KIPS Tr. Software and Data Eng.
Vol.4, No.10 pp.447~454 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2015.4.10.447

Geographical Name Denoising by Machine Learning of Event Detection Based on Twitter

Seungmin Woo† · Byung-Yeon Hwang††

ABSTRACT

This paper proposes geographical name denoising by machine learning of event detection based on twitter. Recently, the increasing number of smart phone users are leading the growing user of SNS. Especially, the functions of short message (less than 140 words) and follow service make twitter has the power of conveying and diffusing the information more quickly. These characteristics and mobile optimised feature make twitter has fast information conveying speed, which can play a role of conveying disasters or events. Related research used the individuals of twitter user as the sensor of event detection to detect events that occur in reality. This research employed geographical name as the keyword by using the characteristic that an event occurs in a specific place. However, it ignored the denoising of relationship between geographical name and homograph, it became an important factor to lower the accuracy of event detection. In this paper, we used removing and forecasting, these two method to applied denoising technique. First after processing the filtering step by using noise related database building, we have determined the existence of geographical name by using the Naive Bayesian classification. Finally by using the experimental data, we earned the probability value of machine learning. On the basis of forecast technique which is proposed in this paper, the reliability of the need for denoising technique has turned out to be 89.6%.

Keywords : SNS, Twitter, Realtime Event Detect, Geographical Name Denoising, Machine Learning

트위터 기반 이벤트 탐지에서의 기계학습을 통한 지명 노이즈제거

우 승 민† · 황 병 연††

요 약

본 논문에서는 트위터 기반 이벤트 탐지에서의 기계학습을 통한 지명 노이즈제거 방식을 제안한다. 최근 스마트폰 이용자의 증가로 소셜 네트워크 서비스(SNS) 이용자가 증가하고 있는 추세이다. 그중 트위터는 140자 이내의 단문서비스와 팔로우 기능으로 정보의 빠른 전달력과 확산성을 가지고 있다. 이러한 특성과 모바일에 최적화된 트위터의 특성상 정보 전달 속도가 매우 빠르기 때문에 재난 상황이나 이벤트 전달의 매개체 역할을 하고 있다. 이와 관련된 연구로는 트위터 사용자 개개인을 이벤트 탐지의 센서로 사용하여 현실에서 발생하는 이벤트를 탐지하였는데 이벤트가 특정 장소에서 발생한다는 특성을 이용해서 지명 키워드를 사용하였다. 그러나 지명과 동형이의어 관계에 관한 노이즈제거에 대한 부분이 누락되어있어서 이벤트 탐지의 정확도를 낮추는 요인이 된다. 이에 본 논문에서는 제거와 예측 두 가지 방식으로 노이즈제거 기법을 적용하였다. 먼저 노이즈 관련 데이터베이스 구축을 이용하여 제거 필터링을 진행한 후에 나이브 베이지안 분류를 이용해서 지명 유무를 결정하였다. 실험 데이터를 이용해서 기계학습을 위한 확률값을 구했으며, 지명마다 본 논문에서 제시하는 예측기법을 검증했을 때 89.6%의 신뢰도로 노이즈제거 기법의 필요성을 보였다.

키워드 : SNS, 트위터, 실시간 이벤트 탐지, 지명 노이즈제거, 기계학습

1. 서 론

최근 스마트폰 이용자의 증가와 무선인터넷 서비스의 확장으로 소셜 네트워크 서비스(SNS)의 이용자가 급증하고 있다. 소셜 네트워크 서비스는 사용자 간의 자유로운 의사소통과 정보 공유, 그리고 인맥 확대 등을 통해 사회적 관계를 생성하고 강화시켜주는 온라인 플랫폼을 의미한다. [1]에 따르면 SNS 중 트위터 이용자 수는 2015년 3월 25일을 기준으로 약 2억 8천만 명에 이르렀으며, 그 수가 지속적으로 증가하고 있다. 그중 트위터는 페이스북, 네이버 블로그, 카카오 스토리 등의 다른 SNS와 구별되는 특징을 가지고 있다. 트위터는 140자 이내 단문메시지 서비스로 장문의 글을 쓰는 블로그와는 달리 신속하게 개인의 의견이나 생각을 공유할 수 있다. 이는 트윗 텍스트가 실시간으로 생성되는

※ 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2011-0009407).
† 준 회 원 : 가톨릭대학교 컴퓨터정보공학부 학사과정
†† 종신회원 : 가톨릭대학교 컴퓨터정보공학부 교수
Manuscript Received : July 7, 2015
First Revision : August 31, 2015
Accepted : September 21, 2015
* Corresponding Author : Byung-Yeon Hwang(byhwang@catholic.ac.kr)

것을 나타낸다. 또한 트위터는 다른 SNS의 친구 맺기와 다르게 팔로우(follow)라는 독특한 방식으로 운영되는데 상대방이 허락하지 않아도 상대방의 최근 활동을 알게 해준다. 따라서 트위터의 사용자는 다른 SNS와 비교할 때 관계의 확산이 빠르고, 트윗과 리트윗을 통해 정보 전달에 유리하다. 마지막으로 개발자는 트위터에서 제공하는 Twitter API를 통해 다양한 애플리케이션 및 웹서비스를 개발할 수 있다. 이러한 특징과 모바일에 최적화된 트위터의 특성상 정보 전달 속도가 매우 빠르기 때문에 급박한 재난 상황이나 이벤트 전달의 매개체 역할을 하고 있다[2-4].

이벤트는 어떤 사건이 특정 장소와 시간에 일어나는 것을 말한다. 트위터 이용자들의 대부분은 스마트폰을 이용해서 글을 올리기 때문에 특정 이벤트가 발생할 경우 발생 장소에 있는 트위터 사용자가 이벤트 내용을 작성함으로써 다른 이용자와 공유한다. 이때 트윗을 작성하는 트위터 사용자 개개인을 이벤트 탐지의 센서로 이용할 수 있다. 이벤트는 물리적 특성을 갖고 있는데, 이러한 물리적인 위치를 갖는다는 사실을 기반으로 지명 키워드를 포함하는 트윗 텍스트를 이벤트 탐지의 도구로 활용하였다. 실시간으로 발생하는 트윗에서 지명 키워드를 이용해서 현실에서 발생한 이벤트를 탐지할 수 있었다. 하지만 한글의 특성상 이벤트 발생 장소로 반환된 키워드가 지명과 형태는 같지만 의미가 다른 동형이의어 관계이거나 다른 구문에 포함되어 지명으로 사용되지 않은 경우가 많았다. 이러한 경우에 시스템에서 반환된 키워드가 실제로 지명이 아닌 경우이므로 이벤트 탐지의 방해요소로 작용하여 탐지 정확도를 떨어뜨린다. 방해요소로 작용한 지명 키워드를 지명 노이즈로 정의하고 시스템의 정확도를 높이기 위해서 지명 노이즈를 제거할 필요성이 있다.

본 논문에서는 지명 키워드로 행정구역명과 지하철역명을 포함하는 170개의 키워드를 사용하였으며, 그중 지명 노이즈가 많은 43개의 지역을 대상으로 노이즈제거 기법을 적용하였다. 노이즈제거 기법은 크게 제거와 예측 2가지로 나뉘며, 제거는 기계학습을 적용하기 위해 지명 노이즈를 제거하는 과정이고 예측은 실험 데이터로 구한 확률값을 통해 기계학습을 이용해서 지명 유무를 판단하는 방법이다. 노이즈제거 기법을 적용하여 시스템에서 이벤트 지명으로 탐지하였지만 실제로는 지명이 아닌 경우를 다수 제거할 수 있고, 제거 필터링 후에 지명 유무를 결정한다. 그 결과 문제가 되었던 시스템의 정확도를 높일 수 있다.

본 논문의 구성은 다음과 같다. 2절에서는 관련 연구에 대해서 소개하고 3절에서는 지명 노이즈제거 방안에 대해서 살펴본다. 이후 4절에서 실험 데이터를 통해 노이즈제거 기법을 적용한 실험 결과를 보이고, 5절에서 결론과 향후 계획을 설명한다.

2. 관련 연구

[5]에서는 트위터 사용자 개개인을 센서로 사용하여 지진, 태풍 등의 자연 이벤트를 탐지하는 시스템을 제안하였다. 지진이 발생했을 때 미리 지정한 이벤트 키워드로 트윗을 수집한 후, 트윗 단어의 개수로 대상 이벤트를 검출하고 트위터에서 제공하는 GPS 데이터와 트위터에 등록된 사용자의 위치를 이용해서 발생 위치를 추적하였다. 그러나 탐지할 수 있는 이벤트는 지진, 태풍과 같은 미리 지정된 이벤트 키워드 종류에 한정되기 때문에 입력된 이벤트 외에 새로운 이벤트 종류가 발생할 경우 이벤트 탐지가 불가능하다. 또한 위치정보를 제공하는 데 동의하지 않은 트위터 사용자의 경우 위치정보를 수집할 수 없으므로 이벤트 탐지에 필요한 정보로 적합하지 않다.

[6]에서는 범죄, 사고, 재해 관련 이벤트를 포함하는 트윗을 실시간으로 수집한 후, 이벤트 발생 공간과 시간 패턴을 분석하여 이벤트 발생 위치와 시간을 검출하는 시스템을 제안하였다. [2]와 마찬가지로 트위터 사용자의 GPS 데이터를 활용하여 이벤트 발생 장소를 알 수 있었고 트위터 반환 시간을 통해 이벤트 발생 시간을 얻을 수 있었다. 최종적으로 이벤트 발생 종류와 장소를 실시간으로 탐지하여 재난과 범죄의 경우, 이벤트 탐지 결과를 효과적으로 전파한다면 피해를 줄일 수 있었다. 하지만 트위터 사용자가 프로필에 입력한 위치정보와 실제 이벤트 발생 위치가 동일하다고 판단하기에는 무리가 있기 때문에 이벤트 발생 위치를 신용하기 어렵다.

[7]에서는 트위터 사용자 개개인을 이벤트 탐지의 센서로 사용하여 실시간으로 이벤트를 탐지하는 시스템을 제안하였다. 대다수의 이벤트들은 특정 장소에서 발생한다는 특성을 기반으로 미리 지정한 지명 키워드에 해당하는 트윗을 수집하고 지역별로 분류한 후에 언급 빈도가 급증한 지역을 대상으로 이벤트 후보지역을 선별하였다. 그다음 지역별로 수치값 계산을 통해 이벤트가 발생한 지역을 결정하였다. 미리 지정한 이벤트 키워드를 사용하지 않기 때문에 다양한 종류의 이벤트를 탐지할 수 있고 지명 키워드를 이용하기 때문에 이벤트 발생 위치를 정확하게 탐지할 수 있는 장점이 있다. 하지만 지명과 동형이의어 관계에 관한 노이즈제거에 대한 부분이 누락되어 있어서 시스템의 정확도를 낮추는 요인이 된다. 이러한 한계점을 보완하기 위해 본 논문에서는 지명과 관련한 노이즈제거를 통해 이벤트 탐지 정확도를 향상시키는 방안을 제시한다.

[8]에서는 한국어에서 동형이의어로 인해 발생하는 단어의 중의성을 해결하기 위한 방안을 제안하였다. 문장에서 어절의 동형이의어는 문맥 정보에 의해 결정된다는 가정을 통해 인접한 두 어절을 기계학습하여 중의성이 있는 어절의 동형이의어를 결정하였다. 어절 단위의 분석에서 자료 부족 문제를 해결하기 위해 자료 부족의 정도에 따라 다른 전이 모델을 적용하였고, 그 결과 높은 정확도를 얻을 수 있었다. 하지만 인접한 두 어절을 분석하기 때문에 어절이 독립적으로 사용될 경우 동형이의어를 결정할 수 없다. 따라서 본 논문에서는 지명이 독립적으로 사용되는 경우와 인접한 두 어절이 아닌 문장에서 연관 단어가 나타날 경우를 고려하여 동형이의어를 결정한다.

3. 기계학습을 통한 노이즈제거 기법

3.1 이벤트 탐지 과정

실시간으로 이벤트를 탐지하기 위해 트위터에서 제공하는 Twitter Streaming API[9]를 이용하여 트윗을 수집하였다. 수집한 트윗을 행정구역명과 지하철역명에 해당하는 지명 키워드별로 분류하고 분류된 데이터에서 언급 빈도가 급증한 지역들을 기반으로 이벤트가 발생한 지역을 이벤트 후보 지역으로 선별하였다. 리트윗으로 중복된 트윗이 반환되는 경우 언급 빈도가 증가하여 이벤트 후보지역으로 선별될 가능성이 있다. 따라서 최종적으로 중복된 트윗을 제거하여 최종 이벤트 지역으로 결정하였다. 제안하는 시스템이 실시간으로 동작하여 실제 발생한 이벤트를 탐지할 수 있음을 확인하였다. 그러나 특정 시간대에 반환된 지역들의 수에 대해서 실제 이벤트가 발생했던 지역의 수의 비율은 실험 결과 43%로 낮은 수치를 보였으며, 잘못 탐지된 지명으로는 강화, 고양, 달성 등이 있었다. 이외에도 시스템에서 지명 키워드로 사용된 170가지의 지역명 중에서 총 43가지의 지역명이 주로 지명 노이즈로 반환되었고 본 논문에서는 이 43개의 지역을 대상으로 노이즈제거 기법을 적용하였다. 전체적인 이벤트 탐지 과정은 Fig. 1과 같다.

3.2 지명 노이즈제거 필터링

지명이 이벤트 탐지 시스템을 통해 지명 노이즈로서 반환될 때 세 가지 특성을 갖는다. 첫째, 반환된 지명이 지명을 뜻하는 단어가 아닌 해당 지명과 같은 형태를 가진 동형이의어인 경우이다. 즉 다른 의미를 갖는 경우로 '달성'을 예로 들면, '달성'이 지명으로 사용될 경우 대구광역시 달성군의 의미를 갖지만 '목적한 것을 이룸'이라는 뜻의 동형이의어로 탐지될 수 있으므로 반환된 키워드가 지명과 동형이의어일 경우에 이를 노이즈로 판단한다. 둘째, 다른 구문에 포함되어 지명으로 사용되지 않는 경우이다. 즉 의미를 갖지 않는 경우로 '진도'를 예로 들면, '진도'가 지명으로 사용될 경우 전라남도 진도군의 의미를 갖지만 '사진도, 승진도, 공효진도'에서와 같이 '진도'가 다른 구문에 포함되는 경우에 '진도'는 단독으로 의미를 갖지 않고 지명이 아니므로 노이즈로 판단한다. 셋째, 노이즈인 경우에 자주 등장하는 연관 단어가 반환되는 경우이다. '강진'을 예로 들면, 지명으로 사용될 경우 전라남도 강진군을 나타내지만 같은 트윗 내에서 '규모, 진원'과 같은 연관 단어가 같이 반환되면 '강진'이 강한 지진이라는 의미로 사용된다. 이러한 경우에 '강진'이 지명으로 사용되지 않으므로 노이즈로 판단한다.

이 세 가지 특성 중에 한 가지 또는 두 가지 특성을 갖고 있는 경우도 있지만 세 가지 특성을 동시에 갖고 있는 경우도 있는데 키워드 '고양'이 이에 해당된다. '고양'이 지명으로 사용될 때 경기도 고양시의 의미를 갖지만, '고양'이 '정신이나 기분 따위를 북돋워서 높임'의 의미인 동형이의어로 사용되거나 구문에 포함되어 동물 '고양이'로 사용된 경우가 있다. 또한 트윗 내에서 키워드 '야옹'이 키워드 '고양'과 동

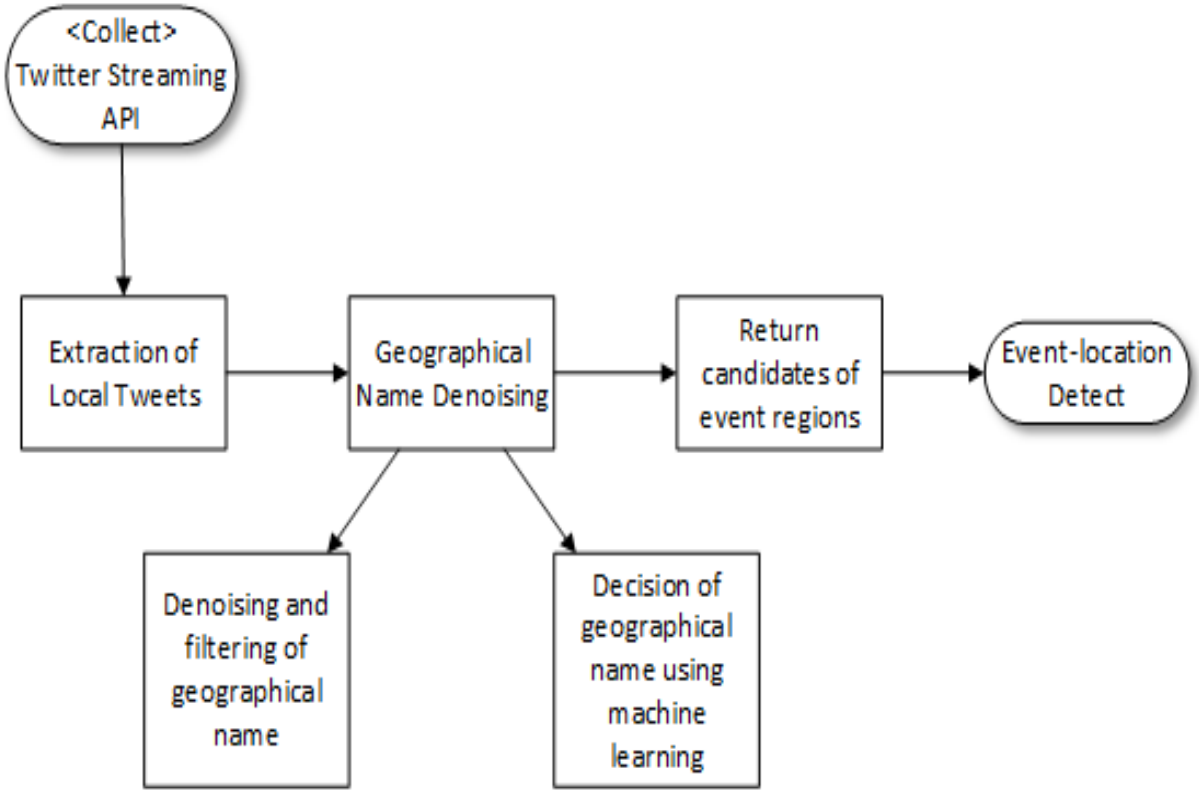

Fig. 1. Event Detection Process

시에 반환되는 경우 '고양'은 지명이 아닌 노이즈 '고양이'의 의미를 가진다. 지명 노이즈인 경우 구분을 쉽게 하기 위해 세 가지 특성으로 각 지명들을 분류하여 데이터베이스에 저장한다. Table 1은 세 가지 특성에 따라 지명 키워드가 노이즈로 반환된 경우를 나타낸 것이다.

Table 1. 3 Types of Denoising and Filtering

지명	지명과 동형이의어 관계	구문에 포함되는 경우	지명 노이즈 연관 단어
진도	진도를 빼다 진도를 나가다 진도를 맞추다	사진도 공효진도 경영진도	진앙 지진 규모
전주	전주곡 전주부분 전주들	예전주소 공전주기 축전주다	x
경주	경주하다 경주하는- 경주할-	월경주기 김경주 신경주사	경마 달리기

본 논문에서 노이즈제거 기법은 제거와 예측 순서로 진행한다. 우선 43개의 지명 키워드로 반환된 트윗이 노이즈인 경우에 특정 구문이나 연관 단어가 트윗 내에서 같이 반환된다. 앞서 설명한 것처럼 지명 키워드가 동형이의어일 때, 키워드가 명사/체언인 경우 '-를/-을/-들'와 같이 특정 조사가 같이 반환되고 키워드가 용언인 경우는 '-하다/-할/-하는'와 같은 특정 어미가 같이 반환된다.

반환된 키워드가 다른 구문에 포함되어 지명으로 사용되지 않는 경우에도 앞서 예로 들었던 지명 키워드 '진도'에서 '공효진도/사진도'와 같은 구문을 데이터베이스에 저장한다. 노이즈로서 반환될 때 같이 등장하는 조사, 어미, 구문, 연관 단어 등은 지명마다 차이를 보이기 때문에 지명별로 구분하여 데이터베이스에 저장해야 한다. 그리고 시스템에서 실시간으로 반환되는 트윗이 데이터베이스의 저장된 키워드와 일치하면 노이즈로 판단하고 필터링하여 제거한다. 제거 필터링 한 후에 일정량의 트윗 텍스트를 학습한 내용을 바탕으로 지명에 관한 새로운 트윗이 반환되면 기계학습을 통해 지명인지 아닌지 예측하는 방법을 사용한다.

3.3 기계학습을 통한 지명의 유무 예측

본 연구에서는 나이브 베이지안 분류 방법을 통해 기계학습을 적용했다. 나이브 베이지안 분류 방법은 하나 이상의 독립적인 속성들로부터 결과를 분류하는 확률적인 분류 방법이다. 각각의 독립적인 속성들이 결합되어 하나의 인스턴스를 이루며 여러 개의 인스턴스를 통해 분류할 결과에 대한 확률을 계산한다. 지명으로 사용될 경우에 나타나는 특정한 규칙을 나이브 베이지안 분류에서 독립적인 속성으로 정하고 분류 결과를 지명의 유무로 나타낸다. 나이브 베이지안 분류는 베이즈 룰을 기본적으로 사용하고 있다. 베이즈 룰은 각각의 속성을 $x_1, x_2, \cdots, x_n$로 나타내고 분류 결과를 클래스 C로 표현하였을 때, C의 부류인 지명 유무에 대한 확률은 값 c로 표현하였다. 확률 c의 값은 Equation (1)로 정의된다[9-10].

$$c = \underset{c \in C}{\arg\max}\ P(x_1, x_2, \cdots, x_n|c)\ P(c) \quad (1)$$

각 부류에 해당하는 조건부 확률을 계산할 수 있으며 가장 높은 확률을 가지는 부류가 지명의 유무를 결정한다. 나이브 베이지안 분류는 베이즈 룰의 속성이 조건부 독립이라는 가정하에 조건부 확률을 구하여 최종 공식은 Equation (2)로 단순화할 수 있다.

$$c = \underset{c \in C}{\arg\max}\ \Pi P(x|c)\ P(c) \quad (2)$$

트윗 텍스트를 나이브 베이지안 분류에 적용했을 때에 해당하는 분류 속성은 세 가지로, 주로 지명인 경우에 나타나는 조건이다. 각 조건은 Table 2와 같다. 분류 결과에 대한 확률 c가 높은 부류가 지명의 유무를 결정한다. 지명의 유무에 해당하는 클래스를 결정하기에 앞서 실험 결과 노이즈 제거 대상 지명의 경우 시스템에서 반환된 트윗에 대해서 실제 지명인 경우의 비율이 매우 낮다. 이는 리트윗을 포함하여 중복된 트윗의 개수가 많고 지명 노이즈로 반환된 경우가 대부분이기 때문이다. 따라서 중복된 트윗을 제거하고 세 가지 특성으로 노이즈 필터링 한 후 나머지 트윗을 대상으로 기계학습을 하였다.

Table 2. Attributes of Naive Bayesian Classifier

Classification	Attribute name	example
지명인 경우	지명 연관 단어가 반환되는 경우	경주 박물관, 불국사, 안압지, 경주역
	지명과 관련된 조사가 반환되는 경우	경주에서, 경주에
	지명이 독립적으로 사용되는 경우	_경주_
지명이 아닌 경우	-	-

나이브 베이지안 확률 모델은 일정량의 트윗 데이터를 통해 미리 구해둔 확률값으로 분류한다. 나이브 베이지안 분류에 사용된 속성은 지명과 관련된 연관 단어나 지명과 관련된 조사가 반환되는 경우, 그리고 지명이 독립적으로 사용되는 경우 세 가지로 지역이었을 때 자주 나타나는 조건들을 이용하였다. 지역명 '경주'를 예로 들면, 첫 번째 속성인 지명과 연관 단어의 경우는 지역명 '경주'를 식별하는 데 목표물로서 랜드마크를 주로 사용했으며 관광지나 학교, 건물 이름 등이 해당된다. 랜드마크 외에도 연관 단어에는 '경주역'도 포함된다. 두 번째 속성인 지명과 함께 나타나는 조사로는 '경주'의 경우 '경주에서/경주에'가 있고 해당 키워드가 나타나면 지명일 확률이 높아진다. 분석 결과, 조사는 각 지명마다 조금씩 차이를 보이기 때문에 지역별로 조사를 따

로 저장하였다. 세 번째 속성인 지명이 독립적으로 사용되는 경우는 지명 키워드의 앞뒤로 공백이 발생하는 것을 말하는데, 실험 데이터로 분석한 결과 지명 키워드가 독립적으로 반환되는 경우 실제로 지명일 확률이 높았다. 이 세 가지 속성을 이용해서 실험 데이터를 통해 각 속성에 해당하는 확률값을 구하고 새로운 트윗이 발생하면 Equation (2)를 이용하여 분류의 확률값이 지명인 경우가 지명이 아닐 경우보다 높으면 지명으로 분류된다.

4. 실험 결과

본 논문에서 제시한 노이즈제거 기법을 적용할 경우 세 가지 특성을 이용한 노이즈 필터링과 기계학습을 이용한 확률값으로 새로운 트윗이 나왔을 때 지명을 검출하는 방법을 동시에 적용한다. 실험을 위해 2014년 1월 1일부터 2014년 12월 31일까지 12개월간 수집한 트윗 데이터를 사용하였다.

4.1 신뢰도를 이용한 성능평가

실험에 앞서 시스템이 실행되고 나서 실시간 동작함을 확인하였고 실험에 사용된 PC의 성능은 Table 3과 같다.

Table 3. Experimental Environment

CPU	INTEL Quad Xeon 3.2GHz
HDD	1TB
RAM	4.00GB
OS	Windows 8

제거 필터링을 위한 데이터베이스 구축과 기계학습 데이터를 얻기 위해서 각 지명마다 수집한 트윗을 대상으로 실험하였다. 우선 중복 제거된 500개의 트윗을 이용해서 제거 필터링을 위한 데이터베이스를 구축하고, 제거된 나머지 트윗을 이용해서 나이브 베이지안 확률에서 분류와 단어의 빈도수 계산을 통해 확률값을 구한다. 이는 일정량의 트윗에 나타나는 키워드 종류가 반복되는 형태이기 때문에 적당량의 표본 데이터 개수로 정하였다. 검증을 위해 같은 개수의 지역별 트윗을 사용하였고 대부분의 지명에서 데이터베이스를 통한 제거 필터링을 통해 50% 이상의 노이즈를 제거하였다.

지명 키워드 '강진'의 경우에도 실험 데이터를 이용해 데이터베이스에 저장한 키워드로 '강진'과 동형이의어 관계에 해당하는 트윗을 우선적으로 제거하고 '강진'이 다른 구문에 포함되어 지명으로 사용되지 않는 경우, 노이즈와 관련된 연관 단어가 같이 반환되는 경우를 순서대로 제거한다. 이 과정에서 검증에 사용된 키워드 '강진'을 포함한 500개의 트윗 중에 295개의 지명이 노이즈로 선 제거되었다.

나머지 205개의 트윗을 대상으로 나이브 베이지안 분류기를 통해 지명의 유무를 판단한다. 결과적으로 205개의 트윗 중 분류기를 통해 정확하게 판단한 경우는 151개의 트윗이 해당되고, 이는 분류기가 지명으로 판단했을 때 실제로 지명인 경우와 노이즈로 판단했을 때 실제로 노이즈인 경우를 합한 개수이다. 분류기를 통해 틀리게 판단한 개수는 54개로, 지명으로 판단했을 때 실제로는 노이즈인 경우와 노이즈로 판단했을 때 실제로 지명인 경우를 합한 개수이다. 제거와 예측을 종합하면 노이즈제거 기법을 적용할 경우 옳게 판단한 경우는 노이즈제거 필터링을 통해 노이즈로 제거한 개수와 나이브 베이지안 분류기를 통해 제대로 분류한 개수의 합이다. 반대로 틀리게 판단한 경우는 나이브 베이지안 분류기를 통해 틀리게 판단한 개수이다. 각 지명마다 분류기가 제대로 동작하는지 신뢰도를 사용해 나타낼 수 있고 Equation (3)과 같다. 여기서 A는 제거 필터링을 통해 실제 지명이 아닐 때 노이즈로 필터링한 개수를 의미하고, B는 분류기를 통해 실제 지명일 때 지명으로 판단한 개수와 실제 지명이 아닐 때 노이즈로 판단한 개수의 합을 나타낸다. 따라서 500-(A+B)는 틀리게 판단한 개수이다.

$$Reliability(\%) = \frac{A+B}{500} \times 100(\%) \quad (3)$$

'강진'의 경우 A는 295개로 제거 필터링 된 개수이고 B는 151개로 분류기를 통해 옳게 분류한 개수이다. 틀리게 분류한 개수는 54개로 500개의 트윗에서 A, B를 뺀 개수이다. Equation (3)을 이용해서 '강진' 지명의 신뢰도를 계산하면 89.2%이다. 이 수치는 '강진' 지명에 한에서 새로운 트윗이 발생할 경우 지명 유무를 판단하는 척도가 될 수 있다. Fig. 2는 '강진, 강화, 진도, 경주' 지명의 신뢰도 값을 나타낸다. 네 가지 지명을 포함한 43개의 전체적인 지명에 대해서 지명 노이즈제거 기법을 적용하지 않았을 경우 각 지명들의 신뢰도

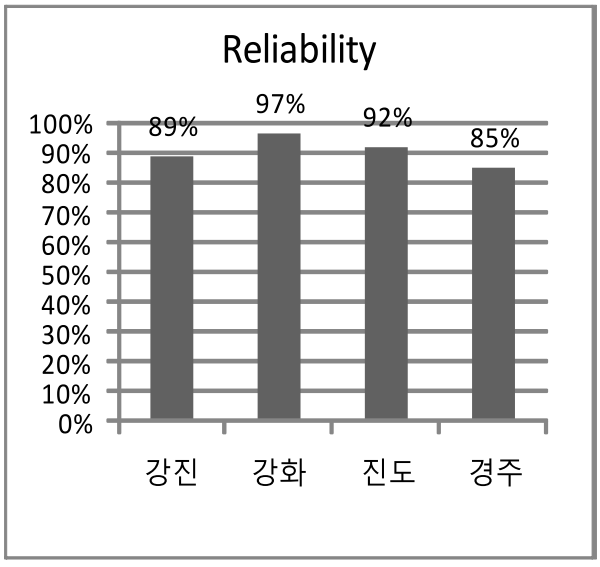

Fig. 2. The Reliability According to Geographical Name

평균은 27%이며, 지명 노이즈제거 기법을 적용할 경우 신뢰도 평균은 89.6%로 62.6% 상승하였다. 이는 43개의 지역이 다른 지역보다 지명 노이즈가 많은 지역이므로 제거 필터링을 적용하여 다수의 노이즈를 제거할 수 있었기 때문이다. 실험 결과 시스템에서 각 지명들의 신뢰도 향상을 통해 본 논문에서 제안한 지명 노이즈제거 기법의 필요성을 확인하였다.

본 논문의 노이즈제거 기법을 적용해서 이벤트 지역을 결정할 경우 나이브 베이지안 분류에서 실제 지명인 경우의 비율이 너무 낮게 나오면 이벤트 지역을 결정할 수 없다. 따라서 대부분의 반환된 트윗이 지명 노이즈일 때 나이즈 베이지안 분류를 적용할 수 없는데, 이런 예외적인 상황으로는 지명 '달성'을 예로 들 수 있다. '달성' 키워드를 포함한 트윗이 시스템에서 반환될 때 트윗은 동형이의어인 '달성하다'로 사용된 경우가 대부분이다. 실험 결과 일정량의 트윗에서 지명으로 사용된 경우의 비율이 약 0.0001%로 나타났다. 이 경우 노이즈 관련 데이터베이스를 구축하여 제거 방법을 사용하는 것보다 반대로 지명인 경우를 지정하여 지명을 검출하는 것이 효율적이다. '달성'은 '달성군'으로 반환될 때 지명으로 사용되는 경우가 대부분이기 때문에 예외적으로 이벤트를 탐지할 때 '달성군'으로 반환되는 경우에만 이벤트 지역으로 판단한다.

'달성'과 마찬가지로 '공주, 기장, 영광, 영양, 예산, 음성' 지명에서 대부분의 트윗이 지명 노이즈로 반환된다. 따라서 위의 6개의 지역들도 '달성'과 같이 실험 데이터에서 구한 지명인 경우 등장하는 연관 단어나 구문이 새로운 트윗에서 반환되면 지명으로 결정하는 방식을 적용한다. 그리고 지명 '남해'의 경우 경상남도 남해군으로 사용되지만 '남해'가 남쪽 해안의 의미로 사용될 경우에는 노이즈로 판단하기 위한 경계가 모호하므로 노이즈제거 대상 지역에서 제외하였다. 또한 지명 '진주'에서는 '진주'가 사람 이름으로 많이 반환되었고 지명 '함양'에서는 '-해야함양'과 같은 인터넷 신조어로 많이 등장하였다. 이러한 경우에 일정 패턴이 없고 계속해서 새롭게 발생되므로 기계학습이 불가능하다.

4.2 시스템 정밀도와 재현율

시스템에서 탐지한 지명들의 정확성을 측정하기 위한 평가 기준으로서 정밀도(Precision)와 재현율(Recall)을 사용하였다. 정밀도와 재현율은 특정 시간대에 발생한 지명을 포함한 트윗 200개에 대해서 측정하였다. 정밀도는 시스템에서 탐지한 다수의 이벤트 중에서 실제 지명인 트윗의 비율로 나타내고 재현율은 지명과 관련된 트윗 중에서 실제 시스템에서 탐지한 이벤트의 비율로 나타낸다. 정밀도는 Equation (4)와 같다. 여기서 tp는 탐지한 지명 중에서 실제 지명인 경우의 개수를 의미하고, fp는 탐지한 지명 중에서 실제 지명이 아닌 경우의 개수를 의미한다. 따라서 tp + fp는 시스템에서 탐지한 지명의 총 개수를 의미한다.

$$Precision(\%) = \frac{tp}{tp+fp} \times 100(\%) \quad (4)$$

지명 노이즈제거 기법을 적용하지 않았을 때 Equation (4)에 의해 계산된 시스템의 정밀도는 약 43%로 낮게 나타났다. 이는 반환된 키워드가 지명과 동형이의어 관계이거나 지명이 아닌 다른 의미인 경우가 많았기 때문이다. 한편 본 논문에서 제안한 지명 노이즈제거 필터링을 적용한 후에 실험 데이터인 200개의 트윗에 대해서 계산된 시스템의 정밀도는 82%로, 이전보다 39% 높게 나타났다. 이렇게 정밀도가 향상된 이유는 주로 많이 발생하는 지명과 관련된 노이즈를 제거 필터링을 통해 없앴기 때문이다. 현재 정밀도가 100%가 되지 않는 이유는 노이즈제거 대상 지명에 해당하는 43개의 지명 외에도 노이즈가 발생한 경우가 있었고, 43개 지명의 실험 데이터에 나타나지 않은 새로운 지명 노이즈가 발견되었기 때문이다. 따라서 시스템의 정밀도 향상을 위해서는 노이즈 관련 데이터베이스와 기계학습을 위한 확률값의 최신화가 필요함을 확인하였다.

재현율은 Equation (5)와 같다. 여기서 tp는 실제 지명을 포함하는 트윗 중에서 시스템이 탐지한 개수를 의미하고, fn은 실제 지명을 포함하는 트윗 중에서 시스템이 탐지하지 못한 개수를 의미한다. 따라서 tp + fn은 실제 지명을 포함하는 트윗의 총 개수를 의미한다.

$$Recall(\%) = \frac{tp}{tp+fn} \times 100(\%) \quad (5)$$

Equation (5)에 의해 계산된 시스템의 재현율은 37%로 계산되었고, 본 논문의 지명 노이즈제거 필터링을 적용한 후의 재현율은 46%로 나타났다. 노이즈제거 필터링을 적용할 경우 재현율의 상승폭은 9%로 정밀도의 상승폭보다 낮게 나타났다. 이는 지명을 포함하는 트윗의 대부분의 경우 시스템에서 탐지하지 못한 이벤트는 노이즈제거 필터링을 적용한 후에도 탐지하지 못했기 때문에 재현율이 비슷하게 나타난 것이다. 시스템에서 이벤트 탐지를 위해 언급 빈도가 급증한 지역들을 이벤트 지역으로 정하기 때문에 적절히 수치를 조정하여 전체적으로 재현율은 낮게 측정되었다.

이러한 정밀도와 재현율을 하나의 지표로 통합하여 정확성을 측정하는 방법으로 F-Measure가 있다. F-Measure는 정밀도와 재현율의 트레이드오프를 잘 통합하여 정확성을 한 번에 나타내는 지표로, 보통 가중치를 가진 조화평균이라고도 한다. F-Measure를 구하기 위해 정밀도와 재현율에 대한 조화 평균에 가중치 알파를 적용하면 Equation (6)과 같다. 여기서 P는 정밀도를 나타내고 R은 재현율을 나타낸다.

$$F(\%)=\frac{1}{\alpha\frac{1}{P}+(1-\alpha)\frac{1}{R}}\times\frac{(\beta^2+1)PR}{\beta^2P+R}(\%) \quad (6)$$

F1-Measure는 정밀도와 재현율의 중요성을 동일하게 부여하여 베타값은 1이 되며 최종 식은 Equation (7)과 같다.

$$F_1(\%)=2\times\frac{P\times R}{P+R}\times 100(\%) \quad (7)$$

정밀도와 재현율을 이용하여 Equation (7)을 통해 계산된 F1-Measure는 40%로 계산되었고 지명 노이즈제거 필터링을 적용한 결과 59%로 이전보다 19% 높게 나타났다.

5. 결론 및 향후 연구

본 논문에서는 트위터를 이용한 이벤트 탐지에서의 지명과 관련된 노이즈를 제거하는 기법을 제시하였다. 현실에서 발생하는 이벤트는 공간과 시간적 특성을 갖고 있기 때문에 지명을 키워드로 트윗을 수집하여 이벤트의 장소와 이벤트 내용을 검출하려고 할 때, 반환된 키워드가 지명과 동형이의어나 구문에 포함된 다른 의미로 사용될 경우, 그리고 지명이 아닐 경우의 연관 단어가 같이 반환되면 지명 노이즈로 판단하고 이를 제거하였다. 제거 필터링 후에 나머지 트윗에서 키워드의 지명 유무를 예측하기 위해 기계학습을 이용하였다. 지명과 관련된 조사와 지명 연관 단어가 같이 반환되는 경우와 지명이 독립으로 사용될 경우를 각각 독립된 속성으로 가정한 나이브 베이지안 확률 모델을 적용했으며, 실험 데이터를 통해 지명 유무에 해당하는 확률값을 구하고 새로운 트윗이 발생할 때 지명 유무를 결정할 수 있었다. 시스템에서 새로운 트윗이 반환될 때 지명 유무를 판단하는 척도가 되는 신뢰도는 약 89.6%로 지명 노이즈제거 기법의 필요성을 보였다. 그리고 지명 노이즈제거 기법을 적용할 경우 시스템의 F1-Measure가 40%에서 59%로 증가함을 확인하였다.

본 논문에서 제시한 지명 노이즈제거 기법을 이용하여 지명과 관련된 동형이의어를 필터링 하여 이벤트 탐지 시스템의 정확도를 높일 수 있었다. 동형이의어의 중의성을 해결하기 위한 기존의 연구는 보통 연속된 어절의 연관성을 이용하여 동형이의어를 결정하였다. 하지만 본 논문에서는 연속된 어절뿐만 아니라 한 문장 내에서 나타나는 지명 연관 단어와 다수의 지명 키워드가 독립적으로 사용되는 특징을 이용하여 기계학습을 통해 지명을 추출하는 방법을 사용하였다. 이러한 지명에 특화된 지명 노이즈제거 기법은 한국 지명과 관련된 다른 연구에서도 사용될 것으로 기대된다.

향후 연구 과제로는 지명 노이즈제거에서 어려움이 있었던, 지명 키워드가 사람 이름이거나 인터넷 신조어에 포함되어 일정한 패턴을 갖지 않는 경우의 노이즈제거 연구가 필요하다. 또한 시스템에서 이벤트 장소와 탐지 시간 이외에도 이벤트 내용을 효과적으로 전파하기 위한 연구를 할 계획이다.

References

[1] Statistic Brain, Twitter Statistics [Internet], http://www.statisticbrain.com.

[2] E. Lee, J. Kim, and D. Baik, “An Evaluation Method for Contents Importance Based on Twitter Characteristics,” *Journal of KIISE*, Vol.41, No.12, pp.1136-1144, 2014.

[3] T. Bayar and K. Lee, “Extracting Core Events Based on Timeline and Retweet Analysis in Twitter Corpus,” *KIPS Transactions on Software and Data Engineering*, Vol.1 No.1, pp.69-74, 2012.

[4] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a Social Network or a News Media?,” *Proc. of the 19th International Conference on World Wide Web*, pp.591-600, 2010.

[5] T. Sakaki, M. Okzaki, and Y. Matsuo, “Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors,” *Proc. of the 19th International Conference on World Wide Web*, pp.851-860, 2010.

[6] R. Li, K. H. Lei, R. Khadiwala, and K. Chang, “TEDAS: a Twitter Based Event Detection and Analysis System,” *Proc. of the IEEE 28th International Conference on Data Engineering*, pp.1273-1276. 2012.

[7] J. Yim, J. Yoon, B. Lee, and B. Hwang, “Designing of Event Decision Module using Twitter,” *Proc. of Korea Computer Congress*, pp.248-250, 2013.

[8] J. Shin and C. Ock, “A Stage Transition Model for Korean Part-of-Speech and Homograph Tagging,” *Journal of KIISE*, Vol.39 No.11, pp.889-901, 2012.

[9] Twitter Streaming API [Internet], http://dev.twitter.com/docs/streaming-apis.

[10] W. lan H, F. Eibe, and H. Mark A, “Data Mining,” 3rd ed., Morgan Kaufmann, pp.594-595, 2011.

[11] J. Yim and B. Hwang, “Predicting Movie Success based on Machine Learning Using Twitter,” *KIPS Transactions on Software and Data Engineering*, Vol.3 No.7, pp.263-270, 2014.

우 승 민

e-mail : simter@catholic.ac.kr

2011년~현 재 가톨릭대학교 컴퓨터정보공학부 학사과정

관심분야 : 소셜네트워크분석, 데이터베이스, 데이터마이닝, 정보검색

황 병 연

e-mail : byhwang@catholic.ac.kr

1986년 서울대학교 컴퓨터공학과(학사)

1989년 KAIST 전산학과(석사)

1994년 KAIST 전산학과(박사)

1994년~현 재 가톨릭대학교 컴퓨터정보공학부 교수

1999년~2000년 (美) 미네소타대학교 방문교수

2007년~2008년 (美) 캘리포니아주립대학교 방문교수

관심분야 : 소셜네트워크분석, XML 데이터베이스, 정보검색, 데이터마이닝, 지리정보시스템