사용자 태그와 중심성 지수를 이용한 블로그 검색 성능 향상에 관한 연구

Enhancing the Performance of Blog Retrieval by User Tagging and Social Network Analysis

김은희(Eun-Hee Kim)*
정영미(Young-Mee Chung)**

초 록

최근 다양한 주제 분야의 블로그가 이용자의 정보요구를 충족시켜주는 웹 정보원 중 하나로 활용되고 있다. 본 연구에서는 블로그 페이지의 검색 성능을 향상시키기 위하여 이용자가 부여한 태그 및 트랙백을 이용하여 블로그 페이지의 검색 실험을 수행하였다. 실험을 위해 4,908개의 블로그 페이지와 각 페이지에 트랙백으로 연결된 다른 블로그 페이지의 URL을 수집하였다. 검색 자질로 본문의 용어에 이용자 태그를 추가하였을 경우와 네트워크 중심성 값을 반영하였을 경우 모두 검색 성능이 향상되었고, 본문 용어와 이용자 태그를 검색 자질로 함께 사용하고 여기에 중심성 값을 반영하였을 경우 가장 좋은 성능을 보였다.

ABSTRACT

Blogs are now one of the major information resources on the web. The purpose of this study is to enhance the performance of blog retrieval by means of user assigned tags and trackback information. To this end, retrieval experiments were performed with a dataset of 4,908 blog pages together with their associated trackback URLs. In the experiments, text terms, user tags, and network centrality values based on trackbacks were variously combined as retrieval features. The experimental results showed that employing user tags and network centrality values as retrieval features in addition to text words could improve the performance of blog retrieval.

키워드: 블로그 검색, 이용자 태그, 트랙백, 네트워크 중심성, 사회연결망분석
blog retrieval, user tags, trackbacks, network centrality, social network analysis

* 연세대학교 문헌정보학과 대학원(frubanara@naver.com) (제1저자)
** 연세대학교 문헌정보학과 교수(ymchung@yonsei.ac.kr) (교신저자)

■ 논문접수일자 : 2010년 2월 1일 ■ 최초심사일자 : 2010년 2월 28일 ■ 게재확정일자 : 2010년 3월 10일
■ 정보관리학회지, 27(1): 61-77, 2010.〔DOI:10.3743/KOSIM.2010.27.1.061〕

1. 서 론

최근 블로그는 이용자가 정보를 생산하고 공유하는 새로운 장이 되고 있다. 블로그는 기존의 웹 사이트와 비교하여 글을 작성하는 방법이 간단하고 이용자들 간에 정보를 공유하기 쉽기 때문에 이용이 급속히 증가하고 있다. 그리고 다양한 주제 분야의 블로그들이 이용자의 정보 요구를 충족시켜주는 중요한 정보원 중 하나로 활용되고 있다. 이에 따라, 블로그 페이지를 효율적으로 검색하여 이용자에게 제공할 필요성이 증가하였다.

블로그는 기존의 웹 사이트와 구분되는 특징적 기능을 가지고 있다. 블로그 페이지 작성자의 주관에 따라 부여하는 메타데이터인 '태그(tag)', 다른 블로그의 글과 자신의 글을 서로 링크시키는 '트랙백(trackback)', 특정 글에 대해 의견을 나누는 '댓글(comment)' 등이 그것이다. 태그는 본문이 텍스트 없이 이미지나 동영상만으로 구성된 경우나 보다 이용자 중심적인 검색 결과를 원하는 경우에 활용될 수 있다. 트랙백은 이용자가 다른 블로그의 글을 읽고 관련된 내용의 글을 자신의 블로그에 작성한 후 두 글을 서로 링크시키는 기능으로, 블로그들 사이의 연결고리를 만들어 주어 소통 네트워크를 만들어 내는 역할을 한다(*Wikipedia*). 정보원으로 활용할 가치가 큰 블로그 페이지일수록 여러 이용자들이 트랙백을 연결하게 되며 이러한 연결 관계를 통해 블로그 페이지 간의 네트워크 구조를 확인할 수 있다.

다양한 블로그 중에는 새로운 아이디어를 제안하며, 특정 주제에 대한 의견이나 사용해 본 상품 및 서비스에 대한 평가를 활발하게 제시하고, 다른 블로그들에게 많은 영향을 미치는 블로그가 존재한다. 영향력 있는 블로그는 정보원으로 활용할 가치가 큰 블로그이기 때문에, 이러한 블로그에 다수의 이용자들이 관심을 갖고 이로부터 정보를 획득한다(Agarwal 2008).

본 연구의 목적은 태그와 트랙백과 같은 블로그의 특징적 기능을 이용하여 블로그 페이지의 검색 성능을 향상시키는 데 있다. 이를 위해 이용자가 직접 부여한 태그를 본문의 용어와 함께 검색 자질로 사용하였다. 또한, 트랙백 연결 정도와 영향력 있는 블로그는 관련성이 있다는 가정 하에 트랙백에 의한 블로그 페이지 네트워크 분석을 통해 영향력 있는 블로그 페이지를 선정하고, 이러한 블로그 페이지에 가중치를 주는 방법으로 검색 성능을 향상시키고자 하였다. 수집된 블로그 페이지들을 대상으로 불용어를 제거하고 형태소를 분석하기 위해 파이썬 프로그램 및 지능형 형태소 분석기를 이용하였다. 그리고 사회연결망 분석 도구인 UCINET 6.0을 사용하여 블로그 페이지의 네트워크를 분석하였으며, Okapi 시스템의 BM25 함수로 블로그 페이지와 질의 간 유사도를 구하였다.

2. 선행 연구

빠른 속도로 생성되고 있는 대량의 블로그 페이지들 중에서, 이용자의 정보요구에 적합한 블로그 페이지를 효율적으로 제공하기 위한 방법들이 연구되고 있다. 이를 위해 블로그의 구조적 특성을 이용하여 블로그 페이지를 자동 분류하는 연구와 블로그 페이지 검색 성능을 향상시키고자 하는 연구들이 진행되었다. 또한,

영향력 있는 블로그를 판별을 통해 블로그 검색 성능을 개선하고자 하는 연구가 이루어졌다.

김기현(2009)은 블로그 이용자가 부여하는 태그를 이용하여 블로그 페이지의 자동 분류 성능을 향상시키고자 하였다. 수집한 블로그 페이지들을 대상으로, 본문의 용어만을 분류 자질로 사용한 실험, 본문의 용어와 이용자 태그를 분류 자질로 이용하는 실험, 태그 간 동시 출현 빈도에 기반한 유사도를 이용하여 확장한 태그와 본문의 용어를 분류 자질로 이용하는 실험, 본문의 태그와 트랙백으로 연결된 블로그 페이지에 출현한 태그 간의 유사도를 기반으로 확장한 태그를 본문의 용어와 함께 분류 자질로 사용하는 실험을 수행하였다. 이러한 실험을 통해 이용자가 부여한 태그나 태그 확장 과정을 적용하는 것이 자동 분류 성능 향상에 효과가 있음을 밝혀냈다.

김정훈 등(2009)은 블로그의 특성상 블로그 페이지의 구조적 평가를 통하여 내용적 평가가 암묵적으로 수행된다고 보고, 블로그의 구조적 속성을 기반으로 블로그 페이지를 평가 및 정렬하는 '블로그-랭크' 알고리즘을 제안하였다. 제안한 알고리즘은 특정 주제에 대한 블로거의 명성, 트랙백 연결성 그리고 댓글을 기반으로 한 사용자 반응성 점수로 구성되었다. 이러한 알고리즘을 적용한 블로그 검색 시스템을 구현하고 기존의 블로그 검색 시스템과 검색 효율성을 비교하였을 때, 기존의 검색 시스템보다 검색 성능이 개선된 것으로 나타났다.

Song, Chi, and Tseng(2007)은 영향력 있는 블로그를 판별함으로써 방대하고 복잡한 블로그 공간의 의견들을 요약하는 시스템을 제안하였다. 이를 위해 주어진 질의에 대해 관련된 블로그들을 검색하고 PageRank와 Song 등(2007)의 InfluenceRank를 이용하여 순위화한 결과를 비교한 후, 네트워크를 그려 상위에 랭크된 블로그들의 네트워크 상 위치를 확인하고 그들의 페이지를 모아 블로그 공간의 의견들을 요약하였다.

Mishne(2007)은 블로그 페이지가 작성된 시간, 댓글 수와 같은 특성을 이용하여 검색 성능을 향상시키고자 하였다. 특정 질의로 블로그 페이지를 검색하고 검색된 블로그 페이지들을 살펴본 결과, 특정 주제에 대해 많은 블로그 페이지가 작성되는 시기가 있고 이는 사회적 이슈와 연관된다는 사실을 알아냈다. 이를 통해 실험에 사용할 질의를 선정하였으며 또한, 사용자 반응성을 나타내는 댓글 수를 검색 결과에 반영하였다.

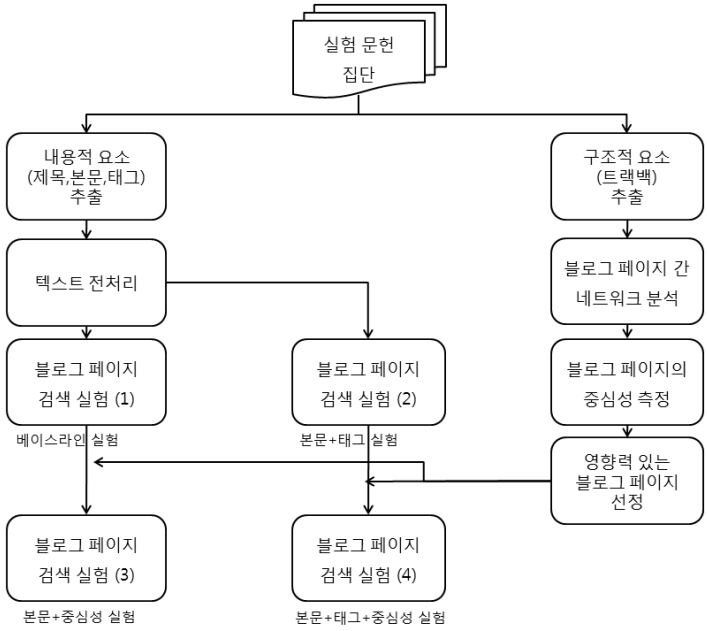
3. 블로그 검색 성능 향상 실험

3.1 실험 설계

3.1.1 실험 개요

본 연구는 기존의 블로그 페이지 검색에 이용자가 직접 부여한 태그와 다른 이용자가 연결한 트랙백을 활용함으로써 블로그 페이지 검색 성능을 향상시키고자 하는 것이다. 전체 실험은 네 단계로 이루어지며, 이러한 실험 과정을 정리하면 〈그림 1〉과 같다.

(1) 첫 번째 실험은 다음 단계 실험의 성능 평가 기준이 되는 베이스라인으로, 블로그 페이지의 본문에 출현한 용어만을 검색 자질로 이용하여 실험한다. 색인어 가중치로는 문헌길이에 의해 정규화를 시킨 Okapi TF를 사용한다.



〈그림 1〉 실험 개요

(2) 두 번째 실험은 첫 번째 단계에서 사용한 본문의 용어와 함께 태그를 검색 자질로 이용하여 실험한다.

(3) 세 번째 실험은 수집된 블로그 페이지의 트랙백 정보를 이용하여 블로그 페이지들의 네트워크를 분석하고, 블로그 페이지의 중심성 값을 구한다. 본 연구에서는 여러 중심성 지수 중, 내향 연결정도 중심성(in-degree centrality)과 외향 연결정도 중심성(out-degree centrality) 및 위세 중심성(eigenvector centrality)을 사용한다.

(4) 앞 단계의 실험을 통해 산출된 블로그 페이지의 중심성 값(중심성 가중치)을 첫 번째와 두 번째 단계의 블로그 페이지 색인어 가중치에 반영하고 재검색한다. 이를 위해 트랙백 수에 따른 검색 성능을 바탕으로 영향력 있는 블로그 페이지를 선정하고, 선정된 영향력 있는 블로그 페이지의 중심성 값과 각 중심성 값을 조합한 값을 색인어 가중치에 반영한다. 이때, 중심성 가중치를 색인어 가중치에 반영하는 비율을 변화시켜 최적의 검색 성능을 보이는 조건을 찾는다.

각 실험 단계에서는 아래와 같은 BM25 함수를 이용하여 블로그 페이지-질의 간 유사도 값을 구하고 n-순위 정확률(n=10, 50, 100)로 검색 성능을 살펴보았다. 그리고 전체 질의의 매크로 평균값을 통해 각 실험 단계의 검색 성능 및 검색 성능 향상률을 살펴보았다.

$$S(Q,D)=\sum_{t\in Q} w^{(1)}\left(\frac{(k_1+1)tf}{K+tf}\right)\left(\frac{(k_3+1)qtf}{k_3+qtf}\right)+\left(k_2|Q|\frac{avdl-dl}{avdl+dl}\right)$$

위 식에서 색인어 가중치(Okapi TF) 부분의 K는 k_1((1-b)+b(dl/avdl))을 나타내며, k_1은 1.2, b는 0.75로 설정하였다. 적합성 가중치 $w^{(1)}$은 역문헌빈도(IDF) 요소인 log(N/df) 값을 사용하였다. 질의어 가중치 부분의 k_3과 마지막 부분의 k_2 값은 다른 TREC 실험에서와 같이 0으로 처리하였다.

3.1.2 실험 집단

본 연구에서는 블로그 서비스가 활발하게 이용되고 있는 티스토리(www.tistory.com)와 메타 블로그 서비스인 믹시(www.mixsh.com)의 카테고리 중에 정치, 영화, 여행, 스포츠, IT 다섯 가지 카테고리를 선정하고 각 카테고리로부터 1,000개씩 블로그 페이지를 수집하였다. 수집된 블로그 페이지는 2009년 8월부터 9월 사이에 작성되었으며, 블로그 페이지에 트랙백으로 연결된 블로그 페이지의 URL도 함께 수집하였다.

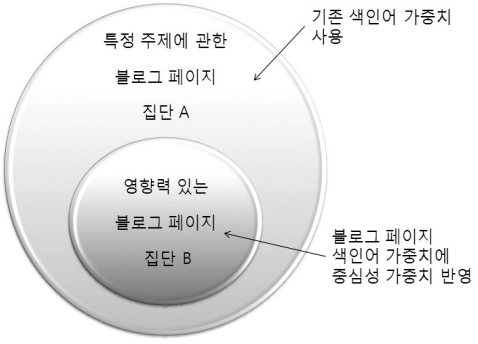
수집한 5,000개의 블로그 페이지들 중, 전처리 과정에서 오류를 일으키는 것을 삭제하고 최종적으로 4,908개의 블로그 페이지를 대상으로 실험하였다. 이와 같은 내용을 정리하면 〈표 1〉과 같다.

또한, 본 연구에서 사용한 질의는 블로그 페이지 수집 기간 동안에 인기 검색어로 선정된 단어들을 수집하여 상기 카테고리 주제에 맞게 분류한 후에 카테고리마다 6개씩, 전체 30개의 질의를 임의로 선정하였다. 〈표 2〉와 같이 카테고리 별 6개의 질의는 짧은 질의 3개와 긴 질의 3개로 구성하였다.

3.1.3 트랙백 정보에 의한 블로그 네트워크 분석

본 연구에서는 사회 네트워크 연구 분야에서 보편적으로 사용되는 도구인 UCINET 6.0으로 각각의 블로그 페이지에서 수집한 트랙백 정보를 이용하여 블로그 페이지들의 네트워크 구조를 분석하고 중심성 값을 산출하였다. 이때, 중심성 지수는 내향 연결정도 중심성, 외향 연결정도 중심성, 위세 중심성 지수를 사용하였으며, 이렇게 산출된 중심성 값과 중심성 값들을 조합한 값으로 영향력 있는 블로그 페이지의 색인어 가중치 값을 증가시켰다. 연결정도 중심성은 한 노드와 직접 연결된 다른 노드들의 합을 가지고 중심성을 측정하며, 외향 연결정도 중심성은 한 노드에서 다른 모든 노드들로 보내는 링크의 수로 측정한다. 반면 위세 중심성은 각 노드와 연결 관계가 있는 다른 노드들의 중심성 값을 이용하는 지수로서 위세가 높은 노드와 많이 연결된 노드일수록 중심성 값이 커진다(김용학 2007).

본 연구에서는 트랙백 수에 따라 블로그 페이지로 구분하여 검색 실험하고 가장 높은 검색 성능을 보이는 집단을 영향력 있는 블로그 페이지를 선정하였다. 이러한 내용을 정리하면 〈그림 2〉와 같다.



〈그림 2〉 블로그 페이지의 중심성 가중치 반영 대상

〈표 1〉 실험 집단 통계

	IT	영화	여행	스포츠	정치	전체
블로그 페이지 수	988	988	977	969	986	4,908
태그 수	10,533	11,134	9,727	7,169	13,630	52,193
트랙백 수	734	906	184	256	821	2,901
태그 평균	10.66	11.27	9.96	7.40	13.82	10.63
트랙백 평균	0.74	0.92	0.19	0.26	0.83	0.59

〈표 2〉 주제 분야별 실험용 질의

카테고리	질 의	구 분
IT	닌텐도	짧은 질의
	아이팟터치	짧은 질의
	트위터	짧은 질의
	효율적인 블로그 활동을 위한 다양한 tip들	긴 질의
	RSS로 블로그 포스트 간편하게 읽기	긴 질의
	새로 출시된 노트북 & 넷북	긴 질의
영화	왓치맨	짧은 질의
	영화 해운대	짧은 질의
	영화 국가대표	짧은 질의
	감동이 있는 다큐멘터리 영화, 워낭소리	긴 질의
	벤자민버튼의 시간은 거꾸로 간다	긴 질의
	인도판 헬렌켈러 이야기, 영화 블랙	긴 질의
여행	제주도 여행	짧은 질의
	일본 여행	짧은 질의
	홍콩 여행	짧은 질의
	예술의 도시 파리로 떠난 여행	긴 질의
	홀로 떠난 인도 배낭여행	긴 질의
	즐거운 추억이 가득한 미국 여행기	긴 질의
스포츠	추신수 활약	짧은 질의
	박지성	짧은 질의
	2009 WBC	짧은 질의
	김연아, 세계선수권대회에서 세계신기록으로 우승	긴 질의
	박철우 폭행 사건으로 본, 체육계 폭력 불감증	긴 질의
	야구를 보는 또 다른 묘미, 연예인 시구	긴 질의
정치	신종플루	짧은 질의
	선덕여왕	짧은 질의
	정운찬 국무총리	짧은 질의
	그룹 2PM 리더 박재범 탈퇴로 본 우리 사회 단상	긴 질의
	사실로 드러난 MBC신경민 아나운서 교체	긴 질의
	구글 유튜브, 정부의 실명제 실시 거부	긴 질의

〈그림 2〉와 같은 블로그 페이지 집단 A 중에서 영향력 있는 블로그 페이지로 선정된 집단 B의 기존 색인어 가중치에 중심성 가중치를 반영하고, 그 외의 블로그 페이지들은 변함없이 기존의 색인어 가중치를 사용하였다.

본 연구에서 영향력 있는 블로그 페이지의 중

심성 가중치를 반영한 색인어 가중치 산출 공식은 다음과 같다. 영향력 있는 블로그 페이지의 중심성 가중치에 파라미터 s 값을 곱하고 이를 기존 색인어 가중치에 더함으로써, 블로그 페이지의 새로운 색인어 가중치를 산출하였다.

$$w(t)_{new} = w(t) + s \cdot w_{cen}$$

$w(t)_{new}$: 중심성 가중치를 반영한 블로그 페이지의 색인어 가중치
$w(t)$: 기존의 블로그 페이지 색인어 가중치
w_{cen} : 중심성 가중치
s : 파라미터

이 공식의 파라미터 s는 기존 블로그 페이지의 색인어 가중치에 중심성 가중치를 반영하는 비율을 나타내는 것으로, 본 실험에서는 s 값 0.0을 시작으로 0.1씩 증가시키며 중심성 가중치를 반영하는 비율을 조절하여 블로그 페이지 검색 성능을 최적화하였다.

3.2 실험 결과 분석 및 평가

임의로 선정한 30개 질의와 실험 집단을 구성하는 블로그 페이지에서 추출한 용어나 용어 및 태그를 대상으로 질의와의 유사도를 구하고 n-순위 정확률을 구하였다. 한편, 블로그 페이지 간 네트워크 분석을 통해 선정한 영향력 있는 블로그 페이지의 중심성 값을 블로그 페이지의 색인어 가중치에 반영한 후 질의와 블로그 페이지 간 유사도를 구한 것을 바탕으로 n-순위 정확률을 구하였다. 이러한 실험들의 결과는 다음과 같다.

3.2.1 베이스라인 실험 결과

실험의 기준이 되는 베이스라인 실험의 경우, BM25 함수를 이용하여 블로그 페이지 본문을 대상으로 질의와의 유사도를 계산하고 각 질의마다 10-순위 정확률 p(10), 50-순위 정확률 p(50), 100-순위 정확률 p(100)을 구한 다음 매크로 평균으로 전체 성능을 살펴보았다. 30개 질의에 대한 평균 n-순위 정확률을 살펴보면 p(10)은 0.69, p(50)은 0.41, p(100)은 0.25로 나타났다.

3.2.2 본문+태그를 이용한 실험 결과

두 번째 단계의 실험에서는 베이스라인 실험의 대상인 블로그 페이지의 본문에 블로그 페이지 작성자가 부여한 태그 전체를 추가한 뒤, BM25 함수를 이용하여 질의와의 유사도를 계산하였다. 그 결과, p(10)은 0.73, p(50)은 0.47, p(100)은 0.28로 산출되었다. 베이스라인 실험과 본문에서 추출한 용어에 태그 전체를 추가하여 검색 자질로 사용한 두 번째 단계의 실험 결과를 비교해 보면 〈표 3〉과 같다.

〈표 3〉 베이스라인과 본문+태그 실험 결과 비교

	본문	본문+태그	성능 향상률
P(10)	0.69	0.73	5.5%
P(50)	0.41	0.47	11.89%
P(100)	0.25	0.28	11.19%

〈표 3〉에서 볼 수 있듯이, 본문과 태그를 함께 사용한 실험의 경우 P(10)이 0.73으로 0.69인 베이스라인 실험에 비해 성능이 5.5% 향상되었다. P(50)은 0.41에서 0.47로 성능이 11.89% 향상되었으며, P(100)은 0.25에서 0.28로 성능이 11.19% 향상되는 것으로 나타났다. 이처럼 본문만을 대상으로 검색하는 것보다 본문과 함께 태그를 사용하는 것이 블로그 페이지 검색 성능을 높이는데 효과적임을 알 수 있다.

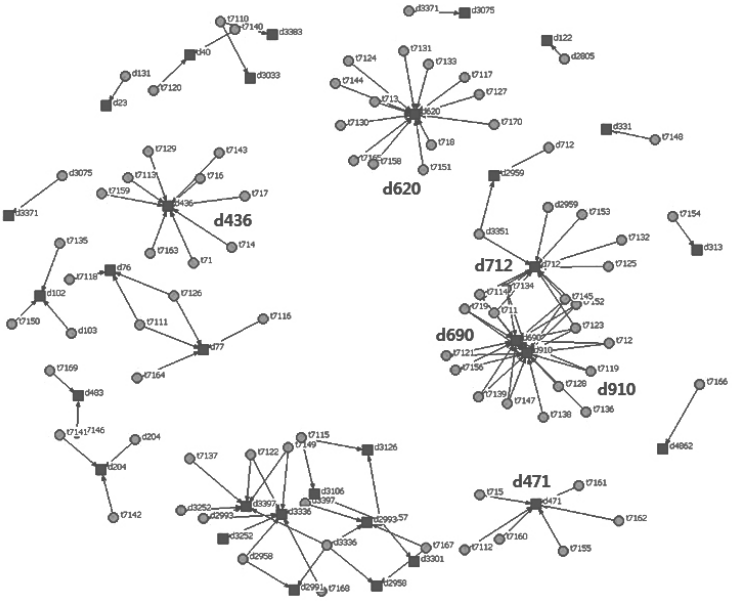
3.2.3 트랙백 정보에 의한 블로그 네트워크 분석 결과

각각의 질의별로 검색된 블로그 페이지들 중에 트랙백으로 연결된 URL이 포함되어 있는 블로그 페이지들을 대상으로 UCINET 6.0을 통해 블로그 네트워크 분석을 수행하였다.

네트워크 분석을 통해 산출된 중심성 값들의 30개 질의에 대한 평균을 구해보면, 내향 연결정도 중심성 값은 0.886, 외향 연결정도 중심성 값은 0.886, 위세 중심성 값은 0.071로 나타났다. 중심성 값의 산출 결과, 위세 중심성 값이 0에서 1 사이인 것과 달리 내향 연결정도 중심성과 외향 연결정도 중심성은 10이상의 값도 보이는 등 중심성 값의 범위가 다르기 때문에, 이어지는 실험에서 중심성 값을 조합하여 색인어 가중치에 반영할 때에 내향 연결정도 중심성과 외향 연결정도 중심성 값은 각 질의별 최대값으로 정규화한 다음 사용하였다.

예를 들어, 특정 질의에 대해 검색된 상위 30위 안 블로그 페이지들의 네트워크 구조를 살펴보면 다음 〈그림 3〉과 같다.

〈그림 3〉 속의 원 형태의 노드는 트랙백을 보내는 블로그 페이지이고, 네모 형태의 노드는 트랙백을 받는 블로그 페이지를 나타낸다. d690,



〈그림 3〉 특정 질의 관련 블로그 페이지들의 네트워크 구조

d620, d910과 같이 굵은 글씨로 표시된 노드들은 다수의 블로그 페이지들로부터 트랙백을 받는 블로그 페이지에 해당한다.

3.2.4 본문에 중심성 가중치를 반영한 실험 결과

본 실험에서는 블로그 페이지 본문에 대해 내향 연결정도 중심성, 외향 연결 중심성, 위세 중심성 값을 반영하여 재검색하고 이에 따른 검색 성능의 변화를 살펴보았다. 이를 위해, 트랙백 수에 따른 검색 실험을 통해 영향력 있는 블로그 페이지를 선정하고, 이러한 블로그 페이지의 중심성 가중치에 곱하는 s 값을 변화시켜 가장 높은 검색 성능을 보이는 s 값이 무엇인지 살펴보았다.

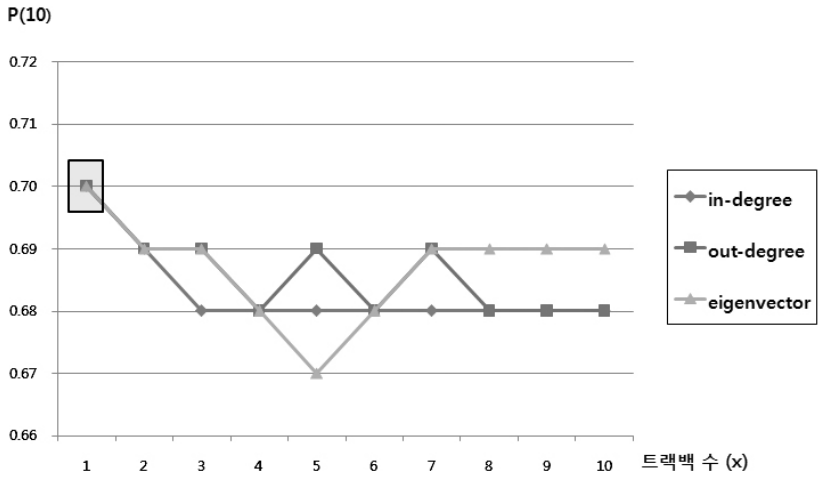
(1) 트랙백 수의 변화에 따른 실험 결과

본 실험에서는 트랙백 수에 따라 블로그 페이지를 구분한 다음, 중심성 가중치를 반영한 새로운 색인어 가중치를 사용하여 검색 성능을 평가해 보았다. 이 때, 중심성 가중치에 곱하는 s 값을 1.0으로 함으로써, 색인어 가중치와 동일한 비율로 중심성 가중치를 반영하여 실험하였다. 이러한 실험 결과를 그래프로 정리하면 〈그림 4〉와 같다.

〈그림 4〉에서 볼 수 있듯이, 내향 연결정도 중심성, 외향 연결정도 중심성, 위세 중심성 지수 모두 트랙백 수가 1개 이상인 블로그 페이지에 중심성 가중치를 반영하였을 때 가장 높은 p(10)을 보였다. 이러한 실험 결과를 바탕으로, 본문을 대상으로 중심성 가중치를 반영하는 실험의 경우는 연결된 트랙백 수가 1개 이상인 블로그 페이지를 영향력 있는 블로그 페이지로 선정하였다.

(2) 파라미터 s 값의 변화에 따른 실험 결과

각각의 질의별로 선정된 영향력 있는 블로그 페이지 즉, 연결된 트랙백 수가 1개 이상인 블로그 페이지를 대상으로, 중심성 지수 유형과 중심성 가중치에 곱해지는 파라미터 s 값의 변



〈그림 4〉 트랙백 수에 따른 중심성 지수별 10-순위 정확률(본문)

화에 따른 검색 성능을 10-순위 정확률을 기준으로 정리하면 〈그림 5〉와 같다.

〈그림 5〉에서 볼 수 있듯이, 중심성 지수별로 비교했을 때 내향 연결정도 중심성의 p(10)이 가장 높았고, 위세 중심성의 p(10)은 그보다 조금 낮으며, 외향 연결정도 중심성의 p(10)이 가장 낮았다. 그리고 전체적으로 내향 연결정도 중심성 가중치에 s값 1.5를 곱하여 색인어 가중치에 반영한 경우의 p(10)이 0.71로 가장 높은 검색 성능을 보였다.

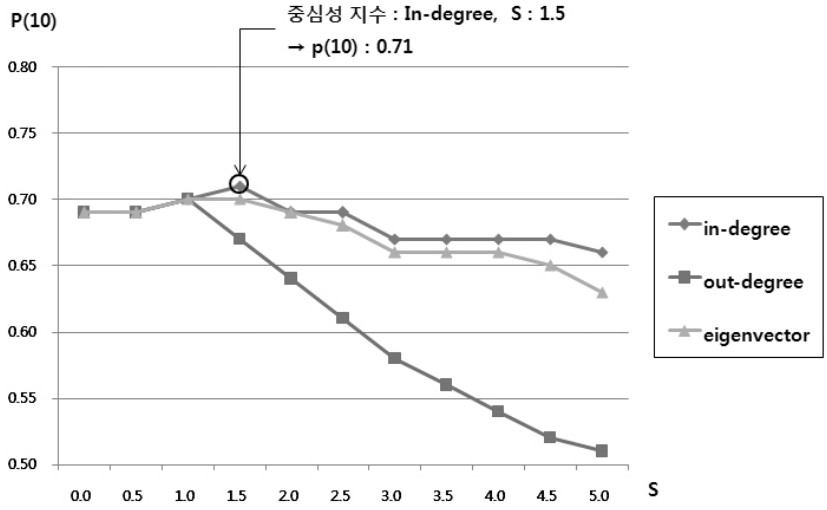
추가로, 두 개 이상의 중심성 값을 조합하여 색인어 가중치에 반영한 실험 결과, 모두 베이스라인보다 낮거나 동일한 성능을 보여 중심성 지수를 조합하는 것보다 각각의 중심성 지수를 단독으로 이용하는 것이 더 효과적인 것으로 확인됐다.

(3) 최적의 조건을 반영한 실험 결과

본 연구에서 중심성 가중치를 적용할 최소 트랙백 수, 중심성 지수 유형 혹은 중심성 지수 조합 유형, 중심성 가중치에 곱하는 파라미터 s 값을 변화시켜 실험한 결과, 블로그 페이지에 연결된 트랙백 수가 한 개 이상인 블로그 페이지를 대상으로 내향 연결정도 중심성 지수를 사용하며 s 값 1.5를 중심성 가중치에 곱하는 경우의 검색 성능이 가장 높은 것으로 나타났다. 이러한 변수 조건으로 블로그 페이지의 색인어 가중치에 중심성 가중치를 반영하여 재검색한 결과를 정리하면, 전체 질의에 대한 평균 n-순위 정확률은 p(10)이 0.71, p(50)이 0.47, p(100)은 0.28로 산출되었다.

본문만을 대상으로 한 베이스라인 실험과 영향력 있는 블로그 페이지로 선정된 블로그 페이지의 내향 연결정도 중심성 값에 s 값 1.5를 곱하여 기존 색인어 가중치에 더해준 실험의 결과를 비교해 보면 〈표 4〉와 같다.

〈표 4〉와 같이, p(10)은 0.69에서 0.71로 3.29%, p(50)은 0.41에서 0.47로 12.14%, p(100)은



〈그림 5〉 s 값 변화에 따른 중심성 지수별 10-순위 정확률(본문)

〈표 4〉 베이스라인 실험과 본문+중심성 실험 결과 비교

	본문	본문+중심성	성능 향상률
P(10)	0.69	0.71	3.29%
P(50)	0.41	0.47	12.14%
P(100)	0.25	0.28	11.50%

0.25에서 0.28로 11.5%로 향상되었다. 즉, 블로그 페이지 검색에 영향력 있는 블로그 페이지의 중심성 가중치를 반영하는 방법을 사용하는 것이 검색 성능 향상에 도움이 되는 것으로 나타났다.

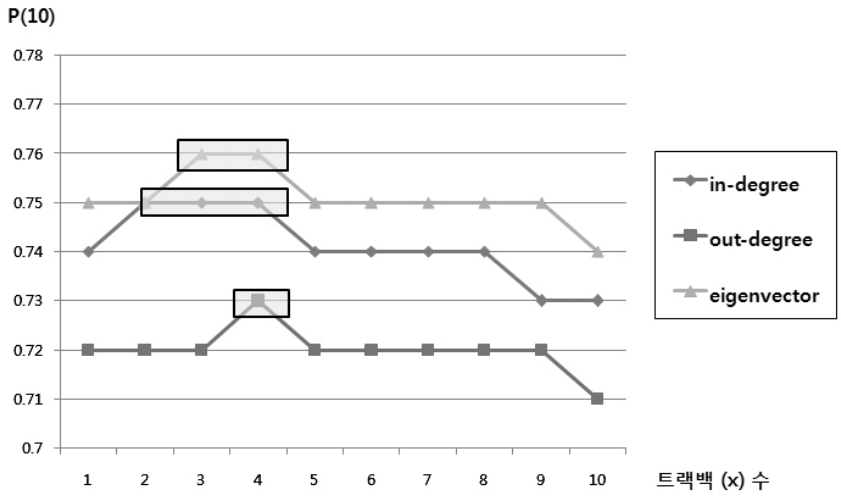
3.2.5 본문+태그에 중심성 가중치를 반영한 실험 결과

마지막으로, 블로그 페이지 본문에 태그를 추가하고 중심성 가중치를 반영한 뒤 재검색하였다.

(1) 트랙백 수의 변화에 따른 실험 결과

본 실험에서는 트랙백 수에 따라 블로그 페이지를 구분한 다음 중심성 가중치를 반영한 새로운 색인어 가중치를 사용하여 검색 성능을 평가해 보았다. 이 실험에서도 중심성 가중치에 곱하는 s 값을 1.0으로 하여, 색인어 가중치와 동일한 비율의 중심성 가중치를 반영하여 실험하였다. 이러한 실험의 결과를 그래프로 정리하면 〈그림 6〉과 같다.

〈그림 6〉에서 볼 수 있듯이, 세 가지 중심성 지수 모두 특정 범위 이후에는 p(10)이 감소하였다. 공통적으로 4개 이상의 트랙백을 가진 블로그 페이지에 중심성 가중치를 반영하였을 때 가장 높은 p(10)을 보였다. 이러한 실험 결과를 바탕으로, 본문과 태그를 대상으로 중심성 가중치를 반영하는 실험의 경우는 연결된 트랙



〈그림 6〉 트랙백 수에 따른 중심성 지수별 10-순위 정확률(본문+태그)

백 수가 4개 이상인 블로그 페이지를 영향력 있는 블로그 페이지로 선정하였다.

(2) 파라미터 s 값의 변화에 따른 실험 결과

각각의 질의에 따라 선정된 영향력 있는 블로그 페이지 즉, 연결된 트랙백 수가 4개 이상인 블로그 페이지를 대상으로, 중심성 지수 유형과 중심성 가중치에 곱해지는 파라미터 s 값의 변화에 따른 검색 성능을 10-순위 정확률을 기준으로 정리하면 〈그림 7〉과 같다.

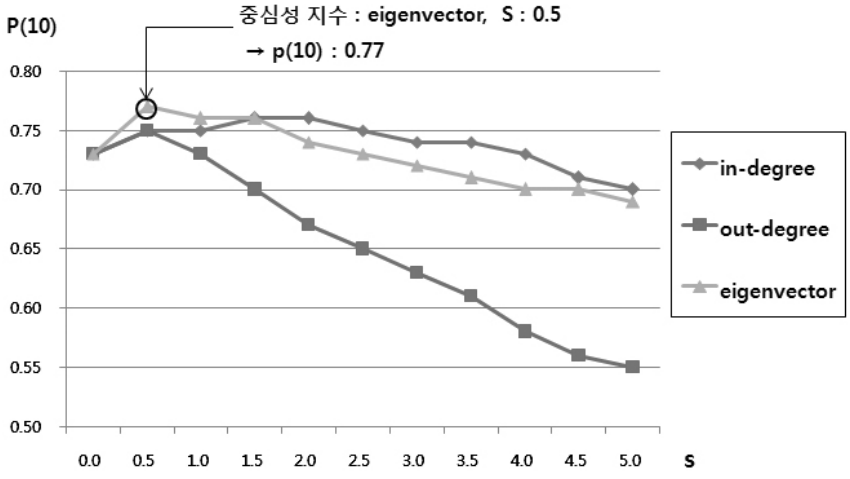
〈그림 7〉에서 볼 수 있듯이, 본문만을 대상으로 s 값을 변화시켜 적용한 결과와 유사하게 세 종류의 중심성 지수 모두 특정 s 값에서 가장 높은 p(10)을 보이고 그 보다 s 값이 커질수록 p(10)은 오히려 감소하였다. 그리고 본문만을 대상으로 실험한 경우와 달리, 본문과 함께 태그를 이용하였을 때 가장 높은 p(10)을 보이는 것은 위세 중심성 가중치에 s 값 0.5를 곱하여 색인어 가중치에 반영한 경우였다.

(3) 최적의 조건을 반영한 실험 결과

블로그 페이지의 본문과 태그를 대상으로 중심성 가중치를 적용할 최소 트랙백 수, 중심성 지수 유형, 중심성 가중치에 곱하는 파라미터(s)를 변화시키며 실험한 결과, 연결된 트랙백 수가 4 개 이상인 블로그 페이지를 대상으로 위세 중심성 지수를 사용하고 s 값으로 0.5를 적용하는 경우가 가장 높은 검색 성능을 보였다. 이러한 변수 조건으로 블로그 페이지 본문과 태그의 색인어 가중치에 중심성 가중치를 반영하여 재검색한 실험의 30개 질의에 대한 평균 n-순위 정확률을 정리해보면, p(10)은 0.77, p(50)은 0.54, p(100)은 0.33으로 산출되었다.

베이스라인 실험과 본문과 태그를 검색 자질로 사용하고 위세 중심성 값을 블로그 페이지의 색인어 가중치에 반영한 실험 결과를 비교해보면 〈표 5〉와 같다.

〈표 5〉와 같이, 본문과 태그를 검색 자질로 사용하고 위세 중심성 값을 반영한 실험의 P(10)은 0.77로 베이스라인 실험의 P(10)인 0.69에



〈그림 7〉 s 값 변화에 따른 중심성 지수별 10-순위 정확률(본문+태그)

〈표 5〉 베이스라인과 본문+태그+중심성 실험 결과 비교

	본문	본문+태그+중심성	성능 향상률
P(10)	0.69	0.77	10.40%
P(50)	0.41	0.54	23.65%
P(100)	0.25	0.33	23.69%

비해 10.4% 성능이 향상되었으며, P(50)은 0.41에서 0.54로 23.65%, P(100)은 0.25에서 0.33으로 23.69% 성능이 크게 향상되었다.

3.2.6 종합평가

본 연구에서는 블로그 페이지의 검색 성능을 향상시키기 위해 네 단계의 실험을 진행하였다. 각 단계별 n-순위 정확률(n=10, 50, 100)과 베이스라인 실험과 비교하였을 때의 성능 향상률을 정리하면 〈표 6〉과 같다.

〈표 6〉의 결과를 10-순위 정확률을 기준으로 살펴보면, 베이스라인 실험에 비해 본문과 태그를 이용한 실험은 5.5%, 본문에 중심성 가중치를 반영한 실험은 3.29% 성능이 향상되었으며, 본문과 태그를 이용하고 중심성 가중치를 반영한 실험은 10.4%로 가장 높은 검색 성능 향상률을 보였다. 이처럼 베이스라인 실험에 비해 본문과 태그를 이용한 실험, 본문에 중심성 가중치를 반영한 실험, 본문과 태그를 이용하고 중심성 가중치를 반영한 실험 모두 성능이 향상되는 것으로 나타났다.

더불어, 질의 길이에 따른 검색 성능 변화를 10-순위 정확률 기준으로 살펴보면 〈그림 8〉과 같다.

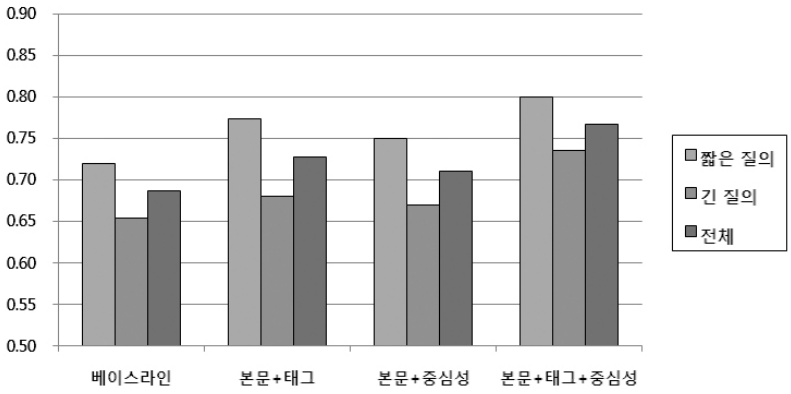
〈그림 8〉에서 볼 수 있듯이, 모든 단계의 실험에서 긴 질의의 P(10)보다 짧은 질의의 P(10)이 높게 나타났다. 그리고 짧은 질의와 긴 질의 모두 본문과 태그를 검색 자질로 사용하고 중심성 가중치를 반영한 마지막 단계 실험의 P(10)이 가장 높은 것으로 나타났다. 그러나 베이스라인 실험과 마지막 단계 실험의 결과를 비교했을 때, 짧은 질의보다 긴 질의의 검색 성능 향상률이 더 높은 것으로 나타났다.

또한, 블로그 페이지를 수집한 카테고리 별 검색 성능과 검색 성능 향상률을 베이스라인 실험과 마지막 단계 실험의 10-순위 정확률을 기준으로 살펴보면 〈표 7〉과 같다.

〈표 7〉과 같이, 10-순위 정확률이 가장 높은 카테고리는 영화 카테고리이며 가장 낮은 것은 여행 카테고리였다. 그리고 카테고리에 따라 검색 성능 향상률에 차이가 있으며, 여행 카테고리의 검색 성능 향상률이 21.21%로 가장 높고

〈표 6〉 종합 성능 비교

	P(10)	성능향상률	P(50)	성능향상률	P(100)	성능향상률
베이스라인	0.69	(비교대상)	0.41	(비교대상)	0.25	(비교대상)
본문+태그	0.73	5.5%	0.47	11.89%	0.28	11.19%
본문+중심성	0.71	3.29%	0.47	12.14%	0.28	11.5%
본문+태그+중심성	0.77	10.40%	0.54	23.65%	0.33	23.69%



〈그림 8〉 질의 길이별 10-순위 정확률

〈표 7〉 각 카테고리별 베이스라인과 본문+태그+중심성 실험 비교

	베이스라인	본문+태그+중심성	성능향상률
IT	0.70	0.78	10.64%
영화	0.82	0.88	7.55%
여행	0.43	0.55	21.21%
스포츠	0.77	0.80	4.17%
정치	0.72	0.82	12.24%
전체	0.69	0.76	10.04%

스포츠 카테고리의 검색 성능 향상률이 4.17%로 가장 낮았다. 이처럼 여행 카테고리는 카테고리 유형 중에 10-순위 정확률이 가장 낮음에도 불구하고, 검색 성능 향상률이 가장 높게 나타났다.

결론적으로 블로그 검색 시 본문 용어 이외에 이용자가 부여한 태그와 트랙백에 기반한 블로그 네트워크의 중심성 지수를 자질로 추가할 경우 검색 성능이 10-순위 정확률에서 최대 10.4%까지 향상되는 것으로 나타났다.

4. 결 론

본 연구에서는 최근 크게 증가하고 있는 블로그 페이지를 대상으로 태그와 트랙백과 같은 블로그의 특징적 기능을 이용하여 블로그 페이지의 검색 성능을 향상시키고자 하였다.

본 연구를 통해 얻어진 결과는 다음과 같다.

첫째, 본문에서 추출한 용어와 태그를 함께 검색 자질로 사용할 경우 본문 용어만을 자질로 사용한 실험에 비해 검색 성능이 10-순위 정확률에서 5.5% 향상되었으며, 본문과 중심성 가중치를 함께 사용할 경우에는 3.29% 향상률을 보여 태그와 중심성 가중치가 모두 검색 성

능을 향상시키는 것으로 나타났다.

둘째, 본문 용어에 태그와 중심성 가중치를 모두 결합하였을 경우 본문 용어만을 사용하였을 때보다 10.40%의 높은 성능 향상률을 보였다.

셋째, 카테고리별 검색 성능 비교에서는 가장 검색 성능이 높은 카테고리는 영화 카테고리이고 가장 검색 성능이 낮은 것은 여행 카테고리였다. 그러나 본문 용어에 태그와 중심성 가중치를 결합할 경우의 검색 성능 향상률은 여행 카테고리가 21.21%로 가장 높았고, 스포츠 카테고리가 4.17%로 가장 낮은 것으로 나타났다.

결론적으로, 블로그 페이지의 검색 성능을 향상시키기 위해 블로그 페이지 작성자가 부여한 태그와 트랙백을 이용한 블로그 페이지 네트워크 분석을 통해 산출한 중심성 값을 블로그 페이지의 색인어 가중치에 반영하는 것이 효과가 있는 것으로 나타났다. 이 연구 결과는 실험집단의 규모를 확대한 실험에서 검증한 다음 실제 블로그 검색 서비스에서 검색 성능을 높이기 위한 방법으로 활용될 수 있을 것으로 보인다.

참 고 문 헌

김기주, 최영식. 2007. 소규모 집단의 웹 사이트들을 위한 사이트 순위 결정 알고리즘. 『한국인터넷정보학회: 학술대회논문집』, 8(1): 379-383.

김기현. 2009. 『이용자 태그 확장을 통한 블로그 자동분류 성능 향상에 관한 연구』. 석사학위논문, 연세대학교대학원, 문헌정보학과.

김영주. 2005. 『블로그: 1인 미디어의 가능성과 한계』. 서울: 한국언론재단.

김용학. 2007. 『사회 연결망 분석』. 서울: 박영사.

김정훈, 윤태복, 이지형. 2009. 블로그의 구조적 특성을 고려한 효율적인 블로그 검색 알고리즘. 『정보과학회논문지: 소프트웨어 및 응용』, 36(7): 580-589.

김지수. 2004. 블로그의 사회문화적 이슈. 『정보통신 정책』, 16(8): 18-36.

박한우. 2007. 블로그에 나타난 정치인 네트워크 - 17대 국회의원을 대상으로. 『한국언론학보』, 51(3): 385-406.

이경희, 김민구, 박승규. 2003. 문서간의 유사도를 이용한 개선된 PageRank 알고리즘. 『한국정보과학회: 학술발표논문집, 30(2): 169-171.

이재현. 2005. 블로그와 저널리즘. 『관훈 저널』, 2005년 봄호: 22-30.

정영미. 2005. 『정보검색연구』. 서울: 구미무역(주) 출판부.

홍성국. 2004. 『블로그의 속성과 이용 동기에 관한 연구』. 석사학위논문, 서강대학교대학원, 신문방송학과.

Adar, E., L. Zhang, L. Adamic, and R. Lukose. 2004. "Implicit structure and the dynamics of blogspace." *Proceed-*

ings of the 13th International World Wide Web Conference.

Adamic, L. A. and N. Glance. 2005. "The political blogosphere and the 2004 U.S. election: divided they blog." *Proceedings of the 3rd International Workshop on Link Discovery*, 36-43.

Agarwal, N. 2008. "A study of communities and Influence in Blogosphere." *Proceedings of the 2nd SIGMOD PhD Workshop on Innovative Database Research*, 19-24.

Agarwal, N. and H. Liu. 2008. "Blogosphere: Research issues, tools, and applications." *KDD Explorations*, 10(1): 19-29.

Agarwal, N., H. Liu. L. Tang, and PS. Yu. 2008. "Identifying the influential bloggers in a community." *Proceedings of the International Conference on Web Search and Web Data Mining*, 207-218.

Akritidis, L., D. Katsaros, and P. Bozanis. 2009. "Identifying influential bloggers: time does matter." *Arxiv Preprint* 0925.2416.

Blood, R. 2003. "Weblogs and journalism: Do they connect?" *Nieman Reports*, 57(3): 61-63.

Borgatti, S.P., M.G. Everett, and L. C. Freeman. 1999. *UCINET 6.0 Version 1.00.* Natick: Analytic Technologies.

Chin, A., and M. Chignell. 2008. "Automatic detection of cohesive subgroups within social hypertext: A heuristic apporoach." *New Review of Hypermedia and Multimedia*, 14(1): 121-143.

Clyde, L. A. 2004. "Library weblogs." *Library Management*, 25(5): 183-189.

Katz, E. and P. Lazarsfeld. 1955. "Personal Influence." *New York: The Free Press.*

Kritikopoulos, A., M. Sideri, and I. Varlamis. 2006. "BlogRank: Ranking weblogs based on connectivity and similarity features." *Proceedings of the 2nd International Workshop on Advanced Architectures and Algorithms for Internet Delivery and Applications.*

Langville, Amy N. and Carl D. Meyer. 2008. "Google's pagerank and beyond: The science of search engine rankings." *Princeton, NJ: Princeton University Press.*

Lin, F., and W.W. Cohen. 2008. "The multi rank bootstrap algorithm: Semi-Supervised political blog classification and ranking using semi-supervised link classification." *Proceedings of the International Conference on Web Search and Web Data Mining Poster*, 206-207.

Macdonald, C., I. Ounis, and I. Soboroff. 2008. "Overview of the TREC-2007 Blog Track." *Proceeding of TREC 2007.*

Mishne, G. 2007. "Using blog properties to improve retrieval." *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management*, 831-840.

Ounis, I., C. Macdonald, G. Mishne, and I. Soboroff. 2007. "Overview of the TREC-2006 Blog Track." *Proceeding of TREC 2006.*

Ounis, I., C. Macdonald, and I. Soboroff. 2008. "On the TREC Blog Track." *Proceeding of TREC 2008.*

Song, X., Y. Chi, K. Hino, and BL. Tseng. 2007. "Identifying opinion leaders in the blogosphere." *Proceedings of the 16th ACM conference on Conference on Information and Knowledge Management*, 971-974.

Song, X., Y. Chi, and B.L. Tseng. 2007. "Summarization system by identifying blogs." *International Conference on Weblogs and Social Media*, March 2007.

Todeva, E. and D. Keskinova. 2009. "Pharmaceutical blogging and on-line distribution of information." *Proceeding of the 42th Hawaii International Conference on System Science, Waikoloa, Hawaii, USA, 5-8 January*, 1-13.

Wikipedia. 〈http://en.wikipedia.org/〉.