트위터와 집단지성(Collective Intelligence)을 이용한 사용자 특성 분석 시스템

백성문, 강신욱, 이은석
성균관대학교 정보통신공학부

e-mail : {wasaby100, dfoot, leees}@skku.edu,

A user profiling system with CI(Collective Intelligence) on SNS(Twitter)

Sungmoon Baek, Shinwook Gahng, Eun seok Lee
Dept. of Electrical and Computer Engineering, Sungkyunkwan University

요 약

Web 2.0 이 도래한 이후 SNS(Social Network Service)에 대한 관심이 널리 퍼짐에 따라 인터넷 사용자들은 SNS 를 통하여 수 많은 정보를 교류하고 있다. SNS 에서는 사용자들을 중심으로 수많은 메시지가 생성되고 있으며, 그러한 메시지에는 사람들의 성향이 그대로 묻어 있다. 수많은 사람들이 만들어내는 메시지들은 매우 방대하며 의미 있고 실속 있는 다양한 개인 정보를 담고 있다. 본 논문에서는 트위터를 이용하여 특정 사용자 중심의 네트워크에서 생성되는 메시지들을 집단지성의 측면에서 수집, 분석하는 시스템을 개발하였다. 이 시스템은 사용자 주변에서 오가는 키워드들을 찾아내고, 그런 키워드를 생성하고 있는 사람들이 누구인지를 알아본다. 그 결과 한 사용자 주변에 분포되어 있는 집단들의 특성을 알아볼 수 있다. 특정 사용자 주변에는 어떠한 집단이 있는지 알 수 있고, 그 집단들의 연관성을 분석한다면 이는 마케팅, 서비스 차원의 사회 여러 분야에서 유용하게 쓰일 수 있을 것이다.

1. 서론

집단지성은 다수의 개체가 서로 협력하거나 경쟁을 통하여 얻게 된 지적 능력의 결과로 얻어진 집단적 능력을 일컫는 용어이다. 이러한 협업으로 인한 시너지 효과는 이전부터 기업, 정부 등에서 많이 사용되고 있다. Pierre Levy 는 사이버 공간의 집단지성을 제시하였는데, 그는 누구나 자신의 공간을 가지고 정보를 형성하는 시대가 오면 어디에나 분포하고, 지속적으로 가치가 부여되며, 실시간으로 조정되고, 역량의 실제적 동원에 이르는 집단지성이 발현될 것이라고 주장하였다[1]. 실제로 현재 사이버상에서 이러한 모습 즉, 지식 정보의 생산자나 소비자가 따로 없이 누구나 생산할 수 있고 모두가 손쉽게 공유하면서도 정체되지 않고 계속 진보하는, 집단지성의 특성을 볼 수 있다. 이러한 특성을 이용해 많은 연구가 이루어지고 있으며, 위키피디아, Web 2.0 등으로 그 결과가 나타나고 있다. 사람들은 쇼핑, 연구, 오락, 웹 사이트 구축을 할 때 항상 인터넷을 사용한다. 사용자에게 질문을 던져 서핑을 방해하지 않으면서도 정보를 이끌어낼 수 있다. 이런 정보를 가공하고 해석하는 방법은 다양하다.

최근에 SNS(Social Network Service) 에 대한 관심이 널리 알려짐에 따라 인터넷 사용자들은 SNS 를 통하여 수많은 정보를 교류하고 있다. SNS 에서 생성되는 수많은 사람들의 정보를 집단지성의 특성을 이용한다면 아주 유용한 정보로 가공할 수 있을 것이다. 다양한 사람들이 만들어내는 데이터를 수집하여 그 의미를 찾아보는 시도를 하기 위해 본 논문에서는 트위터[2]를 이용한다. 트위터는 최근 가장 많은 사용자들에게 이용되고 있는 SNS 이며 짧은 메시지를 남기는 간단한 방법으로 그 메시지를 사람들 사이의 네트워크를 통해 전 세계로 퍼뜨릴 수 있는 힘을 가지고 있다. 트위터에서는 한 사용자를 중심으로 수 많은 메시지들이 오가고 있으며, 그러한 메시지에는 사람들의 성향이 그대로 묻어 있다. 이러한 메시지들을 수집하여 분석한다면 그 사용자를 프로파일링 할 수 있고, 그 결과는 사회 여러 분야에서 유용하게 쓰일 수 있을 것이다.

본 논문에서는 트위터에서 사람들이 주고 받는 메시지를 수집하고 파싱한다. 그 후 메시지들 사이에서 생성되는 키워드를 추출해내고, 그 키워드들을 생성해낸 사람들을 묶어 군집화 한다. 특정 사용자를 중심으로 생성되어 있는 집단들의 특색은 곧 그 사용자의 성향을 나타내며 해당 사용자의 흥미, 관심사를 나타내는 지표가 될 것이다. 본 논문의 구성은 2 장에 관련연구를 설명하고 3 장에서 제안하는 시스템의 설계와 구현에 대해서 설명한다. 4 장에서는 특정 사용

자를 임의로 선정해 테스트를 하고 그 결과를 보여준다. 마지막으로 5 장에서 결론 및 향후 연구에 대해 논한다.

2. 관련 연구

2.1 집단지성

집단지성의 효과를 웹에 접목시켜 서비스화에 성공한 사례로는 위키피디아와 구글 검색엔진 등이 있다. 위키피디아는 완전히 사용자 공헌으로만 만드는 온라인 백과사전이다. 누구나 어떤 페이지도 만들고 수정할 수 있다. 각 표제어가 많은 무리의 사람들에 의해 유지되고 몇몇 조직화한 그룹이 전에 만든 어떤 것보다 큰 백과사전을 만들고 있다. 구글은 인터넷에서 가장 널리 사용되는 검색엔진이다. 이 기업은 웹 페이지 평가에 다른 페이지에서 얼마나 링크를 많이 받았는지를 적용한 첫 검색 엔진을 선보였다. 수천 명이 특정 웹 페이지에 대해 말한 정보를 랭킹 방법에 적용하여 검색 결과 순서를 정하는 데 이용했다. 최근에는 다수의 사용자가 검색했던 키워드를 비슷한 키워드로 추천해 주는 추천 시스템도 도입하여 집단지성의 특색을 잘 살려 서비스화 하고 있다.

2.2 Tweetmix

트위터에서 유통되는 다양한 대화, 뉴스, 동영상, 사진 등의 정보들 속에서 집단지성을 발견하고 유용한 정보를 만들어내려는 시도는 이전부터 행해지고 있다. Tweetmix 에서는 불특정 다수의 사용자들의 메시지 내용을 수집하여 의미 있는 정보를 추출해 내는 시도를 하였다. Tweetmix 의 목적은 사람들이 관심을 가지는 정보를 수집하고 수집된 메시지 중 사람들로부터 가장 인기가 있는 이슈를 찾아내는 것이다. 이 시스템은 각 메시지의 링크를 조사하는데 가장 링크가 많은 메시지들의 랭크를 순차적으로 보여주는 서비스를 하고 있다. 링크가 가장 많은 메시지는 주로 뉴스거리가 되는데, 여러 사람들의 메시지에서 추출되는 뉴스거리는 매스컴에서 보이는 뉴스와는 다른 종류의 새롭고 빠른 정보이기 때문에 사용자들에게 많은 인기를 얻고 있다[3].

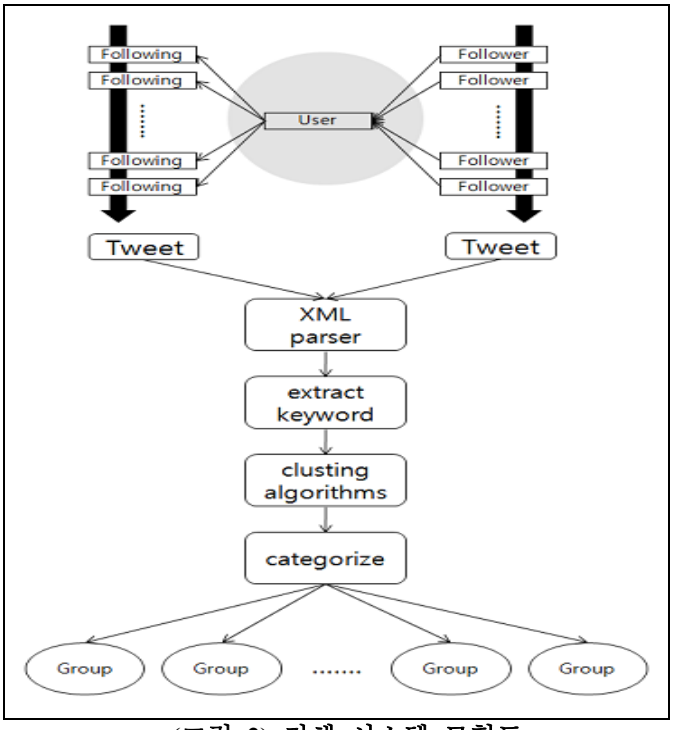
2.3 Mention-map

트위터에서는 팔로윙과 팔로워 라는 관계가 생기게 된다. 트위터에서 팔로우라는 개념으로 서로의 네트워크가 맺어진다. 그 관계에서 사용자들은 자신의 메시지를 입력해서 팔로워들에게 보여주거나, 자신이 팔로윙 하고 있는 또 다른 사용자들이 남긴 메시지를 보게 된다. Mentionmap 은 이러한 메시지들을 누구와 주고받는지를 실시간으로 보여주는 서비스이다. 자신이 팔로우 하고 있는 사용자와의 관계를 화살표로 보여주어 트위터에서 생성되는 인적 네트워크를 그래프로 볼 수 있다. 이 서비스는 자신이 맺고 있는 일차적인 관계와 이, 삼차적인 관계까지 그래프 형식으로 보여주는 장점이 있다. 하지만 이 서비스는 정보의 전달이 관계의 모형화에 그치고 있다[4].

3. 시스템 설계와 구현

3.1 전체 시스템 모델

이 논문에서는 트위터에서 생성되는 전체적인 트랜드가 아니라 개인의 네트워크에서 생성되는 트랜드에 초점을 맞춘다. 본 논문에서 제안하는 시스템은 원하는 특정 사용자의 아이디를 입력 받아 그 사용자의 네트워크를 분석하는 시스템이다. 시스템의 전체적인 흐름은 그림 1과 같이 사용자의 팔로윙, 팔로워의 메시지 내용을 모두 분석하고 키워드를 생성한 뒤 그 사용자의 중심으로 어떠한 집단의 사람들이 분포해 있는지 그리고 그 집단에서 어떠한 트랜드가 생성되는지를 알아보는 것이다.



(그림 2) 전체 시스템 모형도

3.2 데이터 수집

데이터 수집은 트위터에서 제공하는 API 와 GNU 에서 제공하는 XMLParser 1.3.0 를 사용한다[6]. 이 논문에서는 PHP 5.2.12 를 사용하여 데이터 수집 시스템을 구축하였으며 서버는 APACHE 2.2.14, DBMS 로는 MYSQL 5.1.39 을 사용하였다. 전제적인 데이터 수집의 순서는 다음과 같다.

I. 사용자 아이디 입력.
II. 입력된 사용자 아이디의 팔로워, 팔로윙 목록 받음.
III. 팔로워, 팔로윙 목록의 모든 사용자들이 남긴 메시지를 수집.
IV. 메시지 파싱 및 키워드 추출.

프로그램의 시작은 분석을 원하는 아이디를 입력 받으면서 시작된다. 아이디를 입력 받아 API 전달인자로 사용하여 API 를 호출하면 사용자의 팔로워, 팔로윙 목록을 XML 형태로 받아올 수 있다. XML 형태로 받아지는 모든 사용자의 메시지 내용을 파싱한다. 하지만 사용자의 메시지는 주로 자연어이기 때문에 각 단어를 토큰화 할 필요가 있다. 토큰화 하는 과정에서 자연어의 접사, 동사 등을 제거하는 과정을 거친다. 미리 만들어놓은 파일에는 접사, 동사, 관사 등 키워드로서 의미를 부여할 수 없는 단어들의 리스트가 있고 그 리스트에 포함되어 있는 토큰은 키워드로 취급하지 않고 버린다.

그림 2 는 API 를 이용해서 받아온 XML 파일의 일부로 이 논문에서는 테두리 안의 <id>태그와 <text>태그의 내용을 이용한다. 아래의 표와 같은 경우 이 시스템에서는 “I’m not drunk anymore. 2 more 2 go” 의 자연어에서 이미 만들어 놓은 파일시스템에 저장되어 있는 제거 해야 할 리스트인 주어, 동사 등의 단어가 빠져나가고 “drunk” 라는 키워드가 추출되게 된다. 위와 같은 방법을 이용하여 모든 사용자가 남긴 메시지들을 분석하여 키워드화 한다.

```
<status>
  <created_at>Sun Sep 26 23:27:49 +0000 2010</created_at>
  <id>25633785032</id>
  <text>I'm not drunk anymore.2 more 2 go.</text>
  <source><a href="http://twicca.r246.jp/" rel="nofollow">twicca</a></source>
  <truncated>false</truncated>
  <in_reply_to_status_id />
  <in_reply_to_user_id />
  <favorited>false</favorited>
  <in_reply_to_screen_name />
  <retweet_count />
  <retweeted>false</retweeted>
```

(그림 2) 사용자의 정보와 tweet 내용이 담겨있는 XML 파일

3.3 군집화

추출된 키워드는 데이터베이스에 넣는다. 데이터베이스 테이블의 필드는 다음과 같이 3 개로 나뉜다.

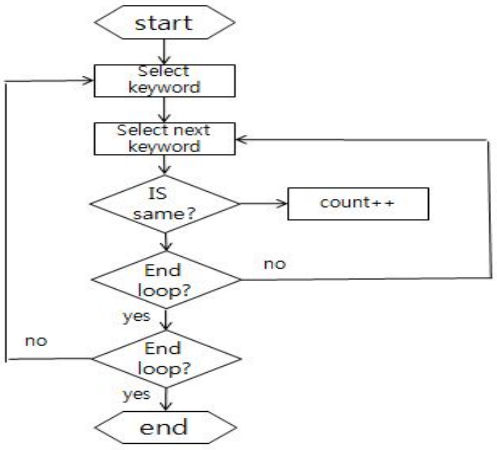
- count: PRIMARY key 로 단지 순차적인 숫자이다.
- id: 메시지를 남긴 사용자 이름을 나타낸다.
- trend: 해당 사용자가 남긴 메시지를 파싱하여 얻어낸 키워드를 나타낸다.

데이터베이스 테이블에는 파싱된 모든 키워드들이 그림 3 과 같은 형태로 들어가게 된다. 그림 3 의 테이블은 전체 추출된 키워드 테이블의 일부분이다. Count 는 키워드의 개수만큼 증가하게 되고 id 와 trend 는 중복을 허용한다. id 에 해당하는 사용자가 생성한 키워드는 모두 trend 에 들어가게 된다. 생성된 테이블은 사용자에 특성에 따라 매우 방대한 양의 데이터를 가지게 될 수도 있다.

count	id	trend
1	Manuel Salgado	OlAge79
2	Manuel Salgado	sexy
3	Manuel Salgado	Damn
4	Manuel Salgado	About
5	Manuel Salgado	Late
6	Manuel Salgado	usual
7	Manuel Salgado	Pandora
8	Manuel Salgado	listening
9	Manuel Salgado	112
10	Manuel Salgado	ime
11	Manuel Salgado	xbox
12	Manuel Salgado	live
13	Manuel Salgado	Still

(그림 3) 추출된 모든 키워드를 데이터베이스에 삽입한 후 테이블

테이블에 들어간 내용 중 트랜드 열의 내용을 카운트하여 출현 횟수를 찾아낸다. 많이 언급되는 키워드는 상대적으로 더 의미를 가진다고 볼 수 있다. 그림 4 는 키워드들의 출현 횟수를 알아내는 알고리즘을 보여준다.



(그림 4) 키워드의 출현 횟수를 구하는 알고리즘

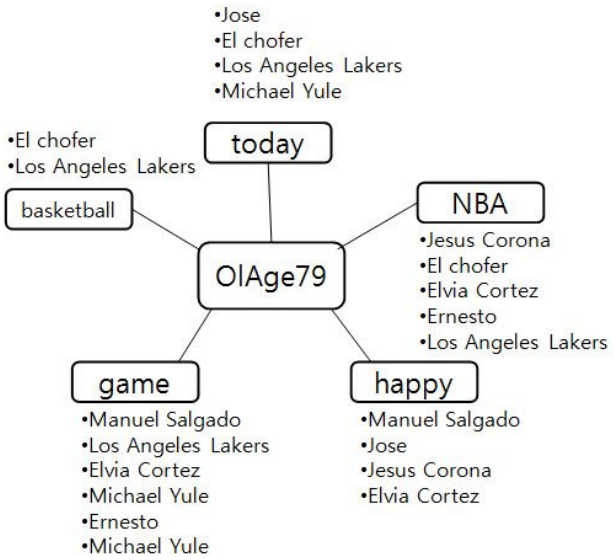
2 회 이상 출현한 키워드들을 모아 키워드를 추출한 근원인 사용자 이름과 함께 데이터베이스 테이블을 생성하면 테이블상에서 어떠한 사용자가 어떠한 키워드를 많이 언급했는지를 확인할 수 있게 된다. 그 후 위의 알고리즘으로 추출된 키워드를 데이터베이스 테이블의 필드로 사용한다. 사용자 이름과 추출된 트랜드들을 필드로 가지는 테이블에 각 사용자들에게서 나온 키워드의 출현 횟수를 내용으로 입력한다. 그 후에는 트랜드를 만들어낸 사용자 이름을 확인할 수 있으므로 그룹화를 할 수 있게 다. 그림 5 는 8 명의 팔로윙들을 가지고 있는 OLAge79 라는 사용자를 분석한 결과이다. 팔로윙들로부터 생기는 키워드는 총 28 개였으며, 사용자 이름 필드를 포함하여 29 개의 필드를 가지는 테이블이 생성되었다.

username	xbox	game	power	campus	happy	train	puppy	group	NBA	OlAge79	ice	California	damn	house	bird	hot	funny	face	mirror	bed	today	ghost	radio	school	pretty	job	Lakers	basketball
Manuel Salgado	2	3	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Jose	0	0	0	2	3	1	1	0	0	5	0	1	1	0	1	0	1	0	1	0	1	0	0	2	1	2	0	0
Jesus Corona	0	0	0	3	2	1	1	2	3	2	0	0	0	1	0	0	0	1	1	0	0	0	1	2	1	2	0	0
El chofer	0	0	0	1	0	0	0	0	3	2	0	0	0	1	0	1	0	0	0	0	2	1	0	0	0	0	2	3
Elvia Cortez	1	2	1	0	3	1	0	2	1	3	0	1	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Los Angeles Lakers	0	3	0	0	0	0	0	2	3	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	5	8
Ernesto	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	1	3	0	0	2	0	1	1	2	1	0	0	0
Michael Yule	1	2	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	2	0	0	1	0	2	0	0	0	0	0

(그림 5) 최종 데이터베이스 테이블 화면

4. 결과 분석 및 평가

본 논문의 시스템에서 사용하는 파싱 알고리즘은 한글을 지원하지 않기 때문에 영어권 지역의 불특정 사용자를 대상으로 테스트를 해 보았다. 총 10 명의 사용자를 선택하여 팔로윙, 팔로워 사이에서 생기는 키워드들을 추출해 보았고 그 결과는 해당 사용자의 기호나 특성을 알아볼 수 있는 지표가 될 수 있었다. 다음은 OLAge79 라는 사용자를 프로파일링 결과이다. 그림 6 은 7 명의 팔로윙들 에서 생성된 키워드 중 가장 빈도수가 높은 5 개를 트랜드로 선정하여 그룹화 한 결과이다.



(그림 6) 그룹화를 시각화한 모습

가장 빈도수가 높았던 5 가지 키워드는 “today”, “NBA”, “basketball”, “game”, “happy” 였다. 이 5 가지 키워드 중 “today”, “happy” 같은 경우는 의미를 찾아내기 힘든 키워드이기 때문에 제대로 추출하였다고 볼 수 없다. 이러한 단어가 생겼을 경우 자연어 처리 부분에서 제외될 대상 리스트파일에 추가하여 다음부터는 파싱 단계에서부터 추출되지 않도록 한다. 수정을 한 후 다시 프로그램을 실행 시켜서 “NBA”, “basketball”, “game”, “xbox”, “campus” 라는 5 개의 새로운 트랜드를 추출해 낼 수 있었다. 이전의 5 개의 트랜드보다 더 의미 있는 그룹화 결과가 나온 것을 확인할 수 있었다.

5. 결론 및 향후 연구 과제

본 논문에서는 SNS 상에서 생기는 수 많은 정보를 수집하여 집단지성의 측면에서 의미 있는 정보로 가공하는 시스템을 개발하는 것이었다. 트위터를 기반으로 사용자들 사이에서 오가는 메시지를 수집하였고, 그러한 메시지들을 정해진 규칙에 따라 파싱하여 해당 사용자의 주변에는 어떠한 트랜드가 오가고 있는지 알아볼 수 있었다. 그 결과 해당 사용자의 기호나 특성을 추출된 키워드를 통하여 찾아볼 수 있었다. 뿐만 아니라 사용자 주변의 인적 네트워크를 그룹화 함으로써 특정 사용자가 속해있는 집단의 특성도 알아볼 수 있는 가능성을 볼 수 있었다.

생성된 집단들의 특성을 살펴보면 전혀 다른 속성의 집단인 경우도 있고 비슷한 속성의 집단인 경우도 있다. 그러한 집단들 사이의 연관성을 찾아내는 것은 더욱 의미 있다. 하지만 키워드를 추출해 내는 과정에서 의미를 부여할 수 없는 키워드들이 많이 나오고 있고, 추출된 키워드를 평가는 방법을 그 키워드의 빈도수에만 의존하고 있어 의미를 부여하는 부분에 신뢰성을 더 높일 수 있는 방법이 필요하다. 많이 등장했던 단어라고 해도 그 사용자가 선호하지 않는 대상일 가능성도 있기 때문이다.

향후 과제로는 자연어 처리의 완성도를 높여 의미 있는 키워드를 추출하는 것과 군집화를 하기 위한 방법 중 하나인 빈도수 측정 이외의 방법을 찾는 것이라고 할 수 있겠다. 나아가 시스템에서 추출된 트랜드를 시각화 하는 기능을 추가하고 사용자 주변의 네트워크인 팔로워, 팔로윙의 특성을 잘 분석해 해당 사용자의 기호에 맞는 서비스를 제공하는 추천시스템으로 발전시킬 계획이다.

참고문헌

[1] Toby Segaran, “Programming Collective Intelligence: Building Smart Web 2.0 Applications”, O’Reilly Media
[2] http://twitter.com, 트위터
[3] http://www.tweetmix.net, tweetmix
[4] http://apps.asterisq.com/mentionmap, mentionmap
[5] http://apiwiki.twitter.com/twitter-API-Documentation, 트위터 API
[6] http://www.criticaldevelopment.net/XML/doc.php, XMLParser