Journal of Korea Multimedia Society Vol. 13, No. 8, August 2010(pp. 1245-1255)

소셜 북마킹 시스템에서의 북마크와 태그 정보를 활용한 웹 콘텐츠 랭킹 알고리즘

박수진[†], 이시화[††], 황대훈[†††]

요 약

현재 웹 2.0 환경에서의 핵심 기술 중 하나는 사용자가 관심 있는 웹페이지를 태깅 및 북마킹 하는 소셜 북마킹 기술이다. 소셜 북마킹은 웹 콘텐츠에 태깅된 북마크 정보 및 태깅 결과를 기반으로 검색, 분류, 공유를 통해 효율적인 정보 제공을 주목적으로 하고 있다. 그러나 현재 소셜 북마킹 시스템들은 웹 콘텐츠의 사용자들의 관심 정도를 측정할 수 있는 북마크 수 및 검색과 분류를 목적으로 하는 태그 정보를 각각 독립적으로 검색에 활용하는 방식을 사용하고 있다. 이는 소셜 북마킹 시스템에서 중요한 특징을 가지는 북마크와 태깅 기술을 효율적으로 활용하지 못하는 결과가 된다. 이에 본 연구에서는 태그 클러스터링을 통한 연관 태그 추출에 관한 선행연구를 기반으로, 북마크 정보와 혼합하기 위한 웹 콘텐츠 랭킹 알고리즘을 제안하였다. 또한 제안 알고리즘의 효율성 분석을 위해 기존 검색 방법론들과의 비교평가를 시행하였으며, 그 결과 본 연구의 핵심적인 특징인 북마크와 태그 정보를 함께 활용한 소셜 북마크 시스템이 기존 시스템보다 효율적인 검색결과를 도출하였다.

A Web Contents Ranking Algorithm using Bookmarks and Tag Information on Social Bookmarking System

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), Clustering(클러스터링), Social Bookmarking(소셜 북마킹), Web Contents(웹 콘텐츠), Ranking(랭킹), Retrieval(검색)

※ 교신저자(Corresponding Author) : 황대훈, 주소: 경기도 성남시 수정구 복정동 산 65번지(461-701), 전화: 031)750-5327, FAX: 031)757-6715, E-mail: hwangdh@kyungwon. ac.kr
접수일 : 2009년 12월 17일, 수정일 : 2010년 3월 26일
완료일 : 2010년 6월 4일

[†] 준회원, 경원대학교 전자계산학과 석사과정
(E-mail : hohivi@gmail.com)
[††] 준회원, 경원대학교 전자계산학과 박사과정
(E-mail : leesihwaman@gmail.com)
[†††] 종신회원, 경원대학교 교수
※ 이 연구는 2010년도 경원대학교 지원에 의한 결과임

1. 서 론

최근에 사용자의 참여와 공유라는 웹 2.0 트렌드에 따라 다양한 웹 서비스와 기술들을 제공하고 있다. 이러한 웹 2.0 트렌드 중 최근 웹 기반 북마킹 서비스와 태깅 및 소셜 기술을 도입한 소셜 북마킹 기술이 대두되고 있다[1,2].

소셜 북마킹 기술은 사용자들이 웹상에 관심 있는 웹 콘텐츠들을 즐겨찾기하는 기술이다. 이를 통해 자신이 북마킹한 웹페이지, 즉 웹 콘텐츠뿐만 아니라 다른 사용자들이 북마킹한 웹 콘텐츠들을 서로 공유할 수 있다는 특징이 있다. 또한 북마킹 시스템에서의 핵심적인 기술 중 하나는 태그 기술이다. 태그는 콘텐츠의 검색, 분류, 공유를 통한 효율적인 정보제공을 목적으로 하고 있다.

그러나 소셜 북마킹 시스템에서의 검색 시 단순한 키워드 매칭 혹은 사용자가 저장할 때 적어두는 메모 등을 바탕으로 검색을 하며 또한 검색 결과는 웹 콘텐츠의 저장순이나 이름순, 북마크순, 이용순 등으로 제공하고 있는 실정이다. 이는 소셜 북마킹 시스템의 특징을 반영하지 않은 검색 방법으로 효율적이지 못하다.

이에 본 논문에서는 소셜 북마킹의 특징 중 하나인 웹 콘텐츠들을 다른 사용자들과 함께 즐겨찾기하여 관심도를 나타내는 북마크 인원수와 태깅된 태그 정보를 이용하여 웹 콘텐츠를 랭킹하여 제공한다. 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 정보를 효율적으로 활용하기위한 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

웹 2.0은 사용자들의 참여와 공유라는 트렌드를 바탕으로 많은 발전을 해 왔으며 그 중 대두되는 기술로 소셜 북마킹과 태그가 있다[1,2]. 웹 상에서 유용한 정보를 발견하면 이전에는 즐겨찾기로 자신의 웹브라우저를 통해 등록해 두었지만, 요즘에는 소셜 북마킹을 이용하여 웹페이지, 블로그 등 다양한 웹 콘텐츠들을 웹상에 저장하고 태그를 붙여 다른 사람들과 공유하는 서비스를 제공하고 있다. 또한 북마킹된 웹페이지에 사용자가 덧붙이고 싶은 태그들을 자유롭게 붙임으로써 좀 더 효율적인 검색을 하는데 사용되고 있으며 대표적인 소셜 북마킹 웹사이트로 딜리셔스(del.icio.us), 빕소노미(BibSonomy), 마가린(mar.gar.in) 등과 같은 사이트가 있다[3-5].

소셜 북마킹의 중요한 특징은 사용자가 자신이 관심 있는 웹 콘텐츠를 북마킹할 때 태그를 붙이고 그것을 공유하는 것이지만, 검색 시 단일 태그 혹은 사용자가 작성한 메모, 제목 등으로만 검색되어 다양하게 태깅된 태그들로 인해 잘못된 검색 결과를 보여주거나 다른 사용자의 관심도를 나타내는 북마크 인원수 등을 포함한 소셜 북마킹의 특징을 반영하지 못하고 있다. 이에 본 논문에서는 북마크 수와 태그 정보를 검색에 효율적으로 반영하기 위해서 태그 클러스터링을 통한 연관 태그 추출에 관한 선행연구를 진행하였으며, 다음 2.2절과 같다.

2.2 태그 클러스터링 시스템

태그 클러스터링 시스템은 북마킹된 웹 콘텐츠들에 태깅된 태그들 중 부정확하게 태깅된 태그들은 제거하고 연관도가 높은 태그들을 클러스터링하기 위한 시스템으로서 선행연구로 진행하였다[6,7].

시스템은 크게 4개의 모듈로 구성되며, 연관 태그 맵핑 모듈(Tag Relation Mapping Module)은 수집된 태그들을 기반으로 연관 태그들 간의 맵핑을 수행하며, 이 과정에서 빈도수 추출 모듈(Frequency Extraction Module)은 유사 태그 및 빈도수를 추출한다. 또한 가중치 매트릭스 생성 모듈(Weight Matrix Generation Module)은 추출된 연관 태그 및 빈도수를 기반으로 가중치 행렬(weight matrix)을 생성한다. 이렇게 생성된 가중치 매트릭스를 기반으로 태그 클러스터링 모듈(Tag Clustering Module)은 연관성이 높은 태그 그룹으로 클러스터링하기 위한 기능을 수행한다.

이와 같이 추출된 연관태그 정보는 북마크 수와 혼합하여 효율적인 웹 콘텐츠 랭킹을 위한 중요한 기초데이터로 활용된다.

2.3 랭킹 알고리즘

본 논문에서 랭킹을 이용한 검색 방법론을 위해 다양한 랭킹 알고리즘의 장단점을 분석하였다. 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[8]와 HITS[9]가 대표적인 구조적 측면의 평가방법이다.

PageRank[8]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 페이지랭크를 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[9]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[10]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.

2.4 검색 분석 기법

제안한 알고리즘의 검색결과에 대한 효율성의 평가를 위해 다양한 검색 분석 기법을 분석하였으며, 그 중 최근에 다양한 연구에서 사용되는 검색 기법들을 분석하였다[11].

먼저, N순위의 적합률(precision at N)은 기존의 precision의 응용으로 5위, 10위, 15위… 등의 검색 순위에서의 정확률을 계산하여 순위화에 대한 요약을 나타내는 분석 기법이다[12]. 다음은 N순위의 적합률을 나타내는 식이다.

$$r = \frac{1}{R}\sum_{i=1}^{R} x_i$$

R : 키워드와 연관된 문서의 수

MAP(Mean Average Precision)은 TREC에서 사용하는 표준 평가방법으로 재현율을 반영한 정확률로서 각각의 재현율 수준에서의 정확률을 모두 더한 뒤 평균을 낸 값이다. 따라서 정답이 예측결과의 상위에 있을수록 높은 값을 가지게 된다[13].

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\mathrm{Precision}(R_{jk})$$

$|Q|$: 모든 쿼리 집합의 사이즈

Normalized Discounted Cumulative Gain(NDCG at K)는 페이지에 포함된 키워드와 연관된 태그의 수를 측정하고 그 수를 K순위까지의 정규화된 적합성 누적 점수이다[14]. NDCG의 점수를 구하기 위해 먼저 DCG를 측정하게 되는데 DCG는 기존의 precision, recall 기반의 검색엔진 평가 방법으로는 순위에 따른 차별점을 부과하기 힘들다는 판단에 따라 나온 방법이다. 이는 50% 이상의 검색 사용자가 검색 결과의 1, 2 페이지 정도만 참고한다는 것을 통해 precision, recall 만으로는 정확한 사용자 패턴에 기반한 성능평가를 하기 힘들다는 것을 반영한 평가 방법이 되겠다.

DCG는 검색 엔진 결과에서 문서의 연관도를 측정을 이용하여, DCG를 측정한다. DCG는 관련성에 따라 0~3의 값을 가지는 웹페이지들의 검색 결과 순서에 패널티를 적용한 점수 값으로 log함수를 사용하였다. 이는 검색 결과의 1, 2위 사이의 간격은 큰 차이를 가지지만 999위와 1000위 사이의 차이는 거의 없다는 가정 하에 사용되었다. 이러한 DCG 점수는 현재 검색 엔진의 결과 상태를 보여주며 가장 이상적인 검색엔진의 점수를 나타내는 정규화된 DCG가 NDCG가 된다. 다음은 NDCG의 식이다.

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

N_q : 쿼리 q에 대한 NDCG

M_q : 정규화 상수(가장 완벽한 랭킹의 경우 NDCG가 1이 될 수 있도록 조정)

J : 검색 결과 내 순위

r(j) : j 순위에서의 결과 적합성

이에 본 논문에서는 최근 많이 사용되며 랭킹 알고리즘에 적합한 NDCG at K를 이용하여 비교평가를 하였다.

3. 시스템 설계

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

태그 클러스터링 시스템은 웹 콘텐츠 내 태그들을 기반으로 연관관계가 높은 태그들을 클러스터링 하기 위한 역할을 수행하여 키워드와 좀 더 유사한 웹 콘텐츠들을 추출하게 된다. 그리고 웹 콘텐츠 랭킹 시스템은 태그 클러스터링 시스템을 통해 연관 태그 쌍으로 추출된 태그들과 웹 콘텐츠 내 북마크 인원수를 기반으로 정규화 및 랭킹모듈을 통해 웹 콘텐츠들을 순위화하여 사용자들에게 제공하는 역할을 수행한다.

본 연구에서의 태그 클러스터링 시스템은 선행연구[6,7]로서 진행하였으며, 웹 콘텐츠 랭킹 시스템을 중심으로 다루었다.

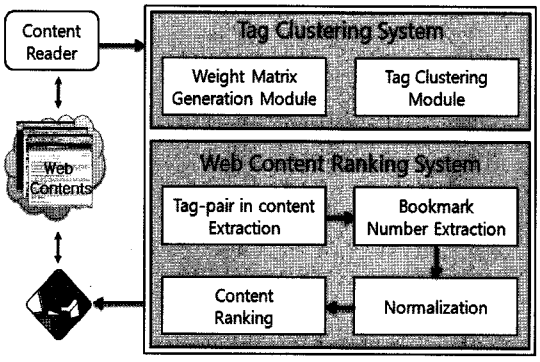

그림 1. 제안 시스템

3.1 태그 클러스터링 시스템

북마크 및 태그를 활용한 효율적인 웹 콘텐츠 랭킹을 위한 첫 번째 과정으로 북마킹된 콘텐츠들에 태깅된 태그들 중 부정확한 태그들은 제거하고 연관도가 높은 태그들을 추출한다.

태그 클러스터링 시스템은 선행연구[6]에서 진행하였으며, 다음 그림 2는 4.1절의 그림 4에서 키워드 "영어공부"를 통해 추출된 클러스터 내의 태그들을 그래프로 표현한 것이다. 그래프의 타원은 태그(영어, study, 토익 등)를, 태그와 태그간의 연결선 및 값(영어공부-41-영어)은 태그 간의 연관정도를 의미한다.

이와 같이 추출된 연관 태그들은 북마크 수와 본 논문에서 제안하는 정규화를 통한 랭킹 알고리즘에 적용하게 되며, 다음 3.2절과 같다.

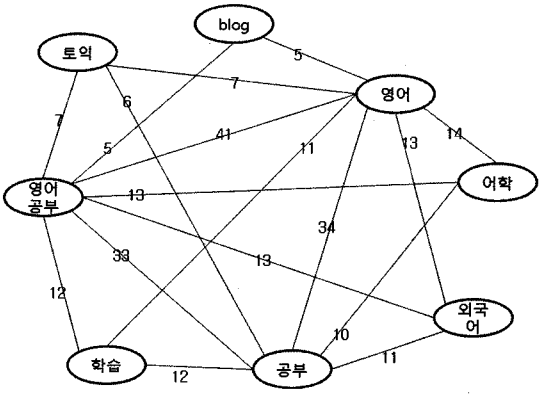

그림 2. '영어공부' 클러스터의 예

3.2 웹 콘텐츠 랭킹 시스템

웹 콘텐츠 랭킹 시스템은 북마크와 태그를 각각 독립적으로 검색에 활용하여 검색하는 기존 북마크 시스템의 문제점을 해결하기 위해 3.1절에서 선행연구로 진행된 클러스터 내의 연관 태그들 간의 가중치 값과 북마킹된 북마크 수를 이용하여 두 가지 핵심 기술의 특징을 반영한 검색 결과를 제공하는 역할을 수행한다. 이를 위해 본 논문에서 제안하는 웹 콘텐츠 랭킹 알고리즘은 다음 그림 3과 같다.

알고리즘의 첫 번째 단계로 클러스터(C) 내 연관 태그의 가중치 총합(CTW)를 구한다. 그 뒤 클러스터 내 연관 태그 쌍을 포함하는 i번째 콘텐츠 연관 태그들의 가중치 값(CTTW(i))를 구하고 i번째 콘텐

```
//num : number of cluster
//Clstr : Cluster
//WC(i) : ith Web contents
//AT : associative tags
//NB : number of bookmarks

 //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 연관태그의
        가중치 합을 계산
      Compute weight summation of AT among
                                    tags of 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 수를 정규화
      Normalize weight of AT and NB on
                                each web contents

      //정규화한 값들을 합산한 결과를 정렬
      Make ranking according to the summation of
                                  normalized values

   } until(WT(i) == empty)
 }until(Clstr(num) == empty)
```

그림 3. 콘텐츠 랭킹 시스템 알고리즘

츠의 북마크 인원수(BN(i))도 추출한다. 그 뒤 콘텐츠 내 태그 쌍 가중치의 합(CTTW(i))과 북마크 인원수를 각각 정규화(TWReg(i), BReg(i))한 뒤, 두 개의 정규화 값을 더하여 순위를 랭킹 한다.

제안한 알고리즘에서 정규화 함수를 사용한 식은 다음 식(1), (2)와 같으며 각 식은 태그 쌍 가중치 값과 북마크 인원수를 정규화 한다.

$$TWReg = \frac{\sum_{i \in CT} TW_i}{\sum_{i \in C} TW_i} \quad (1)$$

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

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

$$BReg = \frac{BN(i)}{MAX(BN(i))} \quad (2)$$

$BN(i)$: i번째 콘텐츠의 북마크 인원수

$MAX(BN(i))$: 콘텐츠들 중 최대 북마크 인원수

식 (1)은 i번째 웹 콘텐츠(CT) 내 태그 쌍들 중 클러스터링(C) 내 태그 쌍들에 포함되면 태그 쌍들의 가중치 값(TW)을 더한 합들을 클러스터링된 태그들의 가중치 합(TW)으로 나눈 값이 정규화된 태그 쌍 가중치 값이 된다.

북마크 인원수의 정규화를 나타내는 식 (2)은 i번째 웹 콘텐츠의 북마크 인원수를 가장 큰 북마크 인원수로 나누어 정규화 값을 구한다.

이러한 정규화 식을 사용함으로 가중치 값과 북마크 수를 0~1 사이의 값으로 제한을 두어 일관성을 유지할 수 있으며 데이터들 간의 빈약성, 오차, 편차 등을 고려할 수 있는 장점을 가진다.

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', '맛집', '영어공부'와 최근 이슈가 되는 키워드인 '아이폰'을 통해 검색의 효율성을 제시하였다. 다음 그림 4는 mar.gar.in 사이트에서 키워드 '영어공부'를 통해 검색된 결과이다.

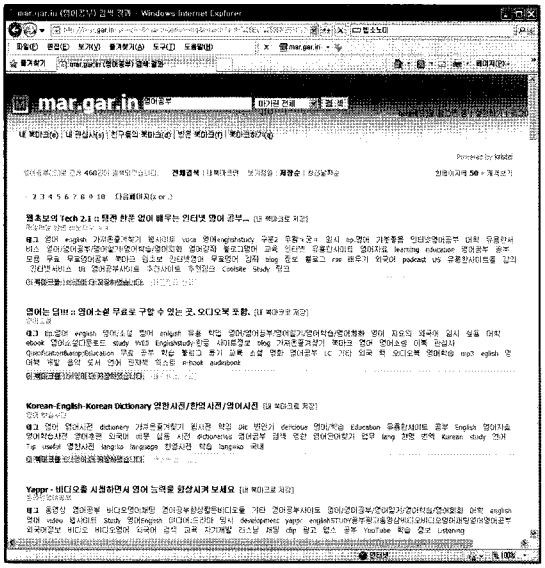

그림 4. 키워드 '영어공부' 검색 결과

다음 아래의 표 1은 mar.gar.in에서 각 키워드 별로 검색되어 추출된 웹 콘텐츠와 태깅된 태그 및 북마크의 수를 나타낸다.

표 1. 추출 데이터

키워드	웹콘텐츠	태그	북마크
공 부	100	2,433	230
영어공부	100	1,012	192
맛 집	100	525	270
아 이 폰	100	505	121

아래의 그림 5는 mar.gar.in 사이트 내 키워드 '영어공부'를 통해 검색된 웹 콘텐츠로 웹 콘텐츠의 제목과 사용자가 작성한 메모, 태그, 북마크 인원수로 구성되며 영어, 공부, 외국어 등 다양한 태그를 가지며 186명의 북마크 수를 가진다.

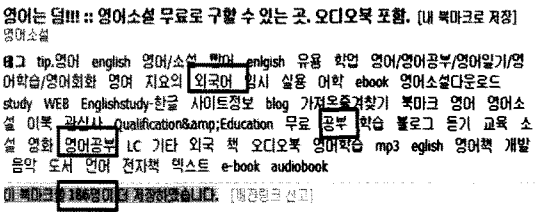

그림 5. '영어공부' 검색 결과 내 임의의 웹 콘텐츠

이러한 웹 콘텐츠를 랭킹 알고리즘에 적용하기 위한 단계로 3.1절에서 생성된 클러스터 그림 2 내에 존재하는 연관태그(예, 영어공부-외국어; 13, 영어공부-공부; 33 …등의 값을 가짐)를 찾아 3.2절의 랭킹 알고리즘을 따라 가중치 값을 더함으로 웹 콘텐츠의 연관 태그 가중치 총 합(그림 5의 경우 총 233의 값을 가짐)을 구하고 186명의 북마크 수 추출하여 3.2절의 정규화 식 (1)과 (2)를 통해 정규화 값(그림 5의 경우 연관태그의 정규화 값; 1, 북마크 수의 정규화 값; 0.96875)을 구한 뒤 합(그림 5의 경우 랭킹 값; 1.96875)을 통해 랭킹을 하여 웹 콘텐츠를 제공한다.

다음 표 2는 기존의 소셜 북마킹 사이트인 mar.gar.in의 키워드 '영어공부'의 상위 25개의 검색 결과이다. 표 2의 행의 경우 기존의 웹사이트에서의 검색 결과로 상위부터의 순서를 No.으로 나타내고, 각 열은 북마킹된 웹페이지들의 웹사이트 이름, 클러스터링을 통해 추출되어 계산된 연관 태그 가중치 값, 가중치 값을 정규화한 값, 웹 콘텐츠의 북마크 인원수, 북마크 인원수를 정규화한 값, 각 정규화 값을 랭킹을 통해 구한 값을 나타낸다.

표 3은 표 2를 기반으로 제안 알고리즘으로 재랭킹된 결과 값을 나타낸다. No.을 보게 되면 기존의 No.의 순서가 변화된 것을 볼 수 있다. 예로 기존에서는 22번째로 검색되었던 No.22가 제안 알고리즘을 통해 1.015491의 값을 가지며 9번째로 검색되었다. 이러한 결과는 다른 사용자의 관심도라 볼 수 있는 북마크 인원수를 랭킹에 적용하여 많은 관심을 받고 있는 웹 콘텐츠를 상위에 검색됨으로써 관심 분야의 트렌드 혹은 다른 사용자들의 관심도를 볼 수가 있다.

4.2 평가

제안 랭킹 알고리즘의 효율성을 제시하기 위해 검색결과의 랭킹 정확성을 측정하는 NDCG at K를 사용하여 기존의 소셜 북마킹 사이트의 검색 결과와 선행연구로 진행하였던 연관 태그 가중치 기반의 검색 결과와 본 논문에서 제안한 알고리즘을 통한 검색 결과의 비교분석을 진행하였다. NDCG at K는 검색 결과의 순위 1에서 K까지의 gain의 합으로 계산된다.

먼저, 다음 그림 6은 키워드 '공부'의 NDCG 값을 그래프로 표현한 것으로 상위 5개의 웹페이지의 NDCG 값을 보게 되면 본 논문에서 제안한 가중치 값과 북마크 인원수를 사용한 알고리즘이 다른 방법론보다 높은 값을 가지고 있다. 이는 상위 5개 웹페이지에서 콘텐츠들의 연관도가 높은 웹페이지들이 먼저 검색되었다는 의미한다. 또한, NDCG의 값이 다른 키워드에 비해 전체적으로 낮은 이유는 '공부'라는 키워드는 다양한 분야에서 사용되어 웹 콘텐츠를 표현할 경우 너무 많은 태그를 태깅하거나 정확하게 태깅된 태그 수가 적어 검색 결과의 정확률 등이 낮기 때문이다.

다음 표 4는 그림 6의 NDCG 값을 표로 나타낸

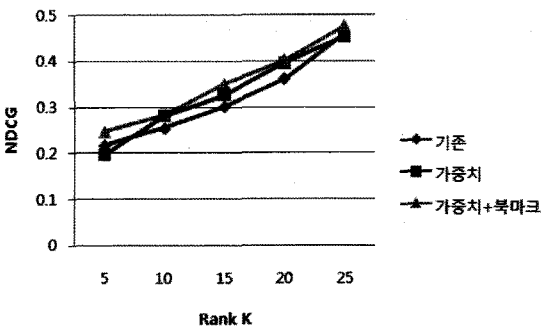

그림 6. 키워드 '공부'의 NDCG 비교

표 2. 기존 소셜 북마킹 웹사이트의 '영어공부' 상위 25개의 검색 결과

No.	Web Site	연관태그 가중치	연관태그 가중치 정규화값	북마크 인원수	북마크 인원수 정규화값	제안랭킹
1	웹초보의 Tech 2.1 :: 땡전 한	233	1	192	1	2
2	영어는 덤!!! :: 영어소설 무료	233	1	186	0.96875	1.96875
3	Korean-English-Korean	180	0.772532	169	0.880208	1.652741
4	Yappr - 비디오를 시청하면	217	0.93133	127	0.661458	1.592789
5	해커스영어 :: No.1 영어 정보	163	0.699571	99	0.515625	1.215196
6	:: Daily English - 대한민국	108	0.463519	83	0.432292	0.895811
7	오마이리딩 닷컴에 오신 것을	180	0.772532	83	0.432292	1.204824
8	만점비법 해커스토플 - 'TOE	165	0.708155	68	0.354167	1.062321
9	♣ 기술, 디자인, 엔터테인먼	41	0.175966	62	0.322917	0.498882
10	The Internet Movie Data	39	0.167382	60	0.3125	0.479882
11	무료 영어소설, 오디오북 Pr	108	0.463519	57	0.296875	0.760394
12	영국의 공영방송 BBC에서 운	216	0.927039	56	0.291667	1.218705
13	YBMsisa.com - 인터넷 영어	136	0.583691	53	0.276042	0.859733
14	LuxCozy(럭스코지) :: 영어속	162	0.695279	48	0.25	0.945279
15	Randall's ESL Cyber Listeni	128	0.549356	44	0.229167	0.778523
16	Wordbreak :: 단어를 외우는	143	0.613734	42	0.21875	0.832484
17	대한민국 No.1 외국어 교육	67	0.287554	43	0.223958	0.511512
18	http--weekstudy.coolschool.c	180	0.772532	42	0.21875	0.991282
19	Listen and Write - Di	145	0.622318	42	0.21875	0.841068
20	토익 전문 - 해커스토	82	0.351931	38	0.197917	0.549848
21	영어에서 관사를 쉽게 파악하	118	0.506438	34	0.177083	0.683521
22	Livemocha:LearnLanguagesO	182	0.781116	45	0.234375	1.015491
23	English Cube - 영어학습을	145	0.622318	32	0.166667	0.788984
24	YBM 어학시험 (TOEIC, JE	149	0.639485	29	0.151042	0.790527
25	[STUDY] 오마이리딩 닷컴-	180	0.772532	29	0.151042	0.923574

것으로 키워드 '공부'의 기존 검색 결과, 연관 태그의 가중치 값만으로 랭킹한 결과와 제안한 알고리즘의 누적 NDCG의 값으로 값이 높을수록 좋은 검색결과라 볼 수 있다. K의 값은 상위 페이지의 누적 수를 나타낸다.

다음 그림 7은 키워드 '아이폰'으로 검색된 결과의 NDCG 값을 나타낸다. 제안한 알고리즘으로 랭킹한 결과의 NDCG 값을 보면 기울기가 급격히 증가하는 것을 볼 수 있는데 이는 기존의 검색 결과나 연관 태그의 가중치 검색 결과보다 더 좋은 누적 점수를 받았다는 것을 의미한다. '아이폰'의 키워드는 최신 이슈어를 반영하기 위한 것으로 키워드의 특징 상 좁은 분야를 가져 연관 태그들 또한 한정된 태그들이

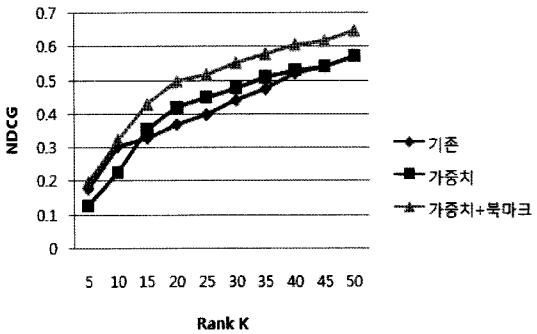

그림 7. 키워드 '아이폰'의 NDCG 비교

반복되어 가중치 값이 높게 나타난다. 북마크 인원수 또한 최근에 다양한 분야에서 사용자들이 북마크하

표 3. 제안한 알고리즘을 이용한 '영어공부' 상위 25개의 검색 결과

No.	Web Site	연관태그 가중치	연관태그 가중치 정규화값	북마크 인원수	북마크 인원수 정규화값	제안랭킹
1	웹초보의 Tech 2.1 :: 땡전 한	233	1	192	1	2
2	영어는 덤!!! :: 영어소설 무료	233	1	186	0.96875	1.96875
3	Korean-English-Korean	180	0.772532	169	0.908602	1.652741
4	Yappr - 비디오를 시청하면	217	0.93133	127	0.682796	1.592789
5	해커스영어::No.1영어정보	163	0.699571	99	0.532258	1.218705
12	영국의 공영방송 BBC에서 운	216	0.927039	56	0.301075	1.215196
7	오마이리딩 닷컴에 오신 것을	180	0.772532	83	0.446237	1.204824
8	만점비법 해커스토플 - 'TOE	165	0.708155	68	0.365591	1.062321
22	Livemocha:LearnLanguagesO	182	0.781116	45	0.241935	1.015491
18	http--weekstudy.coolschool.c	180	0.772532	42	0.225806	0.991282
14	LuxCozy(럭스코지) :: 영어속	162	0.695279	48	0.258065	0.945279
25	[STUDY] 오마이리딩 닷컴-	180	0.772532	29	0.155914	0.923574
6	:: Daily English - 대한민국	108	0.463519	83	0.446237	0.895811
13	YBMsisa.com - 인터넷 영어	136	0.583691	53	0.284946	0.859733
19	Listen and Write - Di	145	0.622318	42	0.225806	0.841068
16	Wordbreak :: 단어를 외우는	143	0.613734	42	0.225806	0.832484
24	YBM 어학시험 (TOEIC, JE	149	0.639485	29	0.155914	0.81753
23	English Cube - 영어학습을	145	0.622318	32	0.172043	0.790527
15	Randall's ESL Cyber Listeni	128	0.549356	44	0.236559	0.788984
11	무료 영어소설, 오디오북 Pr	108	0.463519	57	0.306452	0.778523
21	영어에서 관사를 쉽게 파악하	118	0.506438	34	0.182796	0.760394
26	토익 전문 - 해커스토	108	0.463519	29	0.155914	0.744278
20	대한민국No.1외국어교육	82	0.351931	38	0.204301	0.707484
17	♣ 기술, 디자인, 엔터테인먼	67	0.287554	43	0.231183	0.705651
9	The Internet Movie Data	41	0.175966	62	0.333333	0.683521

표 4. 키워드 '공부의 NDCG 값

K	기 존	가중치	가중치+북마크
5	0.215123	0.195319	0.246307
10	0.252151	0.279325	0.28229
15	0.299698	0.325726	0.349503
20	0.360035	0.39575	0.401054
25	0.459215	0.452162	0.474592

기 때문에 대부분의 웹페이지들이 고르게 북마크 인원수를 가지게 된다. 이로써 제안한 랭킹 알고리즘이 높은 NDCG 값을 가지게 되었으며 이는 상위페이지로 검색된 결과들이 키워드와 연관된 웹페이지들이 검색되었다는 것을 의미한다. 표 5는 키워드 '아이폰'

표 5. 키워드 '아이폰'의 NDCG 값

K	기 존	가중치	가중치+북마크
5	0.179338	0.128129	0.197708
10	0.304347	0.227418	0.325247
15	0.329142	0.352991	0.429719
20	0.369554	0.420675	0.497112
25	0.398526	0.449315	0.517649
30	0.440913	0.476438	0.552535
35	0.474217	0.509552	0.578371
40	0.520405	0.530795	0.606958
45	0.541046	0.541117	0.617279
50	0.57121	0.571284	0.647446

의 NDCG 상위 50개의 누적 값을 표로 나타내었다.

다음 그림 8은 '맛집' 키워드의 NDCG 값을 그래프로 표현하였다. '맛집' 키워드의 특징은 사용자들이 웹페이지를 북마킹 할 시 태그의 수가 적고 한정된 태그를 사용한다는 것이다. 예로, 맛집, 음식점, 요리, 지역이름 등을 사용하여 나타낸다. 또한 개인적으로 추천하고 싶은 맛집을 블로깅한 블로그 웹페이지 등을 표현하기 위한 개인적인 태그들도 많이 달려있다. 예로, 블로거의 이름, 음식이름, 음식점 이름 등을 가진다. 따라서 연관 태그 가중치 값이 작아지며 이를 사용한 제안 알고리즘 또한 값이 작아졌으나 북마크 인원수라는 값을 통해 기존의 검색 결과와 비슷한 결과를 도출할 수 있었다. 표 6은 이러한 NDCG 값을 표로 나타냈다.

그림 9는 키워드 '영어공부'로 위에서 보였던 '공부'라는 키워드에 비해 좀 더 제한을 두고 있어 연관 태그들의 빈도수가 높게 나타난다. 기존의 검색결과와 제안 알고리즘은 상위에서는 비슷한 연관 페이지

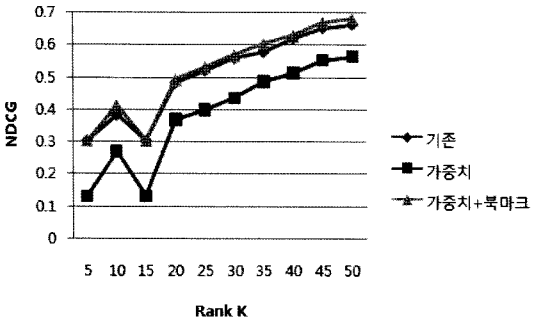

그림 8. 키워드 '맛집'의 NDCG 비교

표 6. 키워드 '맛집''의 NDCG 값

K	기 존	가중치	가중치+북마크
5	0.304358	0.133356	0.304358
10	0.384401	0.270246	0.413032
15	0.304358	0.133356	0.304358
20	0.482974	0.369286	0.493006
25	0.521771	0.399327	0.531929
30	0.558475	0.436289	0.568891
35	0.577945	0.487039	0.603873
40	0.619122	0.513271	0.630106
45	0.651906	0.553311	0.670145
50	0.662488	0.563892	0.680727

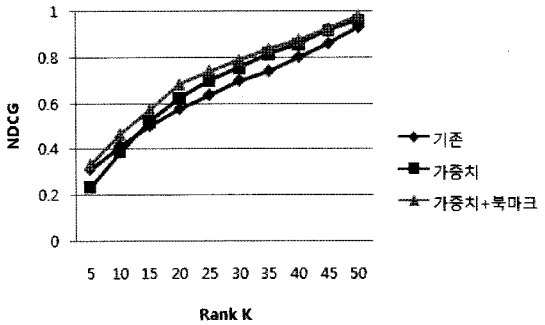

그림 9. 키워드 '영어공부'의 NDCG 비교

를 검색 결과로 도출하였지만 하위 페이지로 내려갈수록 제안 알고리즘이 연관된 페이지를 좀 더 상위에서 보여주고 있다. 이는 사용자들이 많은 웹페이지들을 검색할 시 하위페이지에도 연관도가 높은 웹페이지를 제공함으로서 유용한 정보제공 가능을 의미한다. 연관 태그 가중치 방법 또한 하위의 랭크에서는 비슷한 결과를 도출하였지만 사용자들이 많이 참고하는 첫 페이지에서는 본 논문에서 제안한 알고리즘이 더 좋은 결과를 나타내고 있다. 표 7은 키워드 '영어공부'의 NDCG 값을 나타낸다.

다음 표 8은 기존 랭킹결과, 연관태그 가중치를 통한 랭킹 결과와 제안 알고리즘을 적용한 Avg-NDCG값을 정리하였다. 기존의 검색결과는 사용자

표 7. 키워드 '영어공부'의 NDCG 값

K	기 존	가중치	가중치+북마크
5	0.313074	0.236987	0.3336
10	0.418632	0.388805	0.466838
15	0.502287	0.523337	0.572165
20	0.577109	0.625014	0.684279
25	0.638169	0.703495	0.741275
30	0.700334	0.753797	0.787989
35	0.740909	0.816949	0.83639
40	0.801598	0.85943	0.875918
45	0.86045	0.918029	0.923937
50	0.931115	0.961478	0.98

표 8. 제안 알고리즘의 검색 성능 개선

	기존 검색결과	연관태그 가중치	제안 알고리즘
Avg-NDCG	0.522672	0.495147	0.579095

들의 북마크순을 통해 랭킹된 결과로 연관태그 가중치의 랭킹보다 높은 값을 가진다. 이는 사용자들이 북마킹 시 태그를 정확히 붙이지 않거나 아예 태그를 붙이지 않기 때문에 연관 태그를 통한 랭킹은 NDCG 값이 떨어지게 되었다. 본 논문에서 제안한 알고리즘을 통한 랭킹은 기존 검색결과보다 좋은 성능을 보인다. 이는 소셜 북마킹 사이트 내 중요한 역할을 하는 연관 태그와 북마크 인원수를 동시에 활용함으로 향상된 검색결과를 보여준다.

5. 결론 및 향후연구

본 논문에서는 기존의 소셜 북마킹 시스템에서 독립적으로 검색에 활용되었던 태그 정보와 북마크 인원수를 혼합한 랭킹 알고리즘을 제안하여 소셜 북마킹의 특징을 가진 효율적인 검색 결과를 도출하였다. 이를 위해 선행연구로 진행하였던 태그 클러스터링 시스템을 통해 소셜 북마킹 사이트 내 웹 콘텐츠들에 태깅된 태그들 중 부정확한 태그들을 제거하여 연관도가 높은 태그들을 추출하였다. 또한 웹 콘텐츠 랭킹 시스템을 통해 추출된 연관 태그들을 기반으로 웹 콘텐츠의 관심도를 나타내는 북마크 수를 혼합한 뒤 정규화를 통해 웹 콘텐츠를 랭킹하여 제공하였다. 이로써 기존의 소셜 북마킹 시스템에서는 가지지 못한 다른 사용자들과의 공유를 나타내는 북마크 수, 즉 소셜의 의미를 가지는 검색 결과를 얻을 수 있었다.

향후 연구로는 기존에 존재하는 랭킹 알고리즘들 예로 구글에서 사용하는 PageRank, PageRank를 이용하여 폭소노미의 의미를 담은 FolkRank 등 다양한 랭킹알고리즘들과 비교 평가할 예정이다. 또한 북마크 인원수와 연관태그의 가중치뿐만 아니라 다양하게 존재하는 소셜 북마크 사이트의 특징을 활용하여 더욱 효율적인 검색 알고리즘을 연구할 예정이다.

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," *Internet Computing, IEEE*, Nov.-Dec. 2007.

[3] http://delicious.com/

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

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

[6] 이시화, 무효려, 이만형, 황대훈, "web2.0 환경에서의 Tag Clustering 시스템 설계 및 구현," 한국멀티미디어학회, Vol.10, No.1, pp. 251-254, 2007.

[7] 이시화, 이만형, 황대훈, "web2.0에서의 Tag Clustering을 통한 이미지 검색의 효율성 분석," 한국멀티미디어학회, Vol. 10, No. 2, 2007.

[8] 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.

[9] J. M. Kleinberg, "Authoritative sources in hyperlinked environment," *Journal of the ACM*, Vol.46, No.5, pp. 604-632, Sep, 1999.

[10] 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.

[11] A. Turpin and F. Scholer, "User performance versus precision measures for simple search tasks," in Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Seattle, Washington, USA, August 06-11, 2006). SIGIR '06. ACM, New York, NY, 11-18.

[12] W. Bruce Croft, Donald Metzler, and Trevor Strohman, *Search Engines: Information Retrieval in Practice*, 2009.

[13] Kalervo Järvelin and Jaana Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems (TOIS)*, v.20 n.4, p.422-446, October 2002.

[14] 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 졸업(학사)
2008년~현재 경원대학교 전자계산학과 석사과정
관심분야 : Web2.0, Semantic Web, Tag, Social Bookmarking, Ranking, Retrieval

이 시 화

2005년 서울보건대학 컴퓨터정보과 졸업
2005년 블루M 개발실 연구원
2007년 경원대학교 전자계산학과 석사과정 졸업
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