마이크로 블로그기반의 공간 지식 추출 기법연구

A Technique for Extracting GeoSemantic Knowledge from Micro-blog

하 수 욱* 남 광 우** 류 근 호***
Su Wook Ha Kwang Woo Nam Keun Ho Ryu

요 약 최근 ISO/TC211, OGC, INSPIRE 등 국제기구들을 중심으로 시맨틱 기술을 활용한 공간정보의 공유 노력이 진행되고 있다. 또한 스마트폰의 대중화와 소셜 네트워킹 서비스의 활성화로 인해 온라인 소셜 커뮤니티에서 이슈를 추출하기 위한 연구들이 이루어지고 있다. 그러나 응용 수준에서 가용한 공간정보 온톨로지는 부족한 실정이며, 소셜 네트워크 서비스에서의 공간정보 추출 역시 텍스트 마이닝을 통한 지오코딩 부분에 집중되어 있다. 따라서 소셜 미디어 정보에서 공간 현상을 추출하여 시맨틱 공간 지식으로 변환하기 위한 방법은 매우 유용하게 활용될 수 있다. 또한 공간 현상을 단순한 빈발 키워드가 아닌 연관 이슈의 형태로 사용자에게 제공함으로써 공간상에 발생하는 이슈에 대한 이해도를 향상시킬 수 있을 것이다. 따라서 본 논문에서는 소셜 미디어 서비스의 하나인 마이크로 블로그를 기반으로 데이터를 수집하여 데이터 마이닝 기술을 접목하여 연관 이슈를 추출하고, 이를 시공간 지식으로 변환하기 위한 공간 이슈 온톨로지 모델을 제안하였다. 이를 통해 향후 관련 시스템의 개발을 위한 참조모델 및 공간 온톨로지 구축을 위한 모델로써 유용하게 사용될 수 있을 것으로 기대된다.

키워드 : 시공간 GIS, 지오시맨틱스, 마이크로 블로그, 데이터 마이닝

Abstract Recently international organizations such as ISO/TC211, OGC, INSPIRE (Infrastructure for Spatial Information in Europe) make an effort to share geospatial data using semantic web technologies. In addition, smart phone and social networking services enable community-based opportunities for participants to share issues of a social phenomenon based on geographic area, and many researchers try to find a method of extracting issues from that. However, serviceable spatial ontologies are still insufficient at application level, and studies of spatial information extraction from SNS were focused on user's location finding or geocoding by text mining. Therefore, a study of extracting spatial phenomenon from social media information and converting it into geosemantic knowledge is very usable. In this paper, we propose a framework for extracting keywords from micro-blog, one of the social media services, finding their relationships using data mining technique, and converting it into spatiotemopral knowledge. The result of this study could be used for implementing a related system as a procedure and ontology model for constructing geoseemantic issue. And from this, it is expected to improve the effectiveness of finding, publishing and analysing spatial issues.

Keywords : Spatiotemporal GIS, GeoSemantics, Micro-blog, SNS, Data Mining

1. 서 론

최근 ISO/TC211, OGC(Open Geospatial Consortium) 등 국제 표준화 기구와 INSPIRE (Infrastructure for Spatial Information in Europe)를 중심으로 공간정보와 시맨틱 기술의 접목을 통해 기존의 표준 기반의 문법적 상호운용성 확보와 함께 의미적 상호운용성(Semantic Interoperability)을 보장함으로써 이종의 공간정보들을 연계/공유하기 위한 노력들이 진행되고 있다[22]. 그러나 대부분의 연구가

† 본 연구는 지식경제부의 지원을 받는 정보통신기술력향상사업의 연구결과이며, 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2009-0067958).

* 충북대학교 데이터베이스연구실, 한국전자통신연구원 선임연구원 suwook.ha@etri.re.kr
** 군산대학교 컴퓨터정보공학과 부교수 kwnam@kusan.ac.kr(공동 교신저자)
*** 충북대학교 소프트웨어학과 교수 khryu@dblab.chungbuk.ac.kr(공동 교신저자)

공간정보 도메인 온톨로지, 시맨틱 공간 피처 모델, 시맨틱 검색 등을 활용한 공간 DB 연동에 집중되어 있어 어플리케이션 레벨에서 활용 가능한 개방된 시맨틱 공간지식은 사실상 부족한 현실이다.

또한 스마트폰의 대중화와 소셜 네트워크 서비스(이하 'SNS')의 및 확산은 언제, 어디서, 어떤 디바이스로나 음성과 문자, 사진, 영상통화를 주고받을 수 있는 소셜 중심의 사회로의 변화를 이끌어 가고 있다. 트위터의 경우 2012년 3월 기준 하루 약 3억 4천만 건이 등록[9]됨으로써 다양한 사회 현상들을 반영하고 있으며, 기술 내용이 140 문자로 제한됨으로써 효과적인 사회 현상 분석 도구로 사용될 수 있을 것으로 주목받고 있다. 이에 SNS에서 이슈를 추출하여 시간의 경과에 따른 변화를 모니터링, 활용하기 위한 노력들이 진행되고 있으며, 공간정보 분야에서도 지명, POI 온톨로지 등을 이용하여 사용자 위치를 추정하기 위한 연구들이 시작되고 있다. 그러나 단일 이슈(키워드)를 통해 정보를 유추하기 위해서는 해당 키워드가 발생한 원인에 대한 경험적인 부가 지식이 필요하며, 이를 공간상에 발생하는 현상 정보로 가공하여 활용하기 위한 연구는 부족한 것이 현실이다.

따라서 본 논문에서는 공간적 범위 설정을 통해 수집된 SNS 정보를 바탕으로 이슈들을 추출하고, 데이터 마이닝 기술을 이용하여 이슈들 간의 연관성을 분석함으로써 SNS 기반의 공간 현상 정보를 추출하기 위한 방안을 제시하였다. 또한 이를 시맨틱 공간지식(geoSemantic knowledge)으로 가공함으로써 온라인을 통해 다양한 유저들이 링크, 활용하기 위한 시맨틱 공간 이슈 온톨로지(geospatial issue ontology)를 제안하였다.

2. 관련 연구

2.1 SNS에서의 공간정보 추출

SNS에서 공간정보를 추출하기 위한 두 가지 유형의 연구가 진행되고 있다. 첫 번째는 '공간 연관 단어 기반의 위치 추정' 방법[3, 5]에 대한 것으로 프로필에 포함된 사용자의 거주 지역 또는 GPS 기반의 위치 정보와 사용자가 입력한 정보에서 지명 또는 공간과 연관된 명사를 추출하여 '위치-대표단어'와 같은 시맨틱 연관 관계를 만들어 지오코딩에 활용하는 것이다. 예를 들어 'Casino'라는 단어의 발생 빈도와 위치 값을 분석한 결과 'Las Vegas'의 위치와 연관이 있는 것으로 나타나기 때문에, 향후 Casino라는 단어를 입력한 사용의 위치는 Las Vegas일 확률이 높다는 것이다.

두 번째 방법은 '소셜 그래프 기반의 위치 추정' 방법[1, 2]에 대한 연구로 특정 사용자의 위치정보를 알 수 없을 경우 해당 사용자의 소셜 그래프를 통해 친구 관계를 맺고 있는 사람들의 프로필 정보를 추출하고, 관계의 레벨에 따른 가중치를 부여하여 이를 바탕으로 위치를 추정해 내는 것이다. 두 방법 모두 별도의 부가적인 수단을 사용하지 않는 사용자 위치 추정 방법을 제안하고 있으나 국가 규모(미국의 경우 주 레벨 이상)의 위치 정확도를 보장하기 어렵다는 단점을 가지고 있다.

2.2 SNS에서의 이슈 추출

SNS에서 이슈 추출을 위한 가장 일반적인 방법은 최다 검색어 또는 특정 키워드 검색을 통해 해당 단어가 전체 사용자들이 작성한 정보들 중에서 얼마나 많이 사용되고 있는가를 추출하여 그 빈도의 변화를 모니터링하는 방법이 통용되고 있다. 최근에는 별도의 사용자 입력값 없이 자연어 처리를 통해 등록된 정보들 중에서 단어의 사용 빈도에 따른 가중치를 부여하고, 이들이 시간에 따라 발생하고 소멸하는 패턴을 분석하여 이슈를 추출하거나 또는 온톨로지를 통해 추출된 이슈들 간의 연관성을 분석하는 연구들이 진행되고 있다[4, 14, 16].

또한 시간의 변화에 따라 민감하게 변화하는 SNS 정보의 특성을 고려하여, 슬라이딩 윈도우를 도입함으로써 분석 대상 단어의 모집단이 변하기 때문에 발생하는 재계산을 위해 요구되는 비용을 최소화하기 위한 연구가 추진되었다[14].

2.3 SNS 스팸 필터링

온라인 광고 또는 특정 사안을 이슈화하기 위해 하나의 아이디를 통해 동일한 내용을 여러 번 배포하는 등의 스팸이 SNS에서 성행하고 있다. 이를 필터링하기 위한 방법과 또 이러한 필터링을 피하기 위한 양자 간의 대결을 통해 스팸 필터링 방법이 다양화되고 있다[13, 15, 18, 19]. 스팸 필터링 방법은 세 가지 유형으로 분류되는데 첫 번째 방법은 사용자 가 작성한 정보 내에 포함되어 있는 URL을 이용한 것으로, 동일 사용자 ID를 대상으로 최근 등

록되는 정보들이 URL을 얼마나 포함하고 있는지를 분석하여, 특정 비율 이상일 경우 해당 사용자 ID를 필터링한다. 두 번째는 사용자 ID를 이용한 방법으로, 최근 등록되는 정보들 중에서 사용자 ID가 얼마나 포함되어 있는지에 대한 빈도를 분석하여 일정 비율 이상일 경우 해당 등록 정보를 필터링한다. 이는 스패머들이 URL을 포함한 온라인 광고에서 주로 관계없는 사용자 ID를 포함하여 해당 내용이 마치 다른 사람의 의견에 대한 재배포 또는 회신인 것처럼 포장함으로써 스팸 처리 되는 것을 방지함에 따라 이를 추출해 내기 위한 방법으로 고안되었다. 마지막은 특정 사용자가 유사한 내용의 정보를 지속적으로 등록하는지를 분석하여 필터링 하는 방법으로 작성된 내용을 구성하는 개별 단어들이 일정 비율 이상 동일할 경우 해당 ID를 필터링한다.

2.4 기존연구로 부터의 시사점

이상 살펴본 바와 같이 소셜 미디어 스트림에서 공간/비공간적 이슈를 추출하기 위한 노력들이 진행되고 있으며, 이를 위해 자연어 처리, 도메인 온톨로지 사용, 스팸 필터링 등의 연구들이 진행되어 오고 있다. 그러나 위치 추정의 경우 낮은 정확도와 함께 스마트폰에 탑재된 GPS의 활용도가 높아지고, 유무선 IP를 기반으로 위치를 추정하기 위한 연구들이 진행되면서 높은 효용성을 기대하기 어려우며, 시맨틱 기술을 이용한 경우 일반적으로 통용되는 지식이 아니거나 또는 가용한 도메인 온톨로지가 존재하지 않을 경우 SNS에서 불규칙하게 발생하는 특정 이슈들 간의 연관성을 추출하기 어렵다는 단점을 가지고 있다. 또한 스팸으로 추정되는 정보들을 배제하더라도 높은 빈도로 나타나는 사용자들 간의 일반적인 안부 혹은 대화들은 분석 결과의 품질을 저하시킬 수 있다.

따라서 이 논문에서는 지금까지 진행되어온 마이크로 블로그를 통해 데이터를 수집하여 키워드를 추출하는 SNS 마이닝 기법을 확장한다. 공간적 범위를 기반으로 연관 이슈를 추출하기 위한 절차와 방법, 그리고 이를 시맨틱 공간정보로 변환하기 위해 필요한 공간 온톨로지 모델은 웹 환경에서 공간정보의 자유로운 공유를 돕는데 기여할 수 있을 것이다.

3. 마이크로 블로그에서의 공간 지식 추출

3.1 공간상의 연관 지식 추출 절차

일반적으로 이슈는 '논의의 주제' 로 정의된다. 본 연구에서는 이를 '키워드들의 연관'으로 정의하고, 주어진 공간적 범위에서 시간의 흐름에 따라 발생하는 이슈를 '공간 현상 정보'로 정의한다.

SNS에서 공간 현상 정보를 추출하기 위한 절차는 그림 1과 같다. 먼저 공간적 영역을 기반으로 데이터를 수집할 수 있는 서비스를 선정한다. 그 다음 문장을 이루는 각 단어들을 추출하고 이들의 형태소를 분석하여 연관성 분석에 사용되는 데이터 형태로 변환한다. 수집된 정보 내에 포함되어 있는 단어들 간의 연관성을 데이터 마이닝을 이용하여 분석하고 그 결과를 저장한다. 이를 시간적 범위, 연관도(신뢰도, 지지도) 등을 바탕으로 질의하여 결과를 도출한다.

이러한 과정을 통해 추출된 결과는 수집 공간을 바탕으로 시간적 범위를 갖는 현상정보로 가공할 수 있으며, 따라서 시맨틱 시공간 피처 온톨로지로 변환함으로써 온라인 공간지식으로 활용할 수 있다.

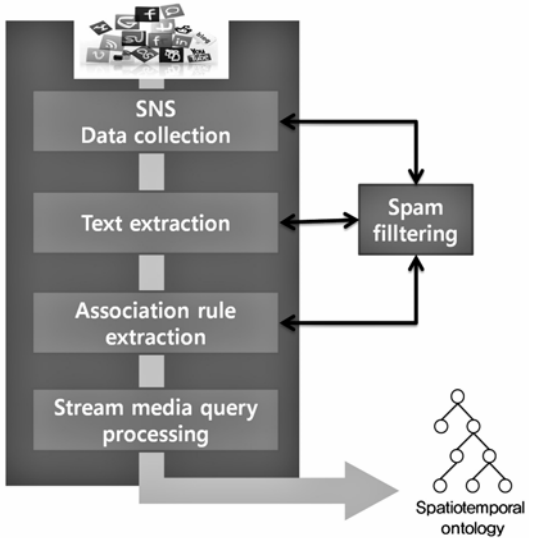

그림 1. 마이크로 블로그에서의 공간 지식 추출 절차

3.2 SNS 데이터 수집/전처리

데이터 수집을 위한 공간적 범위 설정은 분석 결과를 기반으로 공간 피처를 생성하기 위해 반드시

필요하다. 또한 공간적 범위를 제한함으로써, 수집되는 데이터의 규모를 관리할 수 있다. 본 논문에서는 데이터 수집을 위한 두 가지 방법을 제안한다. 첫 번째는 사용자 프로필의 거주 지역을 기반으로 소셜 네트워크를 구성하여, 데이터를 수집하는 방법이다. 트위터의 경우 도시별 사용자 순위를 제공하는 사이트[10]가 운영되고 있으며, 이를 바탕으로 수집 대상 사용자 목록을 팔로잉하여 데이터를 수집할 수 있다.

그림 2. 도시별 인규 규모 및 지역별 트위터 사용자 순위 사이트

그림 3. Open API 기반 SNS 데이터수집

두 번째는 공간적 범위 기반의 데이터 수집 Open API를 이용하는 방법이다. 최근 구글을 중심으로 OpenSocial을 통해 다양한 SNS 정보들을 연계, 공동으로 활용할 수 있는 API가 개발되고 있으며, 여기에 사용자의 위치 정보 또한 하나의 속성으로 포함되어 있다[8]. 또한 트위터는 도시명과 지역 반경을 기반으로 트윗 정보를 수집할 수 있는 API를 제공하고 있다[12]. 이를 통해 사용자 프로필 상의 거주지역과 GPS 값을 통해 설정 지역에 대응하는 정보들을 수집할 수 있다.

수집된 데이터를 바탕으로 연관 분석을 처리하기 위해서는 문자열에서 개별 단어들을 추출, ID를 부여하여 저장하는 과정이 필요하다. 이를 위한 세부 절차는 그림 4와 같다. 먼저 입력된 문자열을 형태소 분석기를 통해 (단어, 형태소) 쌍을 생성한다. 그 다음 해당 단어를 원형으로 복원하고, 각각의 개별 단어들에 ID 값을 부여한다. 이때 원형으로 복원된 단어는 단어 사전에 저장하고, 분석 대상이 되는 정보는 수집된 시간, 위치정보 및 사용자 ID와 함께 다음과 같은 형태로 DB에 저장한다.

$$\{(x, y) \mid x \in \{\text{단어 원형ID}\}, y \in \{\text{형태소 유형}\}\}$$

그림 4. SNS 수집 데이터의 전처리 절차

3.3 마이크로 블로그에서의 연관규칙 탐사

연관규칙이란 $X \Rightarrow Y(s,c)$ 형태의 '조건-결과' 형태의 식으로 표현되는 유용한 패턴을 말한다[20]. 일반적인 연관추출 기법은 고정된 트랜잭션 데이터에서 연관규칙을 추출해내는 것을 목적하며 시간의 흐름에 따른 트랜잭션의 점진적 증가를 반영한 연관규칙 처리 알고리즘들이 개발되고 있다[6, 7]. 특히 아이템들의 관계를 그래프 형태로 구성하여 압축하는 FP-tree는 구조는 정보를 구성하는 개별 인

스턴스들과 이들 간의 관계를 서술 (predicate) 형태로 기술하는 RDF, OWL 등의 시맨틱 언어의 자료 구조와 유사하기 때문에 시맨틱 언어로의 매핑이 보다 용이하다.

마이크로 블로그를 통해 실시간으로 입력되는 정보를 분석하여 연관 규칙의 변화를 추출, 갱신하기 위해서는 다음과 같은 사항을 고려해야 한다. 마이크로 블로그 데이터는 자연어이기 때문에 트랜잭션을 구성하는 항목들의 개수가 제한되어 있지 않으며, 추출되는 결과 패턴 또한 시간의 변화에 민감하게 변화한다. 따라서 수집된 데이터를 실시간으로 처리하고, 그 변화를 모니터링하기 위해서는 이를 위한 시간 관리 방법 및 트랜잭션의 증가에 따라 점진적으로 증가하는 메모리 규모의 유지를 위한 방안이 고려되어야 한다.

3.4 연관 규칙 기반 스팸/분석 결과 필터링

지역을 기반으로 데이터를 수집할 경우 무작위로 배포된 스팸역시 수집 대상에 포함됨으로 이를 추출해 내기 위한 방안이 필요하다. 또한 추출된 연관 규칙에서 의미 없는 일반적인 결과(ex. ‘my’와 ‘friend’, ‘she’와 ‘love’ 등)가 도출될 수 있으므로 이를 배제할 필요가 있다. 따라서 그림 5와 같은 스팸 및 분석 결과 필터링 방법을 제안한다.

연관성 분석을 통해 도출된 결과는 정상 패턴, 스팸 패턴, 그리고 의미가 모호한 패턴으로 분류할 수 있다. 스팸 패턴은 연관 분석 결과 support 값이 비정상적으로 높게 나올 경우 이를 의심 패턴으로 분류하고, 해당 패턴에 대한 기여도가 높은 사용자 ID와 해당 ID를 가진 사용자가 작성한 정보를 분석하여 스팸 패턴으로 결정한다.

스팸으로 의심되는 패턴이 발견된 경우 해당 패턴을 필터링 목록에 저장하고 분석 결과에서 필터링 한다. 또한 해당 정보를 등록한 사용자 ID를 스패머 목록에 추가함으로써 이후 데이터 수집 단계에서 차단한다.

분석 결과로써 의미가 없는 패턴은 이를 필터링 목록에 저장하고 분석 결과에서 삭제한다. 또한 이러한 패턴에서 특정 단어가 지속적으로 사용되며 분석 결과에서 배제해도 될 것으로 판단된다면 필터링 단어 목록에 추가하여 분석 처리 이전에 해당 단어를 필터링한다.

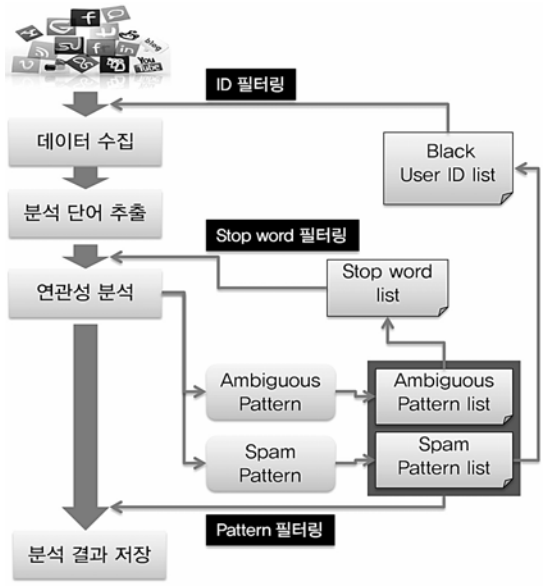

그림 5. 스팸/분석 결과 필터링 절차

4. 공간 현상의 시맨틱 공간정보화

4.1 시공간 이슈 피처 모델

본 연구에서 제시한 공간 현상 정보는 공간을 매개로 다양한 정보들 간의 연관관계로써 표현이 가능하다. 이를 ‘KS X ISO 19109 지리 정보 - 응용 스키마 규칙’을 기반으로 공간 피처 모델로 생성할 경우 그림 6과 같은 시공간 이슈 피처 모델로 나타낼 수 있다.

시공간 이슈 모델은 두 개의 부분으로 구성된다. 하나는 지리적 공간을 표현하기 위한 부분으로 기

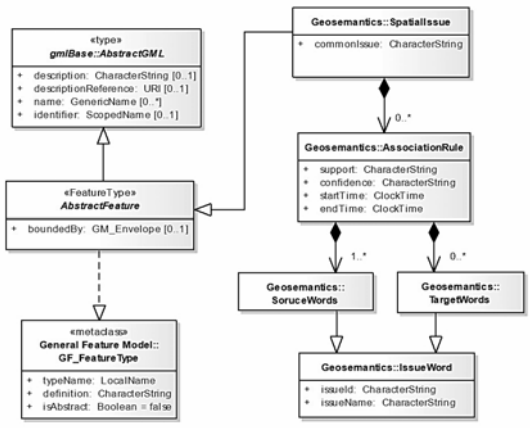

그림 6. UML diagram :
시공간 이슈 피처 모델

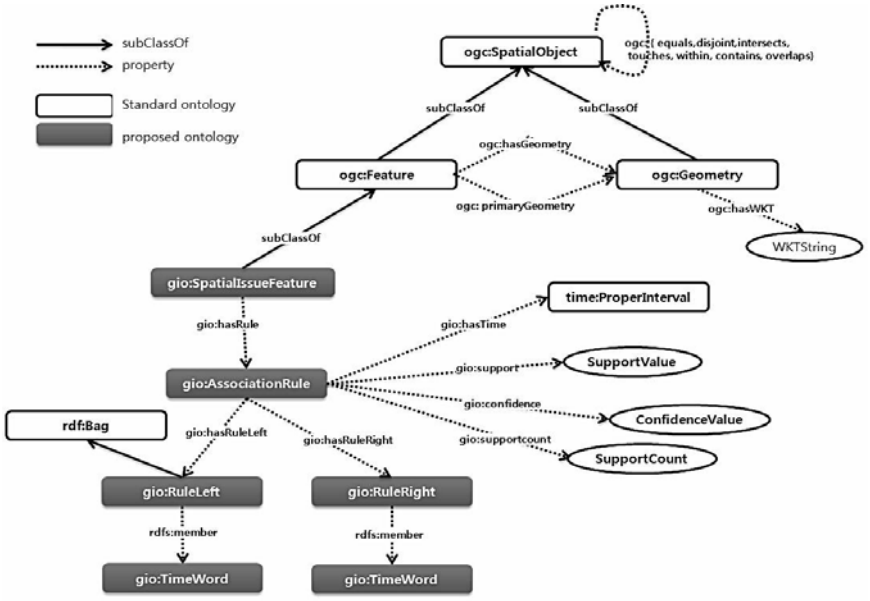

그림 7. 공간 이슈 온톨로지 모델과 표준 온톨로지와의 관계

존의 표준 피처 모델을 상속함으로써 정의되며, 연관 관계를 표현하기 위해 AssociationRule 클래스를 정의하고, 이슈를 구성하는 키워드들 간의 연관 정도와, 발생 시간 등의 메타 정보를 속성 값으로 정의하였다.

4.2 공간 이슈 온톨로지 모델

그림 6의 공간 피처 모델은 다시 그림 7과 같은 온톨로지 모델로 표현할 수 있다. 추출된 키워드들 간의 관계는 공간, 시간, 그리고 의미적 관계로 표현된다. 제안된 공간 이슈 온톨로지는 공간 관계의 표현을 위해 OGC의 GeoSPARQL 표준 [17]을, 시간 관계의 표현을 위해 W3C의 OWL-Time [11]을 활용하여 정의 하였다.

SpatialIssueFeature 클래스는 그림 6의 ogc:Feature 클래스의 하위클래스이다. 따라서 has-Geometry 속성을 상속받아 공간정보를 OGC WKT (Well Known Text) 형태로 표현한다. 또한 이를 통해 equals, disjoint, intersects, within 등과 같은 공간 관계에 대한 표현이 가능하다. Association-Rule 클래스는 LeftRule, RightRule 클래스를 통해 키워드들 간의 연관 규칙을 표현한다. 이 클래스는 W3C의 TemporalThing의 하위클래스를 통해 시간 속성을 기술하며 SupportValue, SupportCount, ConfidenceValue를 속성 값으로 갖는다. 또한 LeftRule, RightRule 클래스는 각각의 키워드들을 나타내는 TimeWord 클래스를 rdf:Bag 형태로 기술된 맴버로 갖는다.

그림 8. 공간이슈 온톨로지 인스턴스 예제

그림 8은 앞서 정의된 공간 이슈 온톨로지 모델을 기반으로 작성된 예제이다. SpatialIssueFeature 인스턴스인 'ST01'은 공간속성으로 데이터가 수집된 지역인 "Polygon(-83.6 34.1, -83.6 34.5, -83.2 34.5, -83.2 34.1, -83.6 34.1)"을 가지며, "rule01"은 추출된 키워드인 "campaign"과 "democracy" 간의 연관 관계를 의미한다. 이때 신뢰도와 지지도 값은 각각 0.2%, 30%이며, 시간 속성으로 데이터가 수집

된 기간을 의미하는 “IntervalDuring (201203310000, 201204010000)”을 갖는다.

4.3 추출된 정보의 활용

관심지역을 바탕으로 시간에 따른 연관 규칙의 변화를 관찰함으로써 공간 현상을 파악하는데 유용한 다양한 분석들이 가능하다.

관심 지역에서 시간의 경과에 따라 어떠한 주제가 가장 많이 논의되고 있으며, 특정 현상의 발생, 소멸과 같은 생명주기를 관찰할 수 있다. 2개 이상의 지역을 대상으로 데이터를 수집/결과를 분석함으로써 지역별 핵심 이슈 및 연관정보를 비교하거나 또한 동일 이슈에 대한 지역별 연관정보를 비교할 수 있다. 데이터 수집 지역을 확장하고 이를 세부 지역 범위로 분류하여 현상을 분석할 경우, 시간의 변화에 따른 이슈의 공간적인 변화(이동, 확산, 소멸) 패턴을 분석 할 수 있다.

또한 추출된 공간 이슈들을 온톨로지로 변환함으로써 시맨틱 웹 환경에서 공간정보 처리 엔진을 사용하지 않고 공간 질의가 가능하며, 다양한 도메인 온톨로지들과 연계하여 활용이 가능하다. 예를 들어 제안된 모델을 바탕으로 기술된 온톨로지 인스턴스들은 시공간 이벤트로써의 의미만을 갖는다. 하지만, 빈발 단어들에 대한 의미적 관계를 정의하고 있는 온톨로지나(ex. DBpedia Ontology API) 지명 온톨로지(Geonames) 를 연계하여 활용할 경우, 시공간적 이벤트에 추가하여, 지역에 대한 지리적 계층 관계 및 해당 단어들 간의 의미적 포함 관계까지 질의를 확장할 수 있다.

5. 결론 및 향후 연구과제

시맨틱 기술을 활용한 공간정보의 공유 노력이 진행되고 있지만, 텍스트 기반의 공간정보를 시맨틱 정보로 변환, 활용하는 방안에 대한 연구는 거의 이루어지지 않고 있다. 본 논문에서는 데이터 마이닝 기법을 이용하여 마이크로 블로그에서 공간 현상정보를 추출하기 위한 절차와 이를 시맨틱 공간 지식으로 변환하기 위해 필요한 온톨로지 모델을 제안하였다. 이를 통해 향후 관련 시스템의 개발을 위한 참조모델 및 공간 온톨로지 구축을 위한 모델로써 유용하게 사용될 수 있을 것으로 기대된다.

향후 연구로는 본 논문에서 제안한 공간 현상 추출을 위한 소셜 미디어 데이터의 특성을 고려한 스트리밍 환경에서의 연관성 분석 처리와 추출된 연관성을 기반으로 결과의 신뢰성을 향상시키기 위한 필터링, 그리고 추출된 연관지식의 공간 이슈 온톨로지 매핑 부분을 구현하고, 실험 및 평가를 통하여 제안된 방법에 대한 유효성을 확인할 예정이다.

참 고 문 헌

[1] Abel F., Gao Q., Houben G. J. and Tao K., 2011, Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web, In Proceedings of Extended Semantic Web Conference 2011.

[2] Abrol S., Khan L., 2010, TweetHood: Agglomerative Clustering on Fuzzy k-Closest Friends with Variable Depth for Location Mining, In Proceedings of the IEEE International Conference on Social Computing / IEEE International Conference on Privacy, Security, Risk and Trust, pp.153-160.

[3] Abrol S., Khan L., 2010, TWinner: Understanding News Queries with Geo-content using Twitter, In Proceedings of the 6th Workshop on Geographic Information Retrieval, February 18-19, 2010, Zurich, Switzerland.

[4] Celik I., Abel F. and Houben G. J., 2011, Learning Semantic Relationships between Entities in Twitter, In Proceedings of the 11th International Conference on Web Engineering.

[5] Cheng Z., Caverlee J. and Lee K., 2010, You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users, In Proceedings of the 19th ACM international conference on Information and knowledge management, October 26-30, 2010, Toronto, ON, Canada, pp. 759-768.

[6] Giannella C., Han I., Pei J., Yan X. and Yu P. S., 2003, Mining frequent patterns in data streams at multiple time granularities, In Kargupta H., Joshi A., Sivakumar K. and Yesha Y. (eds.), Next Generation Data Mining, AAAI/MIT.

[7] Hong T. P., Lin C. W. and Wu Y. L., 2008,

Incrementally fast updated frequent pattern trees, Expert Systems with Application, Vol. 34, issue 4, pp. 2424-2435.
[8] http://code.google.com/p/opensocial-resources/
[9] http://thenextweb.com/.
[10] http://twitaholic.com/.
[11] http://www.w3.org/TR/owl-time/
[12] https://dev.twitter.com/.
[13] Kreibich C., Crowcroft J., 2004, Honeycomb: creating intrusion detection signatures using honeypots. SIGCOMM Comput. Commun. Rev., 34(1), pp. 51 - 56.
[14] Lee C. H., Wu C. H., Chien T. F., 2011, BursT: A Dynamic Term Weighting Scheme for Mining Microblogging Messages, ISNN 2011, Part III, LNCS 6677, pp. 548 - 557.
[15] Lee K., Caverlee J. and Webb S., 2010, Uncovering Social Spammers: Social Honypots + Machine Learning, SIGIR 2010, Special Interest Group on Information Retrieval, July 19 - 23, 2010, Geneva, Switzerland, pp. 435 - 422.
[16] Mathioudakis M., Koudas N., 2010, TwitterMonitor: Trend Detection over the Twitter Stream, In Proceedings of the 2010 international conference on Management of data, June 06-10, 2010, Indianapolis, Indiana, USA.
[17] Perry M., Herring J., 2010, Draft of GeoSPARQL - A geographic query language for RDF data, Open GIS Consortium.
[18] Prince M. B., Dahl B. M., Holloway L., Keller A. M. and Langheinrich E., 2005, Understanding how spammers steal your e-mail address: An analysis of the first six months of data from project honey pot. In Proceedings of the Conference on Email and Anti-Spam.
[19] Spitzner L., 2003, The honeynet project: Trapping the hackers. IEEE Security and Privacy, 1(2), pp. 15-23.
[20] 안성렬, 2009, FP-tree를 이용한 점진적 연관규칙 추출 기법, 숭실대학교 대학원, 컴퓨터학과 석사 학위 논문.
[21] 이헌규, 나동길, 최용훈, 2011, 시간 및 공간마이닝 기술을 이용한 GIS 기반의 홍보우편 시스템 개발, 한국공간정보학회지, 제19권, 제2호 pp. 65-70.
[22] 하수욱, 남광우, 2011, 비구조적 공간정보를 지원하는 개념적 지오시맨틱 웹 서비스 프레임워크의 설계, 한국공간정보학회지, 제19권, 제6호, pp. 91-97.

논문접수 : 2012.03.06
수 정 일 : 2012.04.12
심사완료 : 2012.04.20

하 수 욱
1997년 부산대학교 공학사
2002년 부산대학교 대학원 공학석사
2011년 충북대학교 대학원 박사수료
2002년~2008년 한국정보화진흥원 선임연구원
2008년~현재 한국전자통신연구원 선임연구원
관심분야는 데이터베이스, GIS, 데이터스트림, 지오센서 네트워크, 지오시멘틱 마이닝

남 광 우
1995년 충북대학교 이학사
1997년 충북대학교 대학원 이학석사
2001년 충북대학교 대학원 이학박사
2001년~2004년 한국전자통신연구원 선임연구원
2004년~현재 군산대학교 컴퓨터정보공학과 부교수
관심분야는 데이터베이스, GIS, LBS 정책 및 기술, 데이터스트림, 지오센서 네트워크, 지오시멘틱 마이닝

류 근 호
1976년 숭실대학교 전산학과 이학사
1980년 연세대학교 공업대학원 전산전공 공학석사
1988년 연세대학교 대학원 전산전공 공학박사
1976년~1986년 육군 군수 지원사 전산실(ROTC 장교), 한국전자통신연구원(연구원), 한국방송통신대학교 전산학과(조교수)
1989년~1991년 University of Arizona, Research Staff (TempIS 연구원, Temporal DB)
1986년~현재 충북대학교 소프트웨어학과 교수
관심분야는 시간 데이터베이스, 시공간 데이터베이스, 지식기반 정보검색 시스템, 유비쿼터스컴퓨팅 및 스트림데이터처리, 데이터 마이닝, 데이터베이스보안, 바이오&메디컬인포매틱스