.

예제의 사용자 태그에 대해태그 벡터를 생성해 보면, TS의 다섯 개 태그 중에서 OWL만 있으므로 V_{user} = [0, 0, 1, 0, 0]이 된다. 마찬가지 방법으로 사용자 검색어는 Semantic Web만 있으므로 태그 벡터를 생성하면 V_{query} = [0, 0, 0, 0, 1]이 된다.

검색결과로 온 북마크들에 대한 태그 벡터는 사용자 태그, 사용자 검색어와는 조금 다르게 생성된다. 예를 들어 예제의 첫째 북마크에 대해 태그 벡터를 생성하면, V1 = [1, 1, 0, 0, 1]이 되는데, 이 경우 각 태그의 상대적 중요도가 반영되지 않는다는 단점이 있다. 즉 사람들이 첫째 북마크에 대하여 RDF보다 Ontology라는 태그를 더 많이 달았다면 이를 반영해 줄 필요가 있다. 따라서 각 북마크에 대한 태그 벡터는 다음과 같이 정의한다.

V_j(각 북마크의 태그 벡터) = [v_{j1}, …, v_{ji}, …, v_{jm}], v_{ji} = $\frac{\text{북마크에 } t_i \text{ 태그를 단 사람의 수}}{\text{북마크에 태그를 단 전체인원수}}$

예를 들어 첫째 북마크에서 이 북마크에 태그를 단 전체인원이 1,369명이고, 그 중에서 Ontology 태그를 단 사람의 수가 683명이라면 Ontology 태그에 대한 벡터 값은 683/1369 = 0.50이 된다. 이와 같은 방식으로 두 북마크에 대해 태그 벡터를 계산하면, V_1= [0.5, 0.2, 0, 0, 0.1], V_2 = [0, 0.4, 0, 0.2, 0.25]가 된다.

3.3.2 태그 네트워크를 이용한 태그 간 유사도의 반영

위 단계에서 만들어진 사용자 태그의 태그 벡터와 사용자 검색어 태그벡터의 문제점은 각 태그 간의 유사도가 반영되지 않았다는 점이다. 예를 들어 사용자는 태그로 OWL 하나를 사용하였지만 OWL은 Semantic Web, RDF, Ontology, Web 2.0의 의미를 어느 정도 내포하고 있다. 따라서 이를 사용자 검색어 태그 벡터에 반영하여야 한다. 이 때 앞 과정에서 만든 태그 네트워크를 활용하게 된다. 제 3.2절에서 설명한 바와 같이 OWL은 Web 2.0과 0.0055라는 유사도를 갖는다. 마찬가지의 방법으로 계산하면 OWL과 Ontology와는 0.12, OWL과 RDF와는 0.08, OWL과 Semantic Web과는 0.05의 유사도를 갖게 된다. 이를 반영하여 앞에서 만든 태그 벡터를 수정한 결과는 다음과 같다.

V_{user} = [0.12, 0.08, 1, 0.0055, 0.05],
V_{query} = [0.09, 0.18, 0.05, 0.11, 1]

3.3.3 개인화 북마크 추천을 위한 북마크 재순위화

여기에서 설명되는 내용은 <그림 1>의 “북마크 재순위화” 과정에 해당된다. 재순위화에 앞서 먼저 앞에서 계산한 태그 벡터들을 단위벡터로 변환한다. 변환결과는 다음과 같다.

V_1 = [0.91, 0.36, 0, 0, 0.18],
V_2 = [0, 0.78, 0, 0.39, 0.49]
V_{user} = [0.12, 0.08, 1, 0.0055, 0.05],
V_{query} = [0.09, 0.18, 0.05, 0.11, 0.97]

재순위화를 위해서는 두 북마크 태그 벡터 V_1, V_2에 대해 각각 V_{user}, V_{query}와의 유사도

를 계산하여야 한다. 유사도 계산을 위해, 정보검색 실험에서 일반적으로 널리 쓰이고 있는 코사인 유사계수(Cosine Coefficient)를 사용하였다. 사용된 코사인 유사계수의 공식은 다음과 같다[7].

$$\cos(x, y) = \frac{\Sigma_i (x_i y_i)}{\sqrt{\Sigma_i x_i^2 \Sigma_i y_i^2}}$$

$Sim_{user}(V_i)$를 V_i와 V_{user}와의 코사인 유사도라 하고, $Sim_{query}(V_i)$를 V_i와 V_{query}와의 코사인 유사도라고 할 때 V_i에 대한 최종 유사도는 다음과 같이 정의된다. 딜리셔스가 제공한 북마크들은 이 최종 유사도의 값에 따라 재순위화 된다.

$$Sim_{total}(V_i) = \partial \cdot Sim_{user}(V_i) + (1-\partial) \cdot Sim_{query}(V_i)$$

위 식에서 ∂는 두유사도 간에 어느 쪽에 더 가중치를 둘 것인가를 결정하는 계수이다. 여기서는 동일하게 고려하는 것으로 가정하고 0.5를 사용하였다. 이 식에 따라 예제에 대한 최종유사도를 계산하면 $Sim_{total}(V_1)$는 0.235, $Sim_{total}(V_2)$는 0.375이 나온다. 즉, 둘째 북마크가 더 높은 순위가 됨을 알 수 있다.

4. 개인화 북마크 추천시스템 평가

4.1 개인화 북마크 추천시스템의 실행 예

구현된 개인화 북마크 추천시스템에서는 딜리셔스로부터 10개의 북마크와 각 북마크에 대한 태그 정보를 가져온다. 그리고 이 북마크 리스트와 태그 정보를 기반으로 태그 네트워크를 구성하고, 각 태그 벡터들을 생성하여 10개의 북마크 리스트를 재순위화하게 된다. 보이고자 하는 실행 예에서는 검색어를

〈표 1〉 딜리셔스의 북마크 순위

No	Title	URL
1	The ProtégéOntology Editor and Knowledge Acquisition System	Protege.stanford.edu/
2	Jena Semantic Web Framework	jena.sourceforge.net/
3	Swoogle Semantic Web Search Engine	swoogle.umbc.edu/
4	openRDF.org : Home	www.openrdf.org/
5	OWL Web Ontology Language Overview	www.w3.org/TR/owl-features/
6	SchemaWeb-RDF Schemas Directory	www.schemaweb.info/
7	OWL Web Ontology Language Guide	www.w3.org/TR/owl-guide/
8	SIMILE Project	simile.mit.edu/
9	SchemaWeb-RDF Schemas Directory	www.schemaweb.info/default.aspx
10	Web Ontology Language OWL/W3C Semantic Web Activity	www.w3.org/2004/OWL/

〈표 2〉 추천시스템의 재순위화 결과

No	Title	URL
1	OWL Web Ontology Language Overview	www.w3.org/TR/owl-features/
2	OWL Web Ontology Language Guide	www.w3.org/TR/owl-guide/
3	Web Ontology Language OWL/W3C Semantic Web Activity	www.w3.org/2004/OWL/
4	Jena Semantic Web Framework	jena.sourceforge.net/
5	SchemaWeb-RDF Schemas Directory	www.schemaweb.info/
6	SchemaWeb-RDF Schemas Directory	www.schemaweb.info/default.aspx
7	Swoogle Semantic Web Search Engine	swoogle.umbc.edu/
8	SIMILE Project	simile.mit.edu/
9	openRDF.org : Home	www.openrdf.org/
10	The ProtégéOntology Editor and Knowledge Acquisition System	Protege.stanford.edu/

OWL로 하였으며, 사용자 태그 집합은 {API, Development, Framework, Java, Ontology}로 설정하였다. 딜리셔스로부터 가져온 북마크의 순위는 <표 1>과 같다.

개인화 북마크 추천시스템의 재순위화 결과는 <표 2>와 같다.

4.2 추천시스템의 평가

추천시스템의 평가를 위하여 두 종류의 평가를 실시하였다. 첫째, 사용자들이 검색순위에 만족하는 정도를 설문을 통해 평가하였으며, 둘째, 사용자들이 10개의 북마크에 대해 스스로 순위를 작성하게 한 후에 그 결과와 비교하여 시스템이 제공하는 순위에 대한 정확성 평가를 실시하였다. 각 평가는 공정하게 딜리셔스가 제공한 원래의 순위와 재순위화한 결과에 대해 모두 실시하였다. 즉, 각 사용자들이 작성한 순위에 대하여 딜리셔스가 제공한 순위의 정확성을 평가하여 이를 합산하고, 다시 사용자들이 작성한 순위에 대해 본 논문이 제안하는 시스템이 제공한 순위의 정확성을 평가하여 이를 합산하였다.

사용자들의 검색순위에 대한 만족도 평가를 위해 20명의 사용자들을 대상으로 설문을 받아 평가를 수행하였다. 10점 척도로 하여 제 4.1절에서 제시한 사용자 태그와 검색어 환경에서 딜리셔스와 본 논문에서 제시한 재순위화 결과의 만족도를 평가한 결과, 딜리셔스에 대해서는 평균 5.4점의 만족도를 보였으며, 재순위화 결과에 대해서는 평균 7.7점의 만족도를 보임으로써 재순위화 결과에 대해 대체로 더 만족하는 것으로 나타났다.

둘째로 검색결과에 대한 정확성 평가는 30명의 사용자들로부터 10개의 북마크에 대한 개별 순위를 받은 후에 이를 딜리셔스의 순위 그리고 추천시스템의 순위와 비교하였다. 순위에 대한 평가는 검색결과의 랭크 정확성을 측정하는 NDCG(Normalized Discounted Cumulative Gain) 식을 응용하여 수행하였

다. 이 방식은 변경된 각 순위에서 기존의 순위를 뺀 값에 대해 적절성 레벨을 부여하고 이를 이용해 지표를 계산한다. 본 논문에서는 이를 응용하여 양쪽의 순위 편차를 합산하는 방법으로 결과를 계산하였다. 순위에 대한 정확성 평가식은 다음과 같다.

$$\mathrm{RankMeasure} = \sum_{i=1}^{n} (\mathrm{NewRank_i} - \mathrm{OldRank_i})$$

위의 평가식은 그 값이 작을수록 시스템이 제시한 순위의 정확도가 커지게 된다. 식에 따라 계산한 결과, 딜리셔스의 순위에 대해서는 평균 27.3의 점수가 나왔으며, 본 논문이 제안한 시스템은 평균 18.8의 점수가 나왔다. 만족도에 대한 평가와 마찬가지로 본 논문이 제안하는 시스템의 정확성이 딜리셔스의 원래 순위에 비해 높게 평가되었다.

실험을 수행한 결과, 사용자들은 전반적으로 딜리셔스에 비해 제안된 시스템의 결과에 더 만족하고 있으며, 시스템의 순위가 딜리셔스보다 정확한 것으로 나타났다. 그러나, OWL에 대해 이해하고 있는 사용자들만을 대상으로 실험을 수행했기 때문에 사용자의 수가 충분하지 않은 것이 이 실험의 한계로 작용할 수 있으며 향후 연구에서는 개선되어야 할 부분이다.

5. 결 론

본 논문은 북마크 검색에 대해 개인화된 검색결과를 추천하기 위하여 사용자 태그를 기반으로 하여 딜리셔스가 제공하는 북마크들의 순위를 재순위화 하는 방법론을 제안하였다. 또한 태그 간의 연관성을 이 과정에 반영하기 위하여 태그 유사도를 기반으로 한 태그네트워크를 구성함으로써, 사용자의 검색어에 의미적으로 유사한 다른 태그들도 순위에 반영될 수 있도록 하였다. 이렇게 함으로써 각 사용자가 본래 의도한 바에 따라 북마크들이 순위화 될 수 있도록 하는 시스템을 개발하였으며, 실험을 통하여 딜리셔스가 제시하는 순위에 비해 본 논문에서 제안하는 시스템의 재순위화 결과가 사용자들에게 더 만족스러우며 정확성도 높음을 확인하였다.

본 연구의 한계로 먼저 제안하는 시스템과 딜리셔스 시스템 간의 통합이 현재로는 쉽지 않고, 또한 사용자 수가 급격하게 증가될 경우 속도가 저하될 가능성이 있다. 이는 현재의 시스템에서 사용자가 질의를 할 때마다 태그 출현빈도를 얻어와 태그 네트워크를 생성하고 있기 때문인데, 본 연구에서는 프로토타입의 구현을 위하여 주어진 환경에서 가장 쉽게 구축이 가능한 최선의 방법을 선택했기 때문이다. 이와 같은 한계는 태그 네트워크 생성작업을 질의에 대한 응답과정과 분리하여 주기적으로 실시하고, 각 태그벡터 역시 질의 이전에 미리 생성함으로써 극복될 수 있다. 또 다른 한계로 실험에서 북마크를 10개만 사용하고 이로 인해 관련태그도 그리 많지 않다는 점을 들 수 있다. 그러나 사용자가 주목하는 검색결과는 대부분 상위랭크 10개 정도에 한정되며, 검색이 특정 주제에 한정되는 경우 관련태그도 이에 따라 결정된다는 점에서 볼 때, 연구의 타당성을 해치는 정도까지는 아닌 것으로 판단된다.

향후 연구로는 먼저, 태그간 유사도의 계

산에 의미적 유사도(Semantic Similarity)를 이용함으로써 검색순위를 더 향상시켜보고자 한다. 현재의 연구에서는 태그간 유사도가 딜리셔스에서 북마크들에 태그들이 동시 출현하는 빈도에 의해 결정되고 있으나, 위드넷 등의 외부 온톨로지를 이용하면 태그 간의 의미적 유사도를 측정하는 것이 가능하다. 이 두 방법에 대해 비교해 보고, 더 나아가 두 방법을 혼합하는 연구를 향후 연구로 진행하고자 한다. 또한 본 논문에서는 태그 벡터 간의 유사도 계산에 코사인 유사계수를 사용하고 있으나, SMM(Separable Mixture Model)을 이용함으로써 보다 많은 정보를 유사도 계산에 활용할 수 있을 것으로 기대된다.

참 고 문 헌

[1] Begelman, G., Keller, P., and Smadja, F., "Automated Tag Clustering : Improving search and exploration in the tag space," Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, 2006.

[2] Durao, F. and Dolog, P., "A Personalized Tag-Based Recommendation in Social Web Systems," International Workshop on Adaptation and Personalization for Web 2.0, Trento, Italy, 2009.

[3] Golder, S. A. and Huberman, B. A., "Using Patterns of Collaborative Tagging System," Journal of Information Science, Vol. 32, No. 2, 2006, pp. 198-208.

[4] O'Reilly, T., "What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software," http://oreilly.com/web2/archive/what-is-web-20.html, 2005.

[5] Satokar, K. D. and Gawali, S. Z., "Web Personalization Using Web Mining," International Journal of Engineering Science and Technology, Vol. 2, No 3, 2010, pp. 307-311.

[6] Shepitsen, A., Gemmell, J., Mobasher, B. and Burke, R., "Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering," Proceedings of the 2008 ACM conference on Recommender systems, Lausanne, Switzerland, 2008, pp. 259-266.

[7] Sneath, P. H. A. and Sokal, R. R., Numerical Taxonomy : the Principles and Practice of Numerical Classification. San Francisco : Freeman, 1973.

[8] Tan, P., Steinbach, M. and Kumar, V., Introduction to Data Mining, Addison Wesley, US, 2005.

[9] Xu, S. and Bao, S., "exploring folksonomy for personalized search," Annual ACM Conference on Research and Development in Information, 2008, pp. 155-162.

[10] 강상욱, 이기용, 김현규, 김명호, "태그를 이용한 웹 페이지간의 유사도 측정 방법", 한국정보과학회논문지 : 데이타베이스, 제37권, 제2호, 2010, pp. 104-112.

[11] 김운용, 박석규, "웹 2.0의 참여형 아키텍처 환경에서 그래픽 기반 포크소노미 태그 연관 검색의 설계 및 구현", 한국인터넷정보학회논문지, 제8권, 제5호, 2007, pp. 1-10.

[12] 김은희, 정영미, "사용자 태그와 중심성

지수를 이용한 블로그 검색 성능 향상에 관한 연구", 한국정보관리학회지, 제27권, 제1호, 2010, pp. 61-77.

[13] 김찬주, 황규백, "소셜 북마킹 시스템의 스패머 탐지를 위한 기계학습 기술의 성능 비교", 한국정보과학회논문지 : 컴퓨팅의 실제, 제15권, 제5호, 2009, pp. 345-349.

[14] 심학준, 윤태복, 이지형, "메타정보를 활용한 블로그 추천방법", 한국지능시스템학회 2010년도 춘계학술대회 학술발표논문집, 제20권, 제1호, 2010, pp. 96-97.

[15] 윤기상, 윤광호, 김재광, 이지형, "태그를 이용한 개인화 검색 시스템", 한국정보과학회 2009 가을 학술발표논문집, 제36권, 제2호, 2009, pp. 320-324.

[16] 정재은, 윤정섭, 조근식, "북마크 정보 공유를 통한 협동적 웹 브라우징", 한국정보과학회 2000년도 봄 학술발표논문집, 제27권, 제1호, 2000, pp. 286-288.

[17] 주상훈, 황규백, "로지스틱 회귀분석을 이용한 소셜 북마킹 시스템의 태그 추천 기법", 한국정보과학회 2009 가을 학술발표논문집, 제36권, 제2호, 2009, pp. 338-341.

[18] 최준연, 김용수, "협력적 북마킹의 태깅 행태 분석", 한국콘텐츠학회논문지, 제9권, 제7호, 2009, pp. 193-201.

[19] 한승희, "연관 태그의 군집화를 위한 클러스터링 기법 비교 연구", 한국문헌정보학회지, 제43권, 제3호, 2009, pp. 399-416.

저 자 소 개

엄태영	(E-mail : tyeom@finedigital.com)
2009년	연세대학교 정보산업공학과 (학사)
2010년	연세대학교 정보산업도시공학과 (석사)
2010년~현재	파인디지털
관심분야	시맨틱 웹 검색, U-city, 지능형 웹서비스

김우주	(E-mail : cosmos65@live.co.kr)
1987년	연세대학교 경영학과 (학사)
1989년	한국 과학기술원 경영과학과 (석사)
1994년	한국 과학기술원 경영과학과 (박사)
2004년~현재	연세대학교 정보산업공학과 교수
관심분야	시맨틱 웹, 시맨틱 웹 환경의 의사결정지원 시스템, 시맨틱 웹 마이닝, 지식관리, 인공지능 웹 서비스

박상언	(E-mail : supark@kgu.ac.kr)
1992년	한국과학기술원 전산학과 (학사)
1999년	한국과학기술원 경영공학과 (석사)
2006년	한국과학기술원 경영공학과 (박사)
2007년~현재	경기대학교 경상대학 경영정보학과 조교수
관심분야	웹기반 지능정보시스템, 시맨틱 웹 마이닝, 시맨틱 웹 검색, 온톨로지 매칭, 지능형 웹 서비스, 지능형 전자상거래