Information Processing and Management 56 (2019) 1185–1203

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Understanding the topic evolution of scientific literatures like an evolving city: Using Google Word2Vec model and spatial autocorrelation analysis

Hu Kai[a,b,*], Luo Qing[c,d], Qi Kunlun[e], Yang Siluo[f], Mao Jin[f], Fu Xiaokang[c,d], Zheng Jie[c,d], Wu Huayi[c,d], Guo Ya[a,b], Zhu Qibing[a,b]

[a] Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, Jiangsu, China
[b] School of Internet of Things, Jiangnan University, Wuxi 214122, China
[c] The State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[d] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China
[e] Faculty of Information Engineering, China University of Geosciences (Wuhan), Wuhan 430074, China
[f] School of Information Management, Wuhan University, Wuhan 430072, China

ARTICLE INFO

Keywords:
Semantic relatedness
Topic evolution
Spatial clustering
Spatial autocorrelation
Word2Vec

ABSTRACT

Topic evolution has been described by many approaches from a macro level to a detail level, by extracting topic dynamics from text in literature and other media types. However, why the evolution happens is less studied. In this paper, we focus on whether and how the keyword semantics can invoke or affect the topic evolution. We assume that the semantic relatedness among the keywords can affect topic popularity during literature surveying and citing process, thus invoking evolution. However, the assumption is needed to be confirmed in an approach that fully considers the semantic interactions among topics. Traditional topic evolution analyses in scientometric domains cannot provide such support because of using limited semantic meanings. To address this problem, we apply the Google Word2Vec, a deep learning language model, to enhance the keywords with more complete semantic information. We further develop the semantic space as an urban geographic space. We analyze the topic evolution geographically using the measures of spatial autocorrelation, as if keywords are the changing lands in an evolving city. The keyword citations (keyword citation counts one when the paper containing this keyword obtains a citation) are used as an indicator of keyword popularity. Using the bibliographical datasets of the geographical natural hazard field, experimental results demonstrate that in some local areas, the popularity of keywords is affecting that of the surrounding keywords. However, there are no significant impacts on the evolution of all keywords. The spatial autocorrelation analysis identifies the interaction patterns (including High-High leading, High-Low suppressing) among the keywords in local areas. This approach can be regarded as an analyzing framework borrowed from geospatial modeling. Moreover, the prediction results in local areas are demonstrated to be more accurate if considering the spatial autocorrelations.

* Corresponding author at: Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, Jiangsu, China.
E-mail addresses: hukai_wlw@jiangnan.edu.cn, hukai@whu.edu.cn (K. Hu).

https://doi.org/10.1016/j.ipm.2019.02.014
Received 25 July 2018; Received in revised form 26 January 2019; Accepted 17 February 2019

1. Introduction

With the expansion of the number of bibliographic datasets, more and more scientific fields are applying scientometric methodologies to analyze domain knowledge. Understanding topic dynamics is vital for analyzing the domain knowledge and is often studied as a sub-question of Topic Detection and Tracking (TDT). The overall developments of TDT are well summarized in Tu and Seng (2012). The TDT research can be roughly categorized into three classes: text and data mining (Pons-Porrata, Berlanga-Llavori, & Ruiz-Shulcloper, 2007), time-line burst detection and measurement (Chen, Luesukprasert, & Chou, 2007), and content-based or link-based analysis (Nallapati, Ahmed, Xing, & Cohen, 2008). Methods proposed in these TDT researches are widely used to analyze user-generated content like social media (Hong, Ahmed, Gurumurthy, Smola, & Tsioutsiouliklis, 2012) and news (Zhao, Jin, & Yue, 2015). In these scenarios, popularity of topics is affected by users. Users are interested in varied topics over time, resulting in topic evolution (Zarrinkalam, Kahani, & Bagheri, 2018). These methods are also employed to find topic dynamics in document sequences (Mane & Börner, 2004) and top topics of science in scientific papers (Blei & Lafferty, 2006), providing a macro-level picture for topic evolution. Interesting to notice, Chen, Tsutsui, Ding, and Ma (2017) used an analogy to explain evolution details of topic splitting or topic merging, regarding topics as territories and words as populations. A word is associated with different topics, as assumed in the topic model of Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). Thus, the process of a word disappearing in one topic and then appearing in another topic is defined as “word migration” via topic channels. Topic evolution details have been well described; however, why they happen is often largely unknown.

Semantic relatedness (Chen et al., 2017) of keywords could be an important factor that explains why the topic evolution happens. In a literature search scenario, researchers input certain keywords/topics to find literatures in academic databases, papers semantically related to the used keywords/topics will be returned, as shown in Fig. 1.

In Fig. 1, the author has planned to cite a paper of topic A at the beginning, but after reading the papers of the semantic related topics, authors may turn to cite the paper of topic B or topic N. Thus, keyword citations of semantically related topics like topic B or topic N increase. Note keywords do not obtained citation directly. In this paper, we define keyword citation as the index that counts one when any paper containing this keyword obtains a citation. We use the keyword citation as the indicator of the keyword popularity in this paper. This citing-change scenario frequently happens. Taking a personal experience as an example, we used to do research on Web Mapping Service (WMS) query optimization which we regard as having weak connections with image retrieval at the beginning, but after realizing that WMSs return image-format map data, the idea of combining image retrieval with WMS came to our mind (Hu et al., 2016). Thus, we turned to cite the work in image retrieval field instead of solely attribute-based retrieval. This citing-change scenario took place, because WMS and image are semantically related.

Taking a more popular topic, “machine learning”, as an example, different machine learning methods are promoting each other's fame, i.e., in the leading pattern, or competing with each other, i.e., in the suppressing pattern (Arora et al., 2010; Lee & To, 2010; Singla, Chambayil, Khosla, & Santosh, 2011). Regarding the topic pairs of “Support Vector Machines (SVM)” and “Artificial Neuron Network (ANN)”, both leading and suppressing patterns can be observed in the citation history of papers on these two topics. The research topic of “SVM” has been hot for a long time, but becomes not so hot recently. Because the competitor, Deep Learning,

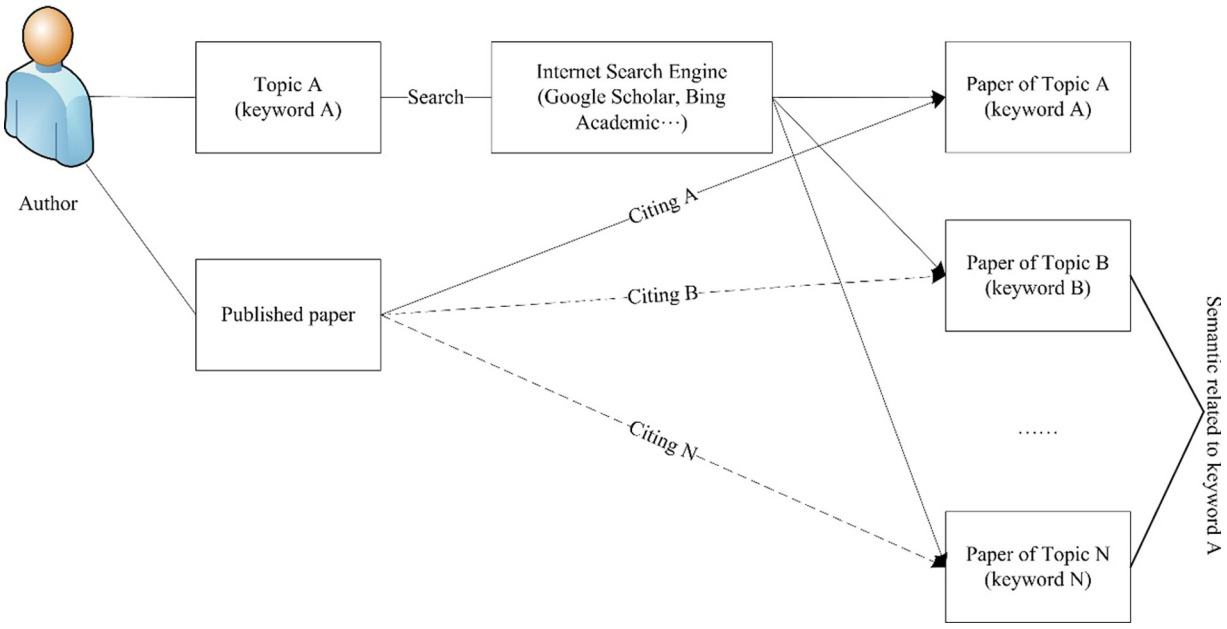

Fig. 1. Citing behaviors affected by semantically related topics resulting in different keyword/topic popularity (author may cite B or cite N rather than cite A).

stemming from ANN, achieved better scores on several artificial intelligent tasks. In this case, the increasing popularity of topic "Deep Learning" makes topic "SVM" seem to be less popular, forming a suppressing patterns. Similarly, leading patterns are also seen when ANN becomes hot again led by the development of Deep Learning. The semantic relatedness among these topics seems to be a potential driver that triggers the topic evolution. However, traditional topic evolution analysis methods such as the frequently used temporal keyword lists or temporal co-word network, cannot provide such support, because semantic relatedness among the keywords is not fully considered.

To be more specific, temporal keyword lists contain the ranking information of the keywords, but not the semantic relations among the keywords. Though the keyword evolution can be easily observed through the temporal changes of keyword frequencies, the semantic relations among the keywords are lacked. Temporal co-word networks provide partial relations among keywords through centrality computation (Newman, 2008). Through different centrality degrees, the functionalities and roles of keywords can be described. Co-word networks, however, still have limitations in modeling semantic meanings. Many semantically linked keywords are often disconnected in different sub-networks, because co-word relations are much sparser in the keyword lists than in the text bodies of papers. Several works have extended the keywords of the co-word network with semantic meanings (Hu et al., 2018a,b; Wang, Li, Li, & Li, 2012). More complete associations among the keywords can be displayed in the topic evolution analysis. However, these research pay little attention to the driving mechanism behind the topic evolution phenomena. Whether the semantic relatedness is the underlying driver or not is still not clear.

Thus, we summarize the research problems of this paper here:

1 Does the semantic relatedness among the topics really affect the topic evolution process?
2 How can we properly model topic evolution caused by semantic interactions among keywords/topics?

To address these problems, we adopt the following two steps. Firstly, we transfer the discrete co-word network to the continuous semantic space with the help of deep learning language model, Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). By using the abstract text of the academic papers as the corpus, the semantic meanings of the keywords are modeled into word vectors. These vectors are reduced as points, using dimension-reduction method of t-Stochastic Neighbor Embedding (t-SNE) (Der Maaten & Hinton, 2008). These reduced points are drawn in a two-dimensional plane and used to build the triangulation network, thus generating a geographical-like space.

Then, we introduce a spatial analysis concept, spatial autocorrelation (Anselin, 2010), to help analyze semantic interaction patterns in the obtained geographical-like space. Spatial autocorrelation was originally designed to quantitatively detect leading or suppressing patterns (or spatial clustering patterns) for geospatial association phenomena or geospatial interaction events. Geographers and economists can explain how the geographic related events or phenomena are auto-correlated with each other through the geographical space. Considering the spatial autocorrelation factor, more accurate geospatial prediction models can be built. Thus, the models can help better policy decision-making for public health (Zhang, Qi, Jiang, Zhou, & Wang, 2013), the ecology system (Kuhn, 2006), spatial econometrics (Anselin, 2003), and urban system (Fan & Myint, 2014).

Economic development in the urban areas is a classical scenario. For example, Beijing in China plays an important economic role; however, geographical nearby cities are economically underdeveloped, because many workers from surrounding cities are absorbed into the Beijing labor market. In this case, Beijing has a negative impact on its neighbors, presenting as a High-Low suppressing impact pattern. On the contrary, Shanghai, another economic center in China, plays a positive impact role in economics of the surrounding cities. Because as a harbor city, Shanghai introduces more entrepreneurship opportunities, thus the economic development is positive correlated, presenting as a High-High leading pattern. Motivated by economic development pattern of cities, the following questions come to our minds: What keywords will be like Shanghai and what keywords will be like Beijing? What keywords will draw more attentions to the related research field like Shanghai creates more opportunities? What keywords will draw attentions too much that the related keywords are ignored like Beijing absorbs the working labors from its neighbors?

To answer these questions in this paper, we regard the keyword semantics as the landscapes and the keyword citations as the attributes of the land parcels. Thus, we can analyze the popularity dynamics of keyword semantics, like land use and land cover change (LUCC) process. The LUCC research may not be familiar to readers from the bibliometric or scientometric field, they are developed theories in the Geographical Information Science (GIS) field. Borrowing these GIS concept and methodology, we aim to provide a new approach to demonstrate whether the semantics affect the evolution process and visually understand the topic evolutions in this paper.

The rest of the paper is organized as follows: Section 2 describes the related work. The data and methodology are detailed in Section 3. By experimenting with the real literature datasets, we detail the results, discussing the identified patterns in concrete keywords and the evolution models in Section 4. Section 5 draws some conclusions and discusses the future work and possible trends.

2. Related work

2.1. "Ghost City" mapping

"Ghost City" is originally used to describe the districts with high building density and low population density (Jendryke, Balz, McClure, & Liao, 2017). In the "Ghost City" Mapping for topic evolution (Hu et al., 2018a), the geographic phenomenon is used as a metaphor to describe the topics that used to be hot but not recently. The idea of using "Ghost City" to describe certain topics is inspired by metaphor like Sleeping Beauty (Ke, Ferrara, Radicchi, & Flammini, 2015), which is used to describe the citing phenomena

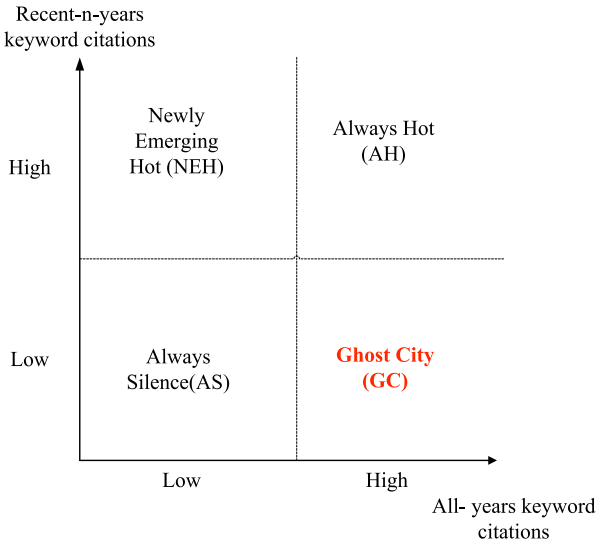

Fig. 2. "Ghost City" and related patterns.

of "delayed recognition" .

In the "Ghost City" frame, the different evolution patterns are analyzed by categorizing the keyword evolutions into four classes, namely the "Ghost City" area (high all-years citations, low recent-n-years citations), the Newly Emerging Hotspot area (low all-years citations, high recent-n-year citations), the Always Silent area (low all-years citations, low recent-n-year citations), and the Always Hot area (high all-years citations, high recent-n-years citations), as shown in Fig. 2.

Moreover, keywords are transformed to the high-dimensional semantic space using the Word2Vec model. Then the keyword vector points are reduced to the two-dimensional space using dimensional reducing method. In this way, semantic relationships in the evolution process are visually depicted in a two-dimensional continuous plane and the analysis is performed in a metaphorical way. This work provides a static evolution pattern analysis framework, which is inspiring. However, the analysis result is relatively primitive. Thus, we develop this idea furtherly by considering the keyword-level semantic interaction as the spatial interactions. Borrowing the concept "spatial autocorrelation" from GIS field, keyword semantic interaction patterns can be explained in a statistical hypothesis-test framing.

2.2. Spatial autocorrelation analysis

The spatial autocorrelation index was introduced in 1950s (Moran, 1950) and became mature in 1980s. Spatial autocorrelation was introduced for mineral detecting in the geology field. Later, many geographic events or phenomena were also demonstrated to be geospatial correlated and spatial autocorrelation can be used to build the prediction models for geographic event and phenomena. Taking landslide risk prediction as an example, the possibility of landslide in one location is often highly correlated to the possibility of the geographical nearby locations, as shown in Fig. 3. If the spatial autocorrelation is considered, the prediction accuracy for landslide risks will be greatly improved (Brenning, 2005).

As Fig. 3 shows, it is an image taken by camera from rightly above. Location A is nearer to the locations of B and C than to the location of D. As the spatial homogeneity suggested, the geographic-nearby locations are more likely to have similar geological formations. Moreover, landslides are also thought to have a knock-on effect, thus the geographical-nearer areas are more vulnerable. Therefore, landslides in one location are often highly correlated to the nearby locations, presenting a positive spatial autocorrelation.

If we regard the keyword semantics as land parcels, spatial autocorrelation can also help figure out spatial homogeneity and spatial heterogeneity regions in the keyword semantic mappings. Thus, the suppressing and leading patterns among the keyword popularity can be identified. With these patterns, we can answer the questions of how semantics affect the topic evolution process, and which keyword semantics have a positive or negative effect on promoting keyword popularity. In this way, we can understand the domain trends from the perspective of keyword semantic interactions. Also, the obtained patterns can be used as the basis of establishing the evolutionary prediction model.

3. Data and methodology

3.1. Data

Our experimental dataset is a typical literature dataset. By refining the searching conditions using topic search, we obtained 10,384 records for articles on the topic of "natural hazards". The time span was constrained to the period from 1985 to 2016. The searched indexes included the Science Citation Index Expanded (SCI-E) and the Social Science Citation Index (SSCI) located in the

Fig. 3. Spatial autocorrelation illustration for landslide possibility (Landslide possibility of A is highly spatially correlated with that of location B and location C, but low spatially correlated with that of location D.).

core collection of Web of Science (WoS). The article types were confined to "Articles" and the language were confined to "English". In addition, out of consideration that we are more familiar with the topic related to geography, we obtained the subdomain dataset of 614 literature of "geographical natural hazard" field by setting the topic search words as "geographic" and "natural hazard". The datasets are described in Table 1.

In Table 1, the column of "accumulated keyword counts" means the raw count of keywords without removing duplicate keywords. We choose to use the 1,791,232-word sentences extracted from the abstract text from "natural hazards" dataset to build up the corpus for training and use the 1667 keywords from "geographic natural hazards" dataset to conduct the analysis. Then, the word vectors of 1667 keywords are computed through the Google Word2Vec model, which is trained using the dataset of 1,791,232-word sentences. A two-dimensional dataset was obtained for subsequent processing and analysis using t-SNE methods to reduce the obtained word vectors, which will be discussed in detail in the methodology section.

3.2. Methodology

Our methodology can be described as an extensional visualization and analysis on the base of the reduced semantic mapping generated by the corpus formed by literature, as illustrated in Fig. 4.

In Fig. 4, the first step is to generate the semantic mapping. In this process, the abstract text of literature records are used as the corpus to construct the mathematical vectors for each of the keywords. The stop words are removed and all the rest words are stemmed. Then the texts in the abstract are sliced into sentence collections. The overall processing workflow is shown in Fig. 5.

In Fig. 5, Word2Vec as the neuro language model takes keywords as the input and the context in the sentences from the abstract texts as the output. After the training, the Word2Vec model can be obtained. Words are represented as points in the high-dimensional space through the Word2Vec model. We then use the simple average values of the several word vectors for each word contained in the keyword as the final semantic vector of the keywords (as keyword often contains more than one word). Though there are disadvantages of lacking the information of the orders of the word in the n-gram keyword, the average value can represent the semantic meaning well if the word count is not too large. The obtained word vectors are often vectors with high dimensions, often 100 dimensions or bigger. To make the obtained vectors easy to understand and analyze, we map the keyword vectors from the high-dimensional space to the two-dimensional space, as shown in Fig. 6.

In Fig. 6, left part of the image describes the keyword points, they are reduced results of the high-dimensional word vectors using the method of t-SNE. The method of t-SNE can help with visualizing the high-dimensional word vectors in a two-dimensional space.

Table 1
The overall view of the collected literature datasets.

Type	Literature records	Keyword counts	Accumulated keyword counts	Abstract word counts
Geographic natural hazard	614	1667	2789	121,535
Natural hazard	10,384	21,109	39,997	1,791,232

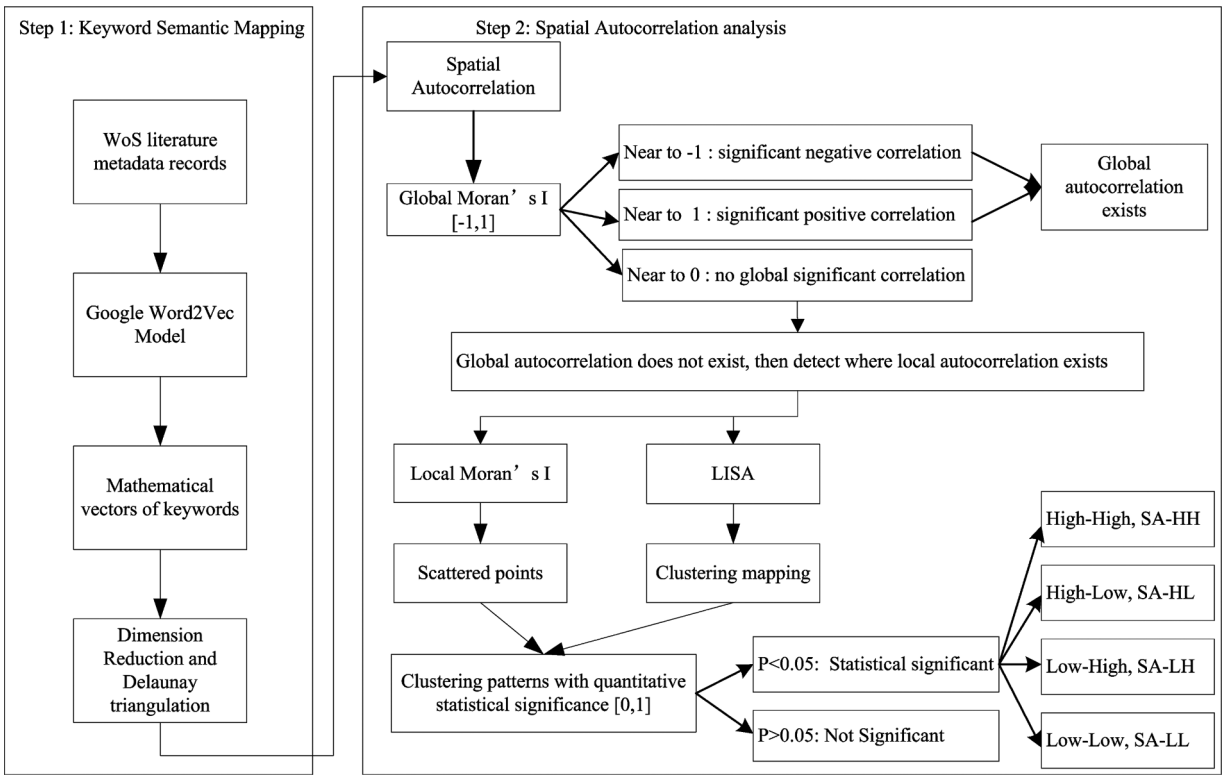

Fig. 4. The spatial autocorrelation analysis based on the reduced semantic space of keywords.

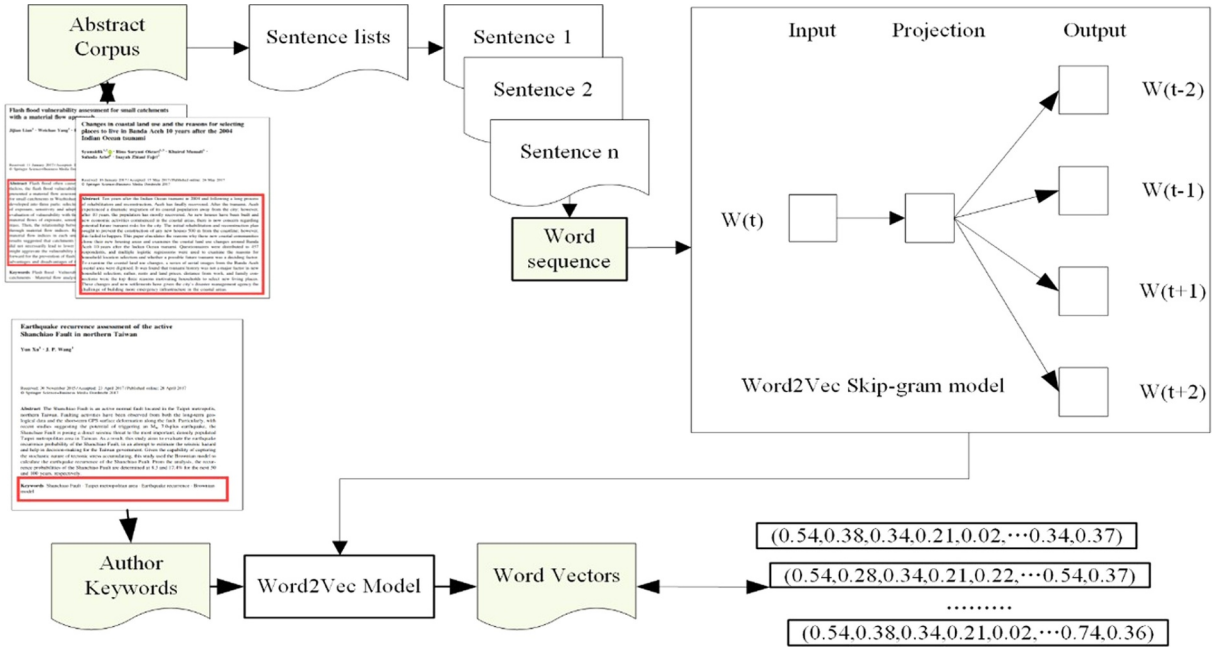

Fig. 5. The Word2Vec working process (Hu et al., 2018a).

The two-dimensional points are processed with the Software GeoDa for generating the triangulation network and Thiessen polygons, as shown in the right part of Fig. 6. Each of the Thiessen polygons contains a keyword point and stands for the keyword semantic region (Hu et al., 2018a). These polygons are processed as land parcels and the keyword citations are regarded as the land attribute.

In the second step, the spatial autocorrelation analysis is performed, including local indicator of spatial association (LISA)

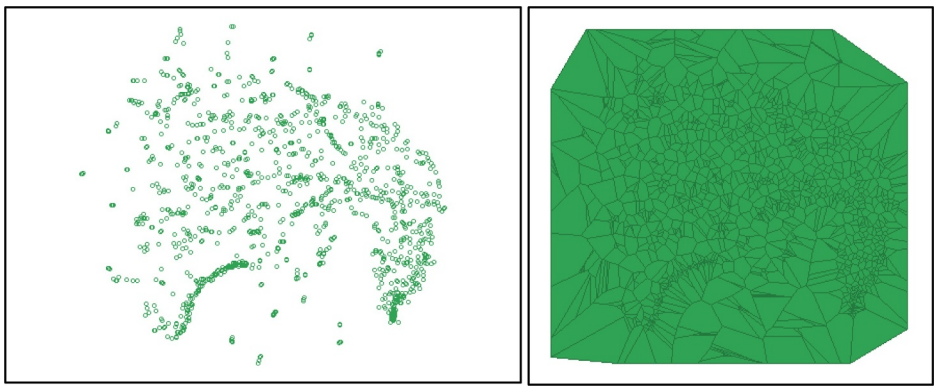

Fig. 6. The geometric points of the reduced two-dimension word vectors and generated Thiessen Polygons.

(Anselin, 2010) and Moran's I (Moran, 1950). The global Moran's I test is used to find whether global spatial autocorrelation exists. If the global spatial autocorrelation does not exist, then the LISA analysis and local Moran's I test are undertaken. Five spatial autocorrelation patterns, including “Not Significant”, “High High” (SA-HH), “High Low” (SA-HL), “Low High” (SA-LH), and “Low Low” (SA-LL) are identified in the analysis results. The threshold for judging whether the spatial autocorrelation is significant or not is set by experience. Often this threshold is set as 0.05 or 0.01. In this paper, the threshold of statistical significance is set as 0.05.

3.2.1. Moran's I computation and statistical measurement

Moran's I is designed to quantify the spatial autocorrelation of an interested variables that distributes across a geographical area. Moran's I is a rational number in the range of $[-1, 1]$. The closer the value approaches to 1, the stronger positive spatial autocorrelation it indicates. While the closer it goes to -1, the stronger negative spatial autocorrelation it shows. Those zero or approximated zero values indicate no spatial autocorrelation, i.e., randomness. The Moran's I can be written as following equation:

$$\mathrm{I} = \frac{N}{W}\frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})} \tag{1}$$

Specially, N is the number of the spatial units, subscript i and j denote the i*th* and j*th* unit of the studied area respectively. x_i is the

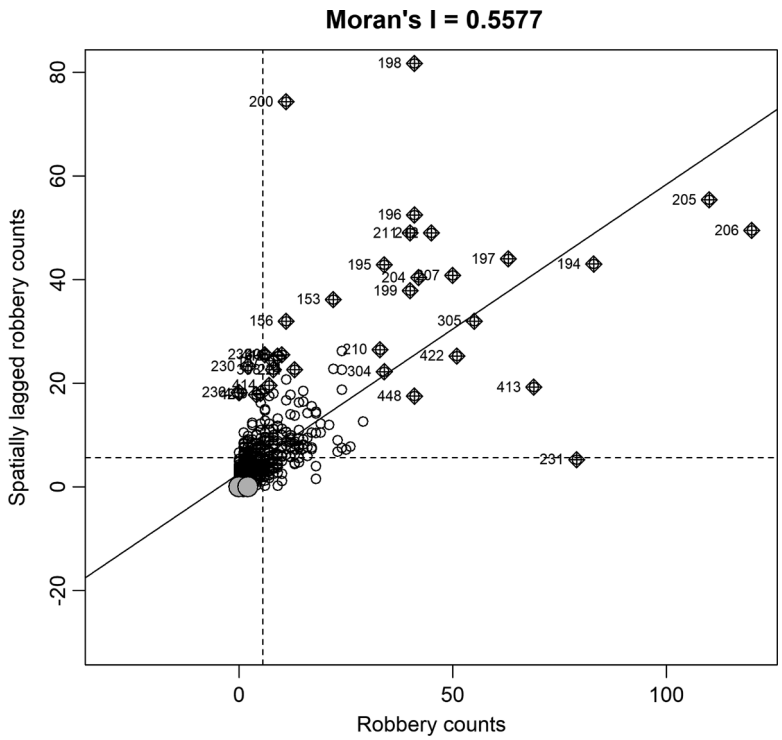

Fig. 7. Example result for Moran's I scattered point plot for describing the spatial autocorrelation for the robbery in the district of San Francisco.

interested variable value of the i*th* unit. In this paper, the interested variable is keyword citations. $\bar{x}$ is the average, w_{ij} is the entry of the i*th* row and j*th* column of the spatial weight matrix, which has zero values on the diagonal line. *W* is the sum of the value of all *wij*. Fig. 7 shows an example spatial autocorrelation for the robbery distribution in San Francisco.

In Fig. 7, the slope of the line in the center is the value of spatial autocorrelation which describes the global relativity. The value of Moran's I, 0.5577 is relatively high, meaning the spatial autocorrelation is existent in the robbery of San Francisco district. That is, the robberies in certain areas in San Francisco are tend to happen very frequently and other places are relatively safe.

However, in many cases, the Moran's I value may be close to 0, meaning no statistical-significant spatial autocorrelation is detected. When the global spatial autocorrelation is weak, Anselin introduced local indicator of spatial association (LISA) for local scale (Anselin, 2010). This work leads to the born of GeoDa[1] and related R packages like "spdep" (Bivand & Wong, 2018), which is used to produce Fig. 7. The horizontal axis stands for the value of target variables, i.e., the robbery counts in the example case. The vertical axis presents the weighted summation of surrounding spatial units. And the plot is divided into four quadrants representing four spatial patterns, upper right region, upper left region, bottom left region, and bottom right region, i.e., High-High (SA-HH), Low-High (SA-LH), Low-Low (SA-LL), and High-Low (SA-HL) patterns respectively. The scattered points in Fig. 7 present corresponding geospatial units, for example, a point in quadrant I implies a SA-HH pattern, i.e. its high variable value is surrounded by high values of its neighbors. A more detailed introduction of LISA is arranged in the following section.

3.2.2. LISA mapping and spatial cluster visualization

Moran's I is a statistic value for evaluating the overall autocorrelation degree of an interested variable distributing within some geographical area. However, a possible situation is that the global spatial autocorrelation is not statistically significant, while the local-level clustering can still be found. To detect this local clustering phenomenon, Anselin devised LISA to achieve this end (Anselin, 2010). Unlike the Moran's I, LISA is a set of values, each of which quantifies the association degree of a point or an area with its surroundings. And a linear combination of these values equal to the global Moran's I. The mathematical equation of LISA is

$$I_i = \frac{(x_i - \bar{x})}{m_2} \sum_j w_{ij}(x_j - \bar{x}) \tag{2}$$

where

$$m_2 = \frac{\sum_i (x_i - \bar{x})^2}{N} \tag{3}$$

In this way, the LISA, as seen in Eq. (4), represents the Moran's I value:

$$I = \frac{\sum_i I_i}{N} \tag{4}$$

where I is the Moran's I measure of global autocorrelation, I_i is local, and N is the number of analyzed units in the study area.

Spatial autocorrelation has two forms, global and local. The global spatial autocorrelation stands for a global trend of autocorrelation pattern, either positive or negative, varying in the range of $[-1, 1]$. If the global spatial autocorrelation is not statistically significant, then the LISA can be used to check whether the local clustering of SA-HH leading or SA-HL suppressing exist. Once a global or a local clustering is detected, an economist can use the clustering patterns to explain the economic phenomenon of interested regions.

In our case, the interested parts are the land parcels standing for keyword semantics. Through spatial autocorrelation analysis, we can figure out how the semantic meanings of certain keywords are affecting the popularity of the semantically related keywords.

4. Results

4.1. Moran's I scattered points and LISA clustering mapping

The construction of word semantic vectors is detailed in the methodology section. Through the generation of the semantic vectors of the keywords, the t-SNE based dimension reduction, and the Delaunay triangulation network construction, we obtained the reduced semantic keyword polygons two-dimensional mapping, as shown in Fig. 8.

In Fig. 8, each of the polygon stands for a keyword and they are stored in the "Shapefile" data format, which is a widely-used spatial dataset format. Considering the attribute of keyword citations, the spatial autocorrelation of the keyword citations in the semantic space can be computed, just as economical geographers dealing with the economic spatial autocorrelations phenomena in the administrative regions. The analysis process is performed with the software GeoDa and R packages to process keyword semantic mapping. Thus, we firstly get the Moran's I scatter points, as shown in Fig. 9.

In Fig. 9, the overall tendency of the spatial autocorrelation is indicated by the Moran's I value. Moran's I as the statistical test offers the quantitative measurement describing whether the clustering is statistical significant or not. If the value of Moran's I is closer to 1 or -1, then the global spatial autocorrelation is positive or negative. However, in this case, the value of Moran's I, 0.03676, is closer to 0. Thus, the global spatial autocorrelation is approximately not existent. We still can find some outliers in local districts as

[1] https://spatial.uchicago.edu/software.

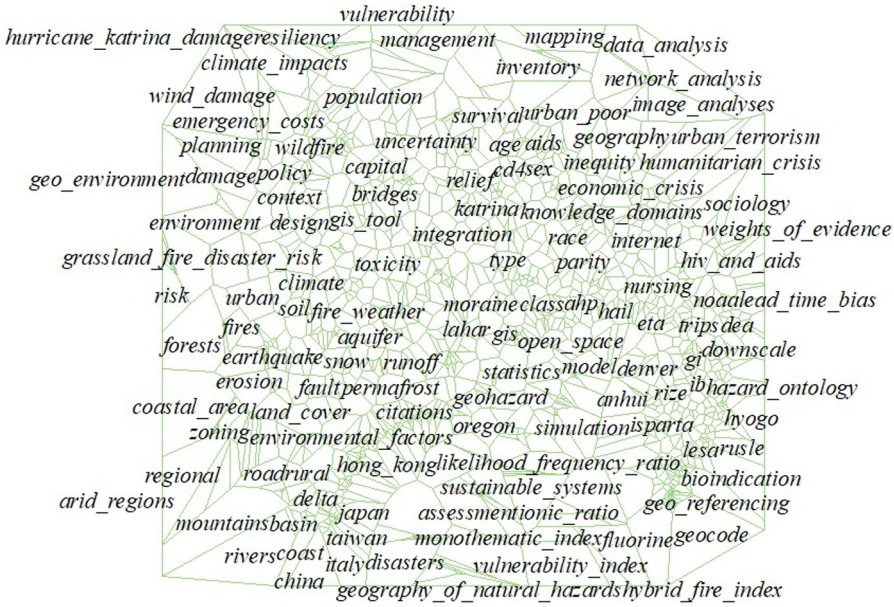

Fig. 8. The obtained keyword polygons generated by the Delaunay triangulation network.

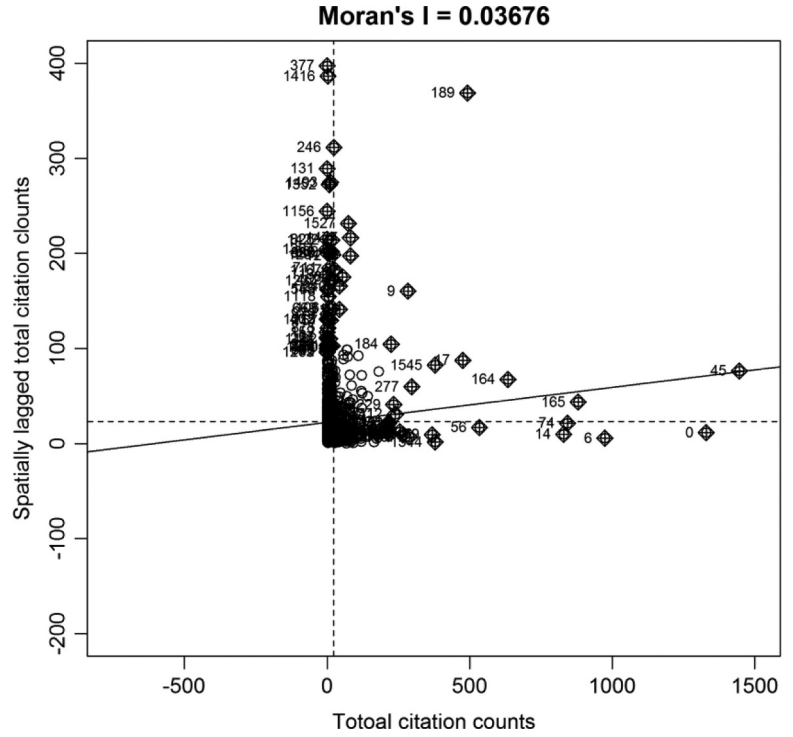

Fig. 9. Moran's I scatter points plot.

scattered point plot indicates.

Each of the points in the plot stands for a keyword polygon in the keyword semantic mapping. Different from the example of robbery spatial autocorrelation in the methodology section, the variants are keyword citations in our results. Keyword citations follow the definition in the introduction part. Considering keyword citation associated to each of the keywords as the variable, points in the upper right region indicate that the corresponding keyword citations are in a SA-HH pattern. The keyword citations of keyword points in this region are high and the keyword citations of the surrounding keyword points are also high. Correspondingly, points in bottom right, upper left, and bottom left regions indicate that the corresponding keyword citations are in a SA-HL, SA-LH, and SA-LL pattern, respectively. However, Moran's I scattered points do not reveal the clustering pattern well when only local spatial

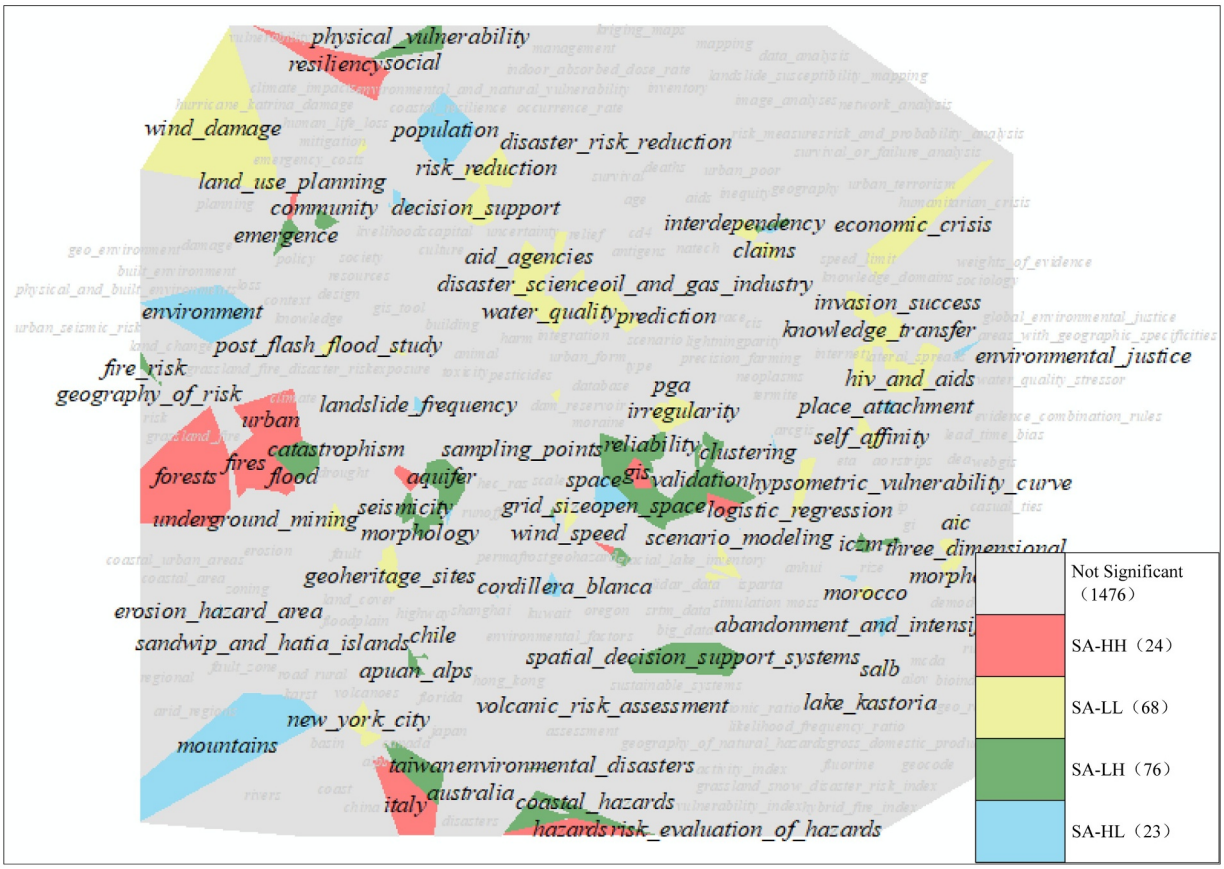

Fig. 10. LISA clustering mapping for the keyword land parcels. (The visualization is made in Software ArcGIS).

autocorrelation exists, because most points are not statistically-significant correlated. In this case, LISA clustering mapping can help identifying these patterns.

LISA clustering results for all the 1667 keywords in the two-dimensional results are shown in Fig. 10. The results are rendered with different colors in the software of ArcGIS, using the obtained field of "LISA_CL" after processing the polygon with GeoDa. The field of "LISA_CL" has five values from 0 to 4, standing for "Not Significant", "SA-HH", "SA-LL", "SA-LH", and "SA-HL", respectively.

In Fig. 10, 1476 keywords do not present significant spatial autocorrelations, but the rest keywords as the outliers are of interest to domain experts. 24 keyword polygons present SA-HH pattern, 68 keyword polygons present SA-LL pattern, 76 keyword polygons present SA-LH pattern, and 23 keywords present SA-LH pattern. To explore the clustering patterns in depth, we collected complete keyword list of different patterns in Appendix A. Some keywords are classified into a High-High (SA-HH) pattern, indicating that they will lead the semantically related keywords. Other keywords are in the High-Low (SA-HL) patterns, meaning they will suppress the surrounding keywords. To furtherly explore how these keywords present the SA patterns, we investigated SA-HH and SA-HL examples with detailed mapping visualization.

4.2. The High-High leading pattern and the High-Low suppressing pattern

To closely investigate the how High-High leading patterns are displayed, we choose the keyword "resiliency" as the observation target. "Social" as the neighbor keyword also presents a SA-HH pattern, as shown in Fig. 11.

In Fig. 11, the keywords "resiliency" and "social" have strong associations from semantic meanings. The resiliency for certain natural hazard often create social impacts, such as affecting the confidence that public has for the government. "Resiliency" has 284 keyword citations, and the related keyword of "social" has 83 keyword citations. Thus, the SA-HH pattern is detected and the popularity of "social" is led by the popularity of "resiliency".

Different from SA-HH, SA-HL stands for the suppressing pattern. Taking the "social vulnerability" from the SA-HL patterns as an example, keyword "social vulnerability" is surrounded by keywords with very few keyword citations, including "wildfire vulnerability", "flood vulnerability", "port vulnerability", "structural vulnerability", and "regional vulnerability", as shown in Fig. 12.

In Fig. 12, we can see that these keywords in the display area are closely related regarding semantic meanings. However, "social vulnerability" is more representative; the surrounding keywords are more detailed, thus obtaining far less attention. "Social vulnerability" has 164 keyword citations, but keywords "port vulnerability" and "wildfire vulnerability" have fewer than four keyword

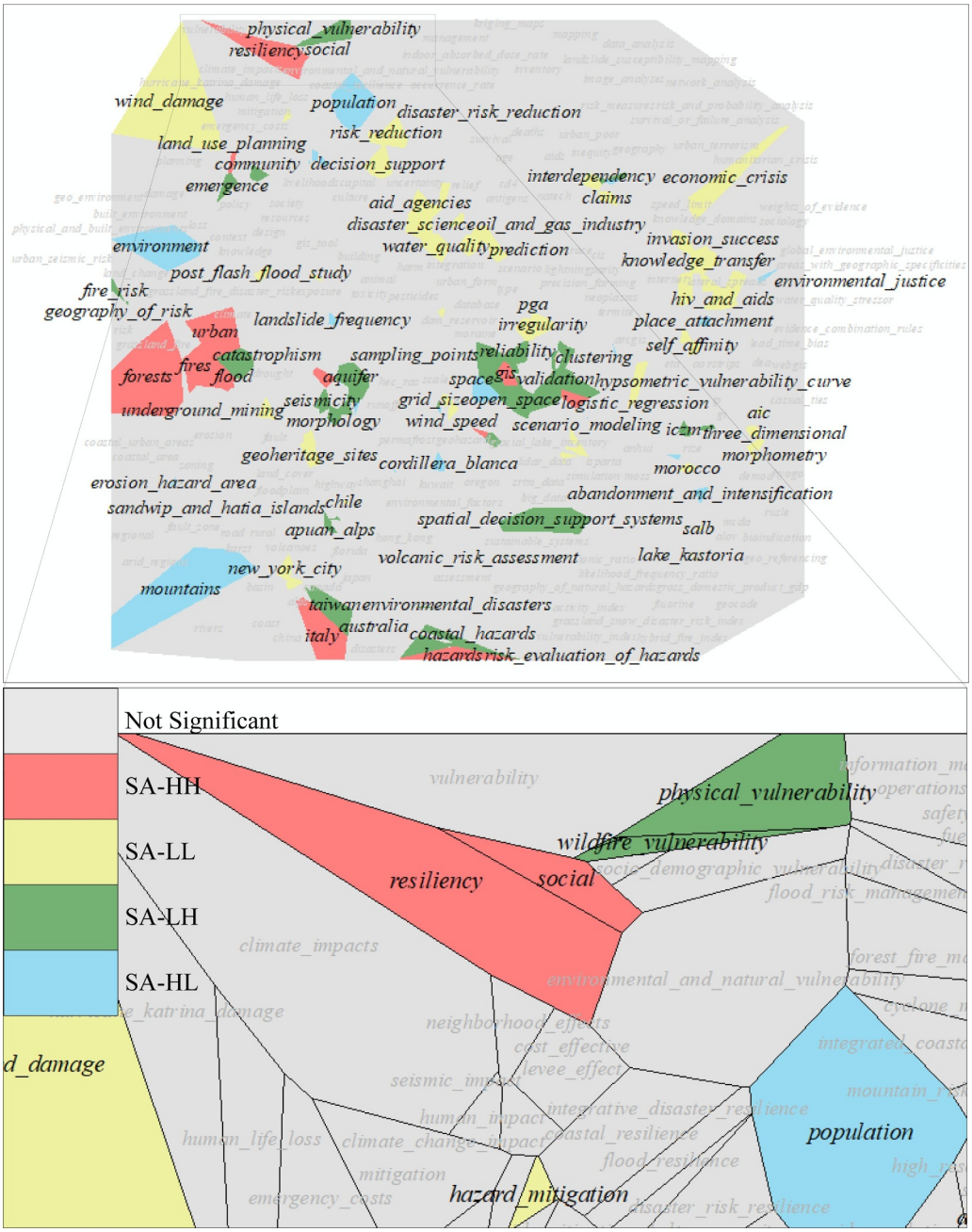

Fig. 11. The SA-HH example of keyword “resiliency” and its surroundings. (The visualization is made in the Software ArcGIS).

citations, respectively. Keywords in SA-HL patterns seem to be belonging to specific sub-topics of the mainstream natural-hazard research. The leading and suppressing patterns at certain moments are described, but with time goes on, how these patterns are changed and how these patterns can help build the prediction models are not clear.

4.3. Static visualization and dynamic prediction for topic evolution

To display the SA patterns in city-like environment, the three-Dimensional visualization of the semantic space of the year 2016 is made, as shown in Fig. 13.

In Fig. 13, the heights of the land parcels are proportional to the corresponding keyword citations of keywords in the year of 2016. Keyword “remote sensing” related regions are selected as the target for exhibiting temporal pattern, as all the typical SA patterns are completely shown in this area. Thus, by considering the timeline, we visualize the evolution dynamics for the years of 2010, 2013, and 2016, as shown in Fig. 14. Through this timeline-based visualization, the temporal changes of keyword citation or keyword

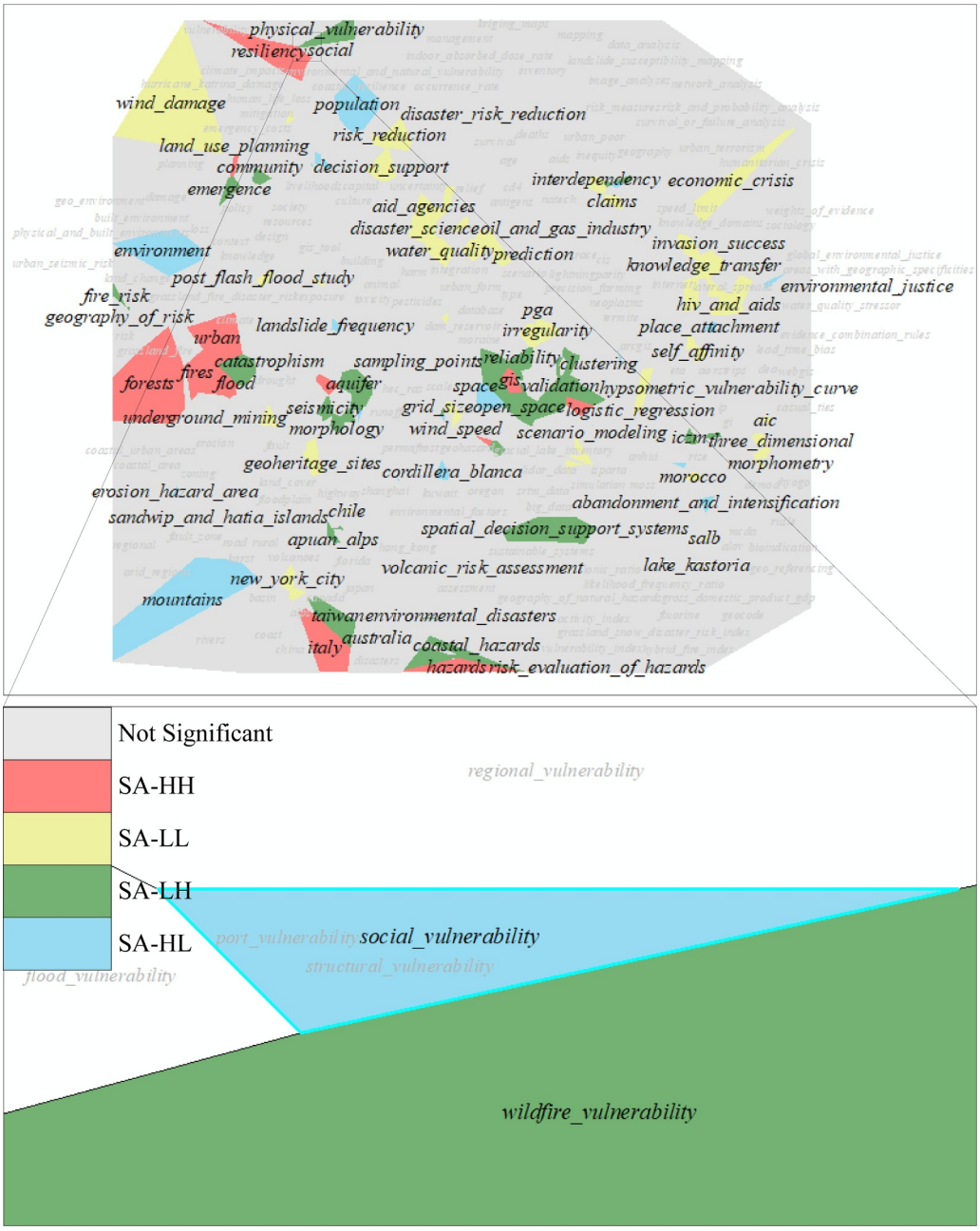

Fig. 12. The SA-HL example of keyword “social vulnerability” and its surroundings. (The visualization is made in Software ArcGIS).

popularity are displayed.

In Fig. 14, the heights of land parcels of the keywords are gradually growing like a developing city. With the visualization, we can explore the topic evolution in a way that geographers perceive a city. The visualization, however, is still a static presentation of the keyword citation at different moments. How the approach can be used to help model the topic evolution is still not clear.

To clarify how the approach can help model the evolution and further verify that the spatial autocorrelation can help build more precise prediction models, we conduct an experiment to compare prediction results by using two models, the Lagrange Multiplier (LM) and the Spatial lagging model (SLM). The difference lying between the two models is that LM does not take the spatial autocorrelations into consideration but SLM does. The filtered keywords are the top 20 keywords with the highest keyword citations. Using these two methods, the keyword citations of the 20 keywords are predicted, as seen in the Table 2. Also, we use the real-world datasets in the year of 2016 to compute the prediction accuracy of the used two methods.

As Table 2 shows, SLM results have smaller RMSE comparing with LM results, i.e., SLM has a better performance than LM. Comparing with the real world dataset, the fluctuation between the prediction values and the real world values can be evaluated

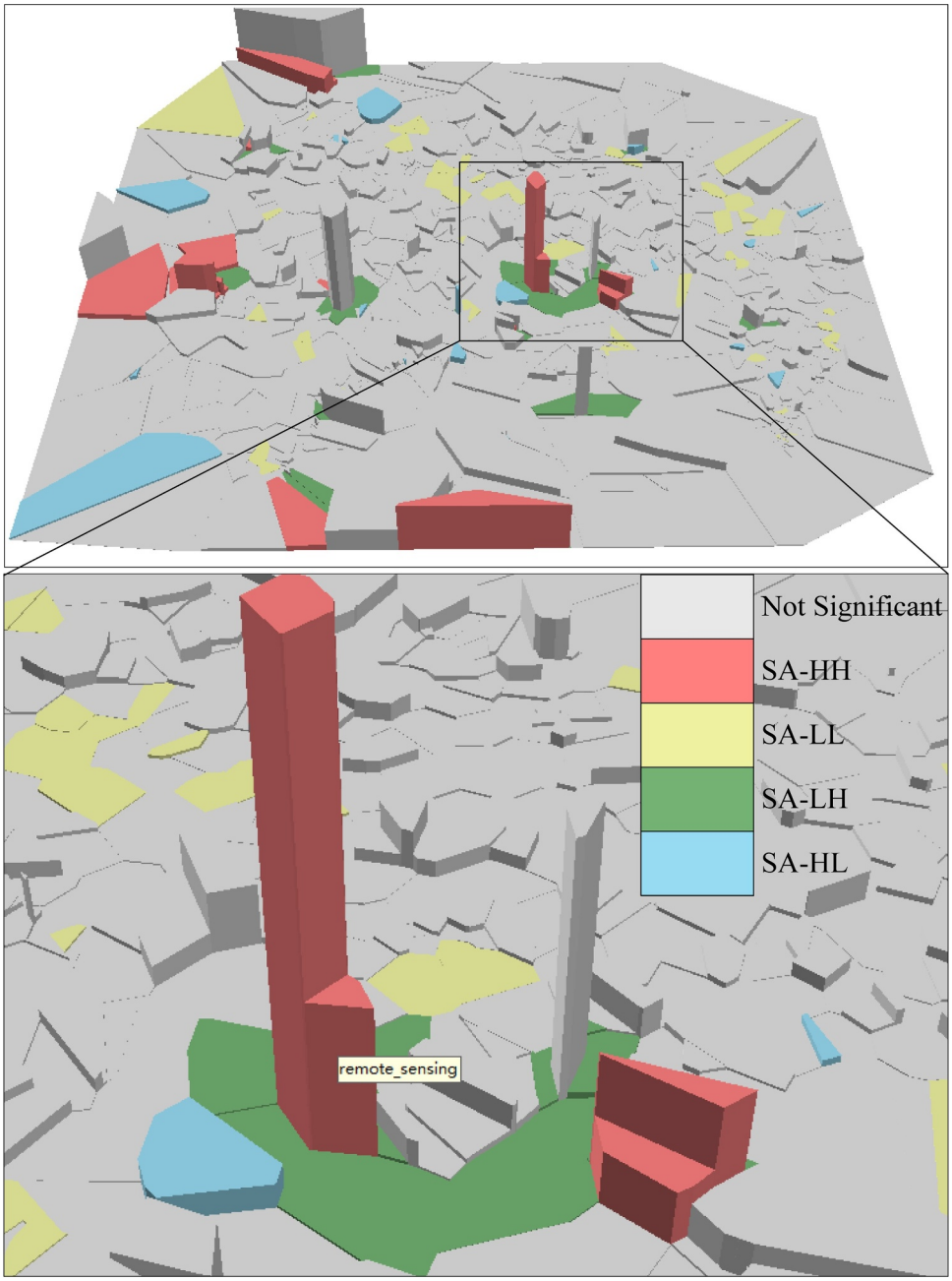

Fig. 13. The 3D visualization for spatial autocorrelation results in the year of 2016. (The visualization is made in Software ArcGIS).

through RMSE. From the table, it can tell that SLM obtains a better result with RMSE of 3.259 smaller than RMSE of the LM result, 3.289. Moreover, different from the work studied topic evolution in a complex network (Pobiedina & Ichise, 2016; Shibata, Kajikawa, & Sakata, 2012; Yu, Gu, Zhou, & Han, 2012), our approach takes the topic evolution as a continuous plane; keywords are treated as land parcels, thus SA based models can be applied and the evolution can be perceived like an evolving city. Our approach can provide at least an alternative choice, especially when the visually illustration and interpretations of the topic evolution are needed. Overall, the conclusion can be drawn that considering the spatial autocorrelation among the keyword semantics can help build more accurate and comprehensive models.

5. Discussion and conclusion

Overall, this paper presents a topic evolution analysis based on the keyword semantic mapping. Inheriting the idea of the "Ghost City" mapping for domain keyword semantics, this idea is further developed, using spatial autocorrelation methods. The SA-HH, SA-

Fig. 14. Spatial autocorrelation results in the years of 2010, 2013, and 2016. (The height corresponds to the count of keyword citations).

HL, SA-LH, and SA-LL patterns are identified in the topic evolution process. How the keyword semantics affect the surrounding keywords is illustrated. The model considering SA impacts is demonstrated to be effective in the topic evolution predictions through the experiment.

5.1. Implications

The implications of this work can be organized from four perspectives: interdisciplinary application, evolution mechanism explanation, new knowledge exploration schema, and different visualizations.

1) The proposed approach is an interdisciplinary application borrowing concepts from geospatial sciences to explore the interaction patterns among keyword popularity. The spatial autocorrelation concept may be unfamiliar to readers from scientometric domain,

Table 2
Prediction for total keyword citations in year 2016, comparing LM and SLM using the years of 2010 and 2013.

Keywords	Total keyword citation in 2016	LM PREDICTED	LM RESIDUAL	SLM PREDICTED	SLM RESIDUAL
landslide	153	158.125	−5.125	157.846	**−4.846**
gis	98	86.630	11.370	86.690	**11.310**
geographic information systems	83	82.336	0.664	82.409	**0.591**
logistic multiple regression	80	74.668	5.332	72.871	**7.129**
malaysia	74	78.944	−4.944	79.036	**−5.036**
geographic information system gis	53	50.008	2.992	49.926	**3.074**
vulnerability	53	52.834	0.166	53.363	**−0.363**
remote sensing	50	48.142	1.858	48.077	**1.923**
hazards	48	48.143	−0.143	47.805	**0.195**
natural hazards	45	44.470	0.530	44.405	**0.595**
risk	43	43.791	−0.791	44.401	**−1.401**
urban geology	29	26.723	2.277	26.599	**2.401**
land use planning	29	26.723	2.277	27.261	**1.739**
disasters	23	24.350	−1.350	24.273	**−1.273**
inequity	21	21.976	−0.976	22.165	**−1.165**
landslide susceptibility	18	21.297	−3.297	21.182	**-3.182**
runout	18	19.093	−1.093	18.298	**−0.298**
flood	13	13.781	−0.781	13.848	**−0.848**
risk assessment	11	10.164	0.836	10.042	**0.958**
resiliency	9	10.616	−1.616	10.957	**−1.957**
flash flood	1	2.760	−1.760	2.466	**−1.466**
Root Mean Square Error (RMSE)			**3.289**		**3.259**

Note: The keyword citations follow the definition in the introduction part.

the identified patterns by this method, however, is important for explaining the intrinsic mechanism that triggers keyword evolution. In the geographical field, many human activities or natural phenomena are closely related to the geographical space. Geographical nearby events are often highly correlated with each other. In this paper, we regarded and analyzed the semantic space as a geographical space, because the semantic similarity in the semantic space presents similar attribute with the geographical relatedness. Semantic similar keywords are more often seen together than those semantic-different keywords. Thus, the keyword-level interaction can be studied like geographical events or phenomena. To the best of our knowledge, this is the first time that the spatial autocorrelation method is used for analyzing the topic semantic space of scientific papers.

2) Our work has proposed a new explaining framework to explain topic evolution mechanism. Comparing with the previous method based on the discrete network built by the co-word relationships, our approach can help identify the associations and the potential semantic relatedness among keywords more completely. Moreover, to what degree the association is statistically significant can be clearly shown. Like the economical phenomena in the geographical space, some keywords or topics play the role of leading popularity of the semantic related keywords; other keywords play the suppressing role. Like the spatial autocorrelation analysis for the crime geography, police can better monitor and identify the crime events and find a safer district for tourism. The SA-HH or SA-HL can also help researchers to figure out what kind of keywords, topics, or research can be the focus of the public and academic communities. These patterns can also help institutes to make a better research plan.
3) With our approach, we can have a new knowledge exploration schema. In past scientometric analysis, only centralities are used to detect which keyword is more central. In the proposed approach, not only centrality is considered, the interaction modes are also included. The SA patterns will often appear in certain types of keywords. HH patterns often emerge in the keywords that represent more general concept, often standing for the collection of terminologies. The low patterns in the HL patterns often appear in the parallel but smaller concepts of the high pattern concept.
4) In addition, our work can also be regarded as the extension of the knowledge visualization with cartography means (Skupin, 2004; Skupin & Fabrikant, 2003), who used the Self Organizing Mapping (SOM) for cartographically depict the concepts. Based on the similar characteristics of manifold learning, the t-SNE adopted in this paper learns the distances in high-dimensional datasets and keep the distance relations among the keywords in the reduced two-dimensional space. Moreover, we also introduce the 3D visualization for interpreting the semantic space like urban areas.

5.2. Advantages, limitations, and future work

Our work has explored the keyword-level semantic interaction patterns. Also future topic evolution prediction modeling can benefit from these interactions. The leading or suppressing patterns provide structural information for different roles of the topics, similar to the work by Xiao, Chen, Sun, Han, and Zhang (2016), who used the k-core decomposition for the domain knowledge structure analysis. Differently, our approach is based on the continuous semantic space, which can help grasp the latent semantic relations among the keywords. Also the revealing patterns are based on hypothesis-test framework.

Moreover, the spatial autocorrelation patterns among the keywords are visualized in the two and three dimensional space. They are proved to be helpful in building better prediction models through comparing SLM and LM experimentally. Our work can be regarded as an analyzing framework; besides the keyword semantics, other factors like the impact of authors, journals, and academic

organizations can also be incorporated in this framework. Thus, we can explore topic evolutions with the manner how the geographers study urbanizations. A new insight for the topic dynamic mechanism is thus provided.

There are still some limitations in this work. Though the semantic meanings are modeled by using Word2Vec, the semantic meanings are counted on the unit of only keywords without considering synonym merging (merging the keywords with similar meanings as one semantic unit). Thus, we may need to automatically merge synonyms in the future, using theories like fuzzy sets. In addition, though the spatial autocorrelation patterns are extracted, the applications based on popularity interaction patterns are still primitive. More prediction models considering the spatial interactions like urbanization simulation can be introduced for the topic evolution in future research.

Acknowledgment

This work is supported by Open Research Fund of State Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University (Grant no. 18I04). The National Natural Science Foundation of China (no. 41371372) partially supports this research. This project is also partially supported by National Natural Science Foundation of China (no: 31771680 and 21706096), Natural Science Foundation of Jiangsu Province (no: BK20160162), Fundamental Research Funds for the Central Universities of China (no: JUSRP51730A), the Modern Agriculture Funds of Jiangsu Province (no: BE2015310), the New Agricultural Engineering of Jiangsu Province (no. SXGC[2016]106) , the 111 Project (B12018) and the Research Funds for New Faculty of Jiangnan University.

Notes: The data (WoS datasets and Shapefiles of the keyword polygons) and codes (Python and R) are available at the following link: https://github.com/hukaiwlw/UnderstandingTopicEvolutionUsingSpatialAutocorrelation.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ipm.2019.02.014.

Appendix A. All keywords of different clustering patterns under the statistical significant threshold of 0.05

Keywords (SA-HH, count: 24)	LISA_P
resiliency	0.011
flood	0.032
environmental hazard	0.040
hurricanes	0.028
gis	0.043
tsunami	0.009
italy	0.045
hazards	0.046
flash flood	0.029
remote sensing	0.004
land use planning	0.026
logistic regression	0.038
seismic hazard	0.013
flood hazard	0.001
urban	0.034
flood susceptibility	0.036
fires	0.026
volcanic hazard	0.033
forests	0.036
hurricane storm surge	0.006
social	0.010
glacier hazards	0.012
potential hazard	0.002
logistic multiple regression	0.032
Keywords (SA-LH, count:76)	**LISA_P**
geomorphology of mars	0.039
health hazards	0.029
environmental disasters	0.035
flash flood hazard	0.031
modis	0.020
reliability	0.002
spatial hazards	0.008
academic performance	0.021
accuracy assessment	0.047
shallow landslides	0.011
landslide volume	0.027
impact assessment	0.049
household income	0.043

north america	0.024
socioeconomic vulnerability	0.021
winter storm hazards	0.042
community	0.038
emergence	0.030
infrastructure vulnerability	0.037
digitizing	0.002
validation	0.043
seismic vulnerabilities	0.017
clustering	0.016
hydrometeorological hazards	0.033
global positioning system	0.019
geoenvironmental hazards	0.047
oceanic hazard	0.011
geomorphologic hazard	0.044
apuan alps	0.026
soil erosion susceptibility	0.032
aster dem	0.017
fire risk	0.038
open space	0.001
differential vulnerability	0.019
hyperspectral remote sensing	0.031
physical vulnerability	0.031
socio ecological systems	0.017
terrestrial lidar	0.011
catastrophism	0.039
iczm	0.038
coastal hazards	0.017
decision support system	0.014
wildfire vulnerability	0.011
hurricane hazards	0.034
integrated hydrological hazards	0.018
database management system	0.014
geography of risk	0.030
terrorism	0.042
geographically weighted regression	0.037
early warning system	0.020
risk evaluation of hazards	0.021
environmental vulnerability	0.028
aquifer	0.010
remote sensing and gis	0.031
australia	0.045
geographic visualization risk	0.039
visualizing risk	0.039
seismicity	0.002
landslide hazard	0.046
qualitative hazard	0.023
sampling points	0.008
disaster management system	0.016
disaster information system	0.024
natural disaster risk	0.044
morphology	0.004
central america	0.035
dominance	0.005
fire hazard	0.047
fluvial hazard	0.031
space	0.001
radar	0.004
sandwip and hatia islands	0.047
spatial decision support systems	0.031
taiwan	0.049
historical seismicity	0.004
earthquake hazard	0.023

Keywords (SA-LL, count:68)	LISA_P
risk communication	0.018
disaster risk reduction	0.038
semi qualitative method	0.041
statistical methods	0.035
algiers	0.038
time use survey	0.041
pga	0.010
multi criteria decision analysis mcda	0.012
chile	0.012

multivariate hazards evaluation method	0.006
index of entropy	0.028
hydrogeology	0.030
hypsometric vulnerability curve	0.044
urban transformation	0.050
prediction	0.002
irregularity	0.048
spatial smoothing	0.033
shannons entropy	0.043
wind damage	0.033
wind speed	0.010
earth system science	0.025
food security	0.049
hiv and aids	0.033
limit equilibrium method	0.009
aic	0.029
new york city	0.036
morphometry	0.029
underground mining	0.028
map collections	0.010
interdependency	0.046
interconnections	0.030
knowledge transfer	0.041
geoheritage sites	0.016
disaster science	0.014
road density	0.037
human ecology	0.036
morocco	0.021
tetouan coast	0.008
risk reduction	0.011
open system approach	0.003
post flash flood study	0.024
spatial reference	0.010
salb	0.029
oil and gas industry	0.023
copulas functions	0.048
fuzzy logic	0.009
natural hazard distribution	0.028
grid size	0.004
self affinity	0.034
aid agencies	0.020
water supply	0.011
decision support	0.002
water quality	0.047
rbf and idw algorithms	0.020
lake kastoria	0.015
hazards education	0.039
unstructured adaptive grid	0.017
settlement suitability	0.027
invasion success	0.026
three dimensional	0.045
claims	0.036
economic crisis	0.042
hazard mitigation	0.036
active faults	0.022
major hazard early alarm	0.029
coursework	0.023
risk vulnerability	0.026
real time	0.014

Keywords(SA-HL, count: 23)	LISA_P
social vulnerability	0.022
flood preparedness	0.026
debris flow	0.040
mountains	0.010
environmental justice	0.045
emanation coefficient	0.020
population	0.040
scenario modeling	0.007
environment	0.017
bushfire	0.044
adaptation planning	0.002
place attachment	0.043
natural disaster preparedness	0.047

erosion hazard area	0.032
abandonment and intensification	0.030
glacier retreat	0.039
cordillera blanca	0.020
attitudes	0.034
runout	0.023
volcanic risk assessment	0.027
etna	0.008
landslide frequency	0.001
tokyo	0.021

References

Anselin, L. (2003). Spatial externalities, spatial multipliers, and spatial econometrics. *International Regional Science Review, 26*, 153–166.

Anselin, L. (2010). Local indicators of spatial association—LISA. *Geographical Analysis, 27*, 93–115.

Arora, S., Bhattacharjee, D., Nasipuri, M., Malik, L. G., Kundu, M., & Basu, D. K. (2010). Performance comparison of SVM and ANN for handwritten devnagari character recognition. *arXiv: Computer Vision and Pattern Recognition*.

Bivand, R. S., & Wong, D. W. S. (2018). Comparing implementations of global and local indicators of spatial association. *Test, 27*, 716–748.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the international conference on machine learning* (pp. 113–120). .

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Brenning, A. (2005). Spatial prediction models for landslide hazards: Review, comparison and evaluation. *Natural Hazards and Earth System Science, 5*, 853–862.

Chen, B., Tsutsui, S., Ding, Y., & Ma, F. (2017). Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *Journal of Informetrics, 11*, 1175–1189.

Chen, K., Luesukprasert, L., & Chou, S.-C. T. (2007). Hot topic extraction based on timeline analysis and multidimensional sentence modeling. *IEEE Transactions on Knowledge and Data Engineering, 19*, 1016–1025.

Der Maaten, L. V., & Hinton, G. E. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605.

Fan, C., & Myint, S. W. (2014). A comparison of spatial autocorrelation indices and landscape metrics in measuring urban landscape fragmentation. *Landscape and Urban Planning, 121*, 117–128.

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., & Tsioutsiouliklis, K. (2012). Discovering geographical topics in the twitter stream. *Proceedings of the 21st international conference on World Wide Web* (pp. 769–778). ACM.

Hu, K., Gui, Z., Cheng, X., Qi, K., Zheng, J., You, L., et al. (2016). Content-based discovery for web map service using support vector machine and user relevance feedback. *PloS One, 11*, e0166098.

Hu, K., Qi, K., Yang, S., Shen, S., Cheng, X., Wu, H., et al. (2018a). Identifying the “Ghost City” of domain topics in a keyword semantic space combining citations. *Scientometrics, 114*, 1141–1157.

Hu, K., Wu, H., Qi, K., Yu, J., Yang, S., Yu, T., et al. (2018b). A domain keyword analysis approach extending term frequency-keyword active index with Google Word2Vec model. *Scientometrics, 114*, 1031–1068.

Jendryke, M., Balz, T., McClure, S. C., & Liao, M. (2017). Putting people in the picture: Combining big location-based social media data and remote sensing imagery for enhanced contextual urban information in Shanghai. *Computers, Environment and Urban Systems, 62*, 99–112.

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences, 112*, 7426–7431.

Kuhn, I. (2006). Incorporating spatial autocorrelation may invert observed patterns. *Diversity and Distributions, 13*, 66–69.

Lee, M., & To, C. (2010). Comparison of support vector machine and back propagation neural network in evaluating the enterprise financial distress. *International Journal of Artificial Intelligence & Applications, 1*, 31–43.

Mane, K. K., & Börner, K. (2004). Mapping topics and topic bursts in PNAS. *Proceedings of the National Academy of Science of the United States of America, 101*, 5287–5290.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Neural information processing systems* (pp. 3111–3119). .

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika, 37*, 17–23.

Nallapati, R. M., Ahmed, A., Xing, E. P., & Cohen, W. W. (2008). Joint latent topic models for text and citations. *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 542–550). *August*.

Newman, M. E. (2008). The mathematics of networks. *The New Palgrave Encyclopedia of Economics, 2*, 1–12.

Pobiedina, N., & Ichise, R. (2016). Citation count prediction as a link prediction problem. *Applied Intelligence, 44*, 252–268.

Pons-Porrata, A., Berlanga-Llavori, R., & Ruiz-Shulcloper, J. (2007). Topic discovery based on text mining techniques. *Information Processing & Management, 43*, 752–768.

Shibata, N., Kajikawa, Y., & Sakata, I. (2012). Link prediction in citation networks. *Journal of the Association for Information Science and Technology, 63*, 78–85.

Singla, R., Chambayil, B., Khosla, A., & Santosh, J. (2011). Comparison of SVM and ANN for classification of eye events in EEG. *Journal of Biomedical Science and Engineering, 04*, 62–69.

Skupin, A. (2004). The world of geography: Visualizing a knowledge domain with cartographic means. *Proceedings of the National Academy of Science of the United States of America, 101*(Supplement 1), 5274–5278.

Skupin, A., & Fabrikant, S. I. (2003). Spatialization methods: A cartographic research agenda for non-geographic information visualization. *Cartography and Geographic Information Science, 30*, 99–119.

Tu, Y., & Seng, J. (2012). Indices of novelty for emerging topic detection. *Information Processing & Management, 48*, 303–325.

Wang, Z., Li, G., Li, C., & Li, A. (2012). Research on the semantic-based co-word analysis. *Scientometrics, 90*(3), 855–875.

Xiao, L., Chen, G., Sun, J., Han, S., & Zhang, C. (2016). Exploring the topic hierarchy of digital library research in China using keyword networks: A K-core decomposition approach. *Scientometrics, 108*(3), 1085–1101.

Yu, X., Gu, Q., Zhou, M., & Han, J. (2012). Citation Prediction in Heterogeneous Bibliographic Networks. *Siam international conference on data mining* (pp. 1119–1130). .

Zarrinkalam, F., Kahani, M., & Bagheri, E. (2018). Mining user interests over active topics on social networks. *Information Processing & Management, 54*(2), 339–357.

Zhang, A., Qi, Q., Jiang, L., Zhou, F., & Wang, J. (2013). Population Exposure to PM2.5 in the Urban Area of Beijing. *PloS One, 8*(5), e63486.

Zhao, X., Jin, P., & Yue, L. (2015). Discovering topic time from web news. *Information Processing & Management, 51*, 869–890.