사용자의 활동과 영향력을 이용한 트위터의 URL 추천 시스템

(URL Recommendation System in Twitter Using User Activity and User Influence)

이 성 윤 † 이 태 휘 † 김 형 주 ††

(Seong-Yun Lee) (Taewhi Lee) (Hyoung-Joo Kim)

요 약 트위터는 최근 가장 각광 받는 마이크로블로깅 서비스로 전세계인들에게 정보 교류의 장으로 활용되고 있다. 그러나 사용자 수가 급증하고 있어 사용자들이 생산하는 막대한 양의 정보로부터 개인이 유용한 내용을 찾아내기가 어렵다. 이러한 문제를 해결하고자 트위터에서의 URL 추천 시스템이 등장하였으나 이전 연구들은 트위터의 특성을 고려하지 않는 단점을 지닌다. 본 논문에서는 사용자의 활동과 영향력을 고려하여 기존 시스템을 개선한다. 추천 결과를 향상시키는 요인들을 탐구하고 사용자 활동과 영향력을 측정하는 방법을 고안한다. 이를 기존 시스템에 결합한 새로운 URL 추천 시스템을 제안하고, 실험을 통해 실제로 사용자의 만족도가 높아짐을 확인한다.

키워드 : 트위터, URL, 추천 시스템

Abstract Twitter is one of the most attractive microblogging services and has been used as information net-work by people worldwide. But since the number of users is increasing rapidly, it is hard for each individual user to find useful contents from the huge amount of information generated by the users. To solve this problem, URL link recommendation system for Twitter has been proposed, but no previous work has considered the characteristics of Twitter itself. In this paper, we improve the existing system by considering user activity and user influence. We explore factors that improve the quality of recommendations and design methods to measure user activity and influence. We propose a novel URL recommendation system that combines these factors with the existing system. The experiments show that our system increases user satisfaction.

Key words : Twitter, URL, Recommendation System

1. 서 론

참여, 공유, 개방을 표방하는 웹 2.0 기술의 등장과 함께 페이스북[1], 트위터[2]와 같은 많은 소셜 네트워크 서비스들이 호응을 얻고 있다. 사회적 관계 개념을 인터넷 공간으로 가져오면서 사용자들이 인터넷 상에서 다른 사람과의 네트워크를 형성하기 쉽게 해주기 때문이다.

그 중에서도 트위터는 간결한 인터페이스라는 장점을 가지고 급속히 성장하고 있다. 2006년 서비스를 시작한 이후 급속히 성장하여 2010년 12월 현재 1억명 이상의 사용자들이 등록되어 있고, 매일 30만명의 사용자들이 늘고 있다. 또한 오픈 API를 적극 지원하는 특성으로 인해 웹이나 모바일 환경에서의 다양한 매쉬업 어플리케이션들이 제공되고 있다.

트위터 상에서 사용자들이 작성하는 메시지를 트윗이라고 한다. 트윗은 사용자가 자신의 상태를 표현하거나 지인들과 대화하고, 다양한 정보를 공유하는 등의 목적으로 사용되고 있으며, 하루 50만개 이상의 트윗이 생성되고 있다. 특히 최근 스마트폰이 보급되면서 트위터 사용자는 급격히 증가하고 있으며, 시간과 공간의 제약이

· 본 연구는 BK-21 정보기술 사업단의 연구결과로 수행되었음

† 비 회 원 : 서울대학교 컴퓨터공학부
sylee@idb.snu.ac.kr
twlee@idb.snu.ac.kr
†† 종신회원 : 서울대학교 컴퓨터공학부 교수
hjk@snu.ac.kr

논문접수 : 2011년 1월 3일
심사완료 : 2011년 6월 30일

Copyright©2011 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

정보과학회논문지: 컴퓨팅의 실제 및 레터 제17권 제8호(2011.8)

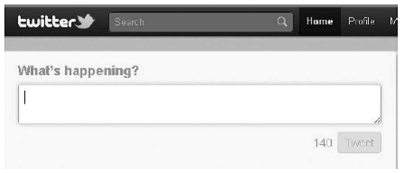

그림 1 현재 트위터 사용자 인터페이스

없는 스마트폰에서의 트위터 사용으로 인해 트윗의 시간과 공간에 관한 데이터가 더욱더 정확하고 풍부해졌다.

이렇게 활발히 생성되는 트윗들을 통해 트위터는 정보 네트워크로서의 역할을 하고 있다. 정보 트윗의 주제와 최신 뉴스 헤드라인이 85% 이상 일치하고[3], 트위터 스스로도 이러한 정보 네트워크로서의 성격을 강화하기 위해 상태 갱신 질문을 “What are you doing?”에서 그림 1과 같이 “What′s happening?”으로 변경하였다.

사용자들이 트위터 상에서 원하는 정보를 얻기 위해서는 사용자가 원하는 정보를 생성하는 사용자들을 팔로우(follow)하고, 제공되는 트윗 스트림 중에서 자신이 원하는 정보를 골라야 한다. 그러나 트윗의 양이 너무나 방대하기 때문에 사용자가 직접 유용한 트윗을 찾는 작업은 쉽지가 않고, 원하는 정보를 남기는 사용자들을 모두 팔로우하는 것도 사실상 불가능하다.

이러한 상황에서 사용자들의 요구에 맞춰, [4]에서는 트위터 스트림상의 컨텐츠를 추천하는 시스템을 연구하였다. 이 시스템은 사용자의 특징을 나타내는 단어와 트윗의 단어들 간의 유사도를 기반으로 한 URL 추천 시스템이다.

하지만 [4]에서는 추천 시스템에 도움이 될 만한 트위터 상의 메타데이터들을 적극적으로 활용하지 못하고 있다. 예를 들어, 비슷한 주제에 관한 트윗을 평범한 학생과 빌 게이츠와 같은 유명인이 남겼다고 가정할 때, 대부분의 사용자들은 빌 게이츠가 남긴 트윗에 더 많은 관심을 보일 것이다. 또한 어떤 트윗이 많이 리트윗 되었다면 그것은 그 트윗이 많은 사람들의 관심을 받고 있다는 근거가 될 수 있다.

이러한 트위터 상의 메타데이터를 적극적으로 활용한다면 트위터 관련 연구들이 좀더 좋은 결과를 얻을 수 있을 것이다. 특히 [4]의 URL 추천 시스템에 사용자 활동에 대한 정보를 더한다면 사용자와 관련 있는 주제를 찾는데 있어 더 좋은 결과를 기대해 볼 수 있다. 또한 사용자 영향력을 측정하여 트윗을 남긴 사용자들에 대해 순위를 매긴다면 좀더 많은 사람들이 원하는 URL을 추천할 수 있을 것이다.

이에 따라, 본 논문에서는 기존의 URL 추천 시스템을 향상시킬 수 있는 트위터 상의 정보들에 대해 탐구하고, 이러한 요소들을 이용한 추천 알고리즘을 통해 기존의 방법을 개선시키는 방안을 연구한다. 본 논문의 구성은 다음과 같다. 2장에서는 트위터 상의 정보를 다루는 연구들에 대해서 살펴보고 3장에서는 본 논문의 기반이 되는 [4]에서의 URL 추천 시스템에 대해서 자세히 설명한다. 4장에서는 추천 시스템의 효과를 향상시킬 수 있는 요소들에 대해 살펴보고, 5장에서는 기존의 시스템에 4장에서 고려한 요소들을 결합한 트위터에서의 URL 추천 시스템을 제안한다. 6장에서는 이 시스템에 대한 실험과 평가에 대해 설명한다. 마지막으로 7장에서는 결론과 향후 연구에 대해서 언급한다.

2. 배경 지식 및 관련 연구

2.1 트위터 배경 지식

트위터에서 다른 사용자의 트윗을 구독하는 것을 팔로우라고 한다. 그림 2는 A가 B를 팔로우한 상황을 표현한 그림이고, A는 B의 트윗을 그림 3처럼 자신의 타임라인(Timeline)에서 실시간으로 확인할 수 있다. 이러한 관계에서 A는 B의 팔로워(follower), B는 A의 팔로위(followee)라고 한다.

리트윗(retweet)은 팔로위가 작성한 트윗을 사용자가 자신의 팔로워들에게 전달하는 기능이다. 리플라이(reply)는 특정 트윗에 대해 자신의 응답을 트윗의 작성자에게 보내는 기능이며, 멘션(mention)은 특정 사용자에게 트윗을 보내는 기능이다. 그림 4는 URL을 포함하는 트윗의 예이다. 사용자들은 URL 정보가 담긴 트윗을 남길 때 다음과 같이 URL 주소와 그에 대한 설명을 함께 남긴다.

그림 2 트위터에서의 팔로우 관계

그림 3 트위터 타임라인

그림 4 URL을 포함한 트윗

2.2 관련 연구

트위터 사용자가 급속히 증가함에 따라 트윗 데이터의 양이 방대해져 이를 분석하는 연구들이 활발히 이루어지고 있다. [3]에서는 트윗들의 주제로 최신 뉴스 또는 꾸준히 회자되는 뉴스가 높은 비중을 차지함을 보이며 트위터가 뉴스 미디어로서의 역할을 한다고 분석했다. 이러한 성질을 활용하여 [5]에서는 트위터를 이용한 뉴스 알람 시스템 “TwitterStand”를 제안하였다. “Twitter-Stand”는 트위터 상의 트윗들과 위치 정보를 결합하여 어떤 지역에서 어떤 뉴스가 발생하였는지 사용자에게 알려준다.

트위터의 실시간성을 활용한 연구들도 진행되었다. [6]은 트위터 스트림을 분석하여 지진 시간과 위치를 예측하는 시스템을 개발하였다. 지진에 관련된 단어를 포함한 트윗들이 많이 올라오게 되면 트윗들의 위치를 분석하고 진행 방향을 예측하여 예상 지점에 있는 사용자들에게 경보를 보내준다. [7]에서는 실시간으로 등록되는 트윗들을 분석하여 현재 어떤 주제가 이슈되고 있으며 그 주제와 관련된 단어들을 그래프 형태로 표현하는 시스템에 대해 연구하였다.

트위터를 이용하여 검색 엔진의 성능을 향상시키는 연구들도 이루어졌다. [8]에서는 트위터의 트윗 정보를 활용하여 뉴스 검색 결과를 향상시킬 수 있음을 보였다. 위치 정보를 활용하여 사용자가 일반적인 정보를 원하는지 아니면 특정한 사건에 대한 뉴스를 원하는지 판단하고, 사용자가 질의한 질의어와 최신 트윗들을 비교하여 최근 가장 이슈된 사건에 대한 단어들을 추출하여 사용자가 질의한 의도를 판단하여 기존의 검색 알고리즘을 향상시켰다. [9]에서는 트윗에 포함된 URL들을 이용하여 실시간 웹 검색 엔진을 강화하는 방안에 대해서 연구하였다. 실시간 웹 검색은 기존의 웹 검색과 달리 실시간으로 자료를 수집해야 하고, 링크과 클릭 정보가 부족한 상태에서 문서를 순위화 한다는 점에 어려움이 있다. [9]에서는 마이크로블로깅 스트림 데이터를 이용하여 실시간으로 새로운 URL을 수집하고, 트위터 상의 요소들을 통해 검출한 URL들을 순위화 한다.

[4]의 연구는 사용자가 남긴 트윗들을 활용해 사용자의 관심에 부합할 만한 URL들을 추천해 주는 시스템을 제안하였다. 본 논문에서는 이 시스템을 향상시키기 위한 요소들을 탐구하고 이를 결합하여 새로운 URL 추천 시스템을 제안한다. [4]는 본 논문의 기반이 되는 중요한 연구이기 때문에 2.3절에서 자세히 설명한다.

[10]은 트위터에서의 사용자 영향력에 대해 연구하였다. 팔로워의 수와 남긴 트윗이 리트윗 되는 횟수, 멘션을 받는 횟수가 관련이 없다는 것을 보이며, 인기와 영향력이 비례 관계에 있지 않음을 밝혔다. 본 논문에서도 사용자 영향력을 평가하는 방법으로 이러한 요소들을 사용해 보았으며, [10]의 결론과 일치하는 실험 결과를 얻었다.

2.3 Short and Tweet

Short and Tweet[4]에서는 트위터 상의 트윗들에 포함된 URL을 사용자에 따라 알맞게 추천해 주는 시스템을 제안하였다. 시스템은 크게 세 단계로 나누어진다. 먼저 트위터 상의 URL을 담고 있는 트윗들 중에 추천할 만한 후보 URL들을 선정하여 수집한다. 그 다음 사용자가 남겼던 글 중 불용어들을 제외한 단어들을 가지고 이들 빈도의 TF-IDF 값으로 이루어진 단어 벡터를 만든다. 수집한 URL에 대해서도 URL이 담겨있던 트윗의 글을 이용하여 같은 작업을 해준다. 마지막으로 두 벡터를 비교하여 유사도를 구하고, 트위터 상의 네트워크를 이용해 계산한 가중치를 더해 순위화하여 높은 값을 가진 URL을 추천한다.

저자는 논문에서 (1)후보 URL들을 어떻게 수집할 것인지, (2)사용자 단어 벡터를 어떤 트윗의 글을 가지고 만들 것인지, 그리고 (3)소셜 점수를 어떻게 이용할 것인지를 바꿔가며 실험하였다. 후보 URL은 인기있는 URL들을 수집하는 것과 사용자의 팔로위의 팔로위들이 남긴 URL들을 수집하는 것 두 가지 경우에 대해 실험하였다. 사용자 단어 벡터는 해당 사용자의 트윗들의 글만 가지고 벡터를 계산하는 경우와 사용자의 팔로위들의 트윗까지 포함하여 벡터를 계산하는 경우 그리고 사용자 단어 벡터를 계산하지 않는 경우 총 세가지 경우에 대해 실험하였다. 소셜 점수는 대상 사용자의 팔로위의 팔로위들 중 해당 URL을 남긴 사용자들의 소셜 점수를 합하여 계산되는데, 각 사용자의 소셜 점수는 대상 사용자의 팔로위가 몇명이나 팔로우 하고 있는지와 트윗을 얼마나 자주 남기는지를 측정하여 계산한다.

모든 경우를 조합하면 총 2(후보 URL 선택) × 3(사용자 단어 벡터 계산) × 2(소셜 점수 사용 여부) = 12가지 경우가 나오게 된다. 12가지 알고리즘을 이용하여 트위터에서의 URL 추천 시스템 “Zerozero88”을 만들었고, 이 시스템을 통해 나온 추천 결과를 웹사이트[11]를 통해 제공하는 방식으로 실험이 진행되었다. 실험 결과 팔로위의 팔로위가 남긴 URL들을 후보로 하고 해당 사용자의 트윗만을 가지고 단어 벡터를 만들며 소셜 점수를 이용하는 경우(FoF-Self-Vote 알고리즘) 가장 좋은 결과를 얻었다. 이를 통해 사용자 단어 벡터를 만들어

사용자와 주제의 관련성을 측정하는 것과 트위터 상의 네트워크를 이용한 소셜 점수의 사용이 추천 결과에 도움을 준다는 것을 밝혔다.

3. URL 추천 시스템의 향상 요소 탐구

이 장에서는 URL 추천 시스템을 향상시킬 만한 요소들에 대해서 살펴보고 어떻게 적용시킬지에 대해 살펴본다. 3.1절에서는 사용자 활동을 고려하여 사용자 단어 벡터를 보정하는 방법을 제안한다. 3.2절에서는 사용자 영향력을 측정하여 추천 시스템을 향상시키는 방법을 제안한다.

3.1 사용자 활동

트위터 이용 시 사용자들은 단순히 트윗을 보고 남기는 것 외에 트윗을 다른 사람에게 전파하거나 트윗에 대한 답글을 보내는 등의 여러가지 활동을 한다. 그 중 사용자의 관심도에 근거가 될 수 있는 활동을 추려 그와 관련된 단어들에 가중치를 주고자 한다.

3.1.1 리트윗한 트윗들의 단어

리트윗(Retweet)이란 자신의 스트림에 나타난 다른 사용자의 트윗을 자신의 팔로워들에게 전파하는 기능이다. 사용자가 어떤 트윗을 리트윗했다는 것은 자신의 팔로워들이 이 트윗을 보기를 바라는 의도가 있다는 의미이고, 따라서 다른 트윗에 비해 리트윗한 트윗에 더 많은 관심을 가지고 있다고 가정할 수 있다.

리트윗된 트윗들이 어떠한 특징을 가지고 있는지 파악하기 위해 무작위로 수집한 1000개의 리트윗된 트윗들을 분석하였다(그림 5). 리트윗된 트윗들 중 가장 많은 부분을 차지한 내용은 유용한 정보나 최신 뉴스를 전달하는 것(44%)이다. 사용자는 리트윗을 통해 자신이 유용하다고 생각하는 정보를 팔로워들에게 전달할 수 있다. 두번째로 많이 차지하는 부분은 다른 사용자의 의견(27%)이다. 사용자는 유명인이 하는 말을 전달하거나

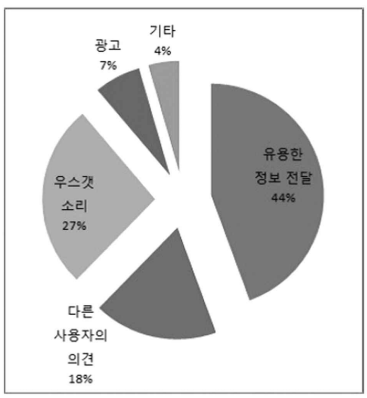

그림 5 리트윗한 트윗의 내용

자신의 생각과 비슷한 사용자의 의견을 다른 사람들에게 보여주는 데에 리트윗을 사용한다. 그 다음으로는 우스갯 소리(18%)가 많은 비중을 차지하였고, 광고(7%)나 기타 내용(4%)은 적은 비중을 차지하였다. 이는 사용자가 남기는 전체 트윗이 쓸데없는 주절거림이 40.55%, 팔로워와의 대화가 37.55%의 내용을 담고있는 것[12]과 비교해 보았을 때 상당히 다른 양상을 보인다.

사용자가 리트윗을 통해 전달하는 정보는 자신이 관심있는 분야의 정보일 가능성이 높고, 사용자가 전달하는 다른 사용자의 의견도 사용자가 관심있어 하는 내용에 대한 의견일 가능성이 높기 때문에 이러한 내용들을 많이 담고 있는 리트윗된 트윗이 다른 트윗들에 비해 사용자의 관심사를 더 잘 표현할 수 있다고 판단하여 가중치를 주었다.

3.1.2 리플라이한 트윗들의 단어

리플라이(Reply)란 자신의 스트림에 나타난 다른 사용자의 트윗에 대한 사용자의 답변을 트윗을 작성한 사용자에게 보내는 기능이다. 사용자는 이 기능을 통해 다른 사용자와 일상적인 대화를 나눌 수도 있고, 특정 트윗에 대한 자신의 의견을 작성자에게 보낼 수도 있다.

사용자가 어떤 트윗에 자신의 의견을 표현했다는 것은 그 트윗의 내용에 관심을 가지고 있는 것이라고 판단할 수 있다. 하지만 사용자들은 리플라이를 다른 사용자들과의 일상적인 대화를 나누는 데에도 많이 사용한다. 리플라이된 트윗 전체를 가중치의 대상으로 선정한다면 일상적인 대화의 내용은 사용자의 관심도를 파악하는 데에 방해 요소가 될 것이다. 이러한 점을 고려하여 가중치를 주는 대상을 일상적인 대화의 내용이 아닌 리플라이한 트윗으로 제한하였다.

일상적인 대화가 내용인 트윗들을 구별하기 위해서 대상이 되는 트윗이 사용자에 대한 멘션으로 시작되었는지를 파악하였다. 만약 리플라이의 대상이 된 트윗이 사용자를 대상으로 쓰여진 트윗이라면 일상적인 대화의 목적을 띈 리플라이라고 판단하였고, 그렇지 않은 트윗은 특정 트윗에 대한 사용자의 의견 표현이라고 판단하였다.

3.1.3 즐겨찾기한 트윗들의 단어

트위터에는 사용자들이 특정한 트윗을 따로 저장할 수 있는 즐겨찾기 기능이 있다(그림 6). 사용자들은 이 기능을 이용하여 자신의 스트림에 있는 트윗들 중 다시 보기를 원하는 트윗들을 따로 보관할 수 있다. 이렇게 즐겨찾기로 보관한 트윗들은 다른 트윗들에 비해 사용자의 관심사를 더 잘 나타낸다고 판단하였다.

3.1.4 트윗을 남긴 장소

스마트폰 보급과 함께 스마트폰 상에서 트위터를 이용하는 사용자들이 많이 늘고 있다. 이에 따라 트윗을

그림 6 트위터의 즐겨찾기 기능

남긴 장소 정보가 트윗과 함께 남길 수 있게 되어 어디서 트윗을 남겼는지 알 수 있다. 사용자들이 주로 생활하는 위치에서 많은 트윗을 남길 것이기 때문에 장소 정보를 분석한다면 직장이나 학교, 집과 같은 주요 생활 장소를 예측할 수 있다.

이러한 장소들에 관련된 단어들을 추출하여 이에 가중치를 준다면 사용자들은 자신의 지역 소식에 관련된 URL들을 추천 받을 수 있다. 또한 지역 정보를 트윗과 함께 남기는 경우는 해당 장소가 트윗의 내용과 관련이 있는 경우가 많기 때문에 장소 정보를 추가하는 것이 트윗의 내용을 고려하는 데 도움이 된다.

트윗을 남긴 주소를 추출하기 위해 우선 트윗에 담긴 GPS좌표 데이터를 야후 거기 OPEN API[13]를 이용하여 한글 주소로 변환하는 작업을 진행하였다. 변환된 주소 중 시·군·구와 읍·면·동 단위의 주소 부분을 추출하여 사용자 단어 벡터에 추가하고 가중치를 주었다.

3.2 사용자 영향력

트위터 스트림은 다양한 사용자들의 트윗으로 채워진다. 그 중에서는 특정 분야에 있어 유명하고 영향력이 높은 사용자도 있고 상대적으로 그렇지 못한 사용자도 있을 것이다. 사용자 영향력을 측정하여 더 영향력이 높은 사용자가 남긴 트윗에 가중치를 준다면 추천 시스템의 성능을 향상시킬 수 있을 것이다. 본 연구에서는 세 가지 요소(팔로워의 수, 하루당 받는 멘션의 개수, 트윗이 하루당 리트윗되는 횟수)를 기준으로 사용자 영향력을 측정하였다.

3.2.1 팔로워의 수

팔로우를 한다는 것은 그 사용자가 남기는 트윗에 관심이 있다는 뜻이므로, 팔로워가 많다는 것은 더 많은 사람들의 관심을 받고 있다는 것을 나타낸다. 이러한 점을 통해 팔로워의 수를 특정 사용자의 인기의 척도라고 볼 수 있다. 특정 사용자의 팔로워의 수와 그 사용자가 남기는 트윗에 대한 관심도가 비례한다고 가정하고 팔로워의 수가 많을수록 사용자 영향력이 더 크다고 측정하였다.

3.2.2 하루당 받는 멘션의 개수

멘션을 많이 받는 사용자를 살펴보면 대부분이 유명 연예인이나 정치인과 같이 많은 사람들에게 알려진 사람들이다. 많은 사람들에게 멘션을 받는다는 것을 그만큼 많은 사람들에게 관심을 받고 있다는 뜻으로 볼 때, 팔로워의 수와 마찬가지로 받는 멘션의 개수도 특정 사용자의 인기의 척도라고 볼 수 있다. 사용자가 받는 멘션의 개수와 그 사용자가 남기는 트윗에 대한 관심도가 비례한다고 가정하고 하루당 받는 멘션의 수가 많을수록 사용자 영향력이 더 크다고 측정하였다.

3.2.3 트윗이 하루당 리트윗되는 횟수

어떤 트윗이 많이 리트윗 되었다면 그 트윗은 많은 사람들이 관심을 갖는 내용을 담고 있다고 판단할 수 있고, 작성한 트윗이 평균적으로 많이 리트윗 된다면 그 사용자는 유용한 내용이 담긴 트윗을 많이 남긴다고 가정할 수 있다. 따라서 어떤 사용자가 남기는 트윗이 리트윗되는 횟수가 평균적으로 높다면 그 사용자가 남기는 트윗이 사람들로 하여금 많은 관심을 끈다고 볼 수 있다. 이러한 점을 고려하여 사용자가 남긴 트윗이 평균적으로 리트윗 되는 횟수와 관심도는 비례한다고 가정하고 하루당 리트윗되는 횟수의 평균값이 클수록 사용자 영향력이 더 크다고 측정하였다.

4. 사용자 활동과 영향력을 고려한 URL 추천

이 장에서는 3장에서 고찰한 요소들을 고려한 새로운 트위터에서의 URL 추천 시스템을 제안한다. 추천 시스템은 그림 7과 같이 구성된다. 먼저 후보 URL을 선정하여 수집하고 그 다음으로 사용자 단어 벡터와 트윗 단어 벡터를 만든다. 후보 트윗들의 단어 벡터와 사용자 단어 벡터를 비교하여 유사도를 측정하고 마지막으로 사용자 영향력을 더해 순위화 하여 높은 순위의 URL들을 추천해준다.

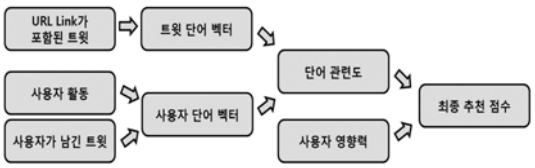

그림 7 URL 추천 시스템 개념 모델

4.1 URL 수집

후보 URL은 Zerozero88에서 가장 좋은 성능을 보였던 팔로워의 팔로워가 남긴 트윗을 대상으로 하였다. 수집은 트위터 API[14]를 이용하였다. 트위터 API를 이용하여 먼저 대상 사용자의 팔로워의 아이디를 구하고, 다시 팔로워들의 팔로워의 아이디를 구한다. 이렇게 수집한 사용자의 팔로워의 팔로워들의 아이디를 가지고 그들이 남긴 트윗들 중 최근 3일간의 트윗을 수집한다.

트윗들을 수집한 뒤 URL이 담긴 트윗을 추려내고 그

트윗에 담긴 URL을 추출하는 작업을 진행하였다. 트윗의 URL은 상당수가 bit.ly[15]나 durl[16]과 같은 사이트들을 통해 줄여진 URL이기 때문에 longurl API[17]를 이용하여 줄여진 URL을 원래의 URL로 바꾸는 작업도 같이 수행하였다.

4.2 사용자 단어 벡터와 트윗 단어 벡터

사용자 단어 벡터는 Zerozero88에서 가장 좋은 성능을 보였던 사용자 본인이 남긴 트윗의 단어들만을 대상으로 하는 방식을 사용하여 만들었다. 사용자 단어 벡터를 만들기 위해 먼저 사용자가 남긴 트윗들 중에서 불용어를 제외한 명사들을 추출하였다. 명사들을 추출하기 위해서 초고속 한국어 형태소 분석기 MACH 1.0[18]을 사용하였다.

그 다음으로 사용자 활동에 따라 추출된 단어들에 가중치를 주었다. 각 트윗의 단어들은 다음과 같은 식을 통해 점수를 계산하였다.

$$weight(w_{ij}) = \gamma^{rt\_flag(w_{ij})} * \delta^{rp\_flag(w_{ij})} * \varepsilon^{fv\_flag(w_{ij})}$$
$$\phi^{geo\_flag(w_{ij})} * (\text{the number of } word_i \text{ in a } Tweet_j)$$

여기서 w_{ij}는 트윗 j에 포함된 단어 i를 뜻한다. γ는 리트윗된 트윗들에 대한 가중치 상수이며 δ는 리플라이 된 트윗에 대한 가중치 상수, ε은 즐겨찾기 된 트윗에 대한 가중치 상수 그리고 ϕ는 지역에 관련된 단어에 대한 가중치 상수이다. rt_flag는 리트윗 되었는지 여부를 나타내고, rp_flag는 리플라이 되었는지 여부, fv_flag는 즐겨찾기 여부, geo_flag는 장소 관련 단어인지 아닌지에 대한 여부를 나타낸다.

이렇게 각각의 트윗과 단어들에 대해서 점수를 계산한 뒤 별개의 단어들에 대해 TF-IDF값을 구하여 사용자 단어 벡터를 만들었다. TF-IDF는 언어 자료 내의 특정 문서에서 어떤 단어의 중요도를 평가하기 위해 사용되는 통계적 수치이다. TF-IDF값은 문서 내에서 해당 단어가 많이 나타날수록 증가하며, 전체 자료 내에서 해당 단어가 많이 나타날수록 감소한다. TF값과 IDF값은 다음과 같은 식[19]을 통해 계산하였다.

$$TF_u(w_i) = \sum weight(w_{ij})$$
$$IDF_u(w_i) = log\left(\frac{\#\text{all users}}{\#\text{users using } w_i \text{ at least once}}\right)$$

이 식에서 나타난 바와 같이 특정 단어의 TF 값은 사용자의 전체 트윗에 나타난 동일한 단어들의 점수의 합으로 계산하였고, IDF 값은 트위터 모든 사용자 수를 해당 단어를 한번이라도 언급한 사용자 수로 나눈 값에 로그를 취하는 방식으로 계산하였다. 트윗 단어 벡터는 특별한 가중치 없이 트윗에 나타난 단어의 빈도 수와 단어를 사용한 사용자의 수를 바탕으로 한 TF-IDF 값을 이용하여 만들었다. 이렇게 계산된 단어 벡터의 값들은 다양한 범위의 값들을 갖게 되므로 모든 값이 0에서 1사이의 값을 가지도록 정규화하여 일정한 범위의 값을 갖도록 하였다.

4.3 유사도 측정

사용자가 특정 트윗과 얼마나 관련이 있는 지를 측정하기 위해 사용자 단어 텍터와 트윗 단어 벡터 간의 유사도를 구하였다. 벡터 간의 유사도는 코사인 유사도를 이용하여 다음과 같은 식을 통해 계산하였다.

$$similarity = \frac{V_u \cdot V_t}{\| V_u \| \| V_t \|}$$

여기서 V_u는 사용자 단어 벡터고, V_t는 트윗 단어 벡터이다.

4.4 사용자 영향력 적용

앞에서 언급한 바와 같이 사용자 영향력은 팔로워의 수, 하루에 리트윗되는 트윗의 수 그리고 하루에 받는 멘션의 수 이렇게 세 가지 요소를 기준으로 하여 측정하였다. 각각의 요소에 대한 측정값은 다음과 같은 식으로 계산하였다.

$$userInfluence_{follower} = log\left(\frac{\text{The number of followers} + 1}{\lambda}\right)$$
$$userInfluence_{retweet} = log\left(\frac{\text{The number of retweets per day} + 1}{\mu}\right)$$
$$userInfluence_{mention} = log\left(\frac{\text{The number of mentions per day} + 1}{\mu}\right)$$

여기서 λ, μ, ν는 각각의 값들을 정규화하기 위한 정규화 상수들이다. 팔로워의 수에 따른 사용자의 분포가 지수 분포의 형태를 띠며[20], 리트윗되는 트윗의 수나 하루에 받는 멘션의 수에 따른 사용자의 분포 역시 마찬가지 형태를 보인다고 판단하여 한 요소에 대해 측정한 영향력이 치우치지 않도록 각 요소들에 대해 로그값을 취하였다.

각 요소들은 서로 연관도가 없으므로[10] 요소별 사용자 영향력 값을 각각 구한 뒤 4.3절에서 구한 유사도와 합하여 최종 유사도 값을 구하며, 이를 기준으로 트윗들을 순위화 하여 사용자에게 추천한다.

$$similarity_{final} = similarity + (\text{each userInfluence})$$

5. 성능 평가

이 장에서는 사용자 활동과 사용자 영향력을 고려하여 URL 추천 시스템의 성능을 얼마나 향상시켰는지를 실험을 통해 살펴본다. 기존 방법과의 비교 실험을 통해 제안한 방법의 추천 성능을 평가한다.

5.1 실험 환경 및 실험 데이터

실험에 사용된 시스템은 Intel(R) Xeon(R) CPU X5450 3.00GHz의 CPU와 16GB의 RAM 사양을 가진다. 운영체제로는 Ubuntu 9.10 버전이 사용되었다. URL 추천 시스템은 Ruby로 구현하였다. 또한 실험 참

가자들의 평가를 위한 UI 프로그래밍은 Rails를 통해 구현되었다. 데이터베이스는 Mysql 5.1.37 버전을 사용하였다.

팔로위의 팔로위의 트윗들을 수집할 때 팔로위가 500명 이상인 사용자는 실질적으로 타임라인을 보지 않는 사용자로 간주하여 수집의 대상에서 제외하였다. 총 34294의 사용자의 69147개 URL 링크를 포함한 트윗을 추천 후보로 선정하였고, IDF 값을 계산하기 위해 Twitter Streaming API[21]를 통해 한 달간 607671개의 트윗을 수집하였다.

5.2 성능 평가 방법

총 20종류의 알고리즘에 대하여 비교하였으며 검색과 추천 엔진 알고리즘의 척도인 NDCG[22]를 성능 평가의 척도로 사용하였다. NDCG는 이상적인 결과 순위와 알고리즘의 결과 순위가 유사할수록 큰 값을 가지며 다음과 같은 방법으로 계산한다.

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

Nq는 쿼리 q에 대한 NDCG 값이며 값이 클수록 더 좋은 성능의 알고리즘이다. Mq는 질의 q에 대해 NDCG 값을 1로 만드는 완벽 정렬을 위한 정규화 상수이고 j는 검색 결과 내 순위, r(j)는 j순위에서의 결과 적합성(값이 클수록 적합)이다. Nq값이 클수록 이상적인 결과 순위와 유사함을 나타낸다.

실험은 50개 이상의 트윗을 남기고 20명 이상의 사용자를 팔로우하는 10명의 트위터 사용자를 대상으로 진행하였다. 실험 대상자들은 각각 알고리즘별로 50개씩 추천되어 총 500개의 URL들을 평가하였다. URL당 관련도 점수 r(j)는 0-3점으로 설정하였다.

Zerozero88에서 가장 좋은 성능을 보였던 FoF-Self-Vote 알고리즘을 비교 대상으로 선정하였고, 먼저 각각의 사용자 활동이 어떤 영향을 미치는지 알아보기 위해 각각의 사용자 활동만을 적용한 알고리즘을 비교 대상으로 하여 실험하였다. 그리고 그 중 가장 좋은 성능을 보이는 알고리즘들을 결합하여 어떤 결과가 나타나는지 실험하였다. 그 다음으로 사용자 영향력이 어떤 영향을 미치는지 알아보기 위해 기존 알고리즘에 사용자 영향력만을 추가한 알고리즘을 비교 대상으로 실험하였다. 그리고 좋은 성능을 보이는 요소들을 결합한 알고리즘 또한 비교 대상으로 실험하였다. 마지막으로 사용자 활동과 사용자 영향력을 모두 고려한 알고리즘을 비교하는 것으로 실험을 마무리 하였다.

5.3 실험 결과

그림 8은 각각의 사용자 활동에 가중치를 적용한 알고리즘들과 리트윗과 리플라이 두 가지 활동에 가중치를 동시에 준 알고리즘을 이용해서 뽑은 상위 50개의 URL에 대한 상위 k개의 NDCG 값을 나타낸 그래프이다. 두 가지 요소를 결합한 경우는 여러 가지 비율 가지고 실험을 하였고, 가장 결과가 좋은 리트윗 상수 γ : 리플라이 상수 δ = 2.5 : 1의 비율을 결과 그래프에 표시하였다. 그림에서 보는 바와 같이 두 가지 요소를 결합한 알고리즘이 가장 좋은 성능을 보였고, 그 아래로는 리트윗에 가중치를 준 알고리즘, 리플라이에 가중치를 준 알고리즘, 마지막으로 기존의 알고리즘 순서대로 좋은 성능을 보였다. 이를 통해 단순히 사용자가 남긴 단어들만을 사용하여 사용자 단어 벡터를 만드는 것보다는 리트윗이나 리플라이한 트윗들의 단어에 가중치를 준 경우에 더 좋은 추천 결과를 얻었음을 알 수 있다.

리트윗한 트윗에 가중치를 주는 알고리즘이 리플라이한 트윗에 가중치를 주는 알고리즘보다 좀더 나은 성능을 보였다. 이러한 결과는 리플라이는 관심있는 주제에 관한 트윗에 응답하는 것 외에도 일상적인 대화를 나누는데 자주 쓰이기 때문에 이러한 트윗들이 노이즈로 작용했을 가능성이 높다. 하지만 그럼에도 불구하고 기존

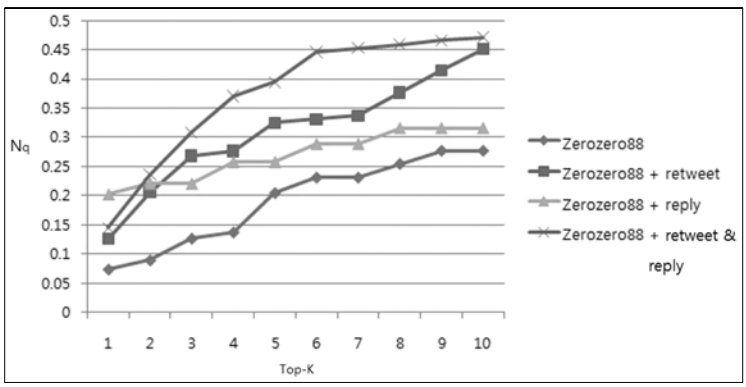

그림 8 사용자 활동에 가중치를 준 알고리즘들의 top-k NDCG

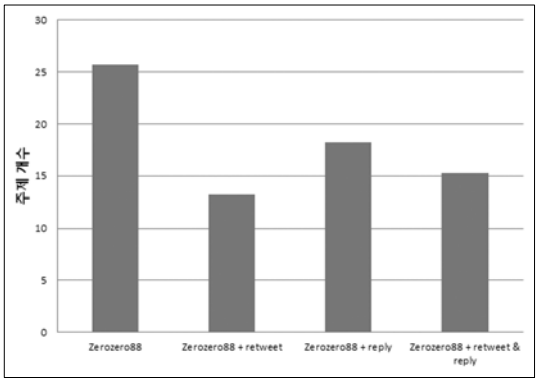

그림 9 사용자 활동에 가중치를 준 알고리즘들의 추천 트윗의 주제 개수

의 연구보다 좋은 성능을 보일 수 있는 이유는 일상 대화에서의 단어들이 URL 트윗들의 단어와 일치하는 부분이 크지 않아 유사도 측정에 있어 큰 영향을 주지 못하기 때문으로 보인다.

사용자 활동을 결합한 알고리즘이 만족도 측면에서는 좋은 결과를 보였지만, 그림 9에서 보듯이 기존의 알고리즘에 비해 추천된 트윗들의 주제의 다양성이 낮다는 단점이 있다. 사용자 활동에 관련된 주제 위주로 추천을 하기 때문에 이러한 결과가 발생한 것으로 보인다.

그림 10은 사용자 영향력을 고려한 세 가지 알고리즘을 이용하여 뽑은 상위 50개의 URL에 대한 상위 k의 NDCG 값을 나타낸 그래프이다. 하루당 리트윗 개수를 사용자 영향력으로 사용한 알고리즘은 성능 향상에 효과가 있었으나, 하루당 멘션 개수와 팔로워의 수를 사용자 영향력으로 사용한 알고리즘은 성능 향상에 도움을 주지 못하였다. 멘션의 개수나 팔로워의 수와 같이 사용자의 인기도를 반영하는 수치들은 성능 향상에 도움을 주지 못한 반면, 사용자가 남기는 트윗 자체에 대한 관심도를 반영하는 리트윗의 빈도를 고려한 알고리즘은 좋은 결과를 가져옴을 알 수 있다.

그림 11은 사용자 활동과 세 가지의 사용자 영향력을 함께 고려한 알고리즘의 결과이다. 리트윗 상수 γ : 리

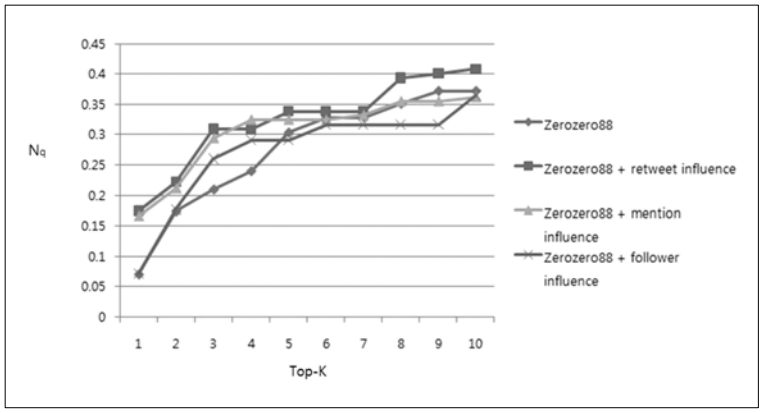

그림 10 사용자 영향력을 고려한 알고리즘들의 top-k NDCG

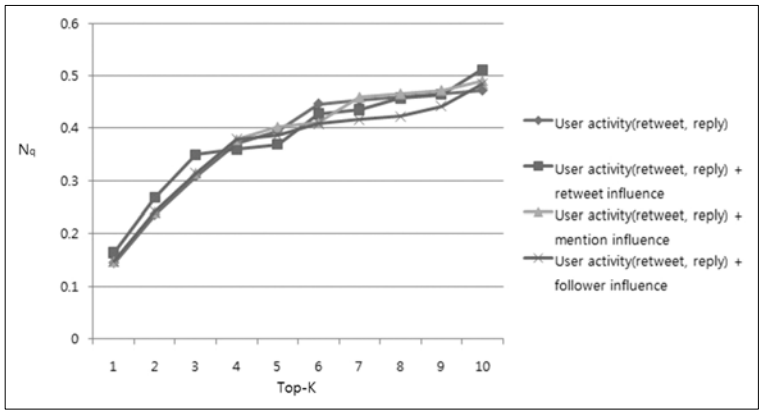

그림 11 사용자 활동과 사용자 영향력을 함께 고려한 알고리즘들의 top-k NDCG

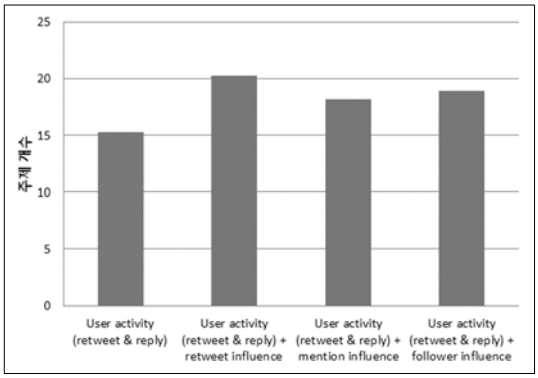

그림 12 사용자 활동과 사용자 영향력을 함께 고려한 알고리즘들의 추천 트윗의 주제 개수

플라이 상수 δ의 비율은 이전 실험과 마찬가지로 실험 결과가 가장 좋은 2.5 : 1로 설정하였다. 각 알고리즘은 사용자 영향력을 결합하지 않은 경우와 결합한 경우들 모두가 비슷한 결과를 보인다. 사용자 활동과 사용자 영향력의 결합이 향상된 결과를 가져오지 못한 것은 각각의 요소가 다른 방향의 주제를 강조하기 때문으로 보인다. 사용자 활동에 대한 가중치는 사용자 개인이 관심을 가지는 트윗을 추천하는 데 도움을 주는 방법인 반면에, 사용자 영향력을 고려하는 것은 일반적으로 대다수의 사람들이 관심을 가지는 트윗을 추천하는 데 도움을 주는 방법이다. 이렇게 두 가지 요소가 강조하는 부분이 다르기 때문에 두 요소의 결합이 시너지 효과를 가져오지는 못한다. 하지만 추천된 트윗들의 주제에 대해 살펴본 결과, 그림 12에서 알 수 있듯이 사용자 활동과 사용자 영향력을 결합한 경우에 좀더 주제의 다양성이 높았다. 이는 두 가지 요소를 결합함으로써 사용자 활동만을 고려했을 때 추천 트윗의 주제가 단조로워진다는 단점을 보완할 수 있음을 나타낸다.

6. 결론 및 향후 연구

본 논문에서는 트위터 내에서의 사용자 활동 중 사용자의 주제에 대한 관심도를 나타낼 수 있는 요소들에 대해서 살펴보고 실험을 통해 추천 알고리즘 안에서의 효과를 비교하였다. 또한 사용자 영향력을 세가지 요소로 구분하여 측정하였고, 그 요소들의 URL 추천 알고리즘에 어떤 영향을 미치는지 비교하였다. 5장의 실험 결과에서 보여주는 바와 같이 제안한 방법인 사용자 활동과 사용자 영향력에 따른 가중치가 기존의 URL 추천 알고리즘의 성능을 향상시켰다. 본 논문에서 제안한 요소들을 마이크로블로깅 서비스를 활용한 검색이나 추천 관련 연구에 활용한다면 더 좋은 결과를 얻을 수 있을 것으로 기대해 본다.

향후 연구로는 트윗이 리트윗된 횟수를 고려한 추천 알고리즘을 구현해 보고자 한다. 트윗의 리트윗된 횟수는 사람들의 관심에 대한 명백한 척도이기 때문에 추천 알고리즘에 이용한다면 좋은 결과를 기대할 수 있다. 트위터에서의 사용자 영향력 측정에 대한 연구도 필요하다. 본 연구에서는 사용자 영향력을 나타내는 요소들을 개별적으로 고려하여 실험하였지만, 이러한 요소들을 복합적으로 고려하여 좀더 근거 있는 사용자 영향력 측정 알고리즘이 나온다면 트위터 관련 연구들에 큰 도움이 될 것이다.

참 고 문 헌

[1] http://www.facebook.com
[2] http://www.twitter.com
[3] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon, "What is Twitter, a Social Network or a News Media?," in *Proc. of the 19th international conference on World wide web (WWW '10)*, pp.591-600, 2010.
[4] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, Ed H. Chi, and Rowan Nairn, "Short and Tweet: Experiments on Recommending Content from Information Streams," in *Proc. of the 28th international conference on Human factors in computing systems (CHI '10)*, pp.1185-1194, 2010.
[5] Jagan Sankaranarayanany, Hanan Samety, Benjamin E. Teitlery, Michael D. Liebermany, and Jon Sperlingz, "TwitterStand: News in Tweets," in *Proc. of the 17th ACM SIGSPATIAL international conference on Advances in Geographic Information Systems (GIS '09)*, pp.42-51, 2009.
[6] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," in *Proc. of the 19th international conference on World wide web (WWW '10)*, pp.851-860, 2010.
[7] Michael Mathioudakis and Nick Koudas, "TwitterMonitor: Trend Detection over the Twitter Stream," in *Proc. of the 36th ACM SIGMOD International Conference on Management of Data (SIGMOD '10)*, pp.1155-1158, 2010.
[8] Satyen Abrol and Latifur Khan, "Twinner: Understanding News Queries with Geo-content using Twitter," in *Proc. of the 6th Workshop on Geographic Information Retrieval (GIR '10)*, article 10, pp.1-8, 2010.
[9] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, and Zhaohui Zheng, "Time is of the Essence: Improving Recency Ranking Using Twitter Data," in *Proceedings of the 19th international conference on World wide web (WWW '10)*, pp.331-340, 2010.
[10] Meeyoung Cha, Hamed Haddadi, Fabr´ıcio Bene-

venuto, and Krishna P. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy," in *Proc. of the 4th international AAAI conference on Weblogs and Social Media (ICWSM '10)*, pp.10-17, 2010.
[11] http://www.zerozero88.com
[12] Twitter Study-pearanalytics. http://www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf
[13] http://kr.open.gugi.yahoo.com/
[14] http://api.twitter.com
[15] http://bit.ly/
[16] http://durl.kr/
[17] http://longurl.org/expand
[18] Kwangseob Shim and Jaehyung Yang, "MACH: A Supersonic Korean Morphological Analyzer," in *Proc. of the 19th International Conference on Computational Linguistics (COLING '02)*, pp.939-945, 2002.
[19] Salton, G. and Buckley, C., "Term Weighting Approaches in Automatic Text Retrieval," *Information Processing and Management*, vol.24, no.5, pp.513-523, 1988.
[20] Alex Cheng, Mark Evans, and Harshdeep Singh, "Inside Twitter: An In-Depth Look Inside the Twitter World," 2009. http://www.sysomos.com/insidetwitter
[21] http://stream.twitter.com
[22] Jarvelin, K. and J. Kekalainen, "Cumulated Gain-Based Evaluation of IR Techniques," *ACM Transactions on Information Systems*, vol.20, no.4, pp.422-446, 2002.

이 성 윤
2009년 2월 서울대학교 컴퓨터공학부 학사. 2011년 2월 서울대학교 컴퓨터공학부 석사. 2011년 2월~현재 KIS채권평가 선임연구원. 관심분야는 시맨틱 웹, 웹 2.0, 소셜 네트워크, 데이터 마이닝

이 태 휘
2004년 2월 서울대학교 컴퓨터공학부 학사. 2004년 2월~현재 서울대학교 컴퓨터공학부 석박사 통합과정 재학중. 2007년 6월~2010년 4월 티맥스소프트 선임연구원. 관심분야는 텍스트/그래프 데이터 검색, 대규모 데이터 처리, 시맨틱 웹

김 형 주
정보과학회논문지 : 컴퓨팅의 실제 및 레터 제 17 권 제 3 호 참조