Scientometrics (2018) 114:1141–1157
https://doi.org/10.1007/s11192-017-2604-7

Identifying the "Ghost City" of domain topics in a keyword semantic space combining citations

Kai Hu[1,2] · Kunlun Qi[3] · Siluo Yang[4] · Shengyu Shen[5] · Xiaoqiang Cheng[6] · Huayi Wu[1,2] · Jie Zheng[1,2] · Stephen McClure[1,2] · Tianxing Yu[1,2]

Received: 24 June 2017 / Published online: 5 December 2017

Abstract As an increasing number of scientific literature dataset are open access, more attention has gravitated to keyword analysis in many scientific fields. Traditional keyword analyses include the frequency based and the network based methods, both providing efficient mining techniques for identifying the representative keywords. The semantic meanings behind the keywords are important for understanding the research content. However, traditional keyword analysis methods pay scant attention to semantic meanings; the network based or frequency based methods as traditionally used, present limited semantic associations among the keywords. Moreover, the ways in which the semantic meanings behind the keywords are associated to the citations are not clear. Thus, we use the Google Word2Vec model to build word vectors and reduce them to a two-dimensional plane in a Voronoi diagram using the *t*-SNE algorithm, to link meanings with citations. The distance between semantic meanings of keywords in two-dimensional plane are similar to distances in geographical space, thus we introduce a geographic metaphor, "Ghost City" to describe the relationship between semantics and citations for hot topics that have recently become not so hot. Along with "Ghost City" zones, "Always Hot", "Newly Emerging

✉ Kai Hu
hukai@whu.edu.cn

✉ Huayi Wu
wuhuayi@whu.edu.cn

[1] The State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

[2] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China

[3] Faculty of Information Engineering, China University of Geosciences (Wuhan), Wuhan 430074, China

[4] School of Information Management, Wuhan University, Wuhan 430072, China

[5] Soil and Water Conservation Department, Yangtze River Scientific Research Institute, Wuhan 430010, China

[6] Faculty of Resources and Environmental Science, Hubei University, Wuhan, China

Hot", and "Always Silent" areas are classified and mapped, describing the spatial heterogeneity and homogeneity of the semantic distribution of keywords cited in a domain database. Using a collection of "geographical natural hazard" literature datasets, we demonstrate that the proposed method and classification scheme can efficiently provide a unique viewpoint for interpreting the interaction between semantics and the citations, as "Ghost City", "Always Hot", "Newly Emerging Hot", and "Always Silent" areas.

Keywords Ghost City · Semantic space · Keyword analysis · Word2Vec · *t*-SNE · Spatial analysis

Introduction

As an increasing number of scientific literature datasets are open access, more attention has gravitated to citation analysis and developing new ways to the mine these datasets. The mining and analysis of literature datasets can help researchers identify important topics, find emerging questions and topics in a field, and thereby make appropriate research plans for future work. Keyword analysis is one of the most commonly used methods among the data mining approaches applied to literature datasets. A collection of keywords encapsulates the knowledge in a certain field, providing a quick look of a field at a particular point in time.

Keyword analysis includes two types of approaches: frequency based and network based methods. In the frequency based methods, the frequency of the keyword is used as an important evaluation index of the semantic importance of the term. Frequency based methods have the advantage of high computational efficiency with low computational complexity. However, these types of methods lack the associational information for keyword pairs. This drawback can be overcome by network based methods. The co-word based network was proposed by Callon et al. (1983). In a co-word network, a keyword is a node and an edge between two keywords represent a co-occurrence of the two keywords. If two keywords appear in one paper, then the co-occurrence of the two keywords will be considered as one. More co-occurrences result in a higher weight of the edges between the keywords. Based on the network generated by the keywords, many analyses can be conducted using the Social Network Analysis (SNA) metrics, including centrality measures such as node degree, betweenness degree, eigenvector degree, and closeness degree (Borgatti and Everett 2006). Different SNA metrics can help describe the roles that keyword nodes play in an entire keyword network, thus providing in-depth analysis of both the network and the topic denoted by a specific keyword. However, there are still some limitations, as inaccurate descriptions of the association occur among keywords. The co-word network is constructed by the co-occurrence of the keywords in scientific papers, and keywords are often sparsely distributed in the papers. Many keywords do not appear in the same paper; however, they may have potentially important connections, which cannot be efficiently described by the only keyword co-occurrence, absent the keyword context. Moreover, in keyword or topic evolution studies, previous work rarely takes the citation into consideration, only occurrences of keywords. The possible association between the semantic meanings and the citations is not effectively revealed, thus creating barriers to understanding and grasping the significance of research foci. We address these research questions in our paper:

1. How to depict the potential relations among the keywords on semantic level rather than word level or sparse co-word network level?
2. How to visualize the evolution patterns of a topic in a continuous semantic space in terms of citations rather than occurrences?

To address these questions, we propose a way to describe potential semantic connections among keywords, depicting the keyword distribution intuitively by introducing spatial analysis and visualization techniques borrowed from the geospatial sciences. To describe the potential semantic connections among the keywords, we introduce the Google Word2Vec word representation model which is capable of depicting the semantic associations among different keywords in a certain corpus. The default word vectors generated by Google Word2Vec are 100-dimensional. To intuitively depict the distribution of the keywords, we introduce the *t*-distributed stochastic neighbor embedding (*t*-SNE) algorithm (van der Maaten and Hinton 2008) to map the keyword onto the two-dimensional plane. We use the "Ghost City" metaphor originally used to describe the phenomena of empty in urban districts in the geographical science field, to explore and cartographically map the spatial-temporal distribution of citations in a semantic space.

A ghost city is a term used in the popular press to describe cities and urban areas with high housing vacancy rates. They were a much discussed phenomena, reputedly occurring in second-echelon cities across China, reflecting uneven development between housing demand and supply. Like ghost cities, key words and terms in the semantic space of keyword citations could also appear, change over time, and vanish. For example, the keyword of "grid computing" was hot with high citation counts. However, with the introduction of "cloud computing" and its success in the commercial field, the frequency of the "grid computing" keyword declined while cloud computing became a new hotspot. Research areas can be always hot, newly hot, or no longer hot; dynamic understanding emerging and receding topics and concepts outlines the temporal development of a domain. A geospatial analysis can be applied to semantic space.

To verify that the proposed idea is efficient and helpful, we conducted an experiment using a dataset retrieved from the core collections of Web of Science (WOS). As we are familiar with the geographic field, we use the dataset of geographic natural hazards to build up the corpus and word vectors. Then the word vectors are reduced to the two-dimensional planes. With the geospatial process, the keyword points are transferred to the Delaunay Triangulation Network and Thiessen Polygons. Based on the Thiessen Polygons, the citation counts of all year and recent n years are collected. By division the citation counts to different levels of high and low, the binary classification can be done and four composed types can be generated: the "Always Hot" area, the "Newly Emerging Hot" areas, "Ghost City", and the "Always Silent" areas. We demonstrate that our proposed method can effectively expose domain topic evolution patterns and suggest possible future study directions.

The rest of the paper is organized as follows: second section introduces the motivations behind this paper. Third section describes the related work and the current problems. The data materials and the methodology are illustrated in fourth section. Middle results and generated "Ghost City" mappings are described and analyzed in fifth section. Sixth section presents the discussion on the obtained results. Final section draws some conclusions on the advantages, limitations, and the future work plans.

Motivations: metaphors and interdisciplinary studies

Our approach is motivated by two types of studies, the metaphor based study and the interdisciplinary studies. Metaphors such as “sleeping beauty” (Ke et al. 2015) and “prince” (Teixeira et al. 2016) describe non-linear patterns in the emergence of research papers in the sciences. In line with intuitive cognition, “sleeping beauty” is a metaphor for scientific articles that at first do not receive attention but latter, suddenly attract attention and many citations. The “sleeping beauty” concept captures how important research hotspots often emerge in the literature, and indicates how the start of the new subfield in a domain can develop in a non-linear, non-cumulative way. Similarly, the concept of the “prince” like the idea of a sleeping beauty, is also drawn from folklore, and acts as a metaphor denoting a critical canonical scientific text that starts a high-citation pattern in a “sleeping beauty” research topic. Sleeping beauties often emerge from self-citations by famous scholars who cite their lesser known early work in their later publications. Concepts such as metaphor borrowed from literature, provide an interpretative framework to understand phenomena, scientifically. As reported in reference (Uzzi et al. 2013), over 170 million articles in the Web of Science (WoS) database show what kinds of papers are highly cited; the most highly cited papers are those which are most accessible and relevant to a wide audience. Papers treating advanced topics may be lost to a wider audience because people cannot understand technical language or topic. A metaphor-based interpretive framework helps to increase understanding.

In addition, the interdisciplinary study generates inspiration for innovation. Text-based mining techniques can be innovatively used to derive semantic information from empirical location data in the geospatial analysis research, considering 2-dimensional coordinate as texts. These kinds of studies can be regarded as interdisciplinary applications. The text mining methods belonging to computer science were borrowed to the geospatial analysis domain. For example, Latent Dirichlet Allocation (LDA) is classical topic modeling model, successfully applied in the field of text mining. Non-negative Matrix Factorization (NMF) has also been used to deal with the text-based data mining. Interesting to note that LDA (Zhang et al. 2016) and NMF (Kang and Qin 2016) algorithms were innovatively applied in analysis of the mobility patterns of the moving vehicles. Vehicle trajectories were treated as sentences and paragraphs to extract the semantic meanings and interpret these data as mobility patterns. These methods can be combined to process the two or three dimensional datasets spatially, as they can be regarded as text and mapped into the high-dimensional semantic space. Classification of traces into mobility patterns was executed using hyperplanes by processing then as a high-dimensional semantic dataset. We regard this research as a measure of the success of the interdisciplinary applications; interdisciplinary thinking engenders new insights and challenging new ideas for studies in a given domain. Conversely, classification results from a purely semantic datasets such as mobility patterns can therefore, be spatially interpreted and visualized. Inspired by these approaches we apply geospatial analysis to explore, interpret, and visualize the semantic space formed by keyword citations.

Thus, we propose to use the geographical metaphor “Ghost City” to depict one pattern in the evolution of topic and introduce geospatial analysis methods to visualize and interpret the associations between keyword semantics and citations.

Related work: topic evolution analysis frame and change patterns

Traditional topic evolution analyses are integral to most scientometric papers. The most common methods slice the time periods into different stages, and calculate quantitative indexes for these periods. The evolution of topics is analyzed and interpreted using these time slice based datasets. Keyword frequency may still be most frequently used metric, but researchers now tend to use more complex indexes to evaluate the topic evolution (Song et al. 2014). Researchers identify hot topics for certain time periods using topic modeling methods like the Latent Semantic Index (LSI) and Latent Dirichlet Allocation (LDA). The evolution trend is analyzed based on the change in topics. For example, LDA method is used to explore the "stem cell" research evolutions (Wu et al. 2014). By regarding the articles as bag of words, every document corresponds to certain topics with certain possibilities, and every topic is corresponding to a set of keywords with certain possibilities. Thus, document clusters can be used to generate a list of topics prevalent in different time periods.

Document clusters can also be generated by community detection methods (Newman and Girvan 2004) based on co-citation or co-author networks. For example, the co-author network standing for the scientific collaboration was used to generate the scientific collaboration communities to explore the topic dynamics in the field of information retrieval (Yan et al. 2012; Zheng et al. 2017). Each community corresponds to several published papers, i.e. the document clusters. Documents in different periods are used to generate the topics. Thus a topic evolution analysis can be performed. Other evolution analysis can also be studied via the different types of the network, such as the co-cited reference network (Small 1973), bibliographic coupling network (Kessler 1963), or author keyword coupling network (Yang et al. 2016). The most frequently used network is the co-word network (Callon et al. 1983).

A co-word network regards each keyword as a node and the co-occurrence of every two keywords is presented as an edge. Many keyword analyses are still based on the co-word network. By incorporating time stamps and keywords, the co-word networks are used to generate different sub-graphs for different time periods. By link-reduction methods, the pivot keywords with high centrality (node centrality degree, eigenvector centrality degree, or closeness centrality degree) (Borgatti and Everett 2006) are preserved. By visualizing the generated sub-graphs, the evolution of topics over time are presented as different clusters during different time periods. Scientometric software like CiteSpace (Chen 2006) and HistCite (Garfield 2009) supports these analyses. Taking CiteSpace as example, by inputing the WoS data record file, co-word or co-citation network output can be organized as a time-line or time-zone view, thus enabling visualization for evolution analysis. Various scientometric analyses of individual domains have employed this method to understand domain topic evolutions, such as anti-cancer studies (Xie 2015) and night-time light remote sensing studies (Hu et al. 2017).

Based on this framework, topic evolution patterns themselves become research foci. Typical evolution patterns receiving attention are emerging sets of related keywords. These "burst words" are keywords suddenly coming into use as detected from the literature (Mane and Borner 2004). These sudden onset keywords can be identified by the bust detection algorithm (Kleinberg 2003). Other change patterns described as the steady, concentrating, diluting, sporadic, transforming, and emerging topics have also been explored (Yan 2014).

These approaches do not directly use the citation in topic evolution analysis but consider the topic or keyword occurrences. A citation however, can be used to provide an effective frame for understanding topic evolutions (He et al. 2009). Thus, we propose a different frame using the semantic space of the keywords instead of network based framing. Moreover, we use citation counts to evaluate the keywords. As distinct from the usage of the term "Ghost Topics" in reference (He et al. 2009), the geographic concept "Ghost City" used in our paper is the foundation for our metaphor based analysis and visualization, and also provides a point of entry for spatial analysis methods and urban planning concepts in keyword analysis.

Methodology and data

Data

Our experimental dataset is the typical literature dataset in many scientometric studies. By refining the searching conditions using topical searches, we obtained 10,384 records for articles about "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 core collection of Web of Science (WoS). The article types were confined to "Articles" and the language to "English". In addition, out of consideration of the familiarity of the subject, we obtained the sub-domain dataset of 614 literature of "geographical natural hazard" field by setting the topic search words as "geographic" and "natural hazard". The dataset is described in Table 1.

As shown in Table 1, the accumulated keyword counts indicate the raw count of keywords without removing duplicate keywords from different articles. From this dataset, we chose to use the 1,791,232 words extracted from the abstract datasets for "natural hazards" to build up a corpus and used the 1868 keywords for "geographic natural hazards" to conduct the analysis. In turn, we created word vectors by processing the 1868 keywords with Google Word2Vec model. A two-dimensional dataset was obtained for subsequent processing and analysis using word vectors and *t*-SNE methods as discussed in the methodology section.

Workflow

Our methodology can be described as a workflow of key processes including keyword vector construction, dimension reduction using the *t*-SNE algorithm, Delaunay

Table 1 The overall view of the collected literature datasets

Type	Literature records	Keyword amount	Accumulated keyword counts	Abstract word counts
Geographic natural hazard	614	1868	2789	121,535
Natural hazard	10,384	21,109	39,997	1,791,232

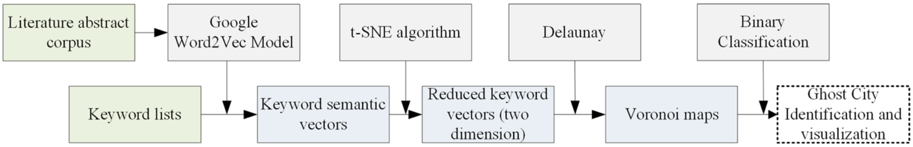


Fig. 1 Workflow for “Ghost City” identification; content in the dashed-line rectangles are the results of our approach

Triangulation Network construction (Voronoi maps), Thiessen polygon generation, binary Ghost City classification as illustrated in Fig. 1.

In Fig. 1, word vector construction preprocessing, is executed with the Google Word2Vec model (Mikolov et al. 2013). The idea of Word2Vec is not new, but only became widely known after 2013 when the paper (Mikolov et al. 2013) has been published. Since then Word2Vec model was widely adopted in many fields. The model maps words found in a corpus to semantic space, represented by vectors such as $(0.34, 0.37, 0.28, \ldots, 0.47)$. A vector length is often set to a hundred dimensions. It is similar to a co-word matrix, expressed as $(0, 1, 0, \ldots, 0)$, but denser and smaller, making vectors more suitable for computation than co-word matrices. Because keywords are unique, a matrix generated by co-occurrence can be very sparse and thus the dimensions can be far larger than the 100 dimensions in the vectors generated by Word2Vec, therefore vectors generated by Word2Vec are computationally efficient. In addition, the Word2Vec model adopts two classical prediction techniques the Continuous Bag of Word (CBOW) and the Skip-gram approaches. Specially, CBoW uses a context word to predict the possibility of the emergence of certain word, while the Skip-gram model uses a current word to predict the surrounding context. These techniques are based on the context in the corpus; prediction quality is associated with the corpus. For example, if only “I love you” and “I hate you” appear in the corpus, the word “love” will be considered as conveying the same meaning as “hate”, as semantic meaning is based on the context computations. In large-volume corpus training, a generated word vector can deliver accurate analogies, as illustrated by the example of “China”-“Beijing”=”Japan”-“Tokyo”. Therefore, the Word2Vec model is regarded as an effective model for obtaining the semantic meanings behind the words.

The generated word vector is highly dimensional, making it impossible to intuitively understand and interpret. The more common quantitative method evaluating texts however, deploys dimension reduction algorithms like Principal Component Analysis (PCA) and *t*-SNE to create visualizations. Different from linear reduction PCA methods, *t*-SNE is a non-linear dimension reduction approach that uses machine learning techniques. The core idea behind *t*-SNE is manifold learning, considering the intrinsic of a dataset distribution providing an intuitive depiction of it. Though the information loss is inevitable in the process of reduction, *t*-SNE can preserve the distance relations between high-dimension points and is relatively fast.

We obtained two-dimensional datasets through dimension reduction depicting keywords as points on a two-dimensional plane, every point is a keyword. The distance between two keywords expresses the semantic similarity between them. Closer keywords are more similar to each other, keywords farther apart are less similar. Merely visualizing the similarities and dissimilarities of keywords is not enough. The extent to which keywords are semantically precise or vague must be described to understand the potential space of meaning. To visualize semantic vagueness, we use a Delaunay Triangulation Network construction algorithm to obtain the Thiessen polygons to express the semantic ranges of keywords.

As the polygons are places in the two-dimensional planes and the distance between the polygons have real-world semantic meanings, we define two additional attribute dimensions of “all year keyword citation count” and “citation counts from recent n years” to help depict the citation patterns of all years and for the recent n year period. The metaphor of Ghost Cities acts as an interpretive frame to understand keywords that become hot and then disappear. By using the Binary classification method, the different types can be obtained including the “Always Hot” area, the “new emerging” area, the “Ghost City” area and the “Always Silent” area, as illustrated in Fig. 2.

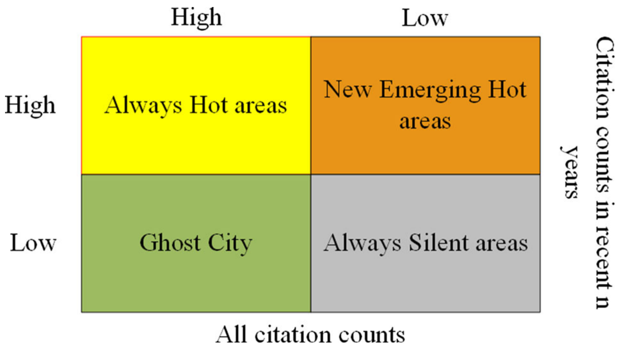
The four types of change patterns shown in Fig. 2 were classified by a binary scheme for citation counts and citation counts in recent years using high or low values. We use the term “Ghost City” to refer to previously hot but recently not so hot topics. The “Always Hot” category has high “all citation counts” and “recent n-year citation counts”, representing core topics in a domain. “New Emerging Hot” category has high values for “recent n-year counts” but low “all citation counts” values, this category indicates sleeping beauties, keywords emerging early but did not gain attentions until only recently; they could also be the newly emerging topics that have never seen before. The “Always Silent” category represents topics that have never received much attention. These four types of topic citation patterns represent different scenarios in the topic evolution process.

Based on the classification, different distribution patterns of citations in different semantic areas can be obtained. Corespondingly, we defined two addition attribute dimensions, “total citation counts” and “citation counts in recent n year”. These two additional attributes were calculated and added to each of the keywords. The equations for the additional attributes can be depicted as follows:

$$C_{\mathrm{T}} = \sum C_{\mathrm{total}} \tag{1}$$

$$C_{\mathrm{n}} = \sum C_{\mathrm{current}} \tag{2}$$

where the C_{total} of the keyword stands for the total citation count of all papers that contain the keyword. The C_{n} of the keyword stands for the citation counts from the papers in the current literature dataset collections that contain the keyword. The value for C_{T} is based on the citation counts from WoS database and C_{n} is based on the citation count from the current dataset, thus C_{n} is far smaller than C_{T}. Based on these characteristics, we designed an easy strategy for binary classification of C_{T} and C_{n} into high and low classes, as shown in the Eqs. (3), (4), and (5).



	High	Low
High	Always Hot areas	New Emerging Hot areas
Low	Ghost City	Always Silent areas

Fig. 2 Detecting the “Ghost City” using the binary classification

$$\text{Type} = \text{Index\_}C_{\mathrm{T}} + \text{Index\_}C_{\mathrm{n}} \tag{3}$$

$$\text{Index\_}C_{\mathrm{T}} = \begin{cases} 1, & \text{if } C_{\mathrm{T}} \geq t \\ 0, & \text{if } C_{\mathrm{T}} < t \end{cases} \tag{4}$$

$$\text{Index\_}C_{\mathrm{n}} = \begin{cases} 2, & \text{if } C_{\mathrm{n}} \geq \text{middle number of recent n year citations} \\ 0, & \text{if } C_{\mathrm{n}} < \text{middle number of recent n year citations} \end{cases} \tag{5}$$

The type of a keyword depends on the total of Index_C_{T} and Index_C_{n}. If the keyword has the total citation number ranked in the front half keywords, Index_C_{T} will be assigned the value of 2, otherwise the value of 0. If the number of recent-n-years citation of the keyword is larger than t, then the Index_C_{n} will be assigned the value of 1 otherwise it will be assigned a value of 0. Therefore, the sum of Index_C_{T} and Index_C_{n} will have four different values, corresponding to four different types: 3 for the "Always Hot" area, 2 for the "New Emerging Hot" area, 1 for the "Ghost City" area, and 0 for the "Always Silent" area.

Experiments and results

Intermediate outputs for secondary analysis

The original dataset of 1868 keywords were mapped onto the semantic space built up by the abstract corpus of more than 170 million words through word vector modeling. Because the default vectors have 100 dimensions, they cannot be intuitively understood and must be reduced to lower dimensions.

With the *t*-SNE algorithms, the geometric points generated by the word vectors can be reduced to the two-dimensional planes, and can be visualized intuitively. Some examples of the vectors after reduction to two-dimensions are shown in Table 2.

Though the unit of the dimensions of *X* and *Y* is not clear, they can help for the two-dimension visualization. As shown in Fig. 3, the geometrics points are depicted on a two-dimension canvas.

From Fig. 3, we can tell that the word vectors are distributed in an unbalanced way, some part of the points has high density, and others have low density. The space between

Table 2 Examples of reduced two-dimensional word vectors

Id	*X*	*Y*	Word
1	1.695132	− 0.72764	Gis
2	− 1.44147	− 1.17728	Landslid
3	2.104787	− 3.59437	geograph_inform_system
4	− 2.10108	5.533459	Vulner
5	0.566331	− 5.76009	natur_hazard
6	2.544471	− 0.37859	geograph_inform_system_gis
7	0.567332	− 5.7768	Hazard
8	− 5.41731	0.642229	Risk
9	1.743038	− 0.77035	remot_sens
10	− 3.56362	− 0.88111	Flood

These keywords are the processing results by the stemming algorithm, thus only basic forms are preserved

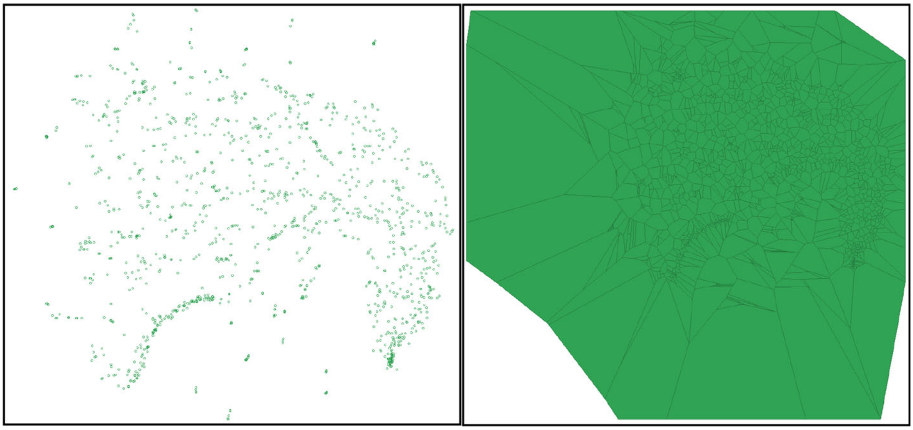


Fig. 3 The geometric points of the reduced two-dimension word vectors and generated Thiessen polygons

the points might reflect the gradually changing meaning of the semantics. The semantic meaning of a word can be vague and vary under different conditions. To help describe the vagueness of the word semantics, an additional process of Thiessen polygons is conducted.

The Thiessen polygons are originally designed by Thiessen for measuring the amount of the rainfall. To use the Thiessen polygons, the amount of rainfall in different districts can be counted statistically. The building process need to construct the Delaunay triangle by connecting the discrete points by using certain strategy that each of the triangle's circumcircle does not contain another triangle. Then based on the Delaunay triangles, the centers of the triangle's circumcircle can be connected, thus forming the Thiessen polygons. Therefore, the polygons have the property that each of the polygon contains one discrete point and points in the polygons is nearest to the discrete point that has been used to construct the Delaunay triangles. The whole map made up by the Thiessen polygons is also called as the Voronoi diagram. Similar to the original usage, points located in the Thiessen polygons stand for certain semantics that are most similar to the corresponding discrete points. Thus the vagueness of the semantics can be depicted.

From depicted Thiessen polygons, we can tell, similar to the discrete points, the polygons also have different distribution patterns. Some of the polygons have larger area and others do not. Note that the area is not corresponding to the importance of the semantic geometric points. Larger area stands for higher sparsity of the semantic meanings.

Parameter setting for n-year Ghost City

In our approach, two values are parameters that can be changed under different analysis demands, namely the n of the recent n-year citation and the threshold *t* for different types of the recent n-year citations. Different people will have different ideas about how many years without many citations can be regarded as "Ghost Cities", but different datasets will also have impacts on how we define a "Ghost City". Thus, we set the n years as a parameter in our approach. To make this observation more clear, we calculated the proportion and counts of the keywords that have no less than 2 citations for the dataset collected from the all recent 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 years; namely, the years 2015, 2014, 2013, 2012, 2011, 2010, 2009, 2008, 2007, and 2006, as shown in Fig. 4.

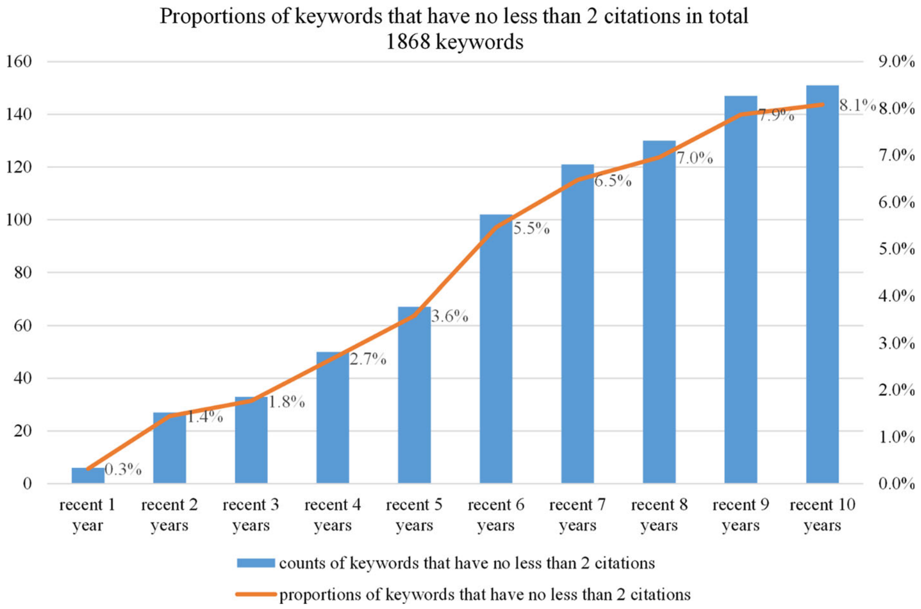


Fig. 4 Proportions of keywords that have no less than 2 citations in total 1868 keywords

From Fig. 4, we can tell keywords that have on less than two citations only comprise very small proportions in all the seven selections. Thus, we take the "recent seven years" as our selecting threshold, as it provides more related keywords (121 keywords having no less than two citations). The "recent seven years" was chosen for cartographical reasons rather than three years or less, because we needed sufficient information to make the mapping. Future studies will explore empirical criterion, for evaluating time period settings. We choose to use the time range of "recent seven years" to make a better visualization because the field of geographic natural hazards has limited influence when compared to other fields like medicine. Moreover, short time windows like three years often do not include enough citations to make a proper visualization. Thus, to highlight the "Newly Emerging Hot" areas, we set the time range longer, to seven years.

In another respect, setting the value for threshold *t* also creates problems. The threshold *t* is used to differentiate the Boolean values of high or low reflects the dataset distribution. The threshold value must be decided by the citation distribution of keyword lists. The keyword citation counts for the recent seven years are depicted Fig. 5.

By ranking the keywords by number of citations from highest to the lowest, we can tell that even when keywords with no less than two citations times were selected, the selected keywords only comprised a very small proportion, about 7% of total keywords. Thus, we use the citation count "2" as the threshold to differentiate the Boolean value for index C_7, to retain more useful information.

Binary classification results of seven-years "Ghost City"

Keywords like "Ghost City" as well as other three types of keywords can be identified using all cumulative citation counts and recent seven year citation counts. Table 3 is an example of tbinary classification results; the column labeled "type" in the table indicates

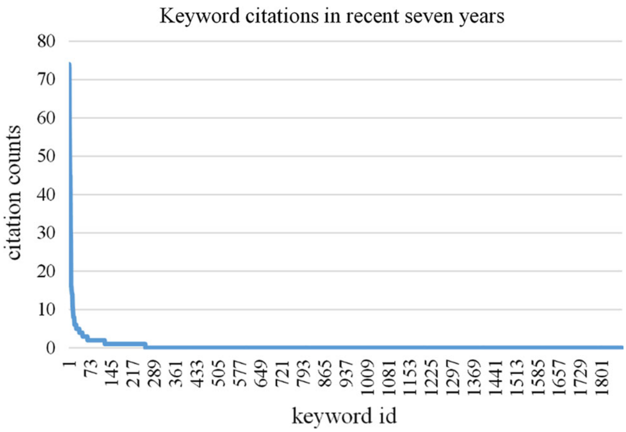


Fig. 5 Keyword citation count distribution

Table 3 The result table with addition attribute C_T and C_7, and the binary classification results

Id	*X*	*Y*	Word	C_T	C_7	Type
1	1.695131627	− 0.72764113	Gis	1446	73	3
2	− 1.441467701	− 1.177282144	Landslid	1330	58	3
3	2.104787485	− 3.594369493	geograph_inform_system	974	14	3
4	− 2.10108472	5.533458848	Vulner	881	11	3
5	0.566330854	− 5.760092308	natur_hazard	844	8	3
6	2.544470792	− 0.378593725	geograph_inform_system_gis	830	10	3
7	0.567331533	− 5.776797385	Hazard	634	1	2
8	− 5.41730994	0.642229183	Risk	535	3	3
9	1.743037938	− 0.770346116	remot_sens	493	45	3
10	− 3.563621004	− 0.881109683	Flood	477	3	3

The keywords are the processing results by the stemming algorithm, thus only basic forms are preserved

the keyword type, 3 for "Always Hot", 2 for "Ghost City", 1 for the "Newly Emerging Hot" areas, and 0 for "Always Silent" areas.

Then with the binary classification results, the "Ghost City" area with other three types of areas can be visualized cartographically, with the different colors corresponding to the four types in Fig. 2 as discussed in "Motivations: metaphors and interdisciplinary studies" section, and illustrated in Fig. 6.

From Fig. 6, we can tell the overall tendency of the distributed keywords that most keywords are belonging to the "Newly Emerging Hot" and the "Always Silent" areas. The "Always Hot" areas colored as yellow and the "Ghost City" areas colored as green are the minority. The proportions of the different types are easy to obtain. In the co-word network visualization generated by the software of VOSviewer (van Eck and Waltman 2009) or CiteSpace (Chen 2006), different proportions can also be detected with the binary classification operation. The unique aspect of our method is that the semantic similarity is counted as the distance between keyword locations on the two-dimensional plane. The

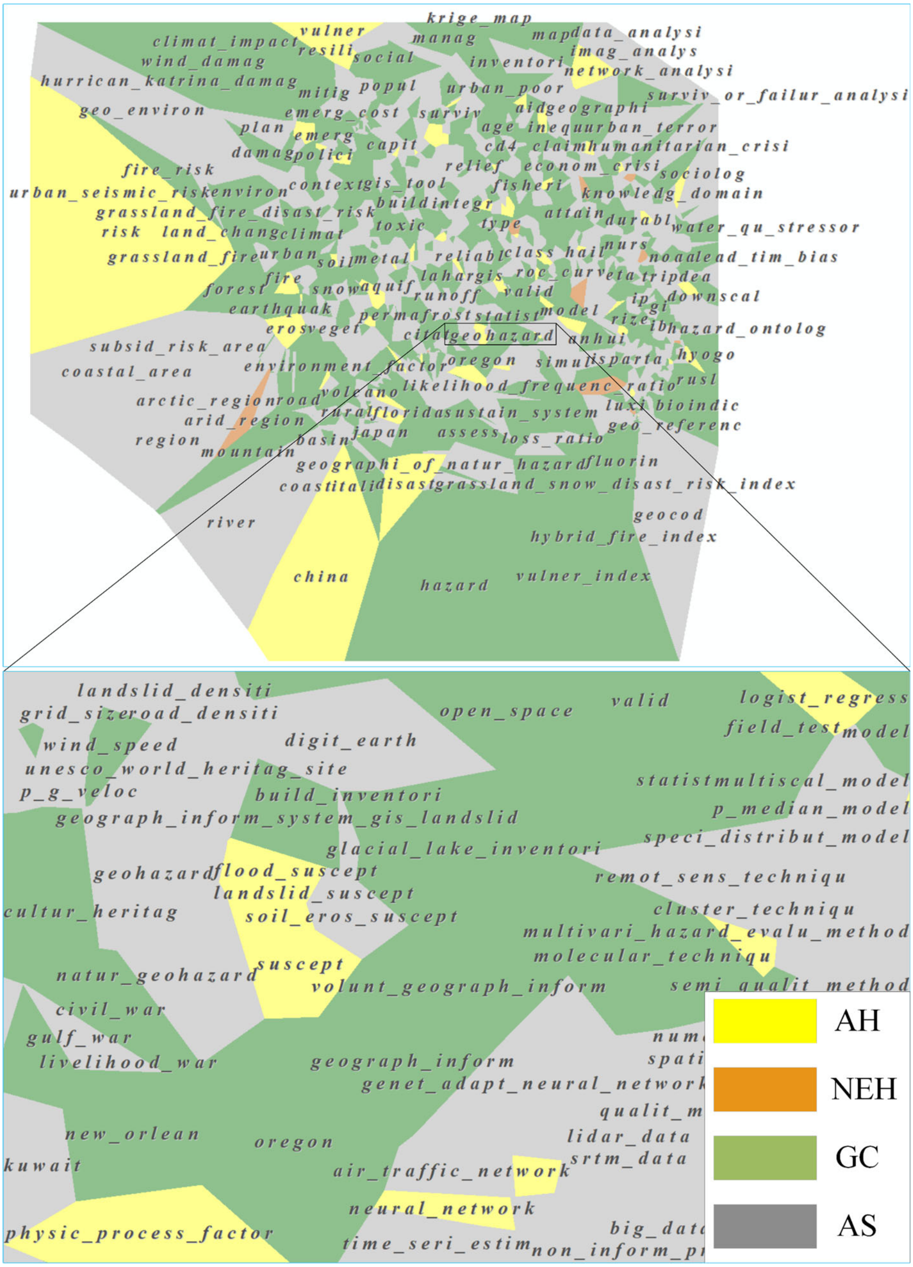


Fig. 6 The semantic maps based on the binary classification (Note AH stands for "Always Hot" area, NEH stands for "Newly Emerging Hot" area, GC stands for the "Ghost City" area, and AS stands for the "Always Silent" area)

“Ghost City” visualization in the semantic can provide a structure-aware view of the keyword semantic distribution, which is more abundant than the general mapping visualization for co-word network. Because co-word visualization only take the co-occurrence of the keywords into consideration, more details may be lost. The potentially semantic connected keywords may lost the associations because of the connectivity of the co-word network.

Discussion

The first advantage of this paper is the conceived idea based on the geospatial metaphor. We introduce the metaphor of “Ghost City” to describe the phenomena of different citation patterns in the continuous two-dimensional semantic space. Moreover, different patterns related to “Ghost City” including “Always Hot”, “Newly Emerging Hot”, and “Always Silent” areas in the semantic space are also identified. As in line with human perceptions, we believe the metaphors can help people to better understand the citation patterns over the semantic space, just like the “Sleeping Beauty” and “Prince” who are both enlightening and meaningful.

The contribution of the metaphor is not only the nominal idea of the geographical concept, but also the mapping workflow from high-dimensional semantic space to the two-dimensional geographical-likewise space. Because by analyzing with the two additional attributes of “total citation counts” and “citation counts of recent seven years”, other approaches can also identify the conceptual “Ghost City”. The uniqueness of our work is by introducing the spatial thinking to the semantic space and present an intuitive illustration by depicting the Voronoi diagrams consisting of Thiessen polygons. Comparing with the previous work of visualization of CiteSpace or VoSviewer, who focus on the visualization on the scale-free network constructed by co-word, our work transfers the scale-free and discrete network to a continuous semantic space, by the help of the Google Word2Vec model. Only based on the continuous semantic space similar to the geographic spatial space, the geographical metaphor can make sense and be close to human understanding.

Moreover, in addition to the intuitive perception of our visualization, we could also introduce quantitative measuring methods in the “Ghost City” mapping that have been originally designed to describe the spatial heterogeneity and homogeneity to describe the semantic heterogeneity and homogeneity quantitatively. The various and ever-lasting changing in the semantics in regards of the citations has been discussed in many previous work (Chen 1999). However, they have been focusing on the scale-free co-word network. Based on the “Ghost City” mapping approach, we can provide a chance to link the spatial quantitative analysis method to the semantic space to describe and evaluate the semantic distribution patterns in regards of citations in the future work.

Conclusion

In this paper, we present metaphor based analysis introducing the geographical idea of “Ghost City” to describe the spatial distribution pattern in the sematic space of the literature data in regards of citation counts. By using the Google Word2Vec model and the abstract corpus, the literature keywords are transferred to high-dimensional semantic space.

Then with the *t*-SNE algorithm, the high-dimensional semantic vectors are mapped to the two-dimensional plane. Through adding two attributes of "total citation counts" and "citation counts in recent seven years", we depicted a Voronoi diagram for the different citation patterns, including "Always Hot", "Newly Emerging Hot", "Ghost City", and "Always Silent" areas. Overall, we believe the metaphor based concept of "Ghost City" is more close to human perceptions and understanding. Comparing with traditional visualization and analyzing method only considering the scale-free co-word relations, our method presents an intuitive structure visualization for the semantic distribution. And to the best of our knowledge, the proposed approach is the first time that introduce the quantitative method to describe the clustering effect in the semantic distribution.

The proposed method is based on the dimension reduction algorithm of *t*-SNE, which stems from manifold learning. The depicted maps are similar to the work present in (Skupin 2004), which is based on the Self-Organizing Mapping (SOM) and also stems from the manifold learning. In our case, we provide more meaning for the distances between the keywords, the closer keywords are more similar with each other. Though the visualization and metaphor interpretation present intuitive semantic related structure, the loss of the semantic information is still inevitable in the dimension reduction process. Therefore, more efficiently or different biased reduction methods can also be a research interest in this metaphor based analysis. Moreover, as we have chosen a relatively narrow field of "geographic natural hazard", we chose the time window for "New Emerging Hot" in the range of "recent seven years". This definition is based on the analysis experience and can be adjusted accordingly. In addition, the additional attributes can also be extended to different dimensions like the citations and the small-world coefficient of the research community. If the citation and the small-world coefficient are both high, meaning the topic may need strong team supports that can only be undertaken by very few research teams, such as the high-energy physics studies. Others with high citation and low small-world coefficient may be the topics that can be publicly accessed and studied, like the social media based big data analysis.

In addition to quantitatively measuring the heterogeneity and homogeneity of the keywords in semantic space with regard to citations, in the future we will also introduce planning concepts borrowed from the field of geospatial analysis. Authors (Zhou et al. 2006) and citations (He et al. 2009) have been used as factors to study and simulate the topic evolution, however rarely has the landscape of keyword evolution been comprehensively mapped in a continuous semantic space. This undertaking might increase contextual understanding of the trends in topic evolution. Thus, mapping keyword evolution in the semantic space as a whole can also be a promising research direction in the later future.

Acknowledgements This research is supported by National Natural Science Foundation of China (Grant No. 41371370) and National Science Foundation for Young Scientists of China (Grant Nos. 41601298, 41501443).

References

Borgatti, S. P., & Everett, M. G. (2006). A graph-theoretic perspective on centrality. *Social Networks, 28*(4), 466–484.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Information (International Social Science Council), 22*(2), 191–235.

Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management, 35*(3), 401–420. https://doi.org/10.1016/S0306-4573(98)00068-5.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359–377.

Garfield, E. (2009). From the science of science to Scientometrics visualizing the history of science with HistCite software. *Journal of Informetrics, 3*(3), 173–179.

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, C. L. (2009). Detecting topic evolution in scientific literature: How can citations help? In *Conference on information and knowledge management, 2009* (pp. 957–966).

Hu, K., Qi, K., Guan, Q., Wu, C., Yu, J., Qing, Y., et al. (2017). A scientometric visualization analysis for night-time light remote sensing research from 1991 to 2016. *Remote Sensing, 9*(8), 802.

Kang, C., & Qin, K. (2016). Understanding operation behaviors of taxicabs in cities by matrix factorization. *Computers, Environment and Urban Systems, 60,* 79–88.

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences, 112*(24), 7426–7431.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*(1), 10–25.

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery, 7*(4), 373–397.

Mane, K. K., & Borner, K. (2004). Mapping topics and topic bursts in PNAS. *Proceedings of the National Academy of Sciences, 101,* 5287–5290.

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

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*(2), 026113.

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

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*(4), 265–269.

Song, M., Heo, G. E., & Kim, S. Y. (2014). Analyzing topic evolution in bioinformatics: Investigation of dynamics of the field with conference data in DBLP. *Scientometrics, 101*(1), 397–428. https://doi.org/10.1007/s11192-014-1246-2.

Teixeira, A. A. C., Vieira, P. C., & Abreu, A. P. (2016). Sleeping Beauties and their princes in innovation studies. *Scientometrics*. https://doi.org/10.1007/s11192-016-2186-9.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science, 342*(6157), 468–472.

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

van Eck, N., & Waltman, L. (2009). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523–538.

Wu, Q., Zhang, C., Hong, Q., & Chen, L. (2014). Topic evolution based on LDA and HMM and its application in stem cell research. *Journal of Information Science, 40*(5), 611–620.

Xie, P. (2015). Study of international anticancer research trends via co-word and document co-citation visualization analysis. *Scientometrics, 105*(1), 611–622.

Yan, E. (2014). Research dynamics: Measuring the continuity and popularity of research topics. *Journal of Informetrics, 8*(1), 98–110.

Yan, E., Ding, Y., Milojević, S., & Sugimoto, C. R. (2012). Topics in dynamic research communities: An exploratory study for the field of information retrieval. *Journal of Informetrics, 6*(1), 140–153. https://doi.org/10.1016/j.joi.2011.10.001.

Yang, S., Han, R., Wolfram, D., & Zhao, Y. (2016). Visualizing the intellectual structure of information science (2006–2015): Introducing author keyword coupling analysis. *Journal of Informetrics, 10*(1), 132–150.

Zhang, F., Zhu, X., Guo, W., Ye, X., Hu, T., & Huang, L. (2016). Analyzing urban human mobility patterns through a thematic model at a finer scale. *ISPRS International Journal of Geo-Information, 5*(6), 78.

Zheng, J., Gong, J., Li, R., Hu, K., Wu, H., & Yang, S. (2017). Community evolution analysis based on co-author network: A case study of academic communities of the journal of "Annals of the Association of American Geographers". *Scientometrics*. https://doi.org/10.1007/s11192-017-2515-7.

Zhou, D., Ji, X., Zha, H., & Giles, C. L. (2006). Topic evolution and social interactions: How authors effect research. In *Conference on information and knowledge management, 2006* (pp. 248–257).