Journal of Korea Multimedia Society Vol. 14, No. 4, April 2011(pp. 567-576)

연관 태그 및 유사 사용자 가중치를 이용한 웹 콘텐츠 랭킹 시스템

박수진†, 이시화††, 황대훈†††

요 약

웹 2.0의 발전에 따라 다양한 기술들이 제공되며 그 중 대두되는 기술로 사용자가 관심 있는 웹페이지를 태깅 및 북마킹하는 소셜 북마킹 기술이다. 그러나 현재 소셜 북마킹 시스템들은 웹 콘텐츠의 중요 정보인 다른 사용자들의 관심 정도를 측정할 수 있는 북마크 수 및 검색과 분류를 목적으로 하는 태그 정보를 각각 독립적으로 검색에 활용하며 또한, 다른 사용자들과의 유사도를 반영하지 못하여 소셜 북마킹 시스템의 특징을 반영하지 못한 검색결과를 도출하고 있는 실정이다. 이에 본 연구에서는 선행 연구를 기반으로 태그 클러스터링을 통한 연관 태그 추출 및 북마크 정보와 다른 사용자의 유사도를 혼합한 웹 콘텐츠 랭킹 알고리즘을 제안하였다. 또한 제안 알고리즘의 효율성 분석을 위해 기존 검색 방법론 및 선행 연구의 방법론과의 비교평가를 시행하였으며, 그 결과 본 연구의 핵심적인 특징인 태그 정보 및 북마크 수와 유사도를 활용한 방법이 기존 방법론보다 효율적인 결과를 도출하였다.

A Web Contents Ranking System using Related Tag & Similar User Weight

Su-Jin Park†, Si-Hwa Lee††, Dae-Hoon Hwang†††

ABSTRACT

In current Web 2.0 environment, one of the most core technology is social bookmarking which users put tags and bookmarks to their interesting Web pages. The main purpose of social bookmarking is an effective information service by use of retrieval, grouping and share based on user's bookmark information and tagging result of their interesting Web pages. But, current social bookmarking system uses the number of bookmarks and tag information separately in information retrieval, where the number of bookmarks stand for user's degree of interest on Web contents, information retrieval, and classification serve the purpose of tag information. Because of above reason, social bookmarking system does not utilize effectively the bookmark information and tagging result. This paper proposes a Web contents ranking algorithm combining bookmarks and tag information, based on preceding research on associative tag extraction by tag clustering. Moreover, we conduct a performance evaluation comparing with existing retrieval methodology for efficiency analysis of our proposed algorithm. As the result, social bookmarking system utilizing bookmark with tag, key point of our research, deduces a effective retrieval results compare with existing systems.

Key words: Web 2.0(웹 2.0), Tag(태그), Web Contents(웹 콘텐츠), Social Bookmark(소셜 북마크), Ranking(랭킹), Social Network(소셜 네트워크)

※ 교신저자(Corresponding Author) : 황대훈, 주소 : 경기도 성남시 수정구 복정동 산 65번지 경원대학교 새롬관 5-14호(461-200), 전화 : 010)3099-4732, FAX : 031)757-6715, E-mail : hwangdh@kyungwon.ac.kr
접수일 : 2010년 8월 24일, 수정일 : 2010년 11월 24일
완료일 : 2010년 11월 24일

† 준회원, 경원대학교 전자계산학과
(E-mail : hohivi@gmail.com)
†† 준회원, 경원대학교 전자계산학과
(E-mail : leesihwaman@gmail.com)
††† 종신회원, 경원대학교 전자계산학과
※ 본 연구는 2010년도 경원대학교 지원에 의한 결과임.

1. 서 론

최근의 웹 서비스는 점차 동적이고 능동적으로 변화하고 있으며, 이러한 웹 서비스의 흐름을 잘 반영하는 것이 웹 2.0이다. 이러한 웹 2.0 기술 중 최근 웹 기반 북마킹 서비스와 태깅 및 소셜 기술을 도입한 소셜 북마킹 기술이 대두되고 있다[1,2].

소셜 북마킹 기술은 사용자들이 관심 있는 웹 콘텐츠들을 웹상에 즐겨찾기하는 기술로 다른 사용자들이 북마킹한 웹 콘텐츠들을 서로 공유할 수 있다는 특징이 있다. 또한 북마킹 시스템에서의 핵심적인 기술 중 하나는 태그 기술로 다양하게 존재하는 웹 콘텐츠의 검색, 분류, 공유를 통한 효율적인 정보제공을 목적을 가진다.

그러나 소셜 북마킹 시스템에서의 검색 시 잘못된 태그들로 인해 만족스럽지 못한 검색결과를 도출하고 일반 검색 알고리즘을 이용하여 단순히 저장순, 가나다순 등으로만 검색결과를 제공하고 있어 다른 사용자 간의 유사성 및 신뢰성을 가지지 못하고 소셜 북마킹의 특징 또한 반영하지 못하고 있다.

이에 본 논문에서는 소셜 북마킹의 특징인 북마크 인원수와 태깅된 태그 정보와 사용자들간의 유사도를 이용하여 웹 콘텐츠 랭킹 알고리즘을 제한한다. 이를 위해 선행 연구로 진행한 태그 클러스터링을 기반으로 연관태그를 추출하고 웹 콘텐츠의 관심도인 북마크 인원수를 이용하여 이를 정규화 및 랭킹을 통해 콘텐츠를 제공하는 시스템을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대하여 기술하고, 3장에서는 태그 정보와 북마킹 정보 및 사용자간의 유사도를 효율적으로 활용하기 위한 웹 콘텐츠 랭킹 방법을 제시한다. 4장에서는 기존 검색 기법과의 비교평가 결과를 기술하고, 마지막 5장에서는 결론 및 향후 계획에 대하여 기술한다.

2. 관련연구

2.1 웹 2.0에서의 소셜 북마킹과 태그

현재 큰 반향을 일으키고 있는 웹 2.0은 이제 모든 인터넷 사이트들의 필수 전략이 되었으며, 웹 2.0을 성공적으로 구현하기 위한 다양한 기법들이 쏟아져 나오고 있다. 이러한 기법들 중 핵심적인 기술이 소셜 북마킹과 태그이다.

소셜 북마킹은 이전에는 웹 상에서 유용한 정보를 발견하면 웹 브라우저의 즐겨찾기로 등록해 두었지만, 요즘에는 소셜 북마크를 이용하여 온라인으로 저장하고 태그를 붙여 다른 사람들과 공유한다[2]. 또한, 태그는 현재 많은 인터넷 사용자들로부터 큰 호응을 얻고 있으며, 블로그와 같은 웹 문서에서부터 이미지, 동영상 등과 같은 멀티미디어 데이터에 이르기까지 폭넓게 활용되고 있다[1,2]. 이러한 대표적인 소셜 북마킹 웹사이트로 딜리셔스(del.icio.us), 빕소노미(BibSonomy), 마가린(mar.gar.in) 등과 같은 사이트가 있다[3-5].

소셜 북마킹은 부정확한 태그들로 인해 만족스럽지 못한 검색결과를 도출하며, 또한 다른 사용자들과의 공유도 혹은 관심도인 북마크 수 및 사용자들 간의 유사성을 고려하지 않은 일반적인 검색 랭킹 알고리즘을 사용함으로써 단순히 저장순, 가나다순으로만 검색 결과를 제공하고 있어 소셜 북마킹의 특징을 제대로 활용하고 있지 못한 실정이다.

이에 본논문에서는 효율적인 소셜북마킹 시스템에서의 랭킹 시스템을 제안하기 위해 현재 웹상에 다양한 리소들을 랭킹하기 위한 대표적인 랭킹 알고리즘의 장단점을 분석하였으며, 다음 2.2와 같다.

2.2 랭킹 알고리즘

랭킹 알고리즘은 크게 내용적 측면과 구조적 측면으로 나눌 수 있다. 내용적인 측면은 키워드와 관련된 단어들의 본문 출현 빈도수 등과 같은 요소들을 기반으로 페이지의 내용을 직접 평가하여 랭킹하는 방법으로 많은 계산량이 요구된다. 반면, 구조적인 측면에서는 다른 페이지에 얼마만큼 많이 연결되어 있는지 혹은 좋은 페이지에 얼마나 많이 연결되어 있는지와 같은 연결성 평가를 기반으로 랭킹한다. 이는 내용적 측면의 방법에 비해 훨씬 적은 계산량을 필요로 한다. 최근 가장 우수한 검색 효율성을 보이는 PageRank[6]와 HITS[7]가 대표적인 구조적 측면의 평가방법이다.

PageRank[6]는 월드 와이드 웹과 같은 하이퍼링크 구조를 가지는 문서에 상대적 중요도에 따라 가중치를 부여하는 방법이다. 이 알고리즘은 웹페이지간의 인용과 참조로 연결된 임의의 묶음에 적용할 수 있다. 현재 구글 사이트 등 다양한 검색 엔진에서 페

이지랭크를 기반한 검색 기술을 사용하고 있다. 하지만 소셜 북마킹 사이트 내에서는 링크된 웹페이지를 기반으로 랭킹하기에는 부적합한 구조를 가지고 있다. 웹페이지, 블로그 글 등 다양하게 존재하는 웹 콘텐츠의 특징상 PageRank에 따른 방법론은 적합하지 않다.

HITS[7]는 웹 페이지들 간의 상호 연결된 링크 정보로부터 웹 문서들의 중요도를 평가하고, 순위 정보에 따른 결과를 제시한다. 이러한 HITS 알고리즘의 문제점은 문서 내의 링크 빈도수만을 고려하고, 입력 값으로 주어지는 웹 문서 집합의 특성에 의존적이라는 것이다.

Adar[8]은 iRank라는 랭킹개념을 제안했다. 이것은 페이지랭크에 존재하는 정보를 포함하는 사이트에 더 높은 점수를 주는 방법이다. 또한 블로그 영역에서의 이슈를 다루며 링크의 동적인 구조의 중요성을 다루고 있다. 이러한 블로그 영역내의 랭킹 알고리즘은 다양한 특징을 반영하여 랭킹함으로 본 논문에서 제안하는 알고리즘에 활용하거나 비교평가하기에 알맞은 알고리즘이라 볼 수 있다.

3. 시스템 설계

제안시스템은 그림 1과 같이 크게 태그 클러스터링 시스템(Tag Clustering System)과 소셜 네트워크 시스템(Social Network System) 및 웹 콘텐츠 랭킹 시스템(Web Content Ranking System)으로 구성된다.

태그 클러스터링 시스템은 태그가 가지는 문제점인 부정확한 태그로 인한 비효율적인 검색의 문제점을 해결하기 위해 부정확한 태그들은 제거하고 연관성이 높은 태그 그룹으로 클러스터링하기 위한 기능을 수행하며, 또한 북마커별 대표 태그를 추출하기 위해 수행된다.

소셜 네트워크 시스템은 사용자에게 신뢰성 및 유사성을 가지는 웹 콘텐츠를 제공하기 위해 북마커별 대표 태그 가중치를 생성한 뒤 이를 기반으로 유사 그룹을 생성하게 된다.

웹 콘텐츠 랭킹 시스템은 웹 콘텐츠 내 북마크 인원수와 클러스터링 시스템을 통해 추출된 연관 태그 가중치 및 소셜 네트워크 시스템에서 계산된 사용자와의 유사도 값을 혼합한 랭킹 알고리즘을 통해 새로운 의미를 가지는 검색결과를 제공한다.

본 연구에서의 태그 클러스터링 시스템은 선행연구[9]로 진행하였으며, 소셜 네트워크 시스템과 웹컨텐츠 랭킹 시스템을 중심으로 다루었다.

3.1 소셜 네트워크 시스템

태그 클러스터링과 소셜 네트워크 및 웹 콘텐츠

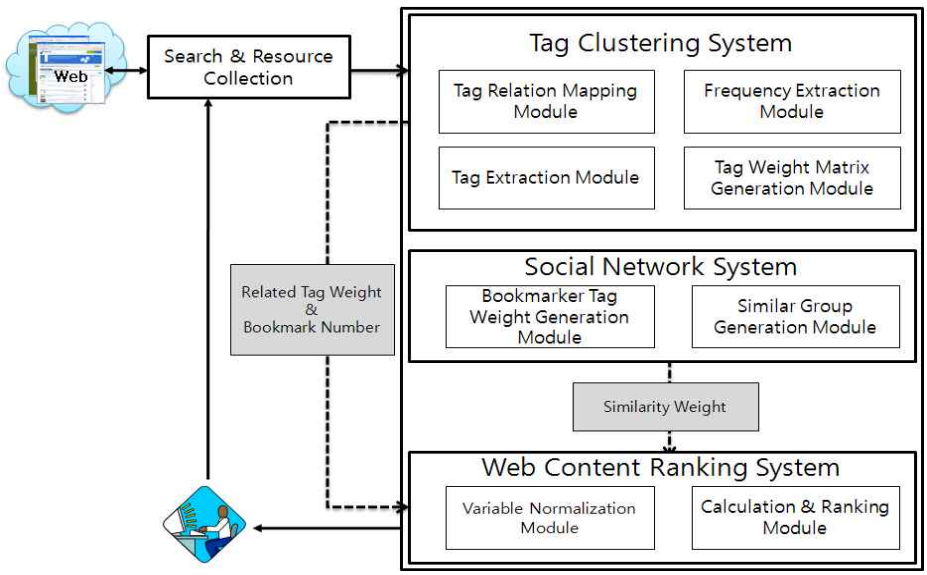


그림 1. 제안 시스템

표 1. 북마커별 대표 태그 가중치 적용 예

u1	태 그	T1	T2	T3	T5	T9	T10	합계
	태깅수	70	50	30	35	25	22	232
	가중치	34.48	17.24	12.93	15.09	10.78	9.48	100
u2	태 그	T2	T4	T6	T3	T10	T11	합계
	태깅수	87	53	40	77	44	35	326
	가중치	26.69	16.26	12.58	23.62	10.12	10.74	100
u3	태 그	T5	T1	T7	T2	T12	T6	합계
	태깅수	57	60	49	34	36	55	291
	가중치	19.59	22.68	14.78	11.68	18.90	12.37	100
u4	태 그	T6	T3	T8	T1	T4	T13	합계
	태깅수	33	28	10	15	13	14	113
	가중치	29.20	24.78	13.27	8.85	11.50	12.39	100

랭킹 시스템을 통해 다른 사용자의 관심도 및 유사도를 반영하여 '소셜'의 의미를 갖는 검색결과를 제공하기 위해 먼저, 태그 클러스터링 시스템을 통해 사용자가 태깅한 태그들을 기반으로 클러스터링하여 사용자별 대표 태그를 추출한다[9].

소셜 네트워크 시스템은 추출된 사용자별 대표 태그를 기반으로 사용자 유사 그룹을 생성하고, 생성된 유사 그룹 내 사용자와의 유사도를 통해 사용자 자신과 유사한 다른 사용자가 북마킹한 웹 콘텐츠를 상위의 랭킹 결과로 얻을 수 있는 장점을 가지게 된다. 이로써 사용자는 자신이 보지 못했던 관심 웹페이지를 제공받을 확률이 높아진다.

유사 사용자 그룹을 생성하기 위한 첫 번째 과정으로 태그 클러스터링에 의해 추출된 사용자 별 대표 태그는 유사 사용자 그룹을 생성하기 위한 데이터로 활용된다. 유사 그룹으로 생성할 시, 단순히 동일 대표 태그 수를 이용할 경우 사용자의 선호 정도가 각각 다르기 때문에 북마커별 대표 태그의 정규화를 위해 가중치를 부여한다.

표 1은 임의의 북마커들의 대표 태그를 추출한 예로 u1을 중심으로 유사 그룹을 구성한다. 이때 단순히 동일 태그 수를 이용하게 되면 많은 수의 태그를 태깅하였다 하더라도 태그에 대한 북마커별 선호도가 다름을 반영하지 못한다. 이는 식 (1)의 사용자 대표 태그별 가중치를 적용하여 해결할 수 있다[10].

$$WU_i = \frac{t_i}{UT_{total}} \times 100 \qquad \text{식(1)}$$

WU_i = user i의 대표 태그별 가중치

UT_{total} = user 대표 태그 i 의 총 개수

t_i = user의 대표 태그들 중 i번째 태그 태깅 수

또한 사용자 대표 태그별 가중치 식을 통해 계산된 값을 기반으로 유사 그룹을 선정하게 된다. 사용자와 선호 태그가 유사한 사용자들을 선정하기 위해 코사인 유사도를 기반으로 수행한다[11].

$$SW = \frac{\sum_{k=1}^{L}(WU_{ij} \times WU_{jk})}{\sqrt{\sum_{k=1}^{L} WU_{ik}^2 \times \sum_{k=1}^{L} WU_{jk}^2}} \qquad \text{식(2)}$$

SW : 북마커별 유사도 값

WUij : user i의 대표태그별 j의 가중치

WUjk : user j의 대표태그별 k의 가중치

다음 식 (2)을 이용하여 각 북마커별 유사도 값을 추출하게 되며, 이러한 유사도 값을 나타낸 것이 표 2이다. 태그의 수가 적기 때문에 유사도 값이 높은 편이지만 대표 태그의 수가 여러 분야를 걸쳐 추출된다면 좀 더 정확한 유사도 값을 추출 가능하다.

본 논문의 랭킹 알고리즘에서는 0.8 이상의 유사도 값을 가지는 사용자들에 한 해 유사그룹이라 정의하며[12], 이와 같이 추출된 유사도 값은 다음 웹 콘

표 2. 북마커별 유사도 추출 값

	u1	u2	u3	u4
u1		0.997	0.955	0.646

텐츠 랭킹 시스템에 데이터로 사용되어 사용자에게 유용한 정보 제공을 가능하게 한다.

3.2 웹 콘텐츠 랭킹 시스템

웹 콘텐츠 랭킹 시스템은 태그 클러스터링 시스템에서 추출된 연관 태그 가중치 값과 소셜 네트워크 시스템에서 추출된 사용자 유사도 값 및 웹 콘텐츠 내 북마크 수를 이용하여 단순 일반 검색 랭킹 알고리즘에 의한 검색결과가 아닌 소셜 및 유사성을 가지는 검색결과를 제공하는 기능을 수행한다.

본 연구에서는 기존의 검색결과보다 정확하고 유용한 정보 제공을 위해 웹 콘텐츠 랭킹 알고리즘을 제안하며, 또한 소셜 북마킹 사이트 내 한계점을 해결함과 동시에 사용자에게 좀 더 효율적이고 의미있는 정보의 제공을 위해 TBS_RANK(Tag, Bookmark and Similarity based Ranking) 알고리즘을 제안한다.

```
// num : number of cluster
// Clstr : Cluster
// WC(i) : I-th Web Contents
// AT : Associative Tags
// NB : number of bookmarks
// BT : Bookmarker's Tag
// SG : Similarity Group
// SW : Similarity Weight

//클러스터 num 개수만큼 반복
for(num=1; num>=n; num++){

    //클러스터 내에 전체 연관 태그의 가중치 총합을 계산
    Compute weight sum total of AT on Clstr

        //i번째 웹 콘텐츠가 empty 될 때까지
    Repeat{
        //i번째 웹 콘텐츠에 태깅된 태그들 중 연관 태그의
        가중치 합을 계산
        Compute weight summation of AT among
                tags of WC(i)

        // i번째 웹 콘텐츠에 북마킹한 유사 그룹의
        유사도 합을 계산
        Compute weight summation of SG's SW among
                                        bookmarking of WC(i)

        //i번째 웹 콘텐츠 내 최대 유사도 값을 선정
        Select maximum SW among WC(i)

        //i번째 웹 콘텐츠의 북마크 수를 추출
        Extract NB on WC(i)

        //최대 북마크 수를 선정
        Select maximum number among bookmarks

        //콘텐츠 별 연관 태그 가중치와 북마크 수, 유사도를 정규화
        Normalize weight of AT,  NB, and SW on
                                        each web contents

        //정규화한 값들을 합산한 결과를 정렬
        Make ranking according to the summation of
                ATReg, BNReg, and SWTeg

    } until(WC(i) == empty)
}
```

그림 2. TBS_RANK 알고리즘

TBS_RANK의 알고리즘의 첫 번째 단계로 클러스터(Clstr) 내 연관 태그의 가중치 합을 구한 뒤, 클러스터 내 연관 태그 쌍을 포함하는 i번째 콘텐츠 연관 태그들의 가중치 값을 구하고 i번째 웹 콘텐츠에 북마킹한 유사 그룹 내 유사도의 합을 계산하며 i번째 콘텐츠의 북마크 수도 추출한다. 콘텐츠 내 연관 태그 가중치의 합과 북마크 수 및 유사도를 각각 정규화한 뒤, 세 개의 정규화 값을 더하여 순위를 랭킹한다.

제안한 TBS_RANK 알고리즘에서 태그 쌍 가중치 값과 북마크 수의 정규화 함수를 사용한 식은 다음 식 (3), 식 (4)와 같으며 유사도의 정규화 식은 식 (5)과 같다.

$$ATReg = \frac{\sum_{i \in CT} AT_i}{\sum_{i \in C} AT_i} \quad \text{식(3)}$$

ATReg : 정규화된 연관 태그 쌍 가중치

$\sum_{i \in CT} AT_i$: 클러스터에 포함된 콘텐츠 내 태그 쌍들의 가중치 합

$\sum_{i \in C} AT_i$: 클러스터 내 태그 쌍 가중치들의 합

$$BNReg = \frac{BN(i)}{MAX(BN(i))} \quad \text{식(4)}$$

BNReg : 정규화된 북마크 수

$BN(i)$: i번째 콘텐츠의 북마크 인원수

$MAX(BN(i))$: 콘텐츠들 중 최대 북마크 인원수

$$SWReg = \frac{SWamong\,WC(i)}{MAX(SWamong\,WC(i))} \quad \text{식(5)}$$

SWReg : 정규화된 유사도 가중치

$SWamong\,WC(i)$: i번째 웹 콘텐츠 내 유사도의 가중치 합

$MAX(SWamong\,WC(i))$: i번째 웹 콘텐츠 내 유사도의 가중치 중 최대 유사도 값

4. 실험 및 평가

4.1 실험

본 논문에서 제안한 알고리즘의 평가를 위해 소셜 북마킹 사이트들 중 mar.gar.in의 웹 콘텐츠의 북마크 및 태그 데이터를 활용하였다. 키워드는 '공부',

표 3. 추출 데이터

키워드	웹콘텐츠	태그	북마크
공 부	100	2,433	230
영어공부	100	1,012	192
아 이 폰	100	505	121

'영어공부'와 최근 이슈가 되는 키워드인 '아이폰'을 통해 검색의 효율성을 제시하였다.

다음 표 3은 mar.gar.in에서 각 키워드 별로 검색되어 추출된 웹 콘텐츠와 태깅된 태그 및 북마크의 수를 나타낸다.

다음 표 4는 제안 알고리즘의 효율성 분석을 위해 사용한 실험데이터 중 mar.gar.in의 키워드 '영어공부'로 검색된 상위 25개의 검색결과를 보여주고 있다.

표 4의 행은 기존의 웹사이트에서의 검색결과로 상위부터의 순서를 No.로 나타내고, 각 열은 북마킹된 웹페이지들의 웹사이트 이름, 클러스터링을 통해 추출되어 계산된 연관 태그 가중치 값, 가중치 값을 정규화한 값, 웹 콘텐츠의 북마크 수, 북마크 수를 정규화한 값, 각 정규화 값을 랭킹을 통해 구한 값을 나타낸다.

다음 표 5는 본 논문에서 제안한 TBS_RANK의 알고리즘을 적용한 랭킹 결과로서 상위 10위까지의 변화를 보기 위해 기존 검색결과는 파랑색 음영으로 표현하였다.

기존 검색결과는 북마크 수와 저장순 등으로만 검색결과를 제공하는 반면에 제안 알고리즘은 연관 태

표 4. 기존 소셜 북마킹 사이트의 '영어공부' 상위 25개의 검색결과

No.	Web Site	연관 태그 가중치	연관 태그 가중치 정규화값	북마크 수	북마크 수 정규화 값	정규화 합
1	웹초보의 Tecg 2.1::땡전	233	1	192	1	2
2	영어는 덤!!!::영어소설 무	233	1	186	0.96875	1.96875
3	Korean-English-Korean	180	0.772532	169	0.880208	1.652741
4	Yappr - 비디오를 시청하	217	0.93133	127	0.661458	1.592789
5	해커스영어::No.1영어정보	163	0.699571	99	0.515625	1.215196
6	::Daily English - 대한민	108	0.463519	83	0.432292	0.895811
7	오마이리딩 닷컴에 오신것	180	0.772532	83	0.432292	1.204824
8	만점비법 해커스토플 - T	165	0.708155	68	0.354167	1.062321
9	♣ 기술, 디자인, 엔터테인	41	0.175966	62	0.322917	0.498882
10	The Internet Movie Da	39	0.167382	60	0.3125	0.479882
11	무료 영어소설, 오디오 북	108	0.463519	57	0.296875	0.760394
12	영국의 공영방송 BBC에	216	0.927039	56	0.291667	1.218705
13	YBMsisa.com - 인터넷	136	0.583691	53	0.276042	0.859733
14	LuxCozy(럭스코지)::영어	162	0.695279	48	0.25	0.945279
15	Randall's ESL Cyber L	128	0.549356	44	0.229167	0.778523
16	Wordbreak ::단어를 외우	143	0.613734	42	0.21875	0.832484
17	대한민국 No.1 외국어 교	67	0.287554	43	0.223958	0.511512
18	http--weekstudy.coolsc	180	0.772532	42	0.21875	0.991282
19	Listen and Write - Di	145	0.622318	42	0.21875	0.841068
20	토익 전문 - 해커스로	82	0.351931	38	0.197917	0.549848
21	영어에서 관사를 쉽게 파	118	0.506438	34	0.177083	0.683521
22	Livemocha:LearnLangua	182	0.781116	45	0.234375	1.015491
23	English Cube - 영어학습	145	0.622318	32	0.166667	0.788984
24	YBM 어학시험 (TOEIC,	149	0.639485	29	0.151042	0.790527
25	[STUDY] 오마이리딩 닷	180	0.772532	29	0.151042	0.923574

표 5. TBS_RANK를 이용한 '영어공부' 상위 25개의 검색결과

No.	Web Site	연관 태그 가중치	연관 태그 가중치 정규화값	북마크 인원수	북마크 인원수 정규화값	유사도 값	유사도정 규화	TBS_ RANK 결과
1	웹초보의 Tech 2.1 :: 땡전	233	1	192	1	3.726	0.821	2.821
4	Yappr - 비디오를 시청하면	217	0.93133	127	0.682796	3.629	0.799	2.392
2	영어는 덤!!! :: 영어소설	233	1	186	0.96875	1.769	0.390	2.358
3	Korean-English-Korean	180	0.772532	169	0.908602	1.873	0.412	2.065
8	만점비법 해커스토플 - 'TOE	165	0.708155	68	0.365591	4.540	1.000	2.062
6	:: Daily English - 대한민	108	0.463519	83	0.446237	4.484	0.988	1.883
19	Listen and Write - Di	145	0.622318	42	0.225806	4.540	1.000	1.841
25	[STUDY] 오마이리딩 닷컴-	180	0.772532	29	0.155914	3.726	0.821	1.744
22	Livemocha:LearnLanguagesO	182	0.781116	45	0.241935	2.792	0.615	1.630
5	해커스영어::No.1영어정보	163	0.699571	99	0.532258	1.874	0.413	1.628
24	YBM 어학시험 (TOEIC, JE	149	0.639485	29	0.155914	3.734	0.822	1.613
18	http--weekstudy.coolschool	180	0.772532	42	0.225806	2.769	0.610	1.601
15	Randall's ESL Cyber Liste	128	0.549356	44	0.236559	3.679	0.810	1.589
12	영국의 공영방송 BBC에서 운	216	0.927039	56	0.301075	0.864	0.190	1.409
7	오마이리딩 닷컴에 오신 것	180	0.772532	83	0.446237	0.918	0.202	1.407
17	대한민국No.1외국어교육	82	0.351931	38	0.204301	3.575	0.787	1.299
16	Wordbreak :: 단어를 외우는	143	0.613734	42	0.225806	1.919	0.423	1.255
23	English Cube - 영어학습을	145	0.622318	32	0.172043	1.883	0.415	1.204
11	무료 영어소설, 오디오북 Pr	108	0.463519	57	0.306452	1.732	0.382	1.142
20	토익 전문 - 해커스토	108	0.463519	29	0.155914	2.678	0.590	1.140
14	LuxCozy(럭스코지) :: 영어	162	0.695279	48	0.258065	0.814	0.179	1.125
10	The Internet Movie Data	41	0.175966	62	0.333333	2.828	0.623	1.103
9	♣ 기술, 디자인, 엔터테인	67	0.287554	43	0.231183	2.691	0.593	1.092
21	영어에서 관사를 쉽게 파악	118	0.506438	34	0.182796	1.829	0.403	1.086
13	YBMsisa.com - 인터넷 영어	136	0.583691	53	0.284946	0.954	0.210	1.070

그 가중치 값을 통해 정확도 및 사용자와 유사한 그룹을 찾아내어 사용자가 관심 있을 것으로 예상되는 웹 콘텐츠를 부각시킴으로써 사용자에게 효율적인 정보 제공이 가능하다.

이러한 정보 제공에 대한 효율성 평가를 위해 기존의 검색결과와 TBS_RANK의 검색결과를 비교 평가 한다.

4.2 비교평가

본 절에서는 랭킹결과의 정확성 평가를 위해 NDCG at K[13]를 적용하여 기존의 소셜 북마킹 사이트의 검색결과와 선행연구에서 진행한 연관 태그 가중치와 본 논문에서 제안한 TBS_RANK의 검색결과를 비교 분석하였다.

NDCG at K는 검색결과의 순위 1에서 K까지의 gain의 합으로 계산된다[13]. 또한, 다양한 키워드 '공부', '아이폰', '영어공부'를 이용함으로써 키워드가 가지는 특징에 대해 기술한다.

각 키워드의 선별 이유로는 먼저, 키워드 '공부'는 포괄적이며 광범위한 분야에서 사용됨으로써 태그 정보 및 북마크 정보의 다양성을 가지고 있으며, 키워드 '아이폰'은 최근 국내 사용자의 관심을 받으며 최신성을 가지는 태그의 대표로써 활용하였으며, 키워드 '맛집'은 한정된 분야 및 좁은 태그 정보를 가지

는 키워드의 경우 검색결과의 효율성을 보이기 위해 사용하였다.

표 6은 키워드 '공부'의 NDCG 값을 표현한 것으로써 값이 높을수록 좋은 랭킹 결과라 하며, K의 값은 상위 페이지의 누적 수를 나타낸다. 검색된 상위 5개의 웹페이지 랭킹 정확도인 NDCG 값은 각각 기존 검색결과의 경우 0.215123, 연관 태그 가중치는 0.195319, TBS_RANK의 경우 0.272634 값을 가짐으로 TBS_RANK가 랭킹 정확도가 높음을 알 수 있다. 상위 5개의 웹페이지의 NDCG 값 중에서 본 논문에서 제안한 TBS_RANK 알고리즘이 가장 높은 값으로 도출되었다. 이는 상위 5개 웹페이지에서 콘텐츠들의 연관도가 높은 웹페이지들이 먼저 검색되었다는 것을 의미한다. 또한, 키워드 '공부'의 NDCG 값이 다른 키워드에 비해 낮은 경우는 '공부'라는 키워드의 경우 다양한 분야에서 사용되어 웹 콘텐츠의 표현 시 너무 많은 태그를 태깅하거나 정확하게 태깅된 태그 수가 적어 기존의 검색결과의 정확률 등이 낮아지만 이러한 문제점을 본 제안 알고리즘에서는 해결할 수 있다.

표 7은 키워드 '아이폰'의 NDCG 상위 50개의 누적 값을 표로 나타내었다. 각 검색 방법론에 따른 상위 5개의 웹페이지의 NDCG 값을 보면 기존 검색결과가 0.304358로 높게 나타나지만 TBS_RANK은 상위 10개, 15개, …, 하위페이지로 갈수록 연관 페이지가 기존 검색 결과보다 더 많이 검색되었다는 것을 볼 수 있다. 또한, '아이폰'의 키워드는 최신 이슈어를 반영하기 위한 것으로 키워드의 특징상 한정된 분야와 한정된 태그인 애플, 맥 등의 태그가 반복되어 가중치 값이 높게 나타난다. 북마크 수 또한 최근에 다양한 분야에서 사용자들이 북마크하기 때문에 대부분의 웹페이지들이 고른 북마크 수를 가지게 된다. 이로써 기존의 검색결과 및 연관 태그 가중치 값을

표 6. 키워드 '공부'의 NDCG 값

K	기존 검색결과	연관 태그 가중치	TBS_ RANK
5	0.215123	0.195319	0.272634
10	0.252151	0.279325	0.359539
15	0.299698	0.325726	0.430396
20	0.360035	0.39575	0.459931
25	0.459215	0.452162	0.482822

표 7. 키워드 '아이폰'의 NDCG 값

K	기존 검색결과	연관 태그 가중치	TBS_ RANK
5	0.149147	0.134856	0.184251
10	0.280719	0.239358	0.265042
15	0.306815	0.351248	0.335765
20	0.34935	0.422486	0.392603
25	0.379843	0.45263	0.431261
30	0.424455	0.481176	0.470427
35	0.459508	0.516029	0.535492
40	0.50812	0.538387	0.554875
45	0.529845	0.549251	0.573031
50	0.561593	0.581002	0.580067

이용한 검색결과는 한 가지의 값을 이용하여 낮은 NDCG 값을 가지게 되었다. 그러나 TBS_RANK의 경우 북마크 수와 태그 정보뿐만 아닌 사용자의 유사도를 혼합하여 사용자에게 유용한 정보를 추천 및 제공 가능하기 때문에 높은 NDCG 값을 가지게 되었으며 이는 제안 알고리즘의 상위 검색결과들이 키워드와 연관된 웹페이지들이 먼저 검색되었다는 것을 의미한다.

표 8은 '맛집' 키워드의 NDCG 값을 표현하였다. '맛집' 키워드의 특징은 사용자들이 웹페이지를 북마킹 할 시 태그의 수가 적고 한정된 태그를 사용한다는 것이다. 예를 들면 맛집, 음식점, 요리, 지역이름 등을 사용하여 나타내며, 또한 개인적으로 추천한 맛집을 블로깅한 블로그 웹페이지 등을 표현하기 위한 개인적인 태그들 예로, 블로거의 이름, 음식이름, 음

표 8. 키워드 '맛집'의 NDCG 값

K	기존 검색결과	연관 태그 가중치	TBS_ RANK
5	0.303358	0.186212	0.303548
10	0.354401	0.254946	0.386942
15	0.304358	0.186212	0.303548
20	0.482974	0.381875	0.505925
25	0.521771	0.495039	0.536872
30	0.558475	0.523672	0.559604
35	0.577945	0.550927	0.580091
40	0.619122	0.592149	0.59947
45	0.651906	0.617634	0.617634
50	0.662488	0.628216	0.628216

표 9. 알고리즘 별 Avg_NDCG 값 비교

	아이폰	맛 집	영어공부	Avg_NDCG
기존 검색결과	0.39494	0.4749	0.68241	0.52
연관 태그 가중치	0.42664	0.44169	0.6913	0.52
TBS_RANK	0.43228	0.50219	0.77104	0.57

식점 이름 등으로 태깅되어 있다. 따라서 연관 태그의 빈도가 낮아지게 되어 연관 태그 가중치 값을 이용한 연관 태그 가중치 알고리즘은 낮은 NDCG 값을 가지게 되었지만 TBS_RANK은 연관 태그 가중치와 북마크 수 및 사용자와의 유사도를 이용하였기 때문에 높은 NDCG 값을 가지게 된다.

표 9는 Avg-NDCG값을 적용한 결과이다. 기존의 검색결과는 사용자들의 북마크 수를 통해 랭킹된 결과이며 연관 태그 가중치는 태그 클러스터링의 선행연구에서 제안한 검색 알고리즘으로 비슷한 Avg-NDCG 값을 가진다. 이는 단순히 북마크 수를 이용해 검색 시 정확하지 않은 검색결과를 도출하며 또한, 사용자들이 북마킹 시 태그를 정확히 붙이지 않거나 아예 태그를 붙이지 않기 때문에 연관 태그만을 통한 검색결과 역시 효율적이지 못한 검색결과를 도출하며 Avg-NDCG 값 또한 낮게 나타난다. 이러한 문제점을 해결하기 위해 제안한 선행연구의 TBS_RANK을 통한 랭킹은 기존 검색결과보다 좋은 성능을 보인다. 그러나 TBS_RANK 또한 태그 정보 및 북마크 정보 한 쪽의 값이 제한되면 성능이 기존의 검색 결과와 비슷한 결과를 도출한다. 이러한 문제점을 해결하기 위해 TBS_RANK은 소셜 북마킹 사이트 내 중요한 역할을 하는 연관 태그와 북마크 수 및 유사 사용자 정보를 동시에 활용함으로 문제점을 해결하며 향상된 검색결과를 보여준다.

5. 결론 및 향후 연구과제

본 논문에서는 기존 소셜 북마킹 시스템에서 가지는 문제점을 해결하기 위해 태그 클러스터링을 통해 부정확한 태그들을 제거하고 연관관계가 높은 태그 그룹으로 구성될 수 있게 하였으며 또한, 태그 정보와 다른 사용자와의 관심도를 나타내는 북마크 수 및 사용자 유사 그룹을 통해 얻은 유사도를 혼합한 랭킹 알고리즘을 제안하여 소셜의 의미를 갖는 검색결과를 제공할 수 있는 TBS_RANK을 제안하였다.

랭킹 알고리즘의 성능평가를 위해 국내의 소셜 북마킹 사이트인 mar.gar.in의 북마킹 정보를 이용하여 NDCG 검색 비교 기법을 통해 비교 평가 하였으며, 그 결과 기존의 소셜 북마킹 사이트는 Avg-NDCG 값이 0.52를 갖으며, 또한 선행 연구로 진행한 연관 태그 가중치 방법론을 이용한 알고리즘은 0.52이지만, 제안한 TBS_RANK 값은 0.57 으로 평균 11.3% 향상된 검색결과를 도출할 수 있었다.

향후 기존에 존재하는 랭킹 알고리즘들을 이용한 폭소노미의 의미를 담은 FolkRank 등 다양한 랭킹 알고리즘들과 제안 알고리즘을 비교 평가할 예정이다. 또한 북마크 수와 태그 정보 및 유사도 값만이 아닌 웹 콘텐츠 내의 다양한 정보를 이용하거나 가중치 값 변화, 웹 검색의 최신성, 인기 북마커 등을 반영한 더욱 효율적인 랭킹 알고리즘으로 발전시키기 위한 계속적인 연구가 필요하다.

참 고 문 헌

[1] 정부연, “2006년 인터넷 화두 웹 2.0(Web2.0),” 기술동향, 2006.

[2] Farooq U, Yang Song, Carroll J.M., and Giles C.L., “Social Bookmarking for Scholarly Digital Libraries,” *IEEE, Internet Computing*, 2007.

[3] http://delicious.com

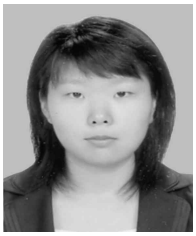
[4] http://www.bibsonomy.org

[5] http://mar.gar.in

[6] S. Brin and L. Page, “The Anatomy of a Largescale Hypertextual Web Search Engine,” In Proceedings of 7th International World Wide Web Conference, Computer Networks and ISDN Systems, Vol.20, No.1-7, pp. 107-117, Apr,1998.

[7] J. M. Kleinberg, “Authoritative Sources in Hyperlinked Environment,” *Journal of the ACM*, Vol.46, No.5, pp. 604-632, 1999.

[8] E. Adar, L.Zhang, L.Adamic, and R. Lucose, “Implicit Structure and the Dynamics of Blogspace,” Workshop on the Weblogging Ecosystem : Aggregation, Analysis and Dynamics, 2004.
[9] 이시화, 이만형, 황대훈, “web2.0에서의 Tag Clustering을 통한 이미지 검색의 효율성 분석,” 멀티미디어학회 논문지, Vol. 11, No. 8, 2008
[10] 이시화, 박수진, 이만형, 황대훈, “콘텐츠 추천을 위한 태그 기반 소셜 네트웍 구축에 관한 연구,” 멀티미디어학회 춘계학술대회, Vol.12, No.1, 2009.
[11] http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html#Cosim
[12] Taek-Hun Kim, Young-Suk Ryu, Seok-In Park, and Sung-Bong Yang, “An Improved Recommendation Algorithm in Collaborative Filtering,” *Lecture notes in Computer Science*, No.2455, pp. 254-261, 2002.
[13] K. Jarvelin and J. Kekalainen, “IR Evaluation Methods for Retrieving Highly Relevant Doc-umnets,” *In Proceedings of the ACM conference on Research and Development on Information Retrieval (SIGIR)*, pp. 41-48, 2000.



박 수 진

2008년 현대전문학교 멀티미디어과(학사)
2010년 경원대학교 전자계산학과(석사)
관심분야 : Web2.0, Semantic Web, Tag, Social Bookmarking, Ranking, Retrieval

이 시 화

2005년 서울보건대학 컴퓨터정보과(학사)
2005년 블루M 개발실 연구원
2007년 경원대학교 전자계산학과(석사)
2008년~현재 경원대학교 전자계산학과 박사과정
관심분야 : e-Learning, Context-Aware, Semantic Web, Web2.0, Tag

황 대 훈

1997년 동국대학교 수학과(학사)
1983년 중앙대학교 전자계산학과(석사)
1991년 중앙대학교 전자계산학과(박사)
1983년~1985년 한국산업경제기술연구원(KIET) 연구원
2009년~2010년 한국멀티미디어학회 회장
1987년~현재 경원대학교 교수
관심분야 : e-러닝, Semantic Web, 유비쿼터스 컴퓨팅