재전송 정보를 활용한 트위터 랭킹의 정확도 평가

An Evaluation of Twitter Ranking Using the Retweet Information

장재영(Jae-Young Chang)*

초 록

최근 들어 트위터나 페이스북과 같은 SNS가 대중화되면서 이에 관련한 연구도 활발히 진행되고 있다. 하지만 SNS가 비교적 최근에 시작된 만큼 관련 연구도 아직 초보적인 수준이다. 특히 포털 사이트와 같은 검색 엔진에서는 트위터에 대한 검색 결과를 최근에 등록된 순으로 보여주는 수준에 머물러 있다. 트위터에서의 검색은 기존의 TF-IDF로 대표되는 웹 검색 방식과는 달라야한다. 본 논문에서는 트위터 환경에서 사용자가 원하는 게시글을 효율적으로 검색하는 방법을 제안한다. 제안된 방법에서는 사용자들의 재전송 빈도를 검색결과의 주요한 평가요소로 활용한다. 재전송 정보는 사용자가 직접 게시글의 가치를 판단하는 중요한 평가 척도가 될 수 있다. 또한 실험을 통하여 제안된 방법이 트위터 검색에 효율적으로 적용될 수 있음을 보여준다.

ABSTRACT

Recently, as Social Network Services(SNS), such as Twitter, Facebook, are becoming more popular, much research has been doing actively. However, since SNS has been launched recently, related researches are also infant level. Especially, search engines serviced in web potals simply show the postings in order of upload time. Searching the postings in Twitter should be different from web search, which is based on traditional TF-IDF. In this paper, we present the new method of searching and ranking the interesting postings in Twitter. In proposed method, we utilize the frequency of retweets as a major factor for estimating the quality of postings. It can be an important criteria since users tend to retweet the valuable postings. Experimental results show that proposed method can be applied successfully in Twitter search system.

키워드 : 트위터, 재전송, 랭킹, 검색
Twitter, Retweet, Ranking, Search

이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(과제번호 : NRF-2011-0022445).
* 한성대학교 컴퓨터공학과 교수
2012년 02월 27일 접수, 2012년 03월 23일 심사완료 후 2012년 04월 03일 게재확정.

1. 서 론

기존의 웹 포털 사이트가 제공하는 블로그(blog)나 검색 기능을 포함한 인터넷 환경을 웹 1.0으로 본다면 개방적인 웹 환경을 기반으로 네티즌들의 정보공유와 참여가 가능하게 된 요즘의 인터넷 환경을 웹 2.0이라고 부른다. 웹 2.0 시대가 도래하면서 사용자들의 편의성과 다양성 그리고 정보의 공유를 추구하는 다양한 서비스들이 등장하고 있다. 이러한 서비스의 대표적인 것이 트위터(Twitter)나 페이스북(Facebook)과 같은 소셜 네트워크 서비스(Social Network Service : SNS)이다. 특히 트위터는 비교적 최근인 2006년도에 서비스가 시작되었지만 사용자가 해마다 기하급수적으로 증가하여 2011년 초에 이미 2억 명을 돌파하였으며 국내에서도 이미 500만 명을 넘어서고 있다. 트위터는 사용자 수뿐만 아니라 정보의 전파력 측면에서 막강한 영향력을 행사하고 있다. 예를 들어 미국 허드슨강 항공기 불시착 사건, 오사마 빈 라덴 사망, 일본 대지진 등 전 세계에서 벌어진 주요 사건들에 대해 실시간으로 정보를 전파였으며, 국내에서도 강남역 침수, 강변 테크노마트 진동 사건 등에서 보는 바와 같이 트위터가 기존 언론에 비해 보다 빠르게 정보를 전달할 수 있다는 사실을 보여주었다. 트위터는 이와 같이 강력한 정보의 전파력을 가지고 있을 뿐만 아니라 특정 주제에 대한 여론이나 동향 등을 파악할 수 있는 주요한 수단이 되고 있다. 최근 정치권에서도 트위터의 영향력을 인정하고 이를 정치나 선거에 적극적으로 이용하려는 움직임도 일고 있다.

트위터는 일반 웹 문서와 비교되는 몇 가지 특징을 갖고 있다. 우선 모든 게시글(posting, tweet)은 140bytes의 단문만이 허용된다. 따라서 트위터 사용자는 모든 게시글을 간단명료하게 표현해야만 한다. 또한 모든 사용자는 시간과 장소에 관계없이 항상 게시글을 등록할 수 있다. 따라서 누구나 트위터에 의견을 피력할 수 있으며 또한 새로운 정보에 대한 제공자가 될 수 있다. 마지막으로 트위터에는 자신이 팔로우(follow)하는 사용자가 등록한 게시글을 자신을 팔로우하는 사용자, 즉 팔로워(follower)들에게 전달하는 재전송(retweet : RT) 기능이 있으며, 각 게시글에 대해 자신의 의견을 응답형식으로 전달하는 멘션(mention) 기능도 있다. 특히 재전송은 정보를 실시간으로 빠르게 전파하는 역할을 하고 있어 트위터에서의 핵심 기능이라고 할 수 있다.

이러한 환경에서 사용자들은 자신이 팔로우하는 사용자들의 게시글을 실시간으로 구독할 수 있다. 따라서 트위터에서 새로운 정보에 대한 출처는 단순히 자신이 구독하는 사용자들에 의존적일 수밖에 없다. 그러나 현재 전 세계적으로 매일 2억 개가 넘은 게시글이 등록되고 있는 상황에서 단순히 관심 있는 사용자들을 팔로우하는 것만으로는 한계가 있다[15]. 자신이 팔로우하지 않더라도 실시간으로 전달되는 주요 이벤트에 대한 게시글을 검색할 수 있는 기능이 필수적이다. 물론 대부분의 트위터 홈페이지에서도 검색기능을 제공하고 있다. 그러나 이들은 단순히 검색어가 포함된 게시글을 등록 시간 순으로 보여주는 형태에 불과하며, 웹 검색에서와 같이 가장 관련성 있는 게시글을 검색하는 기능은 제공하고 있지 않다. 최근 들어 Lauw et al.[10],

Nagmoti and Cock[11], Sarma et al.[12] 등에서 트위터에서의 검색 방법들에 대해서 연구결과가 있었으나 아직까지 기초적인 수준을 벗어나지 못하고 있다.

트위터에서의 검색은 기존의 TF-IDF로 대표되는 웹 검색 방식과는 매우 다르다[1, 2, 3, 7]. 우선 게시글은 140bytes 이하의 단문이므로 내용(contents)만으로 게시글의 중요도를 평가하기에는 제한적이다. 또한 게시글은 중요한 정보에 대한 전달역할도 하지만, 게시자의 감정 상태나 의견과 같이 검색 의도와 부합하지 않은 게시글들도 많아 이들을 검색 대상에서 배제해야한다. 따라서 트위터에서의 검색을 위한 랭킹 방법은 게시글의 내용보다는 관련된 메타(meta) 정보를 이용할 수밖에 없다. 게시글에 대한 메타정보로는 게시자의 팔로워 수, 재전송 빈도, 링크(link) 정보의 포함유무, 멘션 빈도, 해시태그(hashtag) 등이 있다. 이러한 정보들은 각 게시글의 중요도를 평가하는 요소로 활용될 수 있다.

본 논문에서는 게시글의 메타 정보 중에서 재전송 정보를 이용하여 게시글의 중요도를 평가하고 이를 트위터 랭킹에 활용하는 방법을 제안한다. 재전송은 SNS 중에서 트위터만이 갖고 있는 매우 중요한 수단이다. 재전송은 특정 게시글에 대해 사용자가 이 게시글을 다른 사용자에게 전파할 가치가 있다고 판단되면 팔로워에게 이를 전달할 수 있다. 따라서 게시글의 정보력에 대한 가치를 사용자가 직접 판단할 수 있는 가장 좋은 지표가 된다. 제안된 방법은 단순한 재전송의 빈도뿐만 아니라 재전송으로 연쇄적으로 전달되는 상황까지 고려한다. 이와 더불어 재전송한 사용자의 개인정보도 지표에 활용한다. 물론 기존 연구에서도 재전송을 트위터 검색의 지표로 활용한 사례가 있으나[10, 11, 12], 다양한 지표 중에서의 하나로만 취급하였으며 연쇄적으로 전달되는 상황까지 고려하고 있지 않다. 본 논문이 제안한 랭킹 방법에 대한 정확도를 평가하기 위하여 실험을 실시하였다. 실험은 본 논문이 제안한 게시글의 중요도 평가 척도에 따라 여러 단계에 걸쳐 진행하였으며, 최종적으로 본 논문에 제안한 기법이 트위터 검색에 활용될 수 있음을 보여주었다.

본 논문의 구성은 다음과 같다. 제 2장에서는 관련연구에 대해서 논하고 제 3장에서는 트위터에서의 재전송에 대한 중요성에 대해 기술한다. 제 4장에서는 재전송을 이용한 랭킹 방법을 제안하고 제 5장에서는 실험결과를 제시한다. 마지막으로 제 6장에서는 결론을 맺는다.

2. 관련 연구

트위터에서의 게시글은 내용 및 형식면에서 기존 웹문서와 많은 차이점이 존재하므로 기존의 검색 기법을 트위터에 직접적으로 적용하기에는 한계가 있다. 전통적인 검색 기법인 TF-IDF 방식은 문서 내에 단어의 빈도가 가장 중요한 평가 지표가 된다. 하지만 단문으로 구성된 트위터 게시글에서는 문서 내에 주어진 검색어가 두 번 이상 출현하는 경우가 매우 드물다. 기존의 트위터 검색 기법에 관한 연구에서도 게시글의 내용보다는 그 이외의 지표를 활용하여 검색에 활용하고 있다.

우선 Teevan et al.[13]에서는 사용자들의 검색어를 분석하여 트위터에서의 검색 성향

을 웹 검색과 비교 분석하였다. 이 연구에서 웹 검색은 주로 특정 주제에 대한 지식을 얻는 목적으로 사용되나, 트위터에서의 검색은 현재 발생된 특정 이벤트들에 대해 진행 상황이나 시간별 변화를 모니터링하는데 주로 이용된다고 분석하였다. 또한 트위터에서의 검색어는 웹 검색어에 비해 짧고, 반복적이며 인물과 연관된 경우가 많다는 사실을 밝혀냈다.

Sarma et al.[12]에서는 트위터 또는 이와 유사한 환경에서 게시글에 대한 랭킹 기법을 제안하였다. 각 게시글에 대해서 이를 읽은 사용자에게 피드백을 받아 각 게시글에 대한 평가를 상대적으로 분석하여 최종적으로 게시글들에 대한 랭킹을 부여하는 기법을 제안하였다. 그러나 이 방법은 모든 게시글에 대해 사용자들의 피드백을 받아야한다는 점에서 트위터 검색에 직접적으로 적용하기에는 한계가 있다.

Lauw et al.[10]에서는 본 논문에서 제안한 기법과 가장 유사하다. 이 연구에서는 재전송 횟수, 사용자의 팔로워 수, 시간 간격 등을 변수로 게시글의 가치를 평가하였다. 그러나 이 기법에서는 각 게시글에 대해서 재전송된 횟수를 실제 계산하지 않고 팔로워 수를 이용하여 단순히 추정한 수치를 활용하였으며 실험 결과도 제시하지 않고 있다.

Nagmoti and Cock[11]는 게시글에 대한 랭킹을 위해 사용자의 게시글 수, 팔로워 수, URL의 포함 여부, 게시글의 길이를 평가 요소로 활용하였다. 특히 이 연구에서는 제안된 랭킹 기법을 성능을 평가하기 위해서 사용자들의 피드백을 얻기 위한 사이트를 개설하여 그 결과를 활용한 방법을 이용하였다.

이외에도 Char[4], Kwak et al.[9], TunkRank [14], Weng and He[16]들의 연구가 있으나, 이들은 트위터의 검색 기법을 제안하기 보다는 트위터 상에서 영향력 있는 사용자를 탐색하는 기법들에 대한 내용들이다. 특히 TunkRank [14]에서는 TrunkRank라는 기법을 제안하였는데 이는 현재 웹 검색에서 광범위하게 사용되고 있는 PageRank 기법을 트위터의 인적 네트워크에 적용하여 사용자들의 영향력을 평가하였다.

지금까지의 살펴본 본 바와 같이 트위터에서의 검색에 대한 연구는 아직 많지 않다. 트위터 서비스가 최근에야 활성화된 만큼 아직 많은 관심을 받지 못하고 있는 것이 가장 큰 이유라고 할 수 있다. 하지만 트위터의 영향력이 날로 증대되고 있는 상황에서 트위터에서의 검색 기법에 대한 연구는 보다 활발히 이루어질 전망이다.

3. 트위터에서의 재전송 분석

트위터 환경에서 재전송은 정보의 전달력을 극대화하는 가장 강력한 수단이 되고 있다. 본 논문에서는 사용자들의 재전송 성향을 분석하기 위해서 간단한 실험을 실시하였다. 실험을 위해 최근에 발생한 두 사건에 관련된 게시글을 수집하였다. 첫째는 2011년 12월 17일에 발생한 김정일 사망과 관련된 게시글로, 2011년 12월 17일 14시부터 36시간동안 게시된 약 100,000여 개의 게시글을 수집하였으며, 둘째로 2012년 1월 17일에 발생한 케이블 TV에서의 공중파 재전송 중단 후 재개 사건에 맞추어 2012년 1월 17일 09시부터 36시간동안 이와 관련된 3,500여 개의 게시글을 수집하였다. 첫 번째 집합은 최근에 발생한 가

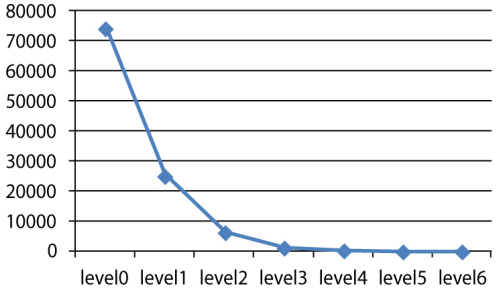
장 큰 사건으로 트위터에서 관련된 게시글들이 가장 활발하게 등록되는 상황이며, 두 번째 집합은 첫 번째에 미치지 못하지만 뉴스에서 크게 이슈화된 사건에 대한 게시글이라고 볼 수 있다. 게시글의 수집은 트위터 API를 이용하였다. 트위터 API는 기능이 제한적이라 수집 기간 동안에 다수의 관련 게시글들이 누락될 가능성도 존재하지만 수집된 자료들만으로도 충분한 분석이 가능하였다.

우선 <표 1>은 각 집합에 대해서 총 게시글 수와 게시자의 수를 보여준다. 이 표에서 볼 수 있듯이 '김정일'에 관련된 게시글에서는 1인당 게시한 게시글의 수가 약 2.8개로 '케이블'에 관련된 평균 게시글 수인 1.4개보다 많은 것을 알 수 있다. 따라서 중요한 이슈일수록 사용자들이 트위터를 자주 이용한다는 사실을 알 수 있다.

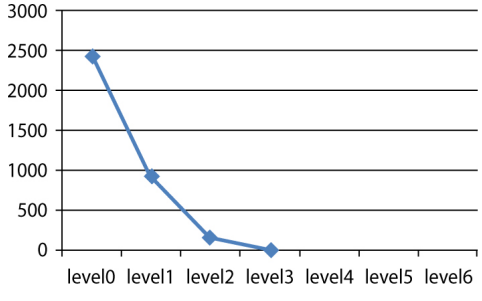
〈표 1〉 수집된 게시글 수와 사용자 수

검색어	'김정일'	'케이블'
게시글 수	108,332	3,575
게시자 수	38,523	2,608
1인당 평균 게시글 수	2.8	1.4

다음으로 재전송의 빈도를 평가하였다. <그림 1>의 (a)와 (b)는 각각 '김정일'과 '케이블' 집합에서 재전송 빈도를 보여준다. 그림에서 $level_0$는 원본 게시글을 의미하며, $level_i$는 i번째 재전송된 게시글을 나타낸다. 즉, $level_1$은 $level_0$로부터 재전송된 수를 나타내며, $level_2$는 $level_1$에서 다시 연쇄적으로 재전송된 수를 나타낸다. 이 그림에서 보듯이 '김정일'의 경우 최대 6회($level_6$)까지 재전송되었으며, '케이블'의 경우 최대 3회까지 연쇄적으로 재전송 되었다. 따라서 사용자들의 관심정도에 따라 재전송 단계가 높다는 것을 확인할 수 있다.



(a) '김정일'관련 게시글

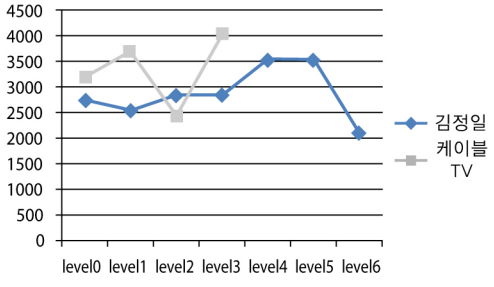


(b) '케이블'관련 게시글

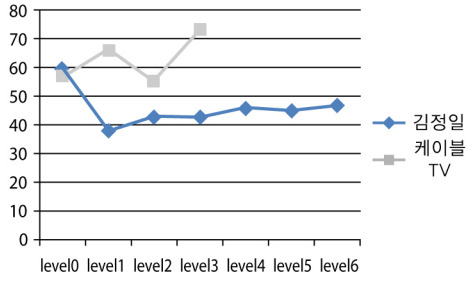
〈그림 1〉 게시글들의 재전송 빈도

<그림 2>의 (a)와 (b)는 각각 재전송한 사용자들에 대한 팔로워 수 및 해당기간 게시한 글의 수를 보여준다. 여기서 게시글 수란 관련 게시글 뿐만 아니라 해당 기간 동안 사용자가 등록한 모든 종류의 게시글을 의미한다. 이 그림에서 보면 팔로워 수(<그림 2>의 (a))는 두 게시글 집합 간에 유의할 만한 큰 차이를 보여주지 않고 있으며, 재전송 단계의 변화에서도 크게 달라지지 않는다는 것을 알 수 있다. 다만,

<그림 2>(b)를 보면 '케이블'에 관한 글을 게시한 사용자들의 평균 게시글 수는 '김정일'에 관한 글을 게시한 사용자들의 평균 게시글 수보다 다소 많다는 것을 알 수 있다. 그 이유를 분석하면 '김정일'에 관한 사건은 최근에 발생한 비교적 중요한 이슈이므로 글을 자주 게시하지 않는 사용자들도 많이 참여한 반면, '케이블'는 상대적으로 중요도가 떨어지므로 주로 트위터 활동이 활발한 사용자들에 의해 게시되거나 재전송된 결과라고 볼 수 있다.



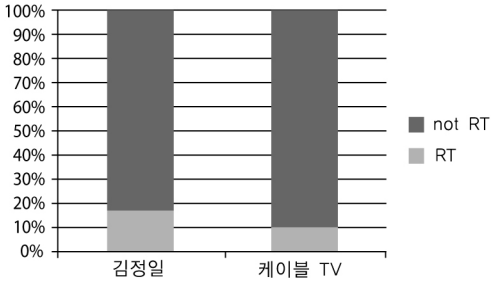
(a) 평균 팔로워 수



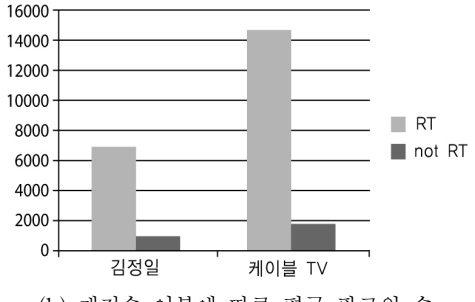
(b) 평균 게시글 수

〈그림 2〉 평균 팔로워 수와 평균 게시글 수

<그림 3>(a)는 원본 게시글 중에서 최소 1회 이상 재전송된 게시글과 그렇지 않은 게시글의 비율을 나타낸다. 이 그림에서 RT는 재전송된 게시글을 의미하며 notRT는 재전송되지 않은 게시글을 나타낸다. 이 그림에서 확인할 수 있듯이 전체 원본 게시글 중에서 재전송되는 비율은 10%~20% 정도이며 이 수치는 이슈의 중요도 여부에 따라 달라질 수 있다. <그림 3>(b)는 이러한 원본 게시글의 재전송 여부에 따라 각각의 게시글 집합에 대한 게시자의 평균 팔로워 수를 계산한 결과이다. 이 결과에서 보듯이 재전송된 글을 게시한 사용자의 팔로워 수가 재전송되지 않는 글을 게시한 사용자의 팔로워 수에 비해 크다는 것을 알 수 있다. 이는 상식적으로 예상할 수 있는 결과로서 팔로워 수가 많은 사용자일수록 이들이 게시한 글이 재전송될 확률이 높아지기 때문이다. 역으로 설명하면 팔로워 수가 적은 사용자일수록 게시한 글들은 재전송될 가능성이 적다고 할 수 있다.



(a) 재전송 비율



(b) 재전송 여부에 따른 평균 팔로워 수

〈그림 3〉 재전송 비율과 재전송 여부에 따른 평균 팔로워 수

4. 재전송을 이용한 검색 방법

제 3장에서의 실험 결과와 더불어 트위터 사용자들의 일반적인 사용 패턴을 고려하면 다음과 같은 가정을 할 수 있다.

- 트위터에서 재전송 기능은 정보에 대한 확산과 공유에 가장 중요한 역할을 한다.
- 게시자의 팔로워 수가 많을수록 게시글이 재전송될 가능성이 많으며 그만큼 게시글의 가치를 높게 평가할 수 있다. 특히 팔로워 수가 많은 사람은 재전송를 이용하여 정보의 확산에 큰 역할을 한다. 그 이유는 팔로워 수가 많은 영향력 있는 사용자는 자신의 재전송으로 인해 미치는 파장을 고려해야하므로 신중히 판단하는 경향이 있기 때문이다.
- 사용자들의 관심이 높은 게시글일수록 연쇄적인 재전송이 발생할 가능성이 많다.

이와 같은 가정 하에 본 절에서는 트위터의 게시글들에 대한 검색 방법을 제안한다. 트위터에서의 검색은 기본적으로 검색어가 포함된 게시글만을 대상으로 한다. 이 때 주어진 검색어가 포함된 게시글에 대한 랭킹이 필요한데, 검색어와 관련이 있으면서 정보로서의 가치가 있는 게시글을 상위에 랭크해야 의미가 있다. 앞서 여러 문헌에서도 밝혀졌듯이 게시글 중에는 사용자가 검색하고자하는 중요한 내용이 포함될 수도 있는 반면, 많은 게시글은 개인적인 감정이나 소소한 일상에 관한 내용 등을 포함하는 등 정보로서의 가치가 떨어지는 것도 있을 수 있다[10]. 이를 구분하기 위해 본 논문에서는 게시글을 접한 사용자들이 해당 게시글을 어떻게 취급하느냐에 초점을 맞춘 검색 기법을 제안한다. 앞서 언급한 바와 같이 게시글들의 가치는 재전송 여부에 따라 판단한다.

수집된 전체 게시글들에 대해서 검색어 q을 포함한 게시물들의 집합을 $\overline{D_q}$라고 하자. $\overline{D_q}$에는 원본 게시글뿐만 아니라 이 게시글로부터 한번 또는 그 이상 연쇄적으로 재전송된 게시글도 포함되어 있다. 따라서 $\overline{D_q^0}$을 원본 게시글의 집합이라 하고 $\overline{D_q^i}$를 i번째 재전송된 게시글의 집합이라 하면 $\overline{D_q} = \bigcup_{i=0} \overline{D_q^i}$가 성립한다. 이와 같은 가정 하에 재전송을 이용한 게시글의 가치를 평가하는 가장 단순한 방법은 각 원본 게시글에 대해서 단계에 관계없이 재전송된 횟수로 평가하는 것이다. 즉, t시간에 동안에 수집된 게시글 중에서 검색어 q가 포함된 각 원본 게시글 d_q^0에 대한 가치 $impact(d_q^0,\ t)$는 다음의 수식으로 평가할 수 있다.

$$impact(d_q^0,\ t) = \sum_{i=0} |\overline{d_q^i}| \qquad (1)$$

이 식에서 $\overline{d_q^i}$는 d_q^0에서 i번째 재전송된 게시글의 집합을 의미하며, $|\overline{d_q^i}|$는 $\overline{d_q^i}$의 크기를 나타낸다. 이와 같은 게시글의 평가 방법은 이미 트위터 관련된 여러 사이트에서도 제공되는 서비스이다. 예를 들어 joinsmsn과 코리안트위터에서는 일정시간 동안 가장 많이 재전송된 게시글을 우선적으로 보여준다[6, 8].

다음 방법은 식 (1)에 연쇄적인 재전송 단계를 고려하는 것이다. 재전송된 게시글을 다시 연쇄적으로 재전송한다는 것은 정보 제공자가 누구인가보다 해당 게시물을 전파할 가

치가 있는 내용인가를 우선적으로 판단할 가능성이 높기 때문이다. 이 방법은 식 (1)을 수정하여 재전송 단계별로 가중치를 부여하는 것으로 해결할 수 있다. 따라서 이를 반영한 $impact(d_q^0, t)$는 다음과 같이 정의할 수 있다.

$$impact(d_q^0, t) = \sum_{i=0} \alpha_i \times |\overline{d_q^i}| \qquad (2)$$

이 식에서 α_i는 i번째 재전송에 대한 가중치를 나타낸다. 따라서 α_i에 큰 값을 부여할수록 연쇄적인 재전송이 발생하는 게시글을 우선적으로 검색하겠다는 것을 의미한다.

식 (2)는 일단 재전송만으로 게시글의 가치를 평가하였다. 그러나 제 3장의 실험에서도 확인할 수 있듯이 정보의 전파력을 생각했을 때 게시글을 재전송하는 사용자의 팔로워 수를 무시할 수는 없다. 팔로워 수는 사용자의 영향력을 간접적으로 나타내기 때문이다. 영향력 있는 사용자는 자신이 게시하는 글에 대한 책임감으로 인해 재전송도 신중히 하는 경향도 있으며, 이 때 게시글 전파에 대한 일종의 '허브(hub)' 역할을 한다. 이와 같은 요소들을 고려한 최종적인 평가 지표는 다음과 같다.

$$impact(d_q^0, t) = \sum_{i=0} \sum_{d \in \overline{d_q^i}} \sum_{u \in user(d)} \alpha_i \times \log(|(follower(u)|) \qquad (3)$$

이 식에서 $user(d)$는 게시글 d를 등록한 사용자를 의미하며 $follower(u)$는 사용자 u의 팔로워 집합을 의미한다. 식 (3)은 식 (2)와 비교하여 재전송을 한 사용자의 영향력 요소를 추가한 것이다. 즉, 재전송을 한 사용자들의 개인 정보에 큰 영향을 받는다. 팔로워 수는 사용자들마다 편차가 매우 크게 나타난다. 따라서 이 값이 전체적인 평가 값에 지나치게 영향을 미치지 않게 하기 위해서 로그를 이용하였다.

5. 실 험

본 논문에서 제안한 트위터 검색 기법의 성능을 검증하기 위하여 실험을 실시하였다. 실험은 '김정일'과 '케이블' 관련된 각 게시글 100,000여 개와 3,500여 개를 대상으로 실시하였고, 각 데이터 집합에 대해서 식 (1)~식 (3)을 이용하여 각 게시글에 가치를 평가한 후 각각의 방법에 대한 랭킹 정확도를 평가하였다. 랭킹 정확도를 평가하기 위해서는 올바르게 랭킹된 참조 대상이 있어야한다. 그러나 일반 문서 검색 시스템에서와 마찬가지로 트위터 환경에서도 검색의 정확도 평가를 위한 참조 대상이 존재하지 않는다. 따라서 각각의 데이터 집합에 대해서 1단계 재전송이 가장 많이 된 상위 100개의 게시글에 대해서 일반인 30여 명을 대상으로 설문을 통해 게시글의 가치를 평가하도록 하였다. 평가는 검색어와의 적합도와 정보로서의 가치를 5단계로 나누어 선택하도록 하였으며, 이 값들의 평균으로 최종적인 참조 대상을 결정하였다. 예를 들어 '케이블'로 검색된 게시글에서 상위 10개의 참조 대상은 <표 2>와 같다.[1)]

1) 원본 게시글 내용 중에 실명이 거론되거나 정치, 사회적으로 논란이 될 만한 내용은 일부 삭제하였다.

〈표 2〉 '케이블'에 대한 상위 10개 참조 게시글

순위	원본 게시글
1	케이블, KBS 2TV 재송신 저녁 7시 정상화 … 이에 따라 종합유선방송사들은 오늘 저녁 7시를 기해 KBS 2TV의 재송신을 정상화했으며 MBC, SBS 등에 대한 순차적인 재송신 중단 방침도 철회했다고 밝혔습니다 ….
2	(속보) KBS2 TV … 오늘 오후 7시부터 볼 수 있습니다. 지상파-케이블 협상 극적으로 타결됐습니다.
3	케이블-지상파 협상 타결 … KBS2 송출 재개(1보) : 케이블TV 종합유선방송사업자(SO)들로 구성된 케이블TV 비상대책위원회는 지상파 방송사와의 협상 타결로 중단했던 KBS ….
4	KBS2 방송 중단사태는 … 시청률 올리기를 위한 … 합작품이라는 주장이 있군요. 케이블쪽에서는 방통위에 수차례 중재를 요청했으나 외면당했고 합니다.
5	케이블TV 사업자, KBS 2TV 방송 이틀째 송출 중단 … 케이블 방송 사업자 KBS 2TV 송출중단 이틀째 이어져, 방송통신위원회 어제 저녁 8시까지 송출 재개하라고 시정명령했지만 케이블 방송 측 거부 ….
6	[속보]케이블・지상파 협상 타결 … KBS2 송출 재개
7	지상파-케이블TV 재송신료 협상 극적 타결 http://t.co/y7OY0DjZ #이데일리
8	[긴급]지상파-케이블 '극적타결' … 방송재개 http://t.co/GoG17Mm4 #zdk
9	[매경속보]지상파-케이블TV 재송신료 협상 극적 타결
10	케이블방송사들이 KBS2TV 재송출을 중단한 비상사태에서 … 덕분에 전체 회의가 2시간 반이나 지나서 열렸고요. 회의 결과는 방송 재개 안 하면 ….

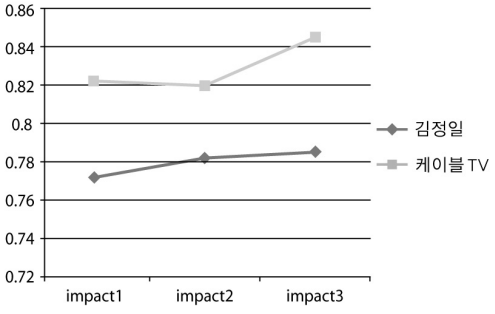
본 실험에서 검색 결과를 비교하기 위한 측정치로서 nDCG(Normalized Discounted Cumulative Gain)[5]를 활용하였다. nDCG는 검색엔진의 효율성을 측정하는 도구로 가장 많이 쓰이는 방법 중 하나이다. DCG는 검색결과로 나온 각 문서에 대해서 검색어와의 실제 관련도에 따라 점수를 부여하는 방식을 사용한다. nDCG는 DCG를 정규화한 것인데, DCG는 관련도가 높은 문서가 검색 결과에 우선적으로 랭킹된다는 사실에 입각해서 그렇지 않을 경우 감점을 부과하는 방식을 사용한다. DCG는 검색결과가 p개일 경우 다음의 수식으로 계산한다.

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log 2(1+i)} \tag{4}$$

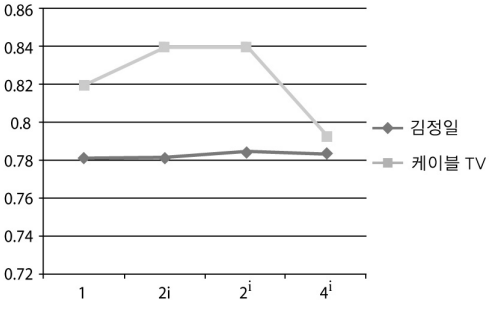
rel_i는 검색 결과에서 i번째 위치한 단어의 관련도를 나타낸다. 일반적으로 DCG는 위의 수식 그대로 사용하지 않고 0부터 1까지의 정규화된 수치인 nDCG로 표현하는데 수식은 다음과 같다.

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{5}$$

$IDCG_p$는 관련도 값에 따라 정확하게 정렬되었다는 가정하의 DCG 값을 의미한다. 따라서 nDCG는 1에 가까울수록 좋은 검색 결과를 나타낸다. 검색어와 게시글 간의 관련도를 평가하기 위한 rel_i는 참조 대상에 존재하는 순위에 따라 0부터 3까지 정수로 고정하였다.



〈그림 4〉 nDCG를 활용한 검색 정확도



〈그림 5〉 α_i의 변화에 따른 검색 정확도 비교

실험 결과는 <그림 4>와 같다. 이 그림에서 $impact_1$, $impact_2$, $impact_3$는 각각 식 (1)~식 (3)으로 계산한 랭킹방식에 의한 결과이다. Y축은 검색 정확도를 nDCG 값으로 나타낸 것으로 앞서 설명한 대로 1에 가까울수록 높은 정확도를 나타낸다. 식 (2)와 식 (3)에서 각 단계가 진행함에 따라 가중치가 두 배씩 증가하도록 α_i를 2^i로 정하였다.

이 그림에서 보듯이 두 게시글 모두에서 식 (3)에 의한 랭킹 방식이 가장 좋은 정확도를 나타냈다. 이는 재전송 빈도와 게시자의 팔로워 수가 게시글을 평가할 때 중요한 역할을 한다는 것을 보여준다. 전반적으로 '김정일'의 검색 정확도에서 $impact_3$가 에 비해 1.7% 정도 향상되었고, '케이블'의 $impact_1$ 경우에는 2.2% 정도 향상되었다.

'케이블'에 대한 검색 결과가 '김정일'에 대한 검색 결과보다 높은 정확도를 보여주었다. '김정일'에 관한 글을 게시할 시점에는 이미 사망 사실이 대중들에게 많이 알려진 상황이다. 따라서 새로운 사실에 대한 정보보다는 게시자들의 의견이 대부분을 차지하기 때문에 새로운 정보에 대한 내용보다는 게시자들의 의견을 재전송하는 경우가 많아 검색 정확도가 다소 떨어진 것으로 해석할 수 있다.

<그림 5>는 식 (3)의 방식을 이용했을 때 α_i의 변화에 따른 랭킹 정확도의 변화를 보여준다. 재전송 단계에 따른 가중치 비중이 점차 높아지도록 α_i를 각각 1, $2i$, 2^i, 4^i로 증가시키며 변화를 관찰하였다. 이 그림에 보듯이 '케이블'의 경우에는 $2i$와 2^i에서 높은 정확도를 나타냈지만 4^i일 때는 정확도가 크게 떨어지는 것으로 나타났다. '김정일'의 경우에는 α_i가 2^i 또는 4^i일 때 높은 정확도를 나타냈지만 그 변화가 '케이블'의 경우보다 민감하지 않았다. 그 이유는 <그림 4>에서의 실험과 같은 이유로 해석할 수 있다. 즉, '김정일'에 관련된 게시글들은 대부분 새로운 사실에 관한 정보보다는 사용자들의 의견이 대부분을 차지하고 있고, 각 의견들에 동조하는 사용자들이 재전송하는 경우가 많았다. 이러한 게시글들이 상대적으로 정보로서의 가치가 떨어지는 것들이다. 따라서 연쇄적인 재전송이 정보로서의 가치를 평가하기에는 상대적으로 큰 역할을 하지 못하는 것으로 해석할 수 있다.

비록 본 실험에서는 두 가지 검색어에 대한 제한적인 실험으로 정확도를 평가하여 과거

연구와 직접적인 비교에는 한계가 있지만 절대적인 정확도만으로 평가한다면 트위터에서의 검색방식으로 충분한 대안이 될 것으로 판단된다. 다만 '김정일'과 '케이블'에 관련된 실험 결과의 차이에서 보듯이 트위터 검색은 속보로서 가치가 있는 새로운 사실에 대한 실시간 정보나 기존 언론에서 접할 수 없는 정보에 대한 내용이 많이 포함되어 있는 경우 높은 정확도를 기대할 수 있는 것으로 판단된다.

6. 결 론

최근 들어 SNS에 대한 관심이 높아지고 있는 상황에서 이를 이용한 정보 검색에 관한 연구가 활성화 되고 있다. 특히 트위터는 최근에 사용자가 폭발적으로 증가하는 가장 주목받고 있는 SNS 중의 하나로 평가받고 있다. 그러나 이러한 성장세에도 불구하고 트위터에서의 검색 방법에 대한 연구는 아직 초보적인 수준을 벗어나지 못하고 있다.

본 논문에서는 이러한 트위터 환경에서 효율적인 검색 방법을 제안하였다. 제안된 방법은 사용자들의 재전송을 주요한 평가요소로 활용했으며, 보조적으로 사용자의 팔로워 수를 사용하였다. 또한 최근에 발생한 중요한 사건에 대한 게시글에 대한 실험결과를 제시하여 제안된 방법이 높은 검색 정확도를 보임을 증명하였다. 비록 본 논문에서 수행한 실험이 광범위하지 않고 두 검색어에 제한적으로 실시한 것이지만 제안한 검색 기법이 트위터 검색에 활용될 가치가 있다는 것을 간접적으로 보여준 결과라고 판단할 수 있다.

아직 트위터 검색에서는 어떠한 요소들이 게시글의 가치를 판단하는 기준이 되는 지에 대한 명확한 연구가 없는 실정이다. 본 논문에서도 멘션이나 링크정보의 유무 등 기타 메타 정보까지 고려하고 있지 않지만 이들도 어느 정도는 게시글의 가치를 판단하는데 간접적인 기준이 될 수 있다. 향후에는 이러한 요소들을 결합하여 보다 세밀한 검색 방식에 다한 연구가 필요할 것으로 보인다. 또한 궁극적으로는 대용량 데이터를 활용하여 기계학습을 이용한 검색기법의 가능성도 검토할 가치가 있다고 하겠다.

참 고 문 헌

[1] 김학래, 김홍기, "시멘틱 웹/온톨로지 기술을 이용한 개인용 전자문서 검색 시스템", 한국전자거래학회지, 제12권, 제1호, pp. 135-149, 2007.

[2] 이경하, 이규철, 김경옥, "키워드 질의를 이용한 순위화된 웹 서비스 검색 기법", 한국전자거래학회지, 제13권, 제2호, pp. 213-223, 2008.

[3] Baeza-Yates, R., Ribeiro-Neto, B., Modern Information Retrieval : The Concepts and Technology behind Search (2nd Edition), ACM, 2011.

[4] Char, M., Haddadi, H., Benevenuto, F., and Gummadi, K., "Measuring User Influence in Twitter : The Million Follower Fallacy," Proc. of International AAAI conference on Weblogs and Social Media, 2010.

[5] http://en.wikipedia.org/wiki/Discounted_

cumulative_gain/.
[6] http://koreantweeters.com/.
[7] http://lucene.apache.org/nutch/.
[8] http://www.joinsmsn.com/.
[9] Kwak, H., Lee, C., Park, H., and Moon, S., "Fiding Influentials Based on Temporal Order of Information adoption in Twitter," Proc. of WWW conference, 2010.
[10] Lauw, H. W., Ntoulas, A., and Kenthapadi, K., "Estimating the Quality of Postings in the Real-time Web," Proc. of SSM conference, 2010.
[11] Nagmoti, R. and Cock, M. D., "Ranking Approach for Microblog Search," Proc. of WI-IAT conference, 2010.
[12] Sarma, A., Sarma, At., Gollapudi, S., and Panigrahy, R., "Ranking Mechanisms in Twitter0like Forums," Proc. of WSDM conference Feb., 2010.
[13] Teevan, J., Ramage, D., and Morris, M. R., "#TwitterSearch : A Comparison of Microblog Search and Web Search," Proc. of WSDM conference, 2011.
[14] TunkRank, http://tunkrank.com, 2009.
[15] TwitterEngineering, "200 million tweet per day," http://blog.twitter.com/2011/06/200-million-tweets-per-day.html.
[16] Weng, J. and He, Q., "TwitterRank : Finding Topic-sensitive Influential Twitterers," Proc. of WSDM conference, 2010.

저 자 소 개

장재영	(E-mail : jychang@hansung.ac.kr)
1992년	서울대학교 계산통계학과 (학사)
1994년	서울대학교 계산통계학과 전산과학전공 대학원 (석사)
1999년	서울대학교 계산통계학과 전산과학전공 대학원 (박사)
2000년~현재	한성대학교 컴퓨터공학과 교수
관심분야	데이터베이스, 데이터마이닝