DETECTING SPATIAL PATTERNS OF NATURAL HAZARDS FROM THE WIKIPEDIA KNOWLEDGE BASE

J. Fan* , K. Stewart

Department of Geographical and Sustainability Science, University of Iowa, USA 52242 {junchuan-fan, kathleen-stewart}@uiowa.edu

KEY WORDS: Volunteered Geographic Information; User-Generated Knowledge; Topic Modeling; Big Geospatial Data; Wildfire

ABSTRACT:

The Wikipedia database is a data source of immense richness and variety. Included in this database are thousands of geo-tagged articles, including, for example, almost real-time updates on current and historic natural hazards. This includes user-contributed information about the location of natural hazards, the extent of the disasters, and many details relating to response, impact, and recovery. In this research, a computational framework is proposed to detect spatial patterns of natural hazards from the Wikipedia database by combining topic modeling methods with spatial analysis techniques. The computation is performed on the Neon Cluster, a high performance-computing cluster at the University of Iowa. This work uses wildfires as the exemplar hazard, but this framework is easily generalizable to other types of hazards, such as hurricanes or flooding. Latent Dirichlet Allocation (LDA) modeling is first employed to train the entire English Wikipedia dump, transforming the database dump into a 500-dimension topic model. Over 230,000 geo-tagged articles are then extracted from the Wikipedia database, spatially covering the contiguous United States. The geo-tagged articles are converted into an LDA topic space based on the topic model, with each article being represented as a weighted multi-dimension topic vector. By treating each article's topic vector as an observed point in geographic space, a probability surface is calculated for each of the topics. In this work, Wikipedia articles about wildfires are extracted from the Wikipedia database, forming a wildfire corpus and creating a basis for the topic vector analysis. The spatial distribution of wildfire outbreaks in the US is estimated by calculating the weighted sum of the topic probability surfaces using a map algebra approach, and mapped using GIS. To provide an evaluation of the approach, the estimation is compared to wildfire hazard potential maps created by the USDA Forest service.

1. INTRODUCTION

As data collection technologies keep evolving, geographic research has progressed from a data-scarce to a data-rich paradigm (Miller and Goodchild 2014). Data from vastly different provenances are being analysed and integrated to provide us with a better understanding of the social and natural processes that are occurring in geographic space, and that drive spatiotemporal dynamics, i.e., different kinds of change and movement that unfold around us every day. Although the massive amount of sensed and collected data provides great opportunities for GIScience researchers to investigate spatiotemporal dynamics from different perspectives, the variety and the volume of the data require GIScience researchers to design new data handling approaches to conquer the challenges associated with big geospatial data.

In this research, we propose a computational framework for detecting the spatial patterns of wildfires from the Wikipedia knowledge base. In this work, spatial patterns refer to the changing characteristics of location and extent associated with such fires. This computational framework will use a probabilistic topic modeling approach to extract the underlying semantic relations (i.e., themes or subject matter that are closely related with wildfires, discussed further in section 4.4) among different *topics* that constitute the Wikipedia knowledge base. The result of the topic modeling process is a topic model that can transform any single Wikipedia article into a weighted collection of *topics* or themes, i.e., each Wikipedia article is transformed from a bag of words into a semantic summary consists of topics with different weights.

A wildfire is a unique natural process whose outbreak is closely associated with the characteristics of the natural environment of a place. Traditional methods to predict the potential of wildfires use environmental factors such as topography, fuel type, fuel condition, and wind speed (Vasilakos *et al.* 2007). In this research, our proposed method characterizes wildfires from a semantic perspective and uses the semantic characteristics of wildfires to estimate spatial patterns of wildfire outbreaks. In other words, the topic model trained from the Wikipedia knowledge base provides us with a high-dimensional semantic space from which wildfire-related articles can be transformed into points in a semantic space. More concretely, in this research the English Wikipedia knowledge base is transformed into a topic model with 500 topics, each of which can be considered as an axis in a semantic space. As a result, we obtain a 500-dimensional semantic space. The *coordinates* for a Wikipedia article in the semantic space are then the set of weights for each constituent topic. In order to estimate the spatial distribution of wildfires, a link between geographic space and semantic space is designed using articles that describe geo-tagged entities (i.e., the Wikipedia articles that describe geographic entities such as city, landmark, administration units, and historical events.) in the Wikipedia knowledge base. We will refer to these geo-tagged Wikipedia articles as *geo-articles* in the following discussions. By treating each geo-article as a point observation at a particular location, we are able to use spatial analysis methods and generate a probability surface for each topic (see details in section 4.3).

Without the underlying support of high-performance computing, extracting geographic knowledge from massive amount of crowdsourced data as well as performing spatial

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
doi:10.5194/isprsannals-II-4-W2-87-2015

analysis on the extracted knowledge is not feasible. In this research, a large part of the computation has to be parallelized in order to generate results within reasonable time duration. In section 2, related works on topic modeling and geographic knowledge extraction are discussed. Terminology and details about the framework are discussed in section 3. In section 4, the computation framework and case study on wildfires is presented. Discussion of the limits and possible future work follow in section 5.

2. RELATED WORK

2.1 Sense of Place

Place has always been an important concept in geographic research and serves as a reference unit in human, economic and cultural geography (Cresswell 2004). However, place as a concept in GIScience research has been plagued by its fundamental vagueness (Goodchild 2011). Formalization of place is still an open research area (Winter *et al.* 2009). As user-generated content and volunteered geographic information become pervasive, researchers have been looking at new ways to exploit these data and extract the meaning of places. Web-based image collections are utilized to represent the conceptualization of a city and measure the popularity of a location (Schlieder and Matyas 2009). Andrienko et al. (2013) use geo-referenced tweets from Seattle-area residents to explore the thematic pattern of people's everyday life. Similar to this research, the natural language description of a place is considered an observation about a phenomenon at a particular location (Adams and Mckenzie 2013). These researchers use a topic modeling approach to create thematic regions from travelogues, trying to identify the most salient topics, features, and associated activities about a place from a tourist's perspective. For GIScience researchers, space plays an important role in shaping different natural and social phenomena. Based on Tobler's first law of geography, we would expect that the natural language description of nearby places should be more similar than places that are far apart, and different places will reveal their unique semantic characteristics due to spatial heterogeneity. Adam and Janowicz (2012) investigate the geographic distribution of non-georeferenced text and find that the unique semantic characteristics of a place can be captured by non-georeferenced text. Worboys and Hornsby (2004) also discuss place in the context of a *setting* for events. Events are described as being inherently spatiotemporal and having spatiotemporal settings. This research is motivated by these previous researches about place and setting. We push further with the idea of semantic characteristics of places, investigating the links between events or *dynamics*, and the locations at which they occur. Based on the massive amount of knowledge encoded in the Wikipedia database, we are able to extract not only semantic characteristics of different places, but also that of natural hazards. If a natural hazard has a close semantic tie with certain places, we may infer that this kind of place is likely to be vulnerable to this type of natural hazard. By comparing the semantic characteristics of places and natural hazards together, we have a new way of gaining insights into places, natural hazards, and the interactions between them.

2.2 Extracting Knowledge from Crowdsourced Data

Crowdsourced data is one of the quintessential characteristics of the Web 2.0 era. The volume and variety of this data source has brought both excitement and challenges to researchers from different communities. Wikipedia is a perfect example of crowdsourced data. As of 2015, Wikipedia has over 4.5 million articles and still increasing. Wikipedia has been used to facilitate information retrieval; as a resource for building ontologies (Medelyan *et al.* 2009) and the vast amount of common-sense and domain-specific knowledge in Wikipedia have been a test bed for machine learning techniques to represent any text as a weighted vector of concepts (Gabrilovich and Markovitch 2007). Wikipedia's category information has been integrated with WordNet, and GeoNames, creating a massive ontology and linked knowledge base (Hoffart *et al.* 2013). In the GIScience community, crowdsourced efforts have been behind the creation of OpenStreetMap (OSM) (www.openstreetmap.org), a very successful effort for creating open source online maps. Estima and Painho (2013) use OSM data for land use classification in Portugal, and achieve a global classification accuracy around 76%, while artificial neural networks and genetic algorithm have been used to estimate urban land-use patterns from OSM data (Hagenauer and Helbich 2012).

2.3 Probabilistic Topic Modeling

Methods for extracting geographic knowledge from natural language data has been studied by machine learning, information retrieval, and geographic information retrieval communities. Traditional methods for extracting geographic location information from text documents uses natural language processing techniques, such as named entity recognition (Etzioni *et al.* 2005). Name entity recognition is a method for identifying and classifying entities in text documents into pre-defined categories (e.g., people, location, expression of times, etc.) The location information associated with identified entities in text documents are then used to estimate the geographic location of the text document.

Topic modeling is a collection of algorithms that can be employed to uncover the hidden topics of a text corpus. Latent Dirichlet Allocation (LDA) is one of the most popular methods for topic modeling (Blei *et al.* 2003; Blei 2012). LDA is a probabilistic generative model that represents each document as a mixture of topics, and a topic as a mixture of words. The distributions of topics over documents and words over topics are modeled by the Dirichlet distribution with given hyper parameters. Griffiths and Steyvers (2002) investigated the feasibility of a probabilistic approach for capturing semantics with a collection of texts. They showed that the semantic representation generated using a probabilistic approach has statistical properties consistent with the large-scale structure of semantic networks constructed by humans. In another application, LDA was applied to analyse abstracts from the Proceedings of the National Academy of Science and established scientific topics that are consistent with class designations provided by authors of the articles (Griffiths and Steyvers 2004). For this research, we apply this probabilistic approach to extract the underlying semantic relations between different kinds of places and wildfire hazards.

3. COMPUTATION FRAMEWORK FOR DETECTING WILDFIRE DYNAMICS

3.1 Notation and Terminology

The Wikipedia knowledgebase is represented as a collection of text documents.

$$\boldsymbol{W} = \{d_1, d_2 \dots d_N\} \quad (1)$$

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
doi:10.5194/isprsannals-II-4-W2-87-2015

where N is the number of documents in the collection. The entire document collection has a set of vocabulary

$$\boldsymbol{V} = \{1: w_1, 2: w_2 \dots, |V|: w_v\} \quad (2)$$

where each word is assigned a unique natural number. $|V|$ is the size the entire vocabulary. Each document is represented as a bag of words

$$\boldsymbol{d} = \{i_1, i_2 \dots i_d\} \quad (3)$$

where each element in this bag of words representation is a natural number from the vocabulary set V .

The entire Wikipedia knowledgebase W is assumed to have K topics.

$$\boldsymbol{T} = \{z_1, z_2 \dots z_K\} \quad (4)$$

Using the LDA modeling framework, each document d is assumed to be generated from a weighted collection of topics, where

$$\boldsymbol{z^d} = \{e_1^d * z_1^d, e_2^d * z_2^d \dots, e_k^d * z_k^d\} \quad (5)$$

z_j^d is one of the topics in T , and e_j^d is the corresponding weight for this topic in document d . $\{e_1^d, e_2^d \dots e_k^d\}$ is assumed to follow a multinomial distribution parameterized by θ^d. More detailed discussion on the mathematical and computational properties of LDA topic modeling approach can be found in (Blei *et al.* 2003).

The words that comprise a topic are assumed to follow multinomial distribution as well, which is parameterized by β^z. During the training process, each word in a document is generated as follows: first, a topic is selected based on the topic distribution for that document (i.e., $multinomial(\theta)$); then, a word can be picked from the chosen topic based on the word distribution of this topic (i.e., $multinomial(\beta)$). Based on observed data, the training process will find the parameters that fit the observed data best. Due to the large volume of training data (over 4 million articles), we used an online training algorithm for LDA (Hoffman *et al.* 2010).

This computational framework is comprised of four components (Figure 1). The first component transforms the Wikipedia knowledge base from a document-word space into a document-topic semantic space. The core function of this component is a probabilistic topic modeling package, which is developed based on a python open source library[1] for topic modeling.

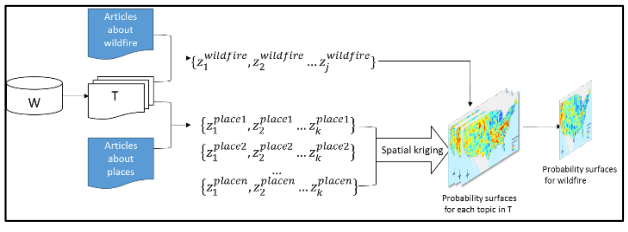

Figure 1. Computational framework for detecting spatial wildfire dynamics

In the second component, geo-articles are extracted from the Wikipedia knowledge base and transformed into their topic representations (i.e., a weighted vector of topic for each article). This is a key step for linking geographic spaces and semantic spaces. By treating a geo-article as an observation at a particular location S, the topic weights $\{e_j^d: 1 \le j \le K\}$ for each article are observed values for location S. A higher topic weight e_i^d means that location S is semantically closer to topic z_i. For this study, we extracted 235,089 articles on places (i.e., *geo-article*) from a Wikipedia database dump obtained in October 2014 for states in the contiguous U.S (Figure 2).

In the third component, a probability surface for each topic in T is created based on the "observed point values" for each topic using spatial kriging. Kriging is a spatial interpolation process that takes into account the autocorrelation between topic weights of near points. The final component extracts wildfire-related articles from the Wikipedia knowledge base and transforms them into weighted topic vector representations. A final probability surface about wildfires is generated using map algebra based on the generated topic probability surfaces.

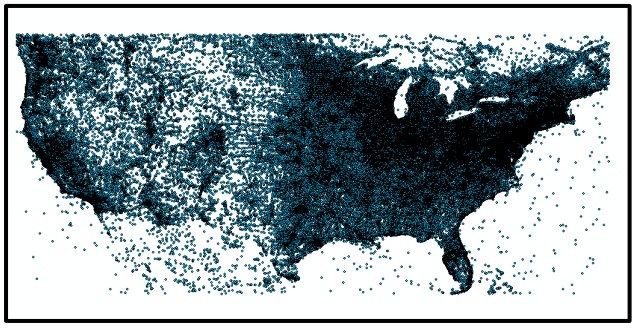

Figure 2. Geo-articles from Wikipedia knowledgebase created from October 2014 database dump.

$$\boldsymbol{WildfireProbabilitySurface} = \Sigma_{i=1}^{K} e_i * PS_{z_i} \quad (6)$$

PS_{z_i} is the probability surface for topic z_i and e_i are the corresponding weights in wildfire articles.

The first key element in this framework is the massive amount of domain knowledge in the entire Wikipedia knowledge base that is extracted using a probabilistic topic modeling approach and represented as a topic model. The second key element is the corpus of geo-articles that cover the U.S, providing a link between geographic space and topic (semantic) space. Using this computational framework, we can also investigate the spatial distributions of other subjects or processes, for example, flooding and soil erosion. In the next section, we will discuss in details the process of detecting spatial patterns of wildfires using this computational framework.

4. DETECTING DYNAMICS OF WILDFIRE

4.1 Data Preprocessing

The Wikipedia project provides a database dump each month that includes all the articles currently in Wikipedia. We choose only the English articles as our training corpus. The English database dump from October 2014 is approximately 40 GB[2]. The provided database dump is in XML format and has many different markup tags that must be filtered before LDA training. In addition, there are many administration pages and page stubs that need to be removed from the database. The reason for the filtering process is that LDA training relies on the frequencies

[1] https://github.com/piskvorky/gensim/
[2] http://dumps.wikimedia.org/enwiki/

[2] http://dumps.wikimedia.org/enwiki/

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
doi:10.5194/isprsannals-II-4-W2-87-2015

of occurrences and co-occurrences between words in the Wikipedia articles. While these markup tags and administration pages do not contain meaningful semantics for our topic modeling, they will be treated equally as other Wikipedia articles during the training process if not removed. This will likely bring in a lot of noises into the topic modeling results. We programmed a Python script to filter these markup tags and administration pages, including *comments, footnotes, links to languages, template, URL in the article, link to images, link to files, outside links, math equations, table formatting, table cell formatting, category information and other tags.* After cleaning up the raw data, we have 3,718,495 documents and 576,588,282 words in the knowledgebase.

The second step for data preprocessing is the extraction of geo-articles. The WikiProject council has been trying to better organize location-based information and knowledge in Wikipedia articles through the WikiProject Geographic Coordinates project. The general goal for this project is to add coordinates to any article about geographic entities such as a city (e.g., San Francisco), building structure (e.g., Golden Gate Bridge), or geographic features that are more or less fixed in geographic space. Articles about events that are associated with a single location (e.g., September 11 attacks) will also be tagged with geographic coordinates. Since this is a crowdsourced project using volunteers, different users often use different types of tags for coordinates. The most frequently used tags are "*lat,long*", "*latitude,longitude*", "*lat_degree,long_degree*". We again use a python script that sifts through the entire database and extracts pages that have coordinate tags. The Python code extracted over 230,000 geo-articles from the English Wikipedia database corresponding to the 48 contiguous states in the US. Figure 2 shows the geographic distribution of these geo-articles. The parallelized preprocessing takes about 3 hours on a 16-core node from the Neon cluster, a high performance computing system at the University of Iowa.

The final step is to extract wildfire-related articles from the Wikipedia database, which is based on the categorization structure of the Wikipedia database. The Wikipedia database organizes articles into different thematic categories. For example, the 2012 High Park fire in Colorado is categorized into four categories: "2012 natural disasters in the United States", "Wildfires in Colorado", "2012 wildfires", "2012 in Colorado". We have developed a program that automates the extraction process based on the category information. We extracted about 200 articles describing wildfire instances from the Wikipedia database, including wildfire instances not only in the U.S but also in other countries (e.g., Australia). Certain articles that may be related to wildfires but not about wildfire instances are not included, for example, articles about wildfire suppression methods, or articles about fire more generally. Future work will consider how to incorporate more specific categories into the training process.

4.2 Topic Modeling

After preprocessing, all the markup tags have been filtered and all the administration pages have been also removed. In order to separate the training corpus with the test corpus, another important step before topic modeling process is to exclude all the wildfire articles extracted in previous step from the knowledge base.

The topic modeling process first generates a dictionary that has all the words in the knowledge base. All the common stop words (e.g., *the, about, and)* are excluded from the dictionary. Based on this dictionary, each document in the knowledgebase will be converted into a bag of words representation. Then a TF-IDF (term frequency) model is trained from the bag of words representation of each article. This TF-IDF model is used as input for LDA training. This part of the computation is parallelizable, since all the word frequencies and word-document counts can be calculated simultaneously using multiprocessors. This data preparation process runs for about 7 hours on the 16-core node in the Neon cluster.

After the Wikipedia knowledgebase is transformed into a TF-IDF representation, the data is ready for LDA training. LDA training may be conducted in online and batch modes. The online model processes a chunk of the documents at a time and then updates the parameters for the topic model, and then repeats this process with another chunk and update to the topic model. In contrast, the batch model LDA will first make a full pass of the whole corpus, and updates the topic model, and then repeats this process. For a relatively stable corpus, the online mode is able to make a reasonably accurate estimation of the topics, but it converges faster than the batch mode (for details, see Hoffman *et al.* 2010). We chose the online mode LDA for the training process and set $K = 500$ as the number of topics to be generated. The training process runs for about 8 hours on a 16-core node.

In Table 1, we show 5 examples from the 500 generated topics. As we have discussed in section 3.1, LDA models each topic as a collection of semantically related words, and the collection of words in a topic are assumed to follow a multinomial distribution. The numeric value in front of each word represents the relative weight of that word in the topic. A higher value means this word has a closer semantic relatedness with the topic. Only the ten most related words for each topic are presented here. We can see that the words that are classified as the same topics are closely related (semantically similar), and it is not difficult to find a thematic summarization from these words. For example, topic 1 refers to mobile-related web technology; topic 2 refers to natural hazards; topic 3 is about farm animals; topic 4 refers to water bodies, and topic 5 is about mountains.

1	0.019*app + 0.012*users + 0.012*internet + 0.011*mobile + 0.010*cloud + 0.008*online + 0.008*security + 0.007*facebook + 0.007*network + 0.007*data
2	0.021*storm + 0.021*earthquake + 0.015*tropical + 0.012*hurricane + 0.009*damage + 0.009*km + 0.009*magnitude + 0.009*tornado + 0.007*winds + 0.007*flood
3	0.015*sheep + 0.014*breed + 0.013*horse + 0.013*zoo + 0.011*cattle + 0.009*horses + 0.009*bred + 0.008*penalties + 0.008*animals + 0.008*animal
4	0.044*river + 0.036*lake + 0.023*creek + 0.023*dam + 0.013*flows + 0.012*water + 0.011*tributary + 0.011*reservoir + 0.008*rivers + 0.008*lakes
5	0.068*mountain + 0.025*glacier + 0.022*mountains + 0.021*peak + 0.020*summit + 0.020*mount + 0.015*lies + 0.014*range + 0.012*ridge + 0.011*spelling

Table 1. Five topics generated from LDA training

Topics and their associated spatial patterns may be mapped and hotspot analysis undertaken to see whether certain locations identify more strongly than others for a topic. For example, for topic 1 from Table 1 (i.e., mobile-related web technology), the Getis-Ord Gi* statistic was applied using ArcMap 10.2 (Figure 3). This analysis identifies statistically significant locations based on a set of weighted point features. In our case, the

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
doi:10.5194/isprsannals-II-4-W2-87-2015

weights are topic weights for mobile and online-related topics in a geo-article.

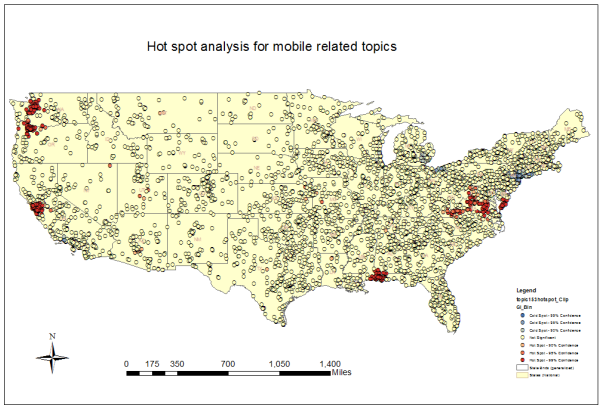

Figure 3. Hotspot analysis for mobile-related topic

Four clusters are identified that are associated with articles on topic 1. Silicon Valley in California, and Seattle in Washington State correspond to two of the western clusters. In the east, Virginia appears as a cluster for mobile-related topic. This may be due to geo-articles discussing mobile technology as used by federal agencies and related businesses located in the Virginia-DC area. Note that Mobile County in Alabama is falsely identified as a hot spot due to the name of the county. The training results demonstrate semantic consistency within topics and semantic distinctions between topics and the hot spot analysis shows how spatial pattern analysis becomes possible. This show how a probabilistic topic modeling approach captures underlying semantic relations between words from a large corpus, such as the Wikipedia knowledge base. Moreover, a link between geographic space and semantic space can be established through the analysis of geo-articles in Wikipedia. Of course, the falsely identified hotspot also reveals a challenge associated with nature language processing—ambiguity. Future work will incorporate gazetteer information into the topic modeling process and explore ways to improve this aspect.

4.3 Generating Topic Probability Surface Using Kriging

As stated above, the link between geographic space and semantic space is through geo-articles. By treating each geo-article as an observation at a particular location, we are able to perform spatial analysis on the observed values at each point (i.e., the topic weights for each geo-article). The comprehensive coverage of Wikipedia knowledge base ensures that each area has enough observed points within it. In other words, the semantic characteristics of a location should be sufficiently captured by the natural language descriptions of this place.

In order to create a continuous probability surface for each topic, we utilize kriging to interpolate a raster surface based on observed points. We assume a spherical semivarigram model for the kriging process. Kriging is processor-intensive. Since we have over 230,000 points and 500 topics, the kriging process takes about 40 computing hours on an 8-core high-end PC. The reason for switching computing environments away from the HPC environment, is the lack of proper visualization capabilities in a clustering computing environment. However, since topics don't interact with each other during kriging, this part of computation can also be parallelized. By running kriging for multiple topics simultaneously, we can greatly decrease the waiting time.

4.4 Analysing Spatial Patterns of Wildfires

In the data preprocessing step, all Wikipedia articles about wildfire instances are extracted. If wildfires have unique semantic characteristics, these articles should comprise a corpus that has captured these aspects of wildfires. Using our computational framework, we expect to be able to discover the links between geographic locations of wildfires and the semantics associated with these hazards. Coupling with a spatial probability surface for each topic, a probability surface for wildfires is generated for in the contiguous states in the US.

The corpus of wildfire articles is input into the topic model, and transformed into a collection of weighted topics. The five topics that scored the highest values in wildfire's weighted topic vector representation are presented in Table 2. The top five topics that the LDA approach returns as being most related to wildfires are response/impact, natural hazards, western US, climate, and water bodies. This result seems to correspond reasonably well with our common sense. The normalized weight for each topic represents the relative percentage content that is devoted to the topic in a Wikipedia wildfire article. It appears that writing on the response/impact-aspects (i.e., the human aspects) of wildfires has the highest share.

Normalized Topic Weight	Topic
0.25	0.007*killed + 0.007*attack + 0.006*battle + 0.006*forces + 0.006*soldiers + 0.006*army + 0.005*troops + 0.004*wounded + 0.004*fire + 0.003*military
0.17	0.021*storm + 0.021*earthquake + 0.015*tropical + 0.012*hurricane + 0.009*damage + 0.009*km + 0.009*magnitude + 0.009*tornado + 0.007*winds + 0.007*flood
0.07	0.051*california + 0.031*san + 0.027*angeles + 0.026*los + 0.015*nevada + 0.015*francisco + 0.012*santa + 0.012*cdp + 0.012*diego + 0.009*vegas
0.05	0.009*climate + 0.009*catchment + 0.006*region + 0.006*hills + 0.006*fault + 0.006*plateau + 0.005*valley + 0.005*vegetation + 0.005*km + 0.005*descends
0.05	0.044*river + 0.036*lake + 0.023*creek + 0.023*dam + 0.013*flows + 0.012*water + 0.011*tributary + 0.011*reservoir + 0.008*rivers + 0.008*lakes

Table 2. Top five topics and their highest value words related to wildfires in Wikipedia.

4.4.1 Spatial Patterns of Wildfires

For GIScience researchers, the spatial context is expected to impact natural and social processes such as those associated with hazards. This understanding extends to natural language descriptions about places that also reflect the influence from spatial contexts. Consequently, we expect that semantic characteristics will be reflected in topic representations of articles about human-environment interactions, such as wildfires. In this work, a probability surface for each topic is created during the kriging process. Based on wildfires' weighted topic vector representation, map algebra is used to calculate a raster surface for wildfires (Figure 4).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
doi:10.5194/isprsannals-II-4-W2-87-2015

This result is the weighted sum of the constituent topics' probability surfaces (*equation 6*), each of which has a distinct pattern. The raster surface for wildfire shows an interesting spatial pattern. The direct interpretation for higher-valued areas on this map is that these areas have more wildfire-related discussions than lower-valued areas. It may also represent locations that share similar semantics to locations that have experienced wildfires historically. In this way, this mapped result may be viewed as showing not only hotspots where wildfires have occurred but also locations that may be at risk for wildfires due to semantic similarity (i.e., the approach may serve as a new method for determining wildfire potential).

Figure 5 shows a map for historical wildfire activities in the US from 1993 to 2014. Comparing this map with our map (Figure 4), higher-value areas in our map appear to correspond well with areas that historically have more wildfire incidents.

Wildfires are highly dynamic natural processes, both spatially and temporally. They can occur any time throughout the year and in a wide variety of locations across the country. However, certain regions have a higher potential for wildfire hazards (Figure 6). From Figure 5, western and southeastern parts of the US suffered more wildfires than the rest of US. By comparing this map (Figure 5) with the 2014 wildfire potential hazard map (Figure 6) created by the USDA Forest Service, we

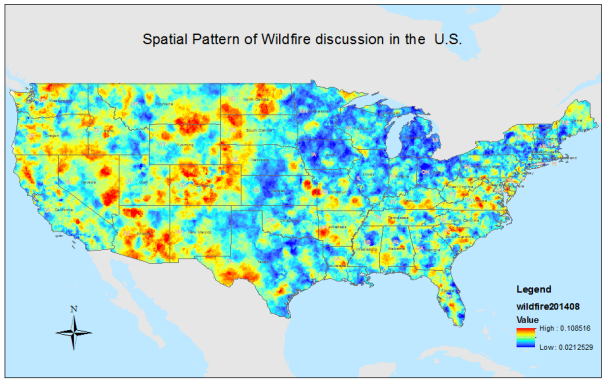

Figure 4. Wildfire potential estimated from Wikipedia 2014 October knowledgebase

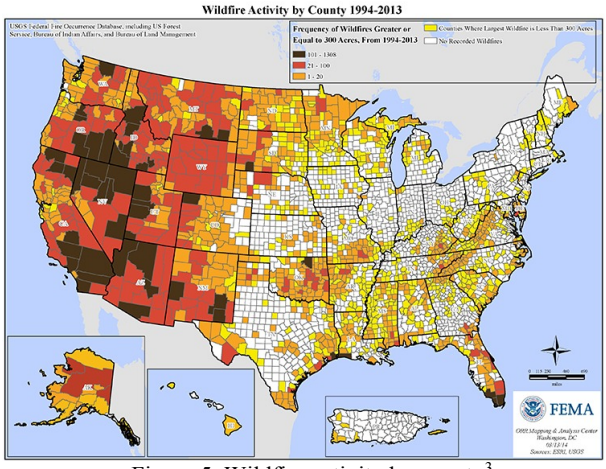

Figure 5. Wildfire activity by county[3]

[3]http://www.community.fema.gov/connect.ti/AmericasPrepareathon/view?objectId=3221840

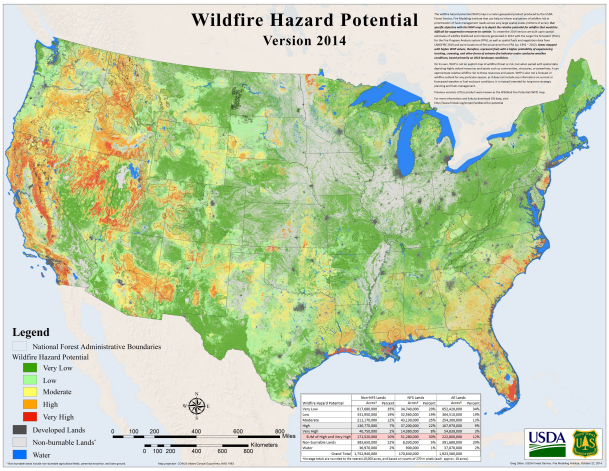

Figure 6. Wildfire potential estimated by USDA Forest Service[4]

can visually determine that areas with high wildfire hazard potential hazard in 2014 (Figure 6) have similar spatial patterns with historical wildfire activity (Figure 5). The differences may be due to wildfire's spatiotemporal dynamics, since each year wildfire outbreaks will vary across the US due to human, climate, and other factors.

The comparison is made based on visual inspection. Further statistical analysis and comparison between these results will be performed in future work. Although more analysis is necessary, the spatial patterns generated from using an LDA-based approach appear to have the potential to be useful as a source for possibly estimating the hazard potential for wildfires for an area. Comparing Figures 4 and 6, the western part of US including California, Nevada, Arizona, and Oregon are estimated to have high wildfire hazard potential in both maps. Similarly, Figure 4 captures similar locations that match the high wildfire potential in Figure 6 for southeastern parts of the US.

5. DISCUSSION AND FUTURE WORK

Wikipedia is a significant data source that contains massive amounts of domain-specific knowledge. It plays an important role as a complementary data source for traditional spatial data (e.g., vector and raster data). This research developed a computational framework that can help researchers explore the Wikipedia knowledge base and leverage the semantic dimensions captured in this database. This work uses wildfires as an example to test the proposed framework, however, we expect that this framework could be extended to other types of natural hazards or geographic processes in order to investigate the semantic relations associated with related topics, as well as spatial patterns of these processes. As the amount of crowdsourced data continues to increase, this framework can also be extended to other kinds of knowledgebase, for example, the entire 2014 tweet collection for the US. The semantics and information encoded in different knowledgebases would be different of course, for example providing more detail about local and real-time events perhaps, but these information could contribute in a meaningful way to applications.

This framework links geographic space with semantic space through geo-articles. Tobler's first law on nearness is applied to the semantic characteristics of places. We expect that places that

[4] http://www.firelab.org/project/wildfire-hazard-potential

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
doi:10.5194/isprsannals-II-4-W2-87-2015

are nearer will be closer in semantic space than places that are far apart, as discussed by Adam and Janowiscz (2012). Using this key spatial understanding, we generated an estimate of locations in the contiguous US that may be viewed as having higher wildfire hazard potential. The result corresponds with the wildfire hazard potential map created by USDA Forest Service for 2014. Of course, natural language is inherently ambiguous and this approach is affected by this constraint as well. We plan to incorporate other data sources into the framework to improve the results in the future. This is in keeping with findings that suggest big data simple models have been shown to outperform small data complex models in many areas (Halevy et al. 2009).

Several possible directions can be followed for future work. Wikipedia is a relatively static corpus; therefore it does not reflect the temporal changes of processes well. By incorporating data that are temporally more responsive, we could potentially detect the temporal dynamics of natural hazards. Wikipedia offers interesting opportunities for data analytics due to the fact that its knowledge base is organized into detailed categories that can be use by topic training models to generate subject areas or themes with better resolution in semantic space, i.e., the generated topics have better consistency and coherence.

REFERENCES

Adams, B. and Janowicz, K., 2003. On the Geo-Indicativeness of Non-Georeferenced Text. *ICWSM*, 375–378.

Adams, B. and Mckenzie, G., 2013. Inferring Thematic Places from Spatially Referenced Natural Language Descriptions. *In*: D. Sui, S. Elwood, and M. Goodchild, eds. *Crowdsourcing Geographic Knowledge*. Dordrecht: Springer Netherlands, 201–221.

Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., and Thom, D., 2013. Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing in Science and Engineering*, 15 (3), 72–82.

Blei, D.M., Ng, A.Y., and Jordan, M.I., 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Blei, D.M., 2012. Introduction to Probabilistic Topic Modeling. Communications of the ACM, 55, 77–84.

Cresswell, T., 2004. *Place: A Short Introduction*. The Professional Geographer.

Estima, J. and Painho, M., 2013. Exploratory Analysis of OpenStreetMap for Land Use Classification. *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, 39–46.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D.S., and Yates, A., 2005. Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence*, 165 (1), 91–134.

Gabrilovich, E. and Markovitch, S., 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. *International Joint Conference on Artificial Intelligence*, 1606–1611.

Goodchild, M.F., 2011. Formalizing place in geographic information systems. *In*: L.M. Burton, S.P. Kemp, M.-C. Leung, S.A. Matthews, and D.. Takeuchi, eds. *Communities, Neighborhoods, and Health: Exapanding the Boundaries of Place*. Springer: New York, 21–35.

Griffiths, T.L. and Steyvers, M., 2002. A probabilistic approach to semantic representation. *In*: *Proceedings of the 24th annual conference of the cognitive science society*. 381–386.

Griffiths, T.L. and Steyvers, M., 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl , 5228–35.

Hagenauer, J. and Helbich, M., 2012. Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks. *International Journal of Geographical Information Science*, 26 (April 2015), 963–982.

Halevy, a., Norvig, P., and Pereira, F. 2009. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, 24(2).8-12

Hoffart, J., Suchanek, F.M., Berberich, K., and Weikum, G., 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194, 28–61.

Hoffman, M., Blei, D., and Bach, F., 2010. Online learning for latent dirichlet allocation. *Nips*, 1–9.

Medelyan, O., Milne, D., Legg, C., and Witten, I.H., 2009. Mining meaning from Wikipedia. *International Journal of Human Computer Studies*, 67, 716–754.

Miller, H.J. and Goodchild, M.F., 2014. Data-driven geography. *GeoJournal*. 1-13

Schlieder, C. and Matyas, C., 2009. Photographing a City: An Analysis of Place Concepts Based on Spatial Choices. *Spatial Cognition Computation*, 9 (April 2015), 5–29.

Vasilakos, C., Kalabokidis, K., Hatzopoulos, J., Kallos, G., and Matsinos, Y., 2007. Integrating new methods and tools in fire danger rating. *International Journal of Wildland Fire*, 16 (3), 306–316.

Winter, S., Kuhn, W., and Krüger, A., 2009. Does Place Have a Place in Geographic Information Science? *Spatial Cognition & Computation*, 171–173.

Worboys, M. and Hornsby, K., 2004. From objects to events: GEM, the geospatial event model. *GIScience*, 1–17.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
doi:10.5194/isprsannals-II-4-W2-87-2015