
모바일 사용자를 위한 컨텍스트 기반 마이크로 블로그 토픽 검출 기법

Context-based Microblog Hot Topic Detection for Mobile Users

한중현, Jonghyun Han*, Xing Xie**, 우운택, Woontack Woo***

요약 최근 모바일 장치를 통한 마이크로 블로그 활용이 늘고 있지만, 모바일 장치가 지닌 하드웨어 제약으로 인해 여전히 모바일 정보 브라우징에 어려움이 있다. 이를 해결하기 위해 모바일 사용자의 컨텍스트 정보를 활용하여 사용자의 관심 정보를 추론하는 연구가 활발히 진행되고 있다. 본 논문에서는 모바일 사용자의 컨텍스트를 이용하여 마이크로 블로그의 토픽을 추천하는 방법을 제안한다. 마이크로 블로그에서 사용자와 연관된 토픽을 추출하기 위해 제안한 방법은 사용자 위치, 행동, 기존에 작성한 블로그 그리고 사회적 관계 등의 사용자 컨텍스트를 모바일 장치로부터 얻어 활용한다. 모바일 장치로부터 얻어온 컨텍스트는 마이크로 블로그 검색 범위를 줄이는데 뿐만 아니라 사용자의 관심을 추론하는 경우에도 활용된다. 추론된 사용자의 선호도를 기반으로 검색된 결과의 우선순위를 다시 결정한다. 제안한 방법을 통해 모바일 사용자들은 사용자가 관심을 가질만한 토픽의 마이크로 블로그 정보를 얻을 수 있을 것으로 기대한다.

Abstract Mobile context-awareness becomes an important research topic since mobile information browsing is still difficult due to the limitations of mobile devices. On the other hand, it is easier to gather more user contexts because mobile devices are equipped with more sensors. In this paper, we introduce a method for detecting local hot topics from microblogs on a mobile device. In order to detect user-related topics from microblogs, it exploits mobile user contexts such as location, activity, blogging history and social relationship. Through taking advantage of these contexts, it retrieves user-related microblogs and also infers user interests. It can filter out unrelated topics based on the inferred interests. Based on our proposed method, a mobile user can be aware of topics related to interests surrounding the user.

핵심어: *Context-awareness, Microblog Topic Detection, Mobile Browser*

본 연구는 2011년 지식경제부와 마이크로소프트 아시아 연구소의 지원을 받아 IT/SW 창의 연구 과정으로 수행된 연구임

*주저자 : 광주과학기술원 정보통신공학과 U-VR 연구실 박사과정; e-mail: jhan@gist.ac.kr

**공동저자 : Microsoft Research Asia Researcher; e-mail: xing.xie@microsoft.com

***교신저자 : 광주과학기술원 정보통신공학과 U-VR 연구실 교수; e-mail: wwoo@gist.ac.kr

1. 서론

스마트 폰의 확산으로 모바일 사용자들의 소셜 네트워크 서비스 이용이 늘어나고 있다.[1] 스마트 폰의 기술 발전으로 데스크톱 환경과 유사한 서비스를 사용자들은 제공받고 있지만, 모바일 장치가 가지는 작은 디스플레이와 낮은 컴퓨팅 파워 등의 한계로 인해 많은 정보를 처리하고 확인하는데 여전히 어려움이 있다.[2] 이러한 어려움을 극복하기 위해 사용자가 어떤 정보를 필요로 하는지 추론하여 적합한 소셜 콘텐츠를 제공해야 할 필요가 있다.

컨텍스트 인지 컴퓨팅은 모바일 사용자들의 정보 수요를 이해하기 위한 연구 분야로 최근 활발히 연구되고 있다. [2-4] 비록 모바일 응용 프로그램은 데스크톱 기반의 응용 프로그램과 비교해 보았을 때 편하지 않은 사용자 인터페이스와 같은 한계들을 가지지만, 사용자의 컨텍스트 정보를 보다 쉽게 활용할 수 있다는 장점을 가진다. 스마트 폰의 경우 데스크톱에 비해 더욱 개인적인 장치일 뿐만 아니라 내장된 센서를 가지고 있기 때문에 사용자 컨텍스트를 획득하기 쉽다. 이러한 컨텍스트 정보들은 웹 사이트나 응용 프로그램에서 사용자 중심의 콘텐츠를 제공하기 위해 사용된다.[2] 따라서 모바일 응용 프로그램의 한계를 극복하기 위해 모바일 폰으로부터 컨텍스트를 획득하여 모바일 사용자에게 적합한 콘텐츠를 제공하는 것이 가능하게 되었다.

본 논문에서는 모바일 사용자의 정보 수요에 적합한 콘텐츠를 찾기 위한 방법을 제안한다. 본 논문에서 제안하는 방법을 보이기 위해, 우리는 모바일 사용자와 연관성이 있는 마이크로 블로그 토픽을 추출하여 제공하는 시스템을 개발하였다. 또한, 컨텍스트 정보 활용의 필요성을 알아보기 위하여, 컨텍스트 정보와 마이크로 블로그 사이의 연관 관계를 데이터 분석을 통해 알아보았다. 우리가 제안한 방법은 지역의 마이크로 블로그 토픽을 검출하는 것으로, 본 논문에서 지역의 토픽은 해당 지역의 사용자들로부터 일정 기간 동안 자주 언급된 토픽으로 정의된다. 제안한 방법은 모바일 폰으로부터 얻은 컨텍스트 정보를 이용하여 마이크로 블로그의 검색 범위를 축소시키고, 검출된 토픽들의 우선순위를 컨텍스트를 활용하여 결정하고 사용자에게 전달한다. 사용자에게 연관성 있는 토픽을 검출하기 위하여, 우리는 사용자의 위치 정보와 행동, 블로그 히스토리, 소셜 관계 정보 등의 컨텍스트를 활용하여 사용자의 관심사를 추론하여 활용하였다. 검출된 토픽들 가운데 사용자의 관심사와 연관성이 가장 높은 k 개의 결과를 선택하는 작업을 통해 최종적으로 사용자에게 전달할 토픽을 결정하게 된다. 제안한 방법을 검증하기 위하여 특정 기간 동안 테스트 데이터 집합의 사용자가 블로그한 마이크로 블로그와 트레이닝 데이터 집합에서 추출되어 제공된 토픽간의 관련성을 계산하여 컨텍스트 정보 활용의 효과를 알아보았다. 제안한 방법을 통해 모바일 사용자는 주변 환경에서의 일어나는 일에 대한 토픽을 쉽게 검색하여 확인할 수 있다.

2. 배경

2.1 관련 연구

제안한 방법과 관련한 연구 분야 중 하나는 유비쿼터스 컴퓨팅 환경에서의 컨텍스트 인지 컴퓨팅이다. 유비쿼터스 컴퓨팅 환경에서 사용자 중심적인 응용 프로그램을 개발하기 위하여 모바일 사용자의 컨텍스트를 활용하는 연구가 진행되어왔다.[2-4] Abowd는 모바일 사용자의 현재 위치와 과거 위치 기록 등의 컨텍스트 정보를 활용하여 개인화된 투어 가이드를 모바일 사용자에게 제공하였다.[2] Xu는 모바일 컨텍스트가 모바일 사용자의 관심에 어떻게 영향을 미치는지 보여주었다.[4] Tamminen은 모바일 행동에 컨텍스트가 어떠한 영향을 미치는지에 대한 조사를 진행하였다.[3]

컨텍스트 인지 컴퓨팅은 모바일 사용자의 정보 수요를 알아내기 위해 모바일 검색 분야에서도 적용되어 오고 있다. 모바일 사용자의 정보 수요를 분석하기 위해 많은 연구들이 진행되고 있는데,[5-7] Kamvar와 Yi는 모바일 검색 쿼리 분석을 통해 모바일 정보 검색의 특성을 제시하였다.[6,7] 그들은 검색엔진의 대규모 검색 쿼리 분석을 통해 모바일 검색어들의 카테고리 분포를 보여주었다. Church는 컨텍스트 기반의 모바일 검색 방법을 제안하였고, 사용자 스테디를 통해 모바일 정보 수요에 대한 결과를 제시하였다.[5] 이 사용자 스테디와 제안한 방법을 통해 Church는 사용자의 위치 정보와 시간 정보와 같은 컨텍스트가 모바일 사용자의 관심사를 이해하기 위한 중요한 요소라고 이야기 하고 있다.

최근에는 모바일 컨텍스트 인지 연구와 정보 검색 연구의 결합을 통해, 모바일 정보 검색을 향상 시키는 연구가 진행되고 있다.[8-11] Brown과 Jones에 따르면, 작은 디스플레이의 한계가 있는 모바일 장치에서는, 정보 검색의 효율성을 향상시킬 필요가 있고, 이를 위해 컨텍스트 인지 기법을 활용한 정보 검색이 필수적이라고 이야기 하고 있다.[9] Context-Aware Browser는 사용자의 컨텍스트 정보를 추론하여 관련된 웹 문서를 추출하여 제공한다.[10] Church는 모바일 사용자를 위한 정보 검출 방법을 제안하였는데, 소셜 컨텍스트는 정보 검출 기법의 향상을 위한 중요한 요소로 활용될 수 있다고 이야기 하고 있다.[11] 본 논문에서는 모바일 사용자를 위한 마이크로 블로그 탐색에 있어서 컨텍스트 인지 기법의 활용 효율성에 대한 연구를 진행하였다.

2.2 용어

2.2.1 마이크로 블로그 (Microblog)

마이크로 블로그는 소셜 네트워크 내에서 또는 공개적으로 다른 사용자들과 의사소통하는 도구로 사용되고 있다. 기존 블로그와의 가장 큰 차이점은 메시지의 길이로, 마이크로 블로그 서비스들은 사용자들의 현재 상태나 의견을 표현할 수 있도록, 짧은 텍스트 메시지를 주고받을 수 있는 기능을 제공한다. 마

이므로 블로그의 메시지는 상대적으로 짧기 때문에 일반적으로 사용자들은 블로그에 비해 보다 자주 업데이트 한다. 이는 실시간 이슈들을 반영하는 내용을 담기도 한다. 그리고 마이크로 블로그 서비스는 보통 소셜 네트워크 기능을 포함하고 있다. 사용자들은 자신이 구독하고 싶은 사용자들을 친구로 추가(follow) 할 수 있다.

2.2.2 토픽 (Topic)

토픽은 일반적으로 하나의 문서를 대표하는 하나의 키워드로 사용된다. 본 논문에서 토픽은 마이크로 블로그 사용자들이 주어진 일정 기간 동안 빈번하게 언급하고 있는 단어로 정의한다. 따라서 지역 토픽의 경우는 해당 지역 사용자들이 포스팅한 마이크로 블로그에서 자주 언급되는 단어로 정의하여 사용한다.

2.2.3 컨텍스트 (Context)

표 1. 도메인 별 사용된 컨텍스트의 예

도메인	예
시간	오전/오후/저녁, 점심식사시간/저녁식사시간, 주중/주말, ...
위치	위도, 경도, 방위 집, 회사, 학교, 쇼핑몰, 레스토랑, ...
행동 정보	일, 공부, 쇼핑, 운동, ...
사회적 관계	following, follower

본 논문에서는 이용되는 컨텍스트는 모바일 장치에 내장된 센서 데이터와, 이 데이터를 통해 추론한 사용자의 행동 정보가 포함되며, 사용자의 블로그 히스토리, 사회적 관계 정보 등과 같은 프로필 정보 역시 포함된다. (표 1) GPS를 통해 얻은 위도와 경도는 사용자의 위치 정보를 결정하고 마이크로 블로그 검색의 범위를 줄이는데 활용된다. 내장된 센서 데이터를 이용하여 사용자의 행동 정보를 추론할 수 있다.

3. 마이크로 블로그와 컨텍스트 관계 분석

마이크로 블로그의 특성을 이해하기 위해 우리는 마이크로 블로그와 컨텍스트 정보 사이의 관계를 대규모 마이크로 블로그 분석을 통해 조사하였다. 이 조사를 통해 우리는 마이크로 블로그와 시간, 장소, 사용자 히스토리, 행동 그리고 사회적 관계 등의 컨텍스트 사이의 관계를 이해한다.

3.1 분석 데이터

이번 조사에서 우리는 트위터¹⁾를 마이크로 블로그 데이터로 활용하였다. 트위터는 사용자들이 그들의 상태 정보를 표현할

수 있는 단문 메시지를 포스팅 할 수 있고, 다른 사용자의 메시지를 구독할 수 있다. 사용자들은 프로필에 자신의 위치 정보를 기술 할 수 있는데, 본 분석에서는 사용자가 입력한 위치 정보를 사용자의 위치 컨텍스트로 활용한다. 이는 필수적인 입력 요소는 아니기 때문에 사용자가 입력한 위치 정보가 정확하다고 가정하고 분석을 진행하였다.

분석을 위한 데이터로 미국에 위치한 사용자들이 2009년 12월 13일부터 2010년 1월 31일까지 트위터에 포스팅한 마이크로 블로그를 활용하였다. 미국 사용자들이 포스팅한 마이크로 블로그 약 5500만개를 전역 데이터로 활용하였다. 이 마이크로 블로그 중 뉴욕 사용자로부터 포스팅 된 마이크로 블로그 개수는 450만, 시카고 사용자로부터 포스팅 된 개수는 190만, 그리고 캘리포니아 사용자로부터 포스팅 된 개수는 약 320만개이다. 본 논문에서는 위의 지역과 해당 지역의 마이크로 블로그를 분석 데이터로 사용하였다. 언어의 문제로 인한 결과의 부정확성을 피하기 위해 영어로 포스팅 된 마이크로 블로그에 한해 진행하였다.

3.2 분석 방법

우리는 마이크로 블로그에 컨텍스트가 미치는 영향을 알아보기 위해, 일반 마이크로 블로그로 구성된 테스트 집합과 특정 컨텍스트에 해당하는 마이크로 블로그 집합 간의 단어 빈도, Term Frequency(TF), 유사도를 계산한다. TF는 주어진 마이크로 블로그 집합 내에서 특정 단어가 포함되어 있는 비율로 계산된다. TF간의 유사도는 문서간의 유사도를 계산하기 위해 이용되는 방법으로, 본 분석에서는 문서 대신 마이크로 블로그 간의 유사도를 계산하기 위해 활용하였다.

먼저, 무작위로 각 날짜별 테스트 집합의 사용자들을 추출한다. 그리고 주어진 기간 동안 이 테스트 사용자들이 포스팅 한 마이크로 블로그들의 TF를 계산한다. 이 분석에서는 하루 단위로 기간을 설정하였다. 단문에서 오는 부정확한 결과를 피하고자, 이 분석에서는 마이크로 블로그 각각의 TF를 활용하는 대신, 주어진 기간 동안 포스팅 된 마이크로 블로그들의 집합을 하나의 문서로 간주하여 TF를 계산한다.

테스트 집합의 TF와 비교하기 위한 마이크로 블로그로 우리는 테스트 사용자의 컨텍스트에 해당하는 마이크로 블로그를 집합으로 하여 TF 계산을 한다. 분석의 기본 대조군으로는 컨텍스트 정보가 포함되지 않은 전역 마이크로 블로그 집합이 이용된다. 이 집합은 주어진 기간 동안 포스팅 된 모든 마이크로 블로그들이 포함된다.

테스트 집합의 사용자 위치에 해당하는 지역의 다른 사용자들이 주어진 기간 동안에 포스팅 한 마이크로 블로그를 집합으로 하여 지역 컨텍스트의 영향을 확인한다. 본 분석에서는 뉴욕, 시카고, 캘리포니아에 거주하는 사용자들의 마이크로 블로그들을 이용하였다. 사용자의 블로그 히스토리 역시 컨텍스트로 간주하여, 테스트 집합의 사용자들이 기존에 포스팅 한 마이크

1) Twitter, <http://www.twitter.com>

로 블로그들을 사용한다. 사용자의 소셜 컨텍스트의 영향을 확인하기 위해 사용자 친구들의 마이크로 블로그 역시 비교 집합으로 이용한다. 그리고 친구들 중 같은 지역에 살고 있는 사용자들을 지역 친구로 간주하고, 이 사용자들이 포스팅 한 마이크로 블로그도 비교 대상으로 고려하였다. 이렇게 나뉜 마이크로 블로그 집합들의 TF를 각각 계산하여 테스트 집합의 TF 값과 코사인 유사도를 계산하여 컨텍스트 정보가 어떻게 영향을 미치는지 확인한다.

사용자 행동 정보에 따른 마이크로 블로그 관심사를 확인하기 위하여, 특정 행동에 해당하는 마이크로 블로그들이 어떤 카테고리리에 주로 속하는지 확인하였다. 마이크로 블로그는 특별히 행동 정보를 지니지 않기 때문에, 특정 상황에서 포스팅 된 마이크로 블로그를 검출해 내는 것이 쉽지 않다. 그래서 특정 상황이나 행동이, 예를 들어 ‘shopping’ 또는 ‘working’, 구체적으로 명시된 마이크로 블로그를 검출하여 사용한다. 이렇게 검출된 마이크로 블로그에서 빈번하게 사용되는 단어들을 찾고, 이 단어들이 어떤 카테고리리에서 주로 사용되는지 확인한다. (4.3.2장 참조) 이렇게 계산된 마이크로 블로그의 카테고리 벡터를 통해서 특정 상황이나 행동이 명시된 상황 (표 2) 에서 어떤 카테고리리의 내용에 사용자들이 관심을 보이는지 조사한다.

3.3 분석 결과

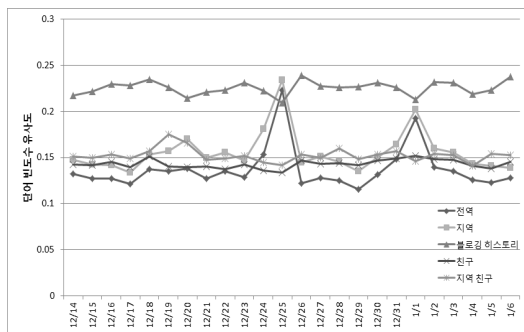


그림 1 테스트 집합과의 TF 유사도 측정 결과

각 컨텍스트들은 마이크로 블로그에 영향을 미치는 각각의 특성을 가지고 있다. 먼저 시간 컨텍스트를 살펴보면, 그림 1에서 볼 수 있듯이 날짜에 크게 종속적인 결과를 보이지는 않는다. 하지만 ‘working’ 과 관련된 단어들은 주중에 비해 주말에는 현저하게 줄어들며, ‘church’의 경우는 반대로 주말에 자주 언급되는 것을 볼 수 있다. 그리고 그림 1에서 볼 수 있듯이, 특정한 날, 예를 들어, 크리스마스 또는 새해 첫 날과 같은 휴일에는 다른 날에 비해 전역 데이터 결과와 지역 데이터의 연관성이 높게 나타나는 것을 볼 수 있다. 이 날들의 마이크로 블로그는 ‘Merry Christmas’ 또는 ‘Happy New Year’와 같은 인사말에 관련된 경우가 많았기 때문이다.

공간 컨텍스트의 경우 전역 마이크로 블로그 결과에 비해 유사

도가 높은 것을 볼 수 있다. 전체적으로 위치 정보가 활용되었을 경우의 유사도 (평균 0.1503) 가 사용되지 않았을 때 (평균 0.1326) 보다 더 높기 때문이다. 분석 결과에서 주목할 사실은, 지역 이벤트가 일어났을 경우에 공간 컨텍스트 정보가 많은 영향을 미치는 것을 볼 수 있다. 예를 들어, 12월 19일의 경우 공간 컨텍스트가 사용된 경우와 사용되지 않은 경우의 유사도 차이 (0.032) 가 가장 큰 것을 볼 수 있는데, 이는 당시 뉴욕에 있었던 눈보라로 인해 뉴욕 지역의 사용자들의 관심이 눈의 영향을 받았기 때문이다. 지역 스포츠 팀의 경기 또는 날씨와 같은 지역적인 영향을 주는 이벤트는 공간 컨텍스트의 영향을 극대화 시켜주는 것을 알 수 있다.

사용자의 블로그 히스토리는 가장 높은 연관성을 보여, 이전에 사용자가 관심을 가졌던 토픽에 다시 관심을 가질 확률이 높다는 것을 그림 1을 통해 볼 수 있다. 이 히스토리 정보는 지역 이벤트 또는 특정한 날의 영향을 크게 받지 않는다. 사용자의 친구들의 마이크로 블로그 정보를 활용한 경우에 유사도가 더 높은 것은 사용자의 소셜 네트워크 정보가 사용자의 관심을 반영한다는 것을 나타낸다. 그리고 평균적으로 같은 지역 내의 소셜 네트워크를 활용한 경우가 더 높은 유사도를 가지는 것은 물론, 지역 이벤트가 있을 시에는 더 많은 영향을 받는 것을 볼 수 있다. 이는 소셜 네트워크에서 친구 사이의 물리적 거리와 우호 정도에 연관 관계가 있다는 사실이 마이크로 블로그에서도 적용되는 것을 알게 해준다.

표 2. 상황에 따른 마이크로 블로그 관심 카테고리

상황	1st	2nd	3rd
working	society	reference	business
studying	reference	society	sports
shopping	shopping	business	recreation
exercising	sports	health	recreation

표 2는 사용자의 행동이 주어졌을 때의 마이크로 블로그들의 관심 카테고리리를 보여주고 있다. 사용자가 마트에 있거나 쇼핑을 가려고 할 때에는 쿠폰, 할인, 선물 등 쇼핑 카테고리에 속하는 내용이 많은 반면, 스포츠 마이크로 블로그는 상대적으로 적은 것을 볼 수 있다. 반대로 사용자가 운동중인 경우에는 스포츠 또는 건강과 관련된 내용이 많은 반면 쇼핑 또는 비즈니스 내용은 적은 것을 볼 수 있다. 이 결과에 따르면 사용자의 행동은 사용자가 흥미를 느끼는 마이크로 블로그에 영향을 미친다는 것을 알 수 있다.

위의 분석 결과를 통해 알 수 있듯이 마이크로 블로그는 사용자 컨텍스트의 영향을 받는다. 따라서 사용자에게 적합한 콘텐츠를 추출하기 위해서는 컨텍스트 정보의 활용이 필수적이다. 본 논문에서는 이와 같은 분석 결과에 따라 컨텍스트를 활용할 때 마이크로 블로그 토픽 추출 방법을 제안하고, 사용자에게 관련된 콘텐츠를 전달하는 시스템을 제안한다.

4. 컨텍스트 기반 마이크로 블로그 토픽 브라우저



그림 2. 컨텍스트 인지 마이크로 블로그 브라우저

우리는 모바일 폰으로부터 컨텍스트를 얻고, 이를 이용해 마이크로 블로그 토픽을 추출하고 모바일 폰에서 브라우징 할 수 있는 시스템을 디자인 및 개발하였다. 그림 2는 구현된 컨텍스트 인지 마이크로 블로그 브라우저로, 추출된 토픽과 함께 해당 마이크로 블로그를 지도 위에 보여준다.

4.1 모바일 마이크로 블로그 토픽 브라우저

우리가 개발한 시스템은 모바일 폰에 장착된 센서로부터 사용자 컨텍스트를 얻고, 마이크로 블로그의 지역 토픽을 찾고, 이를 모바일 폰에 브라우징 한다. 이 시스템은 서버와 모바일 응용프로그램으로 구성되어 있다. 서버는 마이크로 블로그 토픽을 다루며, 모바일 응용프로그램은 사용자 컨텍스트를 얻고, 검출된 토픽을 가시화하는 역할을 한다.

모바일 응용프로그램의 역할은 사용자의 컨텍스트를 얻는 것과 검출된 토픽을 가시화 하는 것이다. 컨텍스트는 모바일 폰에 장착된 센서와 입력된 사용자 프로파일을 통해 얻어진다. 그리고 서버에서 추출된 토픽을 가시화하는데, 이 때, 사용자의 위치와 함께 지역 토픽이 표시된 지도를 화면에 출력한다. 사용자들은 이 모바일 응용프로그램을 통해 최신 지역 토픽을 확인할 수 있게 된다. 사용자들은 표시된 토픽을 클릭하여, 선택된 토픽이 포함되어 있는 마이크로 블로그들을 탐색할 수 있다.

지역 토픽을 검출하는 작업과 사용자가 관심가지는 토픽을 선택하는 작업은 거대한 양의 데이터를 처리하는 많은 계산이 필요하기 때문에 서버에서 작업을 한다. 가장 먼저, 지역 토픽을 추출하기 위해, 사용자 위치 정보를 이용하여 지역 마이크로 블로그를 검색하고 전역 마이크로 블로그와 이전에 포스팅 되었던 데이터와의 비교를 통해 최근 지역 토픽을 검출한다. 두 번째로 서버는 사용자의 관심사를 주어진 컨텍스트를 기반으로 추론한다. 마지막으로 주어진 토픽들을 관심사와 유사한 순서로 다시 순위를 매기고, 높은 순위의 토픽을 선택하여 모바일 응용프로그램에 전달한다.

4.2 지역 마이크로 블로그 토픽 검출 방법

마이크로 블로그의 지역 토픽을 검출하기 위해, 우리가 제안한 시스템은 사용자의 컨텍스트 정보, 특히 사용자의 위치 정보를

활용하였다. GPS를 통해 얻어진 사용자의 위치 (위도와 경도)는 역-지오코딩을 통해 도시의 이름으로 변환된다. 제안된 시스템은 사용자가 위치한 지역에 살고 있는 다른 사용자들이 포스팅 한 마이크로 블로그를 검출한다. 토픽을 검출하기 위해 추출한 마이크로 블로그들은 문서로 간주되어 이용된다.

토픽이 될 단어들의 중요성을 평가하기 위해 우리는 단어의 마이크로 블로그 빈도, Microblog Frequency(MF)를 시간 t 에 주어진 마이크로 블로그 집합 B_t 의 개수와 주어진 단어 $term$ 이 포함된 마이크로 블로그의 개수의 비율로 정의하여 활용한다. 이는 다음 수식 (1)과 같이 정의된다.

$$MF(term, B_t) = \frac{|\{b_t : term \in b_t\}|}{B_t} \quad (1)$$

계산된 MF값이 임계값보다 작은 단어들은 후보에서 제거되며, 중요하지 않은 단어와 시제, 단수, 복수형의 문제를 없애기 위해 stop-word를 제거하고, stemming을 수행하였다.

최근에 이슈가 되는 토픽을 찾기 위해 우리는 이전 시간대의 데이터 집합과 단어 빈도수를 비교한다. 이 비교를 통해 항상 높은 MF값을 가지는 단어들, 예를 들어 'love' 또는 'day', 등의 중요하지 않은 단어들을 토픽의 후보에서 제거할 수 있게 된다. 수식 (2)를 통해 우리가 제안한 Temporal Novelty (TN)을 통해 시간 t 에 따른 단어 $term$ 의 중요도를 계산한다. 측정하고자 하는 시간 t 이전에 언급이 많이 된 단어일수록 TN의 값은 작아지게끔 모델링 하였다.

$$TN(term, B_t) = \frac{MF(term, B_t)}{\sum_{i=0}^n MF(term, B_{(t-\Delta t \cdot i)})} \quad (2)$$

추가적으로, 지역적인 특징을 가속화하기 위해 단어의 전역 MF값과 비교를 하여 최종적으로 토픽의 점수 $TS(\theta)$ 를 수식 (3)과 같이 계산한다.

$$TS(term) = TN(term, LB_t) \cdot \frac{MF(term, LB_t)}{MF(term, GB_t)} \quad (3)$$

수식 (3)에서 LB_t 와 GB_t 는 시간 t 일 때의 지역, 전역 마이크로 블로그를 나타낸다. 예를 들어, 그림 3에서 볼 수 있듯이, 'love'의 경우 비슷한 수준의 MF값을 가지기 때문에 시간적인 중요도를 가지지 않게 되지만 'snow'의 경우에는 12월 19일 급격히 MF값이 증가하여 토픽으로의 중요도가 증가하게 된다. 그리고 'snow'는 다른 지역에 비해 뉴욕의 사용자들로부터 많이 언급 되어 뉴욕의 지역 토픽 중요도가 증가한다. 마지막으로, 위의 시간적 공간적 비교를 통해 얻은 토픽들 중 k 개가 최종적으로 선택되어진다.

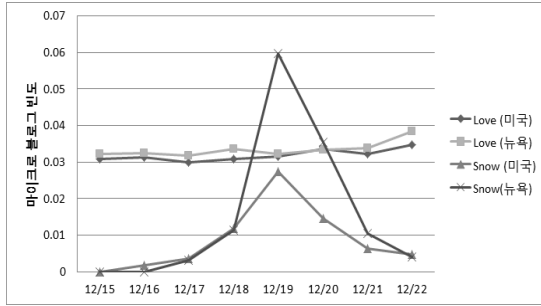


그림 3 뉴욕과 미국에서의 'snow'와 'love'의 빈도

4.3 컨텍스트 기반 사용자 관심 토픽 선택 방법

지역 토픽을 검출한 뒤, 사용자에게 연관된 토픽을 찾기 위한 작업을 수행한다. 이를 위해, 사용자의 관심사를 이용하여 유사도를 계산하고, 이 값을 통해 순위를 다시 매기는 방법을 이용하였다. 추론된 사용자의 관심사와 추출된 토픽 간의 유사도를 측정하기 위해 토픽과 관심사를 카테고리 벡터 형식으로 표현하고, 카테고리 벡터 공간에서 비교하였다. 토픽과 관심사 사이의 유사도 값에 따라 토픽의 순위를 다시 결정한 뒤, 결과를 모바일 응용프로그램에 전달한다.

4.3.1 사용자 관심사

사용자에게 연관된 토픽을 선택하기 위해서, 사용자의 행동 정보를 통해 관심사를 추론한다. 모바일 폰의 센서로부터 입력된 컨텍스트를 통해 사용자의 행동을 인지하게 되고, 인지된 행동을 언급하고 있는 마이크로 블로그를 검색하고 분석한다. 예를 들어, 학교에 있을 때 사용자들의 마이크로 블로그 관심사를 알아보기 위해 'in the school'과 같은 상황 표현 어구를 포함하는 마이크로 블로그를 검색한다. 검색된 결과에서 중요하게 사용되는 단어들을 수식 (1)을 통해 찾아내게 되는데 정규화 과정을 거친 MF(term, B)값은 특정 상황 situ이 주어졌을 때 단어 term의 중요도 p(term/situ)를 나타내게 된다. 이렇게 얻은 중요한 단어들을 이용하여, 주어진 상황에서 어떤 관심을 보이는지 추론할 수 있다.

사용자 행동뿐만 아니라 사용자의 블로깅 히스토리와 사회적 관계와 같은 프로파일 컨텍스트 정보 역시 관심사를 추론하기 위해 사용하였다. 모바일 폰의 경우 매우 개인적인 장치이기 때문에, 사용자의 동의만 있다면 개인적인 정보를 사용하기에 쉽다는 장점이 있다. 사용자가 이전에 포스팅 했던 마이크로 블로그를 검색하여 수식 (1)을 통해 중요한 단어들을 추출한다. 그리고 마이크로 블로그의 소셜 네트워크를 통해, 사용자가 친구들이 이전에 포스팅 한 마이크로 블로그를 검색하여, 중요하게 사용된 단어들 역시 추출하게 된다. 이렇게 얻은 중요한 단어들을 이용하여, 사용자와 그 친구들이 어떤 카테고리에 관심을 보이는지 알아낸다.

4.3.2 토픽과 사용자 관심사의 카테고리 벡터 표현 방법

토픽과 사용자의 관심사 사이의 직접적인 비교가 어렵기 때문에, 각각을 카테고리 벡터 형식으로 변환하여 비교하였다. 이 카테고리 벡터는 토픽 또는 관심사가 특정 카테고리에 속할 확률로 표현된다. 본 논문에서는 Open Directory Project¹⁾의 카테고리 정보를 이용하여 카테고리 벡터로 표현하는 방법을 제안하였다.

ODP의 카테고리 기술 문서를 통해 우리는 특정 카테고리에서 어떤 단어가 중요한지 알 수 있다. 먼저, 특정 카테고리에 해당하는 기술된 내용의 단어들의 빈도를 계산하고 정규화되는데, 이 값은 p(term/ctg)로 특정 카테고리 ctg에서의 단어 term의 확률로 표현된다. 여러 단어가 주어졌을 때 어떤 카테고리에 속하는지 알아보기 위해 p(ctg/t₁, t₂, ..., t_n)을 계산한다. 단어들의 사용은 독립적이기 때문에, naive 베이저안을 통해 수식 (4)와 같이 추산한다.

$$p(ctg|t_1, t_2, \dots, t_n) = \frac{p(ctg) \prod_{j=1}^m p(t_j|ctg)}{p(t_1, t_2, \dots, t_n)} \quad (4)$$

$$\propto p(ctg) \prod_{j=1}^m p(t_j|ctg)$$

토픽과 관심사를 표현하기 위한 카테고리 벡터는 $ctg = [c_1, c_2, \dots, c_n]^T$ 로 표현되며 n은 카테고리의 개수이다.

$$c_i = p(ctg_i) \prod_{j=1}^m p(t_j|ctg_i) \cdot w(t_j) \quad (5)$$

수식 (5)는 c_i를 계산하는 방법을 나타낸 것으로, 수식 (4)의 단어가 주어졌을 때 특정 카테고리에 속할 확률과 주어진 단어의 중요한 정도에 따라 w(t_j)를 부여하여 계산하는데, 이 값은 수식 (1)을 통해 구한 MF값이 할당된다. 예를 들어, 토픽을 카테고리 벡터로 표현하기 위해, 토픽을 포함하는 마이크로 블로그 집합에서의 단어들의 MF값들이 정규화되고, 이 값이 w(t_j)로 이용된다. 최종적으로, c_i의 정규화를 통해 카테고리 벡터 ctg를 얻게 된다.

4.3.3 유사도에 따른 토픽 우선순위 재결정

추출된 토픽들은 추론된 사용자의 관심사에 따라 순위가 다시 결정된다. 제안하고 있는 방법은 토픽과 관심사의 유사도를 계산하고, 이 값에 따라 순위를 다시 결정하게 한다. 사용자의 관심사와 토픽 모두 카테고리 벡터로 변환하고, 코사인 유사도 계산 방법을 이용하여 유사도를 측정한다. 최종적으로 이 유사도 값과 수식 (3)을 통해 얻은 토픽 점수의 곱으로 얻은 값을 바탕으로 순위를 다시 결정하게 된다.

1) Open Directory Project (ODP), <http://www.dmoz.org/>

5. 평가

본 논문에서 우리는 모바일 사용자에게 사용자와 관련된 마이크로 블로그 토픽을 추출하는 방법을 제안하였다. 제안된 방법의 효율성을 평가하기 위해 추출된 토픽과 사용자가 포스팅 한 마이크로 블로그 사이의 연관성에 기반 하여 평가를 진행하였다. 본 실험을 위해 우리는 섹션 3에서 활용한 데이터 집합을 동일하게 이용하였다. 이 실험에서, 미국 사용자들이 올린 마이크로 블로그의 집합을 전역 데이터로, 뉴욕 사용자들의 마이크로 블로그를 지역 데이터로 활용하였다. 우리는 표 2에 기술한 4가지 상황을 사용자의 행동 모델로 이용하였으며, 사용자의 following 정보를 사회적 관계 정보로 활용하였다. 우리는 드물게 나타나는 단어들을 제거하기 위해 0.5%로 임계값을 설정하였다.

5.1 평가 방법

5.1.1 평가 프레임워크

먼저, 하루 동안의 마이크로 블로그를 임의로 테스트 집합과 트레이닝 집합으로 분류하였다. 75%의 마이크로 블로그들은 트레이닝 집합으로 활용되며, 그 외의 마이크로 블로그들은 테스트 집합으로 활용된다. 토픽은 트레이닝 집합의 마이크로 블로그로부터 추출되며, 추출된 토픽은 테스트 집합에 포함된 각각의 마이크로 블로그와 비교되어진다. 만약 마이크로 블로그가 토픽 단어를 포함하고 있는 경우에는 연관성이 1이 되며, 그렇지 않을 경우 마이크로 블로그와 토픽의 유사도를 계산하여 연관성을 측정하였다.

5.1.2 기준 방법

우리는 제안한 방법의 효율성을 비교하기 위해 컨텍스트를 사용하지 않고 전역 마이크로 블로그만 이용하여 토픽을 추출하는 방법을 기본이 되는 기준으로 삼는다. 이는 수식 (2)를 통해 얻은 토픽으로, 가장 큰 TN 값을 가지는 k 개의 토픽을 추출한다. 컨텍스트 정보를 활용하지 않는 방법이기 때문에 컨텍스트를 기반으로 추출하는 제안한 방법과 비교하여 효율성을 비교할 수 있다.

5.2 평가 결과

그림 5는 추출된 토픽의 개수에 따른 연관성 점수 결과를 보여준다. 결과를 보면, 일반적으로 사용자 컨텍스트 정보를 활용했을 경우의 연관성 점수가 기준 방법보다 높은 것을 알 수 있다. 그리고 사용자의 관심에 따른 토픽 선택을 하게 되면 위치 컨텍스트만 활용했을 경우에 비해 전체적으로 높은 연관성을 가지게 된다. 그림 5를 보면, 사용자의 행동 정보를 활용했을 때 가장 큰 연관성을 보이는 것을 알 수 있으며, 사용자의 블로그 히스토리나 사회적 관계 정보 모두 연관성을 높이는 것을 볼 수 있다. 사용자의 관심사가 사용자의 친구들의 관심사와 크게 다르지 않기 때문에, 블로그 히스토리를 활용했을 때와 사회적 관계를 이용했을 때의 차이가 크지 않았다.

제안한 방법의 효율성은 12월 19일의 결과 예를 통해 쉽게 볼 수 있다. 뉴욕에서의 추출된 토픽 중 대부분은 'snow', 'snowstorm' 그리고 'weather'와 같이 눈과 관련되어 있다. 그러나 컨텍스트를 활용하지 않은 전역 데이터의 경우는 '#followfriday' 그리고 'Avatar' 토픽이 추출되었다. 이 날, 지역 토픽의 연관성이 전역 토픽의 연관성 보다 훨씬 높은 것을 볼 수 있다. 12월 25일의 토픽의 경우, 전역 토픽과 지역 토픽이 'Christmas', 'Jeju', '#celebgift'로 유사하였다. 하지만 사용자의 관심사에 따라 토픽 선택 작업을 한 경우에는 결과가 크게 달라진다. 만약 사용자가 쇼핑을 하고 있는 경우에는 '#celebgift'에 가장 높은 우선순위가 부여되고 쇼핑과 관련된 마이크로 블로그를 탐색하게 된다. 이러한 예제를 통해, 우리는 알고리즘에 따라 선택된 토픽의 연관성이 어떻게 향상되는지 알 수 있다.

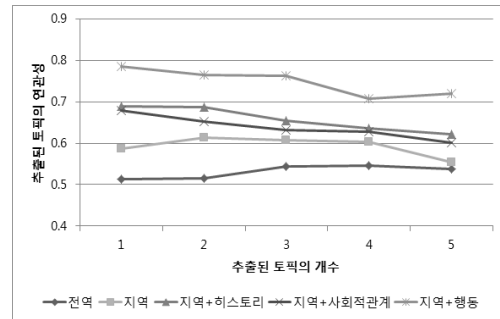


그림 4 검출된 토픽의 유사도 실험 결과

6. 결론 및 추후 연구

본 논문에서 우리는 모바일 사용자에게 사용자와 관련된 마이크로 블로그 토픽을 제공하는 시스템을 소개하였다. 그리고 마이크로 블로그와 컨텍스트 사이의 관계를 데이터 분석을 통해 알아 보았다. 우리의 연구 목적은 효과적으로 모바일 사용자에게 적합한 콘텐츠를 추출해 내는 것으로, 이는 실험을 통해 컨텍스트를 활용한 토픽의 추출과 선택이 모바일 콘텐츠 추천에 도움을 주는 것을 확인하였다. 제안한 방법을 통해 모바일 장치에서 정보를 효과적으로 추출하고 브라우징 할 수 있을 것으로 기대한다. 추후 연구로는 토픽을 추출하는 연구에서 연관성을 증진시킬 수 있는 방법이 필요하다. 본 연구에서는 하나의 컨텍스트 정보를 이용하여 사용자의 관심사를 추론하여 활용하였으나, 사용자의 블로그 히스토리, 행동 정보, 그리고 사회적 관계 모델의 결합을 통한 하이브리드 추론 방식을 통해 추출된 토픽의 연관성을 높일 수 있는 방법의 제안이 필요하다. 그리고 개발된 시스템의 사용성을 검증하기 위하여, 사용자의 정성적 평가를 고려해야 할 필요가 있다.

접수 : 2011,00,00. / 심사 : 2011,00,00. / 게재확정 : 2011,00,00.

참고문헌

- [1] A. Java, X. Song, T. Finin and B. Tseng, "Why we twitter: understanding microblogging usage and communities", In Proceedings of WebKDD/SNA-KDD 2007, pp. 56-65, New York, 2007.
- [2] G. D. Abowd, C. G. Atkeson, J. Hong, S. Long, R. Kooper and M. Pinkerton, "Cyberguide: a mobile context-aware tour guide", Wireless Networks, vol. 3, no. 5, pp. 421-433, 1997.
- [3] S. Tamminen, A. Oulasvirta, K. Toiskallio and A. Kankainen, "Understanding mobile contexts", Personal Ubiquitous Computing, vol. 8, no. 2, pp. 135-143, 2004.
- [4] Z. Xu and Y. Yuan, "The impact of context and incentives on mobile service adoption", International Journal of Mobile Communications, vol. 7, no. 3, pp. 363-381, 2009.
- [5] K. Church and B. Smyth, "Understanding the intent behind mobile information needs", In Proceedings of UII' 09, pp. 247-256, 2009.
- [6] M. Kamvar and S. Baluja, "A large scale study of wireless search behavior: Google mobile search", In Proceedings of CHI' 06, pp. 701-709, 2006.
- [7] J. Yi, F. Maghoul and J. Pedersen, "Deciphering mobile search patterns: a study of yahoo! mobile search queries", In Proceedings of WWW' 08, pp. 257-266, 2008.
- [8] F. S. Tsai, M. Etoh, X. Xie, W.-C. Lee and Q. Yang, "Introduction to mobile information retrieval", Intelligent Systems, IEEE, vol. 25, no. 1, pp. 11-15, 2010.
- [9] P. J. Brown and G. J. F. Jones, "Context-aware retrieval: Exploring a new environment for information retrieval and information filtering", Personal Ubiquitous Computing, vol. 5, no. 4, pp. 253-263, 2001.
- [10] P. Coppola, V. Della Mea, L. Di Gaspero, D. Menegon, D. Mischis, S. Mizzaro, I. Scagnetto and L. Vassena, "The context-aware browser", Intelligent Systems, IEEE, vol. 25, no. 1, pp. 38-47, 2010.
- [11] K. Church, J. Neumann, M. Cherubini and N. Oliver, "Socialsearchbrowser: a novel mobile search and information discovery tool", In Proceedings of UII' 10 pp. 101-110, 2010.



한 종 현

2001년 3월 ~ 2005년 8월 아주대학교 정보 및 컴퓨터공학부 졸업(공학사). 2006년 3월 ~ 2007년 8월 광주과학기술원 대학원 석사 졸업(공학석사). 2007년 9월 ~ 현재 광주과학기술원 대학원 박사과정. 관심분야는 컨

텍스트 인지 컴퓨팅, 소셜 네트워크, 유비쿼터스 컴퓨팅 등임.



Xing Xie

1992년 ~ 1996년 2월 University of Science and Technology of China 졸업(공학사). 1996년 ~ 2001년 University of Science and Technology of China 대학원 박사 졸업(공학박사). 2001년 7월~현재 마이크로 소프트 리서치 아시아 연구원. 관심분야는 위치 기반 서비스, 유비쿼터스 컴퓨팅 등임.



우 윤 택

1985년 ~ 1989년 경북대학교 전자공학과 학사 졸업(공학사). 1989년 ~ 1991년 포항공과대학교 전기전자공학과 석사 졸업(공학석사). 1993년 ~ 1998년 Univ. of Southern California (USC) Electrical Engineering Systems 박사 졸업(공학박사). 1991년 ~ 1992년 삼성종합기술원 연구원. 1999년 ~ 2001년 ATR MIC Lab. 초빙 연구원. 2001년 ~ 현재 광주 과학 기술원 정보기전공학부 정보통신공학과 교수. 2005년 ~ 현재 광주과학기술원 문화콘텐츠기술연구소장. 관심분야는 문화콘텐츠기술, 3D 컴퓨터 비전, 증강/혼합현실, 인간 컴퓨터 상호작용, 컨텍스트 인지 컴퓨팅, 유비쿼터스 컴퓨팅 등임.