사용자 정보 및 활동에 기반한 트위터 사용자 리스트 생성

(User List Generation based on User Information and Activity in Twitter)

김 소 민 † 임 혜 원 † 김 형 주 ††
(Somin Kim) (Hyewon Lim) (Hyoung-Joo Kim)

요 약 트위터는 2012년 현재 활발한 사용자만 1억 4천명 이상인 인기 있는 마이크로블로깅 서비스이다. 트위터는 소셜 네트워크적인 특성을 가지고 있음과 동시에 정보 공유의 장으로서의 역할도 하고 있다. 트위터에서의 친구 관계는 단방향이라는 특성을 가지고 있다. 이러한 특성 때문에 사용자는 다른 사용자들을 다양한 목적으로 팔로우 할 수 있게 된다. 서로 다른 목적으로 이루어진 개인의 팔로위들의 트윗을 구분해서 보기 위해서 팔로위들을 군집화 할 필요성이 제기되었다. 이를 위해 트위터에서 리스트라는 기능을 제공하는데, 리스트를 만드는 과정은 사용자의 노력을 전적으로 필요로 한다. 본 논문에서는 트위터의 리스트가 사회적, 주제, 속성 리스트로 구분된다는 점을 고려하여 시드 사용자 없이 사용자 리스트를 생성하는 연구를 진행한다. 각 군집의 특성에 적합한 사용자 정보와 활동을 이용하여 계층적 군집화 알고리즘을 활용하여 팔로위들을 군집화 하였다. 구성된 군집의 이름을 정하기 위해 기존에 존재하는 리스트 이름을 활용하였다. 실험을 통해 본 논문에서 제안한 알고리즘이 군집 구성원 적합성, 군집 커버리지, 군집 이름 적합성 측면에서 좋은 성능을 보임을 증명하였다.

키워드 : 트위터, 리스트, 사용자 군집화, 소셜 네트워킹 서비스

Abstract Twitter is one of the most popular micro blogging services and has more than 140 million active user in 2012. Twitter plays a role of both social networking service and information sharing. Twitter following is a directed relationship. As a result, Twitter users follow other accounts for various purposes and that makes grouping feature is needed. Twitter has serviced a functionality named “List” which allows users to group users. However, the process of creating and managing lists requires significant efforts from users. In this paper, we propose a method that automatically generates Twitter user lists without requiring seed users. The proposed method considers that there exist social lists, topical lists, and property lists in Twitter. To generate Twitter user lists, we utilize hierarchical clustering algorithms with appropriate user information and activities for each grouping method. Names of existing lists are utilized for group name recommendation. Experimental results show that our approach is effective in aspects of group members, group coverage, and validity of group naming.

Key words : Twitter, List, User Clustering, Social Networking Service

· 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 20120005695). 본 연구는 BK-21 정보기술 사업단의 연구결과로 수행되었음

† 비 회 원 : 서울대학교 컴퓨터공학부
smkim@idb.snu.ac.kr
hwlim@idb.snu.ac.kr
(Corresponding author임)
†† 종신회원 : 서울대학교 컴퓨터공학부 교수
hjk@snu.ac.kr

논문접수 : 2012년 7월 2일
심사완료 : 2012년 9월 18일

정보과학회논문지: 데이타베이스 제39권 제6호(2012.12)

1. 서 론

소셜 네트워킹 서비스는 공통된 관심사를 갖고 있거나 활동을 하고 있는 사람들, 혹은 실제로 관계를 맺고 있는 사람들의 인간 관계를 온라인 환경에 반영한 서비스이다. 이러한 소셜 네트워크 서비스는 모바일 디바이스의 발전과 더불어 웹 2.0에서 표방하는 참여, 공유, 개방의 목표대로 누구나 손쉽게 정보를 생산하고 공유 할 수 있는데 일조하고 있다.

그 중에서도 트위터는 최근 가장 인기 있는 마이크로블로깅 서비스로, 많은 사람들과 관계를 맺고 소통하는 소셜 네트워크적인 특성과 정보를 얻을 수 있는 장으로서의 특성을 동시에 지니고 있다. 이러한 특성에 힘입어 2006년 3월에 서비스를 시작하여 현재까지 6년 동안 꾸준히 성장하였으며, 2012년 3월 기준으로 활발하게 활동하는 사용자만 1억 4천만 명 이상이다.

페이스북(Facebook)[1)]과 같은 일반적인 소셜 네트워크 서비스에서는 서로가 원할 때만 친구 관계를 맺을 수 있도록 하고 있기 때문에, 친구들은 대개 실생활의 친구들로 이루어져 있다. 그러나 트위터에서는 단방향의 친구 관계를 허용하므로, 상대방의 허가 없이도 사용자가 자유롭게 친구 관계를 맺을 수 있다. 따라서 트위터 상에서는 보통 한 사용자의 실제 지인뿐만이 아니라, 사용자가 직접 알지는 못하는 유명인들이나 특정 주제에 대해서 이야기하는 뉴스 미디어 등의 계정도 팔로우하여 의사소통을 하고 정보를 주고받을 수 있다. 이렇게 한 사용자의 팔로위가 여러 목적으로 구성되다 보니 이들을 잘 분류해야할 필요성이 대두되어 트위터는 2009년 말 공식적으로 리스트(List)라는 기능을 제공하기 시작했다.

리스트는 트위터 상에서 사용자들을 군집화하여 각 군집에 적절한 이름을 붙인 것을 의미한다. 트위터 사용자들은 일반적으로 유사한 특성을 가지거나 비슷한 주제에 대해서 이야기하는 사용자들을 묶어 하나의 리스트로 만든다. 한 사용자가 리스트의 이름을 정하여 리스트를 생성하고 다른 사용자들을 리스트에 등록시키면, 사용자는 해당 리스트에 속한 구성원들이 쓴 트윗 메시지들만 따로 모아 볼 수 있게 된다. 예를 들어 '정치인'이라는 리스트를 만들고 정치인들의 트위터 계정을 등록해놓으면, 모든 팔로위들이 쓴 트윗 메시지들 대신에, 리스트에 등록된 정치인들이 쓴 트윗 메시지들만 모아서 볼 수 있다.

트위터에서 리스트를 이용하게 되면 다음과 같은 장점들이 있다. 첫 번째로 개인의 팔로위가 많을 경우에 팔로위들을 편리하게 관리 할 수 있다. 두 번째로 타임라인에서 다양한 특성을 가진 팔로위들이 쓰는 트윗 메시지들을 단순히 시간 순서대로만 볼 수 있었지만, 리스트를 이용하면 어떤 의미 있는 주제를 기준으로 트윗 메시지들을 구분하여 볼 수 있게 된다. 이를 통해 특정 주제나 특정 집단 내에서의 정보의 흐름을 쉽게 파악할 수 있다. 마지막으로, 리스트의 이름이 리스트에 등록된 사용자들에 대한 태그(Tag) 역할을 할 수 있다. 한 사용자가 다른 사용자를 팔로우한다는 것은 그 사용자에게 관심이 있고 그 사용자가 작성하는 트윗 메시지에 관심이 있다는 것을 의미한다. 하지만 팔로우 자체만으로는 어떤 이유로 그 사용자에게 관심을 가지게 되었는지 알 수 없다. 만약 팔로위들을 리스트를 이용해 구분하게 되면 그들이 속한 리스트의 이름이 리스트 구성원들을 설명하는 태그 정보로 활용 될 수 있다.

현재 트위터에서 리스트 기능을 이용하고 있는 사용자는 전체 사용자의 24%에 불과하다[1]. 이렇게 사용자가 적은 이유 중의 하나는 트위터에서 팔로위들을 분류할 수 있는 기능을 제공해주기는 하지만 리스트를 생성하고 관리하는 과정은 전적으로 사용자의 몫이기 때문이다. 사용자는 리스트를 만들기 위해 어떤 리스트를 생성해야할지, 각 사용자 집단에 어떤 사용자들을 구성원으로 등록시킬지, 각 사용자 집단에 적합한 이름은 무언인지 등을 결정해야한다. 팔로위가 많은 경우에 이 과정은 많은 노력을 요하는 지루한 작업이다. 이러한 리스트 생성 과정을 돕기 위해 페이스북에서는 '스마트 리스트'라는 기능을 제공하여 사용자가 입력한 개인 정보들을 바탕으로 자동으로 리스트를 생성해 주고 있다. 하지만 트위터 사용자는 다양한 목적과 특성을 지니기 때문에 분류가 까다롭다. 페이스북과 같은 소셜 네트워크 서비스의 대부분의 사용자들은 실제로 서로 알고 있는 사용자들이 친구를 맺고 활동하지만, 트위터의 경우에는 실제로 알지 못하더라도 뉴스 미디어, 연예인, 정치인과 같은 계정을 팔로우한다. 그렇기 때문에 트위터에서 리스트를 생성하기 위해 사용자들을 분류할 때에는 다양한 요소를 고려해야 한다. 또한 한 사용자가 하나의 리스트로만 분류되는 것이 아니라 여러 리스트에 중복되어 분류될 수도 있다. 무엇보다 사용자 분류의 기준이 될 수 있는 개인정보가 부족하기 때문에 다른 소셜 네트워크에 비해 분류가 어렵다.

트위터 상에 존재하는 리스트는 크게 세 종류로 나눌 수 있다[1]. 첫 번째 종류는 각 사용자의 관점에서 의미 있는 이름 – 예를 들어 "친구", "가족" – 을 사용하여 자신의 지인들의 계정을 묶어놓는 경우이다. 두 번째는 리스트의 이름이 사용자의 주제를 나타내는 경우로, "음악", "스포츠" 등의 이름을 이용해 사용자들이 작성하는 트윗 메시지의 주제가 유사한 사람들을 묶어 놓은 리스

1) http://www.facebook.com

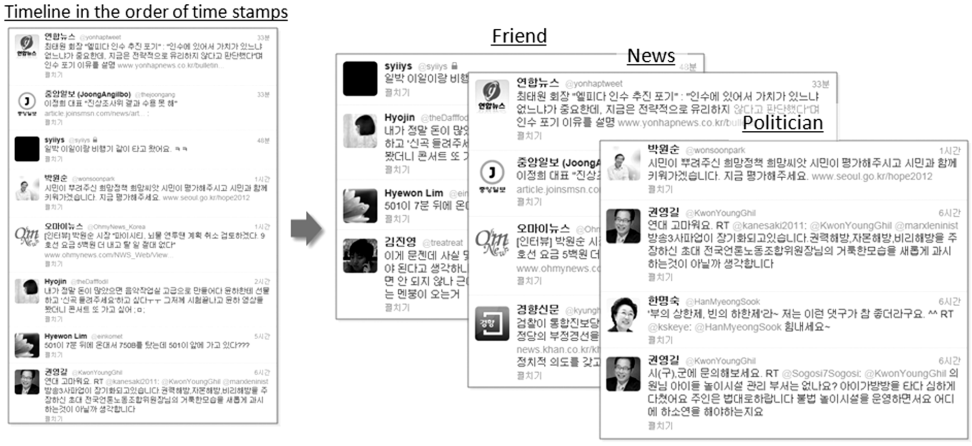


그림 1 리스트를 기준으로 나누어진 타임라인

트이다. 세 번째는 사용자의 속성을 나타내주는 경우로 “Famous”, “정치인”와 같은 리스트가 있다. 본 논문에서는 이와 같이 트위터 상에 다양한 종류의 사용자 리스트가 존재한다는 점을 고려해서 사용자 리스트를 제안할 수 있는 방안에 대해 연구하였다. 세 가지 리스트를 각각 사회적 리스트, 주제 리스트, 속성 리스트로 명명하고, 각 종류의 사용자 리스트를 구성할 수 있도록 사용자의 정보 및 활동을 활용하였다.

본 연구에서는 기존의 리스트와 구분하기 위하여, 본 연구의 시스템에서 찾은 사용자 집합을 “군집”이라 하고, 찾은 군집에 이름이 추천 되어 리스트의 형태를 갖춘 것을 “리스트”라고 한다.

본 논문의 구성은 다음과 같다. 1장에서는 트위터의 리스트에 대해 소개하고 사용자 리스트를 자동으로 제안해주는 방안의 필요성을 제기하였다. 2장에서는 소셜 네트워크 서비스에서 사용자 군집을 자동으로 찾는 연구들에 대해서 살펴보고 3장에서는 개인의 팔로위들을 사회적 군집, 주제 군집, 속성 군집으로 나누어 리스트를 제안하는 본 시스템의 알고리즘에 대해 설명한다. 4장에서는 본 연구에서 제안한 시스템의 성능에 대한 실험과 평가에 대해 설명한다. 마지막으로 5장에서는 제안한 리스트 자동생성 방법에 대한 결론과 향후 연구에 대해서 언급한다.

2. 관련 연구

최근 몇 년 간 트위터 사용자가 급증함에 따라 많은 연구자들이 트위터를 다양한 관점에서 연구, 분석하였다. [2]에서는 사용자들의 트위터 이용 행태를 분석하여 트위터 네트워크 상에서 나타나는 특성들을 전반적으로 분석하여 소셜 네트워크인 트위터가 정보 공유를 할 수 있는 새로운 형태의 미디어로서의 역할도 가지고 있음을 밝혔다.

트위터 리스트가 가지는 특성을 활용한 연구들도 있다. [1]에서는 트위터 사용자들의 리스트 사용 행태 등 트위터 리스트의 전반적인 특성을 살펴보았다. 또한, 개인이 속한 리스트의 이름에 포함된 단어를 태그로 활용하여 사용자가 주로 언급하는 주제를 파악하였다. [3]에서는 트위터 리스트를 이용하여 네트워크 상에서 정보 전달에 핵심 역할을 하는 사용자를 추천하였다.

사용자가 보다 쉽게 트위터 리스트 기능을 이용할 수 있도록 트위터 리스트 기능을 지원하기 위한 연구들도 진행되었다. [4]에서는 트위터의 뉴스 미디어적인 특성을 높이 평가하여 뉴스거리가 될 만한 내용의 트윗만 걸러내기 위한 시도를 하였다. 특정 뉴스에 대해서 트윗을 작성한 사용자들을 시드(Seed)로 하여 그들의 주변 네트워크, 메시지 전달 여부 등을 분석하여 네트워크를 확장시켜가며 비슷한 뉴스에 대해서 이야기하는 사용자들을 군집화 하였다. [5]에서는 사용자의 팔로위 중에서 시드 사용자를 입력받아 시드 사용자와 한 군집으로 묶일 가능성이 높은 팔로위들을 사용자 간 관계의 긴밀도 또는 각 사용자가 작성한 트윗의 유사도를 바탕으로 순서화 하여 나타내었다.

이들 외에 리스트와 직접적으로 연관되지는 않지만 네트워크 상에 존재하는 사용자들의 속성을 분석하는 연구도 있었다. [6]에서는 영상 데이터를 공유하고 의견을 나누는 소셜 네트워크 서비스인 유투브[2]의 사용자 군집에 대하여 분석하였다. 어떤 점이 유투브 상에서 사용자 군집을 형성하게 되는 동기가 되는지, 사용자 군집이 어떻게 형성되고 변화되어 가는지 등에 대해 연구하

2) http://www.youtube.com

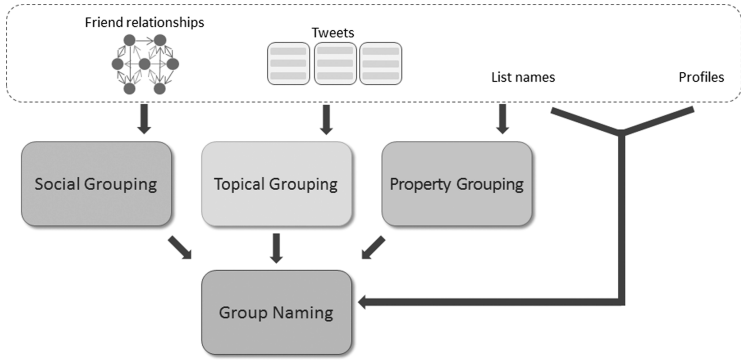


그림 2 시스템 구성도

였다. [7]에서는 기계학습 방법론을 이용하여 트위터 사용자들을 여러 가지 기준으로 분류하였다. 사용자의 프로필, 행동, 작성하는 메시지, 메시지를 보내는 사람 등을 기준으로 하여 사용자들의 정치적 성향이나 관심사를 분석하였다. 또 사용자들의 속성을 분석하는 것에서 그치지 않고, 속성이 유사한 사용자들을 군집화 하는 연구들도 많이 있다. [8]에서는 일반적으로 하듯이 사용자들이 만들어내는 컨텐츠의 속성을 기준으로 사용자들을 군집화 한 것이 아니라, 사용자들의 행동을 분석하여 유사한 사용자들을 군집화 하였다. [9]에서도 사용자들의 시간에 따른 행동 변화에 따라 사용자들을 군집화 하였다. 일정한 시간 간격 동안 특정 행동을 얼마나 반복하는지를 바탕으로 군집화 하여 다양한 소셜 네트워크 서비스에서 나타나는 특징을 분석하였다. [10]에서는 지리정보가 저장 되어있는 트윗 메시지들을 이용하여 전체 네트워크가 아닌 개인의 주변에서 실시간으로 일어나고 있는 여러 가지 일들을 파악할 수 있게 하였다. 주변에서 트윗 메시지를 작성하고 있는 사용자들을 트윗 메시지의 주제를 기준으로 군집화 하였다. [11]에서는 온톨로지로 표현된 사용자의 프로필을 분석하여 사용자들의 관심사나 취향에 따라 사용자들을 군집화 하였다.

본 연구에서는 [1]에서 언급한 다양한 종류의 트위터 리스트를 고려하여 자동적으로 사용자 리스트를 생성하고자 한다. [4,5]에서는 시드 사용자를 입력으로 받아야 한다는 단점을 가지고 있지만 본 연구에서는 한 사용자의 팔로위를 시드 사용자 없이 분류하여 리스트를 자동으로 생성하는 방법을 제안하였다.

3. 사용자 정보 및 활동에 기반한 사용자 리스트 생성 시스템

이 장에서는 본 연구에서 제안한 사용자 정보 및 활동을 이용한 트위터 사용자 리스트 생성 시스템에 대해 설명한다. 본 연구에서 제안한 시스템에서는 세 가지 종류의 군집을 찾아 각각에 이름을 붙여 리스트를 생성한다. 첫 번째는 사회적 군집으로 개인의 입장에서 의미있는 개인의 지인들로 구성된 것이다. 이들은 주로 친구 관계가 서로 긴밀하게 엮여있다는 특징이 있다. 우리는 이 특징을 기반으로 사용자의 팔로위 간에 친구 관계가 어떻게 구성 되어있는지를 파악하여 사회적 군집을 찾았다. 두 번째는 주제 군집으로 유사한 주제의 트윗 메시지를 작성하는 사람들을 찾는 것으로 팔로위들이 작성하는 트윗 메시지들을 분석하여 유사한 사용자들을 군집화하였다. 마지막 속성 군집은 속성이 비슷한 사용자들을 찾는 것이다. 이들은 다른 사람들이 해당 사용자들을 바라보는 관점이 비슷하다는 특징이 있다. 각 사용자들이 등록된 리스트의 이름은 다른 사용자들이 그 사용자에게 붙여주는 태그와 같은 역할을 하는 정보이므로 이를 활용하여 속성 군집을 찾도록 하였다.

시스템은 크게 세 단계로 구성된다. 우선 사용자의 팔로위들의 정보 및 활동을 수집한다. 두 번째 단계에서는 사용자의 팔로위 중에서 사회적 군집, 주제 군집, 속성 군집을 찾는다. 마지막으로 찾아낸 군집의 이름을 추천하여 리스트를 생성한다.

3.1 사용자 정보 및 활동 수집

사용자의 팔로위들을 리스트로 만들기 위해서 팔로위들의 정보 및 활동을 분석해야 한다. 사용자 정보 및 활동 수집은 트위터 API를 이용할 수 있는 Java 라이브러리인 Twitter4J[3]를 이용하였다.

3.1.1 사용자 활동

사용자 리스트를 자동으로 생성하기 위하여 다음과 같은 사용자 활동 데이터를 수집한다. 먼저, 사용자의 팔로위들 간에 친구 관계가 어떻게 구성되어 있는지를 수집하여 이 정보를 바탕으로 사용자의 실제 지인들로 구성되어 있는 사회적 군집을 찾는다. 이를 위해, 사용

3) http://twitter4j.org

자가 팔로잉(Following)하고 있는 팔로위의 아이디를 가지고 팔로위들 간에 친구 관계가 어떻게 구성되어 있는지 파악한다.

사용자들의 트윗 내용을 기반으로 리스트를 자동으로 생성하기 위해서는 사용자들이 남긴 최근 150개의 트윗 메시지를 수집한다. 트윗 메시지의 텍스트 간 유사도를 분석하여 유사한 주제에 대해서 이야기하는 팔로위들을 주제 군집으로 추출해낸다.

3.1.2 사용자 정보

트위터에서 제공하는 리스트 메뉴에서는 자신이 생성한 리스트 외에도 자신이 구성원으로 속해 있는 다른 사용자가 만들어 놓은 리스트 정보도 확인 할 수 있다. 개인이 등록된 리스트의 이름은 개인에게 붙이는 태그와 같은 역할을 한다고 볼 수 있다. 때문에 이를 이용하면 사용자의 팔로위들이 등록된 리스트 이름을 수집하여 친구 관계가 긴밀히 엮여있지도 않고, 작성하는 트윗 메시지의 주제가 비슷하지도 않을 수 있는 속성 군집을 찾아낼 수 있다. 또한 추출된 각 군집의 이름을 추천하는데도 사용된다.

사용자들은 보통 프로필에 사용자가 소속된 집단이나 사용자의 관심사, 간단한 인사말을 등록해놓는다. 사용자가 직접 등록한 이러한 프로필 정보도 각 군집의 이름을 추천하는데 사용한다.

3.2 사용자 군집 찾기

이 단계에서는 앞에서 수집한 사용자의 팔로위들의 정보 및 활동을 이용하여 사회적 군집, 주제 군집, 속성 군집을 찾는다. 세 가지 군집이 가지는 특성을 고려하여 각 군집을 찾아내기 위한 방법을 제안하였다. 그리고 이러한 방법으로 찾아낸 군집의 사용자들을 이용하여 리스트를 생성한다.

3.2.1 사회적 군집 찾기

사회적 군집은 사용자의 팔로위들 간에 어떤 관계가 있는지를 바탕으로 사용자 군집을 찾는다. 트위터에서 친구 관계인 팔로우는 단방향의 관계이기 때문에 트위터에서 사용자 간의 친구 관계는 여러 형태가 존재한다. 서로 팔로우 하고 있는 양방향의 관계도 있을 수 있고 한 사람만 팔로우 하고 있는 일방적인 단방향 관계가 있을 수도 있고 아무 관계가 없을 수도 있다.

이런 사용자 간의 다양한 친구 관계를 그래프로 표현할 수 있고[12], 이 그래프를 인접 행렬 형태로 표현하여 사용자 친구 관계 친밀도 행렬을 만들 수 있다. 본 연구에서는 행렬의 각 요소를 양방향의 친구 관계인 경우 1, 친구 관계가 없을 경우 0으로 표현하여 나타내었다. 단방향의 친구 관계는 실험을 통하여 가장 좋은 결과가 나오는 0.4로 나타내었다. 이 때 각 행은 사용자 벡터를 의미하며, 각 사용자 벡터는 그 사용자가 다른 사용자들과 얼마나 친한가를 나타내준다.

$$u_i^f = (f_{i,0}, f_{i,1}, \cdots, f_{i,n})$$

u_i^f는 사용자 i와 다른 사용자 간의 친구 관계 긴밀도를 나타내는 벡터로 인접행렬에서 i번째 행에 해당한다. $f_{i,j}$은 사용자 i와 사용자 j가 어느 정도로 강한 친구 관계를 갖고 있는지 나타내 주는 요소이다.

사회적 군집화의 성능 개선을 위하여 친구 관계 긴밀도 행렬을 바탕으로 군집화를 하는 것이 아니라 추가적인 단계를 거친다. 이 단계에서는 각 사용자의 친구 관계 친밀도 벡터에 대해 코사인 유사도(Cosine Similarity)를 계산하여 그 값을 요소로 하는 새로운 행렬을 만든다. 이 행렬의 각 사용자 벡터는 한 사용자가 다른 사용자와 친구 관계가 얼마나 유사한지를 나타내게 된다. 이 벡터를 사용자 친구 관계 유사도 벡터라 하겠다. 여기에서 $s_{i,j}$는 사용자 i와 사용자 j 간의 친구 관계 유사도이고, u_i^s는 사용자 i와 다른 사용자들과의 친구 관계 유사도를 나타낸 벡터이다.

$$s_{i,j} = similarity(u_i^f, u_j^f) = \cos(\theta) = \frac{u_i^f \cdot u_j^f}{\| u_i^f \| \, \| u_j^f \|}$$

$$u_i^s = (s_{i,0}, s_{i,1}, \cdots, s_{i,n})$$

사용자 친구 관계 유사도 행렬의 각 사용자 벡터 중 크기가 큰 요소들은 일종의 사회적 군집의 후보로 생각할 수 있다. 이 후보 군집들에 계층적 군집화 알고리즘[13]을 적용하여 비슷한 후보 군집들을 하나로 합쳐가면서 사회적 군집들을 생성하게 된다. 이 때 몇 개의 군집이 존재하는지를 미리 알기 어려우므로 본 연구에서는 사용자 군집을 생성하는데 상향식 계층적 군집화 방식을 이용하였다. 코사인 유사도를 이용하여 각 군집 간의 거리를 측정하여 가장 가까운 개체들부터 합쳐나갔으며 각 개체 쌍의 거리의 평균을 연결 기준으로 삼았다. 그리고 군집화하는 과정 중에 모든 군집 간의 거리가 기준 이상으로 멀어지면 군집화를 중단하도록 하였다.

3.2.2 주제 군집 찾기

주제 군집은 사용자의 팔로위들 중에 비슷한 주제에 대해서 트윗 메시지를 작성하는 사용자들을 모아 군집화 한 것이다. 이때 앞서 수집한 각 사용자들의 트윗 메시지에 대한 전처리 과정이 필요하다. 대부분의 사용자들은 트위터를 통해 서로의 소식을 주고 받거나, 친교를 나누기도 하고 특정 주제에 대해 이야기를 나누고, 공통 관심사에 관해 이야기하는 등 트위터를 여러 가지 목적으로 이용한다. 주제 군집에서는 사용자가 나타내고자 하는 주제가 담겨있는 트윗 메시지만 수집하기 위하여 사용자의 대표 주제를 나타내지 못하는 부분은 제외하였다. 우선 다른 사용자의 이름 앞에 '@'을 붙인 멘션

(Mention)이 있는 메시지는 대부분 개인적인 대화이기 때문에 주제를 담고 있지 않은 경우가 많아 제외하였다. 또 트윗 메시지에 포함된 URL이나 여러 특수 문자들은 어떤 주제를 나타내지 못하므로 노이즈로 생각하여 제외시켰다. 본 연구에서는 고려하지 않았지만 트윗 메시지에 포함된 URL이 가리키는 문서는 결국 사용자가 자신의 팔로워들에게 전달하고 싶었던 정보성이 있는 주제를 포함하고 있을 것이므로 해당 문서에 있는 정보들도 포함하면 사용자의 관심 주제를 파악하는데 더욱 도움이 될 것이다. 다른 사람이 작성한 트윗 메시지를 자신의 팔로위들에게 전파 시킬 수 있도록 재전송하는 기능인 리트윗(Retweet)으로 작성된 메시지는 본인이 작성한 메시지가 아니더라도 사용자의 관심 주제와 연관이 있을 가능성이 높아 제외시키지 않았다. 이렇게 처리한 각 사용자의 문서에 대해 한국어의 형태소를 분석해주는 한나눔 형태소 분석기[4]를 적용시켜 명사만 추출하였다.

또한, 트윗 메시지는 140자의 제한을 가지고 있기 때문에, 각 트윗 메시지는 길이가 짧아 그 자체가 어떤 주제를 담고 있다고 보기가 힘들다. 그래서 각 사용자의 최근 트윗 메시지 150개를 모두 모아 하나의 문서로 간주하였다. 즉 한 사용자가 하나의 문서처럼 간주될 수 있다.

각 사용자가 쓴 트윗 메시지에서 명사를 추출하여 사용자-단어 벡터를 만든다. 이 때 각 사용자가 어떤 단어를 몇 번 사용하였는지를 바탕으로 구한 TF-IDF (Term Frequency-Inverse Document Frequency)[14] 값으로 벡터를 구성한다. TF-IDF는 정보검색이나 텍스트 마이닝 분야에서 많이 사용되는 기법이다. 이러한 TF-IDF 가중치는 한 문서 안에서 특정 단어의 중요도를 측정하는데 사용된다. TF는 어떤 단어가 한 문서 안에서 얼마나 자주 나타나는지를 의미하는 값으로, 이 값이 높으면 그 단어는 해당 문서 안에서 중요한 위치를 갖는다는 것을 나타낸다. DF는 전체 문서군 내에서 어떤 단어가 얼마나 자주 등장하는지를 나타내는데, 이 값이 높으면 그 단어는 여러 문서에 흔하게 사용되는 단어이므로 한 문서 안에서 갖는 가치가 떨어지게 된다. 이 값의 역수인 IDF를 TF에 곱하여 여러 문서 집합이 있을 때 어떤 단어가 한 문서에 얼마나 중요한지를 나타내준다. TF-IDF 값은 다음과 같이 계산한다.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$$tf_{i,j} = Number\ of\ occurences\ of\ word\ i \in document\ j$$

$$df_i = Number\ of\ documents\ containing\ word\ i$$

$$N = Total\ number\ of\ documents$$

이렇게 구성한 TF-IDF 벡터는 크기는 크지만 밀도가 낮아 효율이 떨어진다. 이 벡터에 LSA 알고리즘[15]을 적용하여 단어-문서 벡터를 보다 작은 크기인 k 차원으로 근사시킨다. 이렇게 얻은 의미 관계가 더 명확한 작은 벡터들을 이용하면 유사도 연산 수행 시간이 줄어들고 노이즈를 제거하는 효과도 얻을 수 있다. SVD로 얻어진 대각 행렬과 우측 특이 벡터의 곱으로 각 문서, 즉 각 사용자를 표현할 수 있다. 이렇게 각 사용자들을 k차원으로 표현한 행렬에 계층적 군집화 알고리즘을 적용하여 작성하는 트윗 메시지의 주제가 비슷한 사용자들을 모아 군집을 만든다.

3.2.3 속성 군집 찾기

속성 군집은 다른 사람들이 보았을 때 비슷한 속성을 가지는 사용자들을 묶어놓은 군집을 말한다. 속성 군집의 예로는 연예인, 정치인 등이 있다. 속성 군집의 구성원들은 긴밀한 친구 관계를 맺지 않을 수도 있고 비슷한 주제에 대해서 이야기하지 않을 수도 있다. 이런 속성 군집을 찾기 위해서, 다른 사용자들이 어떤 사용자에게 달아놓은 태그로 볼 수 있는 사용자가 등록된 리스트의 이름을 활용하기로 한다. 등록된 리스트의 이름에 속한 단어들이 유사한 사용자들이 한 군집으로 묶이게 된다.

속성 군집을 찾아내기 위해서 사용자의 전체 팔로위들 중에 속성 군집에 속할 가능성이 높은 사용자들을 두 가지 조건으로 추려낸다. 속성 군집에 속하게 되는 사용자들은 주로 일반적인 사용자가 아닌 유명인이나 회사의 공식 트위터 계정이다. 이런 특수한 트위터 계정은 보통 자신의 계정이 팔로우하는 팔로위보다 해당 계정을 팔로우 하는 팔로워들이 훨씬 많다는 특징이 있다. 그래서 속성 군집화를 위해 전체 팔로위 중에서 속성 군집의 대상자들을 찾는 첫 번째 조건으로 팔로워 수가 팔로위 수보다 2배 이상 높은 사용자들을 모았다. 또 홍보나 마케팅 혹은 특수한 목적으로 트위터 계정을 이용하는 경우에 그 계정을 팔로우 해주는 사람들을 팔로우하는 경우도 있다. 이 경우에는 팔로위의 수와 팔로워의 수가 비슷한 수준이 되지만, 팔로위, 팔로워 수가 모두 일반적인 사용자들보다 훨씬 많다. 트위터에서는 비정상적으로 팔로우를 많이 하는 행위를 규제하기 위해 2000명 이상을 팔로우하기 위해서는 팔로워가 팔로위 수의 90% 이상이 되어야 하는 제한을 두었다. 그래서 일반적이지 않은 팔로위 수의 기준을 2000명으로 보고, 팔로위의 수가 2000명 이상인 사용자들의 계정을 대상으로 하였다.

이렇게 모아진 속성 군집 대상자가 추려지면, 각 사용자가 등록된 리스트의 이름을 모아 그 사용자에 대한

4) http://sourceforge.net/projects/hannanum/

문서로 간주하였다. 사용자 리스트의 이름은 문자, 숫자, 하이픈(-)으로만 이루어져 있다. 이 규칙은 트위터에서 공식적으로 인정하는 리스트 이름에 대한 규칙이다. 실제로는 이 외에도 기타 특수 문자들도 입력하여 사용할 수 있지만, 트위터 API에서는 공식적인 규칙에 따른 리스트 이름을 제공해준다. 보통 사용자들은 단어와 단어 사이를 구분할 때 하이픈을 이용한다. 때문에 리스트 이름의 형식이 '단어1-단어2'의 형태일 경우 단어1과 단어2가 합쳐져서 의미가 있는 경우도 있지만 단어1, 단어2가 각각 유의미한 단어인 경우도 있다. 그래서 리스트의 이름이 하이픈을 포함한 경우에 '단어1-단어2' 뿐만 아니라 '단어1', '단어2'의 형태 모두 유의미한 단어로 정하였다.

속성 군집화에서도 한 사용자가 어떤 단어로 리스트에 등록되어 있는지의 관계가 나타나고 주제 군집화에서와 마찬가지로 한 사용자를 하나의 문서로 간주할 수 있다. 앞서 주제 군집화에서는 단어-사용자 행렬을 단어의 출현 빈도수를 기반으로 구성하였다면, 속성 군집화에서는 단어의 가중치를 기반으로 구성하도록 한다. 주제 군집화에서는 한 사용자가 작성한 최근 150개의 트윗 메시지를 한 문서로 간주하였기 때문에 각 사용자 문서의 길이 차이가 크지 않다. 하지만 속성 군집화에서는 한 사용자가 등록된 리스트의 이름을 문서로 사용하게 되는데, 앞서 언급했듯이 사용자마다 등록된 리스트의 단어 개수가 편차가 심하기 때문에 이를 보완하기 위해 한 사용자가 등록된 전체 단어 수를 고려하여 정규화 시킨 가중치를 기반으로 단어-사용자 행렬을 구성하였다.

이 행렬에 전체 문서 집합을 고려하여 단어의 중요도를 측정할 수 있도록 TF-IDF를 적용한다. 이후의 과정은 주제 군집화와 동일한데, TF-IDF 행렬에 LSA 알고리즘을 적용하여 각 사용자를 k차원으로 표현한다. 계층적 군집화 알고리즘을 적용하여 속성이 비슷한 사용자들을 군집화 하였다.

3.3 사용자 군집 이름 추천

앞의 단계에서 사회적 군집, 주제 군집, 속성 군집을 찾아내었다면, 이 절에서는 찾아낸 각 군집의 이름을 추천한다. 군집 구성원들이 등록되어있는 기존의 다른 리스트의 이름과 각 구성원이 직접 등록한 개인의 프로필을 바탕으로 각 군집을 대표할 수 있는 이름을 찾는다.

각 사용자가 속해 있는 리스트 이름 단어들은 기존에 이미 리스트의 이름으로 쓰이고 있는 단어들이므로 사용자 군집 이름을 정하는데 중요한 정보이다. 개인의 프로필에 사용자들은 간단한 인사말을 등록해 놓는 경우도 있지만, 스스로를 소개하는 단어들을 작성하기도 한다. 전자의 경우 군집의 이름을 정하는데 노이즈로 작용할 수 있지만 후자의 경우 스스로가 자신에 대해 설명해 놓은 글이므로 매우 신뢰성 높은 정보이다. 그러므로 이 두 정보를 적절한 비율로 합해 군집 이름 추천에 사용하였다.

이 때 각 사용자가 갖고 있는 리스트 이름 단어 수의 차이가 굉장히 클 수 있다. 이를 정규화 하지 않으면 군집의 이름을 정하는 과정에서, 군집 내에서 리스트로 등록된 적이 많은 구성원이 가진 단어 위주로 치우쳐서 군집의 이름이 정해지게 된다. 그래서 프로필 단어와 리스트 이름 단어를 각각 정규화하여 Normalized term frequency[16] 가중치 값을 구하고 각 사용자가 가진 단어들의 가중치의 합이 동일하도록 해준다. 군집 내 각 구성원의 관점에서 각 단어의 가중치를 구하고, 각 단어에 대해서 구성원들의 가중치를 모두 합한다. 이를 수식화 하면 아래와 같다.

$$weight(w_i) = \sum_{u_j \in group_j} \left\{ \begin{array}{l} \alpha \times \dfrac{frequency\,of\,w_i\ in\ u_j's\,profile}{number\,of\,words\ in\ u_j's\,profile} \\ +(1-\alpha) \times \dfrac{frequency\,of\,w_i\ in\ u_j's\,list\,names}{number\,of\,words\ in\ u_j's\,list\,names} \end{array} \right\}$$

$(0 \le \alpha \le 1)$

위 식에서 u_j는 그룹 내의 각 사용자를 의미하고, w_i는 그룹 내 사용자들이 프로필에 등록한 단어들을 의미한다. 이러한 과정을 거쳐 군집 내 단어들 중 가장 높은 가중치를 갖는 단어부터 추천해준다.

4. 성능 평가

4.1 실험 환경 및 데이터

본 연구에서 제안한 방법을 실험하기 위해 사용된 시스템의 사양은 다음과 같다. CPU는 Intel(R) Core(TM) i7 870 2.93Ghz, RAM은 8GB, 운영체제는 Windows 7 Enterprise K 64bit인 시스템을 사용하였다. 트위터의 리스트를 자동으로 생성하기 위한 시스템은 Java를 이용하여 구현하였다. 데이터베이스는 MySQL을 사용하여 트위터 API를 이용해 가져온 데이터를 저장하고 리스트 자동 생성 시스템에서 MySQL에 접근하여 데이터를 사용하였다.

팔로위가 500명 이하인 사용자들을 일반적인 사용자들로 간주하고 해당 사용자들을 기준으로 하여 데이터를 수집하였다. 총 40명의 사용자들을 대상으로 그들의 10,136명의 팔로위 정보를 수집하였다. 이 40명이 팔로잉 하고 있는 팔로위의 수는 고르게 분포되어 있다. 팔로위들 간의 친구 관계 정보는 총 13,206,247개이고, 사용한 팔로위들의 트윗 메시지 수는 총 144,593개이다.

4.2 성능 평가 기준

4.2.1 군집 구성원의 적합성

본 연구에서 제안한 방법으로 하나의 군집이 만들어졌을 때, 해당 군집의 구성원이 얼마나 적합한지를 측정

하였다. 이를 위해, 기존에 존재하는 리스트를 정답으로 간주하고 제안한 군집의 구성원들과 비교하여 구성원들이 얼마나 정확한지를 측정하였다. 자카드 계수(Jaccard Coefficient)는 두 집합 사이에 겹치는 정도를 나타내는 수치인데, 기존에 있던 한 리스트와 제안한 군집의 구성원을 비교하였을 때 자카드 계수가 0.5 이상이면 두 군집이 일치한다고 판단하여 구성원들의 정확도(Precision), 재현율(Recall), F값(F-measure)을 측정하였다.

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

이때 각 군집화 방식을 기본선의 성능과 비교해보았다. 본 연구에서는 사용자 간의 양방향, 단방향 관계를 모두 고려하여 군집을 찾았는데, 사회적 군집의 기본선은 [5]에서처럼 양방향의 관계만 고려하도록 했다. 주제 군집과 속성 군집에서는 TF-IDF 벡터의 차원을 줄여 계산 속도를 개선하고 노이즈를 제거하기 위해 LSA 알고리즘을 적용하였다. LSA 알고리즘을 적용하지 않고 TF-IDF 벡터로 군집화 한 것을 주제 군집과 속성 군집의 기본선으로 삼았다.

4.2.2 군집들의 커버리지

본 논문에서는 한 사용자의 팔로위들에게서 나올 수 있는 다양한 군집을 제안하였다. 이렇게 제안한 여러 군집들은 사용자가 원하는 다양한 군집들을 포함할 수 있어야 한다. 이를 입증하기 위해 사용자가 원하는 리스트들과 본 연구에서 제안한 여러 군집들을 비교하여야 한다. 하지만 현재 트위터 실제 사용자 중에서 리스트를 사용하는 사용자도 일부일뿐더러, 리스트를 사용하더라도 대부분의 사용자들은 팔로위 전체를 고려하지 않은 소수의 리스트만 가지고 있어 비교하기가 어렵다. 그래서 트위터를 사용하는 컴퓨터 공학부 대학원생 10명을 대상으로 개인의 팔로위 전체를 고려하여 유의미하다고 여겨지는 리스트들을 최대한으로 만들어보도록 하였다. 특별한 제한 없이 원래 리스트를 만들 수 있는 조건과 동일하게, 한 사람이 여러 리스트에 등록 될 수도 있고, 어떤 사람은 아무 리스트에 등록되지 않을 수도 있다.

이렇게 제안한 리스트와 제안한 군집을 비교하여 각 군집화 방식별로 정확도, 재현율을 측정하였다. 이 실험에서는 정확도보다는 재현율이 더욱 중요하다. 재현율이 높다는 것은 사용자가 원하는 다양한 리스트들을 포함할 수 있다는 의미이기 때문이다.

4.2.3 군집 이름의 적합성

기존에 존재하는 리스트의 이름을 정답으로 간주하여 정답 리스트와 제안한 군집의 구성원들이 매칭 되는 경우에 리스트의 이름과 제안한 군집의 이름을 비교해 보았다. 리스트 이름이 군집 이름에 포함되거나 군집 이름이 리스트 이름에 포함 되는 경우에 두 이름이 일치한다고 판단하였다. 단어들의 자연어 처리에 있어서 영어-국문 단어(예, 'Musical'과 '뮤지컬'), 동의어(예, 'Cinema', 'Film', 'Movie'), 줄임말(예, 'Celebrities'와 'Celebs')의 매칭과 관련한 문제가 있는데, 이 같은 경우에 수동적으로 찾아서 매칭 되는 경우로 처리하였다.

군집 이름을 정하기 위해서 프로필 단어와 리스트 이름 단어를 이용한다. 그래서 프로필 단어와 리스트 이름 단어를 적절하게 섞어서 사용해야한다. 프로필 단어와 리스트 이름 단어의 비율이 어느 정도일 때가 가장 좋은 성능을 보이는지 α 값을 변경시켜 가며 S@k(Success at rank k)를 측정해보았다. S@k는 순위 1위에서 k위까지의 결과 중에 정답이 포함되어있을 확률을 의미한다. 본 실험에서는 S@1부터 S@5까지 측정하였다. 또 가장 좋은 결과가 나오는 α를 찾아서, 각 종류의 군집화 방법별로 S@1부터 S@5를 측정하여 비교하였다.

4.3 실험 결과

4.3.1 군집 구성원의 적합성

그림 3은 기존에 존재하는 리스트를 정답이라고 생각하고 제안한 군집 중에서 그 리스트와 같다고 여겨지는 군집과 비교한 결과이다. 기존 리스트의 구성원과 비교했을 때 제안한 군집의 구성원들이 얼마나 적절하게 들어갔는지를 알아보기 위한 것으로 성능을 뒷받침하기 위해서 기본선과도 비교하였다. 그림에서 보이는 바와 같이 사회적 군집화 방식에서 트위터에 존재하는 다양한 관계를 반영한 것이 기본선에 비해 정확도, 재현율, F값 모두 더 좋은 성능을 보였다. 주제 군집화, 속성 군집화에서도 LSA 알고리즘을 적용한 방식이 기본선보다 성능이 더 좋은 것을 확인할 수 있다. 또 행렬의 크기가 줄어들므로 LSA 알고리즘을 적용한 방식이 수행시간이 훨씬 적게 걸렸다.

주제, 속성 군집은 텍스트 형태에 내포된 정보를 바탕으로 군집화된 반면, 사회적 군집의 경우는 친구 관계 정보가 명시적이기 때문에 더 나은 성능을 보였다. 주제

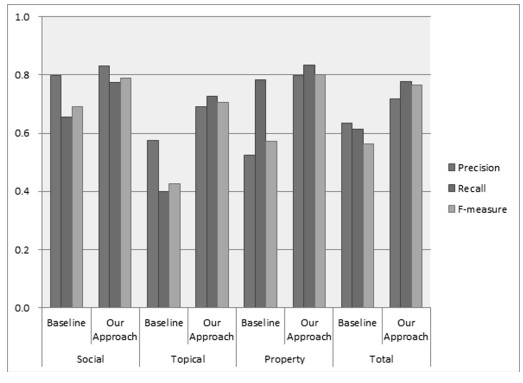


그림 3 군집 구성원의 적합성

군집은 다른 군집화 방식에 비해 낮은 성능을 보인다. 뉴스 미디어나 회사 등 목적성을 가진 트위터 계정에서는 주로 한 가지 주제에 대한 트윗 메시지를 작성한다. 하지만 대부분의 일반 사용자들의 경우 특정 주제의 이야기를 하는 것이 아니라 신변잡기적인 메시지를 작성하기 때문에 이런 트윗 메시지들이 주제 군집을 찾는데 노이즈로 작용한 것으로 보인다.

4.3.2 군집들의 커버리지

표 1은 사용자에게 제안해 준 여러 개의 그룹이 실제로 사용자가 원하는 다양한 그룹을 포함하고 있는가에 대한 실험 결과이다. 이 실험에서는 정확도보다 재현율이 높은 것이 더 중요하다. 10명의 사용자들이 팔로위의 일부가 아닌 전체를 고려하여 만든 여러 개의 리스트와 비교한 것이기 때문에, 사용자들이 만든 리스트들을 최대한 많이 포함하는 것이 좋다.

표 1 제안한 군집들의 커버리지

	정확도	재현율
사회적 군집	0.703	0.854
주제 군집	0.325	0.750
속성 군집	0.556	0.676

사회적 군집화의 경우 정확도, 재현율 모두 가장 높다. 앞서 언급했듯이 사회적 군집화에 쓰이는 정보들이 명확한 정보들이기 때문에 다른 군집화 방식에 비해 좋은 성능을 보인다. 속성 군집의 정확도 및 재현율이 비교적 나쁘지 않지만, 주제 군집화에서는 정확도가 유독 떨어지는 것을 볼 수 있다. 뉴스 미디어나 특정 회사의 계정 등 목적성을 가지고 있어 특정한 주제에 대해 주로 이야기하는 계정들이 있는가 하면, 일상적인 이야기를 나누는 일반적인 사용자들도 있다. 시스템에서는 단어들이 함께 쓰인 관계를 바탕으로 군집을 찾기 때문에 두 형태의 주제 군집을 모두 제안하게 된다. 하지만 일반적으로 사용자들이 유용하다고 여기는 주제 리스트는 전자에 해당하는 경우가 절대 다수이다. 그래서 제안한 군집들이 사용자가 유의미하다고 생각하여 만들어 놓은 주제 리스트들을 포함하지만, 그 외의 다른 군집들도 제안하기 때문에 정확도는 떨어지지만 재현율은 높다. 하지만 후자 경우의 주제 군집도 사건 탐지적(Event Detection) 측면에서 접근하면 유용하게 쓰일 수 있다. 예를 들어 스포츠 경기나 선거 등 일반적인 경우가 아닌 특수한 일이 일어나는 때에는 특별한 목적성을 가지고 있지 않은 사용자들도 비슷한 주제의 이야기를 공통적으로 하게 될 것이다.

4.3.3 군집 이름의 적합성

그림 4는 군집 이름을 정할 때, 프로필 단어와 리스트 이름 단어의 비율을 나타내는 α 값을 변화시켜 가며

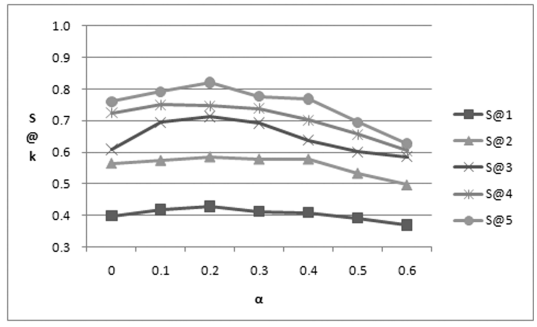


그림 4 α 값 변화에 따른 군집 이름 적합성

S@1부터 S@5를 측정한 결과이다. 이 두 종류의 정보를 어떤 비율로 사용해야 좋은 성능을 보일지 실험한 결과 α가 2일 때, 즉 프로필 단어와 리스트 이름 단어를 2:8로 사용할 때 S@5가 가장 높게 나왔다. 이 경우 제안한 군집의 이름 82%가 실제로 비슷한 리스트의 이름과 매칭 되었다.

사용자들이 프로필에 간단한 인사말 등을 등록해 놓는 경우도 있지만, 자신의 소속기관이나 특성에 대해 작성한 경우에는 사용자가 스스로를 설명한 것이기 때문에 유의미한 정보이다. 리스트 이름 단어들은 이미 기존에 다른 사용자들이 리스트 이름으로 등록해 놓은 단어이기 때문에 더욱 신뢰도가 높은 정보이다. 그러므로 프로필 단어보다는 리스트 이름 단어를 더 높은 비율로 이용하는 경우에 성능이 좋게 나온 것으로 보인다. 프로필 단어를 리스트 이름 단어보다 많이 쓰기 시작하면 성능이 급격하게 나빠져 5개의 이름을 제안했을 때 정답이 포함되어 있을 확률이 62.7%로 떨어진다.

그림 5는 $\alpha = 0.2$ 일 때, 각 군집화 방식 별로 군집 이름 적합성을 비교한 것이다. 대체적으로 상위 5개의 결과를 제안해 주었을 때 80% 전후의 이름이 정답과 일치한다. 특이한 점은 사회적 집단의 경우 S@1이 다른 군집화 방식보다 유독 높다는 것이다. 사람들은 보통 어떤 단체나 집단에 속하게 되면서 다른 사람들을 만나

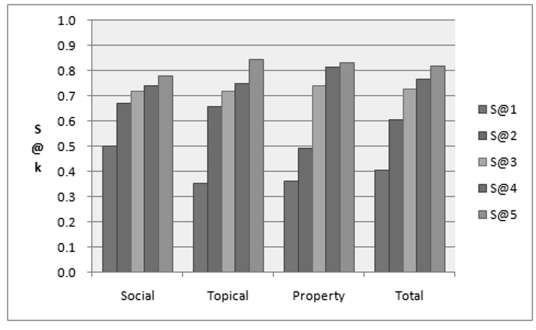


그림 5 각 군집화 방식별 군집 이름 적합성

게 된다. 그래서 대부분 개인이 속한 집단이나 단체 등을 기준으로 하여 묶여서 사회적 군집이 형성된다. 대개 어떤 집단이나 단체가 있으면 그것을 나타내는 공식적인 명칭이 존재하기 마련이다. 그렇기 때문에 사회적 군집의 이름 같은 경우는 사람들이 공통적으로 이용하는 경우가 훨씬 많기 때문에 주제 군집이나 속성 군집에 비해 S@1 값이 높게 나온다.

5. 결론 및 향후 연구

본 논문에서는 트위터 사용자가 팔로윙하고 있는 팔로위들을 자동으로 군집화하여 리스트를 생성하고, 그 리스트의 이름을 부여하는 방법을 제안하였다. 트위터의 리스트는 사회적 군집, 주제 군집, 속성 군집과 같이 크게 세 가지로 구분할 수 있기 때문에 각 군집에 맞는 리스트 자동 생성하였다. 사회적 군집을 찾아내기 위해서는 사용자 사이의 친구 관계를 이용하고, 주제 군집을 찾아내기 위해서는 트윗 메시지를 이용하였다. 마지막으로 속성 군집을 찾아내기 위해서는 해당 사용자가 등록된 리스트의 이름을 사용하였다.

본 논문에서 제안한 리스트 자동 생성 방법론은 트위터 사용자의 정보를 기반으로 계층적 군집화, LSA 등의 기법을 활용하여 리스트를 생성할 수 있으므로 트위터 시스템에 그대로 적용이 가능하다. 리스트 자동 생성 기능 적용을 통해, 사용자가 노력해서 직접 리스트를 만드는 수고를 줄일 수 있을 것으로 기대된다.

향후 연구로는, 사용자 네트워크 확장을 이용한 리스트 자동 생성 방법과 주제 군집 방법론의 성능 향상이 가능할 것으로 보인다. 본 논문에서 제안한 리스트 자동 생성 알고리즘은 현재 사용자가 팔로우하고 있는 팔로위들만을 대상으로 리스트를 생성하는 방법론이다. 여기에 추가적으로 사용자 네트워크 확장을 통해서 현재 사용자가 직접 팔로우하지 않는 사용자도 대상으로 한 리스트 자동 생성 알고리즘이 가능할 것이다. 현재 사용자의 친구 관계를 활용하여 친구의 친구와 같은 팔로우 네트워크 상의 사용자에 대해 랭킹을 매겨서 해당 리스트의 구성원으로 적합하다면 리스트에 포함시키는 방법을 사용할 수 있다. 이러한 방법은 현재 트위터에서 제공하고 있는 친구 추천 기능을 포함하는 광범위한 리스트 자동 생성 방법론이 될 것이다.

참 고 문 헌

[1] Yuto Yamaguchi, Toshiyuki Amagasa, Hiroyuki Kitagawa, "Tag-based User Topic Discovery using Twitter Lists," in *International Conference on Advances in Social Network Analysis and Mining*, 2011.

[2] Haewoon Kwak, Changhyun Lee, Hosung Park, Sue Moon, "What is Twitter, a Social Network or a News Media?," in *19th International Conference on World Wide Web*, 2010.

[3] Oeyman Nasirifard, Conor Hayes, "Tadvise: A Twitter Assistant based on Twitter Lists," in *Social Informatics*, 2011.

[4] Derek Greene, Fergal Ried, Gavin Sheridan, Pá draig Cunningham, "Supporting the Curation of Twitter User Lists," in *The Computing Research Repository*, 2011.

[5] Zhonghua Qu, Yang Liu, "Interactive Group Suggesting for Twitter," in *The 49th Annual Meeting of the Association for Computational Linguistics*, 2011.

[6] Michael Steve Stanley Laine, Gunes Ercal, Bo Luo, "User Groups in Social Networks: An Experimental Study on YouTube," in *44th Hawaii International Conference on Systems Science*, 2011.

[7] Marco Pennacchiotti, Ana-Maria Popescu, "A Machine Learning Approach to Twitter User Classification," in *5th International Conference on Weblogs and Social Media*, 2011.

[8] Mihaela Cocea, George D. Magoulas, "User Behaviour-driven Group Formation through Case-based Reasoning and Clustering," in *Expert Systems with Applications*, vol.39, 2012.

[9] Lipika Dey, Bhakti Gaonkar, "Wavelet-Based Clustering of Social-Network Users using Temporal and Activity Profiles," in *Pattern Recognition and Machine Intelligence*, 2011.

[10] Taehyun Kim, Gonzalo Huerta Canepa, Jongheon Park, Soon J. Hyun, Dongman Lee, "What's Happening: Finding Spontaneous User Clusters Nearby using Twitter," in *3rd International Conference on Social Computing*, 2011.

[11] Ivan Cantador, Pablo Castells, "Extracting Multilayered Communities of Interest from Semantic User Profiles: Application to Group Modeling and Hybrid Recommendations," in *Computers in Human Behavior*, vol.27, 2011.

[12] Barry Wellman and Stephen Berkowits, Social Structures: A Network Approach, Cambridge University Press, 1988.

[13] Rui Xu, Donald Wunsch II, "Survey of Clustering Algorithms," in *IEEE Transactions on Neural Networks*, vol.16, 2005.

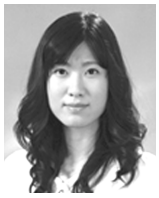
[14] Karen Sparck Jones, "A Statistical Interpretation of Term Specificity and Its Application in Retrieval," in *Journal of Documentation*, 1972.

[15] Thomas K Landauer, Peter W. Foltz, Darrell Laham, "An Introduction to Latent Semantic Analysis," in *Discourse Processes*, vol.25, Issue.2-3, 1998.

[16] Ricardo Baeza-Yates, Berthiere Ribeiro-Neto, Mordern Information Retreival, Addison Wesley, 1999.

김 소 민

2010년 서울대학교 컴퓨터공학부 학사
2012년 서울대학교 컴퓨터공학부 석사
2012년~현재 삼성전자 재직 중. 관심분야는 데이터베이스, 소셜 네트워크 분석



임 혜 원

2008년 숙명여자대학교 컴퓨터과학과 학사. 2008년~현재 서울대학교 컴퓨터공학부 석박사통합과정. 관심분야는 데이터베이스, 소셜 네트워크 분석

김 형 주

정보과학회논문지 : 데이타베이스
제 39 권 제 4 호 참조