


A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word2Vec model

Kai Hu^{1,2}  · Huayi Wu^{1,2} · Kunlun Qi³ · Jingmin Yu⁴ · Siluo Yang⁵ · Tianxing Yu^{1,2} · Jie Zheng^{1,2} · Bo Liu⁶

Received: 26 May 2017 / Published online: 18 November 2017
© Akadémiai Kiadó, Budapest, Hungary 2017

Abstract In bibliometric research, keyword analysis of publications provides an effective way not only to investigate the knowledge structure of research domains, but also to explore the developing trends within domains. To identify the most representative keywords, many approaches have been proposed. Most of them focus on using statistical regularities, syntax, grammar, or network-based characteristics to select representative keywords for the domain analysis. In this paper, we argue that the domain knowledge is reflected by the semantic meanings behind keywords rather than the keywords themselves. We apply the Google Word2Vec model, a model of a word distribution using deep learning, to represent the semantic meanings of the keywords. Based on this work, we propose a new domain knowledge approach, the Semantic Frequency-Semantic Active Index, similar to Term Frequency-Inverse Document Frequency, to link domain and

✉ Kai Hu
hukai@whu.edu.cn

Huayi Wu
wuhuayi@whu.edu.cn

Kunlun Qi
qikunlun@cug.edu.cn

Bo Liu
bliu@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 Changjiang Spatial Information Technology Engineering CO., LTD, Wuhan 430000, China

⁵ School of Information Management, Wuhan University, Wuhan 430072, China

⁶ Faculty of Geomatics, East China Institute of Technology, Nanchang 330013, China

background information and identify infrequent but important keywords. We adopt a semantic similarity measuring process before statistical computation to compute the frequencies of “semantic units” rather than keyword frequencies. Semantic units are generated by word vector clustering, while the Inverse Document Frequency is extended to include the semantic inverse document frequency; thus only words in the inverse documents with a certain similarity will be counted. Taking geographical natural hazards as the domain and natural hazards as the background discipline, we identify the domain-specific knowledge that distinguishes geographical natural hazards from other types of natural hazards. We compare and discuss the advantages and disadvantages of the proposed method in relation to existing methods, finding that by introducing the semantic meaning of the keywords, our method supports more effective domain knowledge analysis.

Keywords Keyword extraction · Word2Vec · Semantic clustering · Semantic similarity · Frequency · Domain knowledge

Introduction

Keyword extraction is an important process in a domain knowledge analysis. Keywords can be regarded as the knowledge generalization of the full text in a corresponding literature and help readers to quickly grasp the core idea, core technique, or core methodology. However, when the literature of a certain field accumulates to a large amount, selecting the most representative keywords of the domain turns out to be a hard problem (Su and Lee 2010). A number of researches have already conducted research on keyword extraction. Many methods like network-based analysis (Newman 2008) or the top-n high-frequency keywords (Zhao and Wang 2010) are proposed. These methods however, still focus on the high-frequency keywords, even methods using the network characteristics are demonstrated to be highly correlated with the frequency-based results. To analyze domain knowledge, methods using high-frequency words may not be sufficient because the high-frequency words are often general and can hardly describe the distinguishing details and boundaries of domain knowledge. In contrast, low-frequency keywords might be related to new and innovative concepts emerging in a field (Quoniam et al. 1998). Therefore, locating low-frequency keywords can identify dynamic areas especially in the sub-domains. Thus, analyzing the domain knowledge with the low-frequency keywords rather than the high-frequency words has become to a research hotspot.

With this idea that low-frequency keywords can also be important specially in domain knowledge analysis, Term Frequency-Keyword Active Index (TF-KAI) method (Chen and Xiao 2016) is proposed. This method can be regarded as extension based on the method of Term Frequency-Inversed Document Frequency (TF-IDF) (Salton and Buckley 1988), which measures the *popularity* and *discrimination* capacity of words in target documents and the background documents, where TF is used to measure the *popularity* and the IDF is used to measure the *discrimination* between the target documents and background documents. As in a keyword analysis approach, the analysis targets are the keyword lists, in which a keyword is more sparsely distributed than in traditional full-text documents, thus the possibility of a keyword emerging in the background documents is much decreased. The TF-KAI method introduced the Active Index (AI) to enlarge the *discrimination* factor, thus the TF-KAI can help identify the representative keywords more efficiently than the traditional TF-IDF methods.

The TF-KAI method uses keyword rankings as the basis for importance evaluation; thus the frequencies of a keyword found in current domain literature and background literature collections, determines how important the keyword is in the domain. Keywords of similar meaning (synonym) however, can be written in different forms. Take the term “slope collapse” from the field of geographic natural hazards as an example. In many cases, “slope collapse” can be written as “slope disintegration”, “hill collapse”, or “gully slump”. Similarly, in the remote sensing domain, night-time light has been used to study the power outages caused by rainstorms. In this area, “night-time lights” can also be written as “nightlight” or “nocturnal light”. Therefore, the semantic meanings behind keywords must be considered; otherwise, keywords with similar meanings could be counted separately (Wang et al. 2012). Very different keyword lists could be obtained, thus affecting analysis. Keywords with similar meanings (synonyms) must be merged before evaluating the importance of any specific keyword.

Synonyms or semantic similarity problems have been research foci in the Natural Language Processing (NLP) field for a long time. Many mature methodologies such as corpus-based and knowledge-based processing have been proposed to deal with this problem. However, in many instances of keyword analysis in the bibliometric field, semantic similarity problems have not been addressed with automatic methods. Nevertheless, authors have noted the disadvantages stemming from this issue (Yang et al. 2016), while synonym problems are often solved by manual methods. For example, the bibliometric analysis software CiteSpace (Chen 2006) asks users to manually select words for an alias list to merge the synonyms. Manual processes that merge synonyms often give high-quality results, but manual processing however, requires certain expert knowledge in the domain. Moreover, manual processing is often overwhelming when the amount of literature is large. Existing automatic methods for merging synonyms in keyword analysis use static knowledge databases and do not account for dynamic keyword contexts that might affect the semantic meanings of the keywords. Therefore, a method to handle and depict dynamic meaning of words in domain-specific contexts is needed.

To deal with the lack of domain-specific synonym merging and the inaccuracy problem stemming from the use of word frequencies as a measure of importance of a term, we introduce the Google Word2Vec model (Mikolov et al. 2013b), which uses word contexts to model the semantic meaning of a word when merging synonyms. Specially, we propose the use of “semantic units” to represent the results of the Word2Vec based synonyms merges. A semantic unit is a collection of keywords in which each of the keywords has a similarity value with other keywords in the collection. The similarity value of every two keyword in the collection must surpass a certain threshold to be included in the semantic unit. In addition, we also adopted the idea of tuning and applied it to *popularity* and *discrimination* measurements in TF-KAI, and extend it to the Semantic Frequency-Semantic Active Index (SF-SAI) to get the most representative keywords. To verify the effectiveness of these proposed methods, we selected the “natural hazard” topical literature as the background corpus and the “geographical natural hazard” topical literature as the target domain field. After the synonym merging process and obtaining the semantic units, we determined representative keywords using SF-SAI methods. By comparing with the original TF-KAI results qualitatively and quantitatively, we demonstrate the advantages of the proposed SF-SAI methods.

The rest of this paper is organized as follows: “[Related work](#)” section introduces related studies, “[Data and materials](#)” section describes the experimental dataset and collection dates. “[Methodology](#)” section introduces the proposed methodology. “[Results analysis](#)” section presents experimental results, discusses the domain analysis results and the

advantages and disadvantages of the proposed method. “Conclusions” section draws some conclusions.

Related work

Identifying representative keywords using frequency method

Term Frequency and Term Frequency-Inverse Document Frequency (TF-IDF)

Term frequency (TF) is the most common method for quantitative analysis of the literatures. Highly frequent keywords can be regarded as indicators of research hot spots. The frequency can indicate the *popularity* of the term or keyword, however, in the domain keyword extraction task, TF method does not provide efficient results, because it lacks the ability to discriminate when filtering out general and not specific words from the keyword list. TF-IDF extends TF to resolve this issue (Salton and Buckley 1988).

$$\text{TF-IDF} = n(i, j) \times \log \left(\frac{n(\text{all})}{n(i, \text{all})} \right) \quad (1)$$

where $n(i, j)$ stands for the TF of word i in the corpus j . The $\log(n(\text{all})/n(i, \text{all}))$ stands for the IDF, specially, $n(\text{all})$ stands for the count of all documents, $n(i, \text{all})$ stands for the counts of all the documents contain the word i .

TF-IDF combines *popularity* and *discrimination* measurements; through the IDF process most irrelative words are quickly eliminated. This function of IDF will be strengthened when the count of a keyword decreases in the original body of documents, or background corpus. In keyword extraction tasks, the results of TF and TF-IDF methods behave very similarly. The TF-IDF does not provide accurate domain-specific representative results, because the processed documents are different from traditional ones who often process the full-text documents. Therefore, same words have relatively higher possibility of emerging in both a target document and the background documents (the so called “Inverse Document” in an Inverse Document Frequency) than the document composed of keyword lists in our case. Therefore, the functionality of IDF has little effect and the discrimination ability seems absent in the process, Term Frequency-Keyword Active Index resolves these issues by introducing new discrimination factor, Keyword Active Index.

Term Frequency-Keyword Active Index (TF-KAI)

Active Index (AI) is an concept used to describe whether a country/region has comparative advantages in a particular field according to the share in total world publication (Chen et al. 2015). $\text{AI} > 1$ means that the country/region emphasizes a given domain comparing with its average research level, and $\text{AI} < 1$ means that the country/region has loose research in the field comparing to its average research level. Therefore the research interest of the country/region can be depicted. Borrowing this idea, Chen et al. used the AI to describe the research interest of certain domains and introduced the KAI instead of IDF to provide the discrimination between the domain and the background.

$$\text{TF-KAI} \sim n(i,j)^2 \times \frac{n(\text{all})}{n(i, \text{all})} \quad (2)$$

Here, the TF-KAI method eliminates IDF log computation and thus greatly enhances the discrimination factors. In reference (Chen and Xiao 2016), top 97 keywords of three methods, TF, TF-IDF, and TF-KAI, are selected as the analysis target. 58 unique keywords are identified by TF-KAI methods. These keywords do not overlap with those from TF and TF-IDF and turn out to be more representative for the domain knowledge in digital library field in contrast to the background of information science. The TF-KAI provides a solution for simple and efficient keyword extraction. However, it does not consider the relatedness among the keywords.

Identifying representative keywords using network methods

Network based methods can be regarded as the most common methods that consider the relatedness between the keywords. They are applied to identify the keywords inherit graph characteristics from the co-word occurrence networks. A co-word network takes the keywords as nodes and co-occurrences of keywords as the edges connecting the nodes. This network property can be leveraged to evaluate the keyword nodes. Many mature metrics measure keyword behaviors in complex co-word networks. For example, the node centrality metrics, betweenness centrality and eigenvector centrality, can be used to measure the importance of keywords in the network (Borgatti 2005). High node centrality indicates the keywords emerge frequently in the text; high betweenness centrality signifies that a keyword plays a connecting role between different sub-networks. Through an analysis of different patterns in a co-word network constructed from keyword lists, meaningful keywords extracted from the literature can be discriminated (Ding et al. 2001). However, most of these centrality metrics are highly correlated with the frequency-based methods. Moreover, the computation load is high; and semantic meanings behind the keywords are not considered. Therefore, the semantic based methods are needed.

Identifying synonyms using semantic similarity measurement

Semantic similarity methods have been applied in Natural Language Processing (NLP), Artificial Intelligence, Cognition Science, and Psychology. These similarity measuring methods can be categorized into two main types: knowledge-based and corpus-based methods.

Before 2013, the most common semantic similarity methods were knowledge-based, such as WordNet (Miller 1995), a well-known human-curated lexical database. The static and pre-built structures are stored in the lexical databased found in WordNet, semantic similarity can be measured through path-based, information content-based, feature-based, and hybrid measures (Meng et al. 2013). WordNet builds upon the relatedness among words, including synonyms, hyponyms, meronyms, hypernyms, and holonyms. It provides a general language ontology enabling high accuracy similarity test results. There is a limitation however, because the knowledge base is pre-built and static, thus many newly emerging words appearing in the information explosion over the internet do not included in the knowledge base.

Corpus-based methods address this situation by modeling semantic meanings from the existing corpus. For example, Latent Semantic Analysis (LSA) and Pointwise Mutual Information (PMI) are classical similarity measures derived from the existing corpora

(Mihalcea et al. 2006); however, the computation load is often high. The Word2Vec model is also based on a corpus but extends the n-gram linguistic model (Mikolov et al. 2013b), which can help decide the semantic distance between two words without supervised information. Word2Vec is also an extension of the work regarding the neural-probability language model (NLM) developed by (Huang et al. 2012), but has higher efficiency when computing the word vectors. Word2Vec models semantic meaning based on the relations between words and the surrounding context word collection. The Continuous Bag of Words (CBOW) and Skip-gram models (Mikolov et al. 2013a) are supported in the Word2Vec model. More specifically, the CBOW model predicts target words using previous and the subsequent context words. The Skip-gram model predicts the words surrounding a target word. With these model implementations, Word2Vec exhibits highly efficiency in semantic similarity testing with relatively high accuracy. Although when comparing Word2Vec with the knowledge-based similarity measurements, the accuracy of Word2Vec based results is slightly lower, nevertheless it has higher recall. Because of these characteristics, Word2Vec has been widely adopted in many similarity measure related studies (Handler 2014).

All these similarity measure methods have been studied and applied in various applications, but the synonym problem in the keyword analysis field has only a limited set of solutions. For example, Wang dealt with the synonym problems in a co-word analysis by using a thesaurus to merge keywords with similar meanings (Wang et al. 2012). Feng improved co-word analysis results by applying the ontological concept mapping (Feng et al. 2017). The methods using a thesaurus and ontological concept mapping are knowledge-based methods; the accuracy of synonym identification can be assured, but some details like newly emerging terms or low-frequency keywords, and very common in research articles, might be missed. Because knowledge based similarity computation is based on expertise knowledge and historical records, these approaches cannot deal with synonym problems among low-frequency keywords. Therefore, we choose this Word2Vec model, a corpus-based method, to dynamically model the semantic meanings behind the low-frequency keywords, lessening the impact of synonym problems.

Data and materials

We collected our experimental data from the well-known scientific database: the core collections of Web of Science (WOS). Because the background of our team in geographic information science, we choose the most familiar field of natural hazards as the background. Natural hazards can have enormous impact on the living conditions and economic development of countries and districts. Research on or related to this topic can be conducted from many different vantage points including urban planning, government policy, or the economic development planning. Therefore, many research goals are clustered in this area for solving problems and supporting decision-making. Geographic natural hazards are related to geological structures, and vegetation coverage. Many natural hazards however, are not so closely related to geography, such as climate change, greenhouse effects, and extreme weather. So it is meaningful to filter the background information from current domain corpus to identify ways in which geographic related nature-hazard research is different from general research on natural hazards.

To refine the search conditions, we set the search index to the range of Social Science Citation Index (SSCI), Science Citation Index-Expanded (SCIE) and English articles. We

also set the search time range to “1985–2016”. For the environmental natural-hazard publication corpus, we set the topic words as “natural hazard”. For the corresponding domain publication corpus, we set the topic words as “natural hazard and geography”. Finally, we get a background corpus of 10,384 records and a domain corpus of 614 records. The search date is 2017-01-01.

The basic descriptive statistics of the dataset appear in Table 1. The keyword count is the amount of all unique keywords. While the accumulated keyword count indicates that the total keyword count includes duplicate keywords. Thus, we can tell that most of the keywords have low frequencies, most appearing not more than twice. All the words in the abstract were the input to the Word2Vec model. These plain text words found in the abstract were the corpus for building up the semantic space; more details about this procedure are discussed in the “Methodology” section of this paper. The words in the keywords collection were evaluated by the original TF-KAI and our proposed SF-SAI methods.

Methodology

In this paper, we applied the ideas behind the TF-IDF method to express *popularity* and *discrimination* between a domain and background field, thus highlighting the domain-specific characteristics. The focus of this paper is to elucidate an approach that extends keyword frequency statistics and manual word disambiguation for automatic semantic unit generation, which is presented as the workflow in Fig. 1. Specially, we define “semantic unit” as a collection of keywords, every two keywords have a high similarity value exceeding a user defined similarity threshold.

In Fig. 1, the part outside the dashed rectangle has been described in previous work of Chen and Xiao 2016. The TF-KAI method can achieve better results than TF and TF-IDF when finding domain specific keywords. However, as we argued, similarity and ambiguity of words should be considered for a more accurate and sufficient analysis, we introduce the word-embedding model to express contextual semantic information. In this paper, three extraction methods, TF, TF-IDF, and TF-KAI are extended to Semantic Frequency (SF), Semantic Frequency-Semantic Inverse Document Frequency (SF-SIDF), and Semantic Frequency-Semantic Active Index (SF-SAI), respectively. The details of how to compute the values of these metrics are illustrated as follows:

Computing a Semantic Frequency (SF) value

In the TF methods, term frequency is expressed as $n(i, j)$. When extending to SF, a keyword frequency is replaced with the frequency of a semantic unit. We use the words in the domain abstract as the corpus to generate the Word2Vec models; Word2Vec provides

Table 1 Geography domain corpus and background corpus of natural hazard

Types	Documents count	Keywords count	Accumulated keyword count	Words count in abstract
Domain	614	1868	2789	121,535
Background	10,384	21,109	39,997	1,791,232

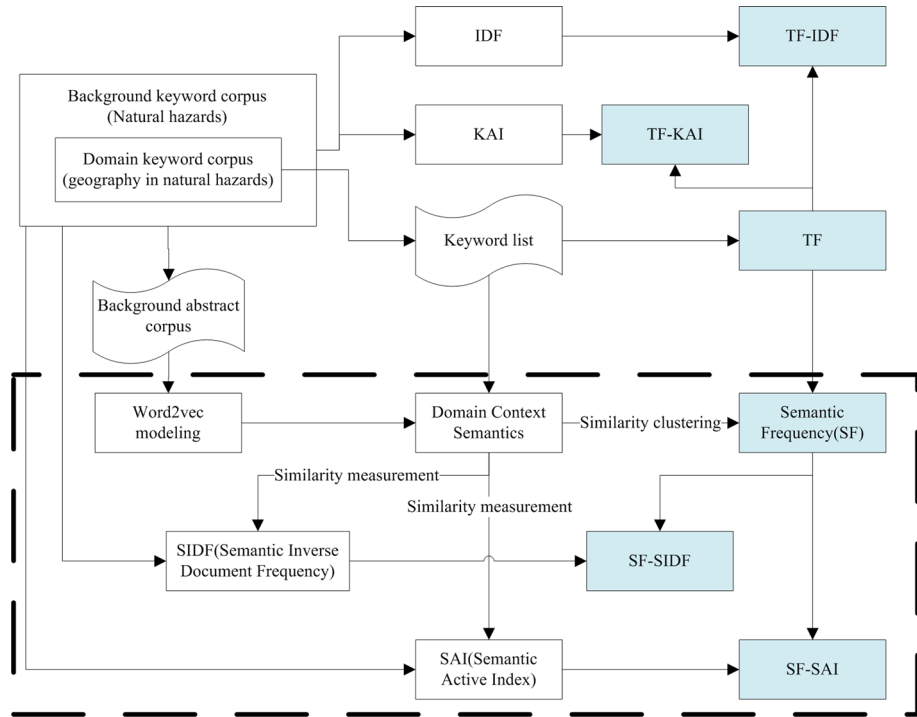


Fig. 1 Workflow of the publication keyword extractions

extraction at a more granular word level. The co-word network can be represented as a collection of 0 or 1, such as (0, 0, 0, 1, 1, 1, 0, . . . 1, 1), also called a “one-hot” representation, which can be a very large and sparse matrix. The mathematical vector generated by Word2Vec is much denser, represented as (0.54, 0.38, 0.34, 0.21, 0.02, . . . 0.34, 0.37), also called a “distributed representation”. Thus, in word vector space, the keywords can be mapped to computable mathematical vectors. To be more specific, the Word2Vec works with plain texts from abstract of literatures as illustrated in Fig. 2.

From Fig. 2, we can tell that the inputs to the Word2Vec are word sequences, generated from the plain text extracted from the literature records. The plain text in the abstracts can be obtained from files downloaded from the Web of Science (WoS) database. Considering computing efficiency and similarity modeling accuracy, we selected the Skip-gram model rather than the CBoW model for training. Skip-gram and CBow are often used interchangeably, they represent different ways of modeling but both can train a Word2Vec model. The each of the author keywords can be mapped to a geometry point in a 100-dimensional semantic space using the Word2Vec model, represented as a 100-dimensional mathematical vector. Using the cosine similarity computation method, the similarity between two words or multi-word terms can be generated. We illustrate how semantic similarity is computed by cosine similarity, in Fig. 3.

Figure 3 shows two vectors, $a(x_1, y_1)$ and $b(x_2, y_2)$ in a two-dimensional space. The similarity will be computed as illustrated in formula (3):

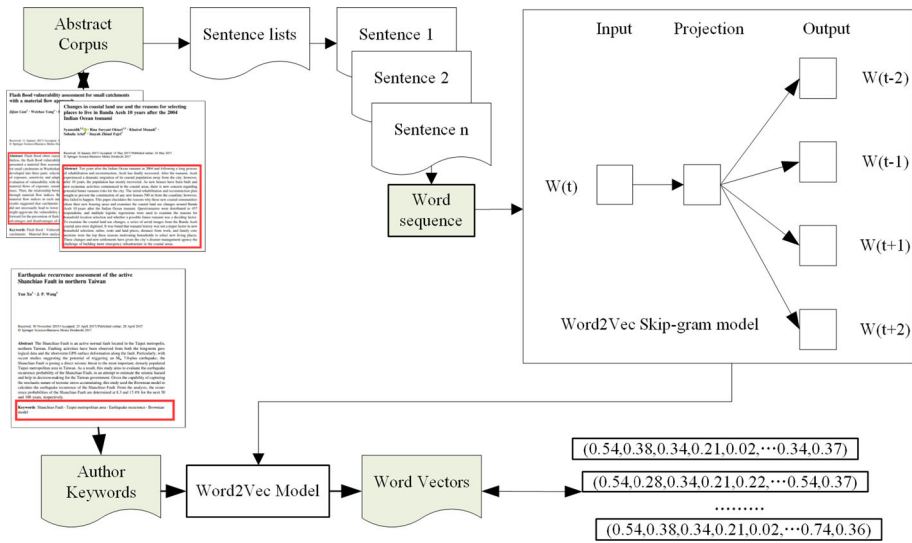
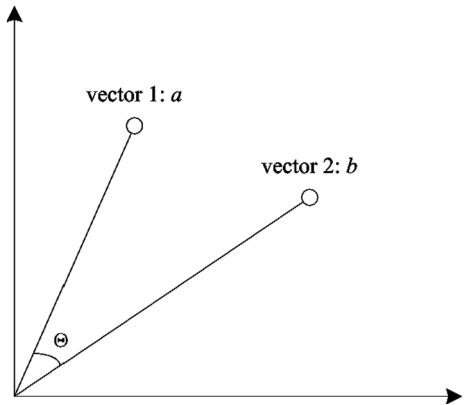


Fig. 2 The Word2Vec working process

Fig. 3 The Cosine similarity diagram for two-dimensional vectors



$$\cos \theta = \cos(a, b) = \frac{a \cdot b}{\|a\| \|b\|} = \frac{x_1x_2 + y_1y_2}{\sqrt{x_1^2 + x_2^2} \times \sqrt{y_1^2 + y_2^2}} \tag{3}$$

When the dimension is extended to higher levels, such as 100 dimensions in our case, vector a and b will be $a(a_1, a_2, a_3, \dots, a_n)$ and $b(b_1, b_2, b_3, \dots, b_n)$. The corresponding equation can be written as formula (4):

$$\cos \theta = \cos(a, b) = \frac{\sum_1^n (a_i \times b_i)}{\sqrt{\sum_1^n a_i^2} \times \sqrt{\sum_1^n b_i^2}} \tag{4}$$

Through the computation, similarity between every two word vectors can be obtained. Note that the similarity value $\cos \theta$ is belongs to $[0, 1]$, 0 means that there is no semantic overlapping between the two words. 1 means that the two words have the same semantic

meaning. In our paper, we use an experienced value of similarity threshold, thus the similarity value of every two words exceeding the threshold can be regarded as belonging to the same “semantic unit”. Therefore SF can be written as $n(i_{\text{cluster},j})$. To be more specific, $n(i_{\text{cluster},j})$ can be described by formula (5):

$$n(i_{\text{cluster},j}) = \sum \{\text{keyword_freq}(k) \mid \cos(i, k) > t, k \in j, j \in \text{t_corpus}\} \quad (5)$$

where k stands a random keyword in the j document in “t_corpus”, the target corpus; t stands for the similarity threshold set by experience. Thus, SF is the sum of all the keywords that have a higher semantic similarity with keyword i than the threshold t .

Computing SF-SIDF and SF-SAI values

In the TF-IDF methods, IDF means the frequency of a certain word that appears in the inversed documents, or background corpus in this paper, as seen in formula (1). Similarly, when extended to semantic method, the TF-IDF is extended to SF-SIDF, as seen in the formula (3). When computing the SIDF, we use the generated Word2Vec model. And when a document contains a word that has the similarity value higher than similar threshold t with the semantic unit in the SF results, the $n(i_{\text{sim}}, \text{all})$ of the SIDF counts one. To be more specific, the equation of $n(i_{\text{sim}}, \text{all})$ can be written as formula (6):

$$n(i_{\text{sim}}, \text{all}) = \sum \{\text{document_freq}(k) \mid \cos(i, k) > t, k \in \text{all}, \text{all} \in \text{b_corpus}\} \quad (6)$$

where k is also a random keyword but in the background corpus; “all” stands for documents from the background corpus; “b_corpus” stands for the background corpus; $n(i_{\text{sim}}, \text{all})$ stands for the number of the documents that contain the words or similar words. Thus the semantic inversed document frequency can be written as $\log(n(\text{all})/n(i_{\text{sim}}, \text{all}))$. The SF-SIDF can be written as formula (7):

$$\text{SF-SIDF} = n(i_{\text{cluster},j}) \times \log\left(\frac{n(\text{all})}{n(i_{\text{sim}}, \text{all})}\right) \quad (7)$$

The KAI value is used to describe the active degree of keyword to the domain and can highlight the domain preferences. When extending the TF-KAI to SF-SAI, the process use the similarity computation methods based on the Word2Vec model generated from the domain corpus. SF-SAI can be written as formula (8).

$$\text{SF-SAI} = n(i_{\text{cluster},j})^2 \times \frac{n(\text{all})}{n(i_{\text{sim}}, \text{all})} \quad (8)$$

Note that the semantic similarity threshold t is assigned as 0.97 by experience learnt from multiple times of experiments, which can obtain a relatively accurate semantic similarity. The process of the Word2Vec modeling is implemented based on python packages for NLP and machine learning, including NLTK (Bird 2006) and Gensim.

Results Analysis

Limitation in TF-KAI results

The extracted keywords from TF, TF-IDF and TF-KAI are listed in “Appendix 1”. As a routine for selecting domain keywords, most keyword analyses take less than 100 keywords for the analysis task (Chen and Xiao 2016); in the selected top 99 keywords for all three methods, 33 keywords are overlapped. The TF and TF-IDF results have 89 overlapping keywords; TF-KAI results and TF-IDF results have 40 overlapping results. TF-KAI identifies 59 keywords that are different from TF and TF-IDF methods, which is a similar result reported in Chen and Xiao (2016). It’s evident that TF-KAI method is more efficient for identifying domain-specific keywords than the TF or TF-IDF method.

The TF-KAI method has indeed provided comparatively better results in our experiment. However, the limitation of this method of neglecting the semantics behind the keywords is visible. From “Appendix 1”, we can tell that the top three keywords in the TF-KAI list are “geograph_inform_system_gis”, “geograph_inform_system”, and “gis”, which represent the same or similar meaning but written in different forms. Therefore, the semantic meanings must be considered. More synonym examples are collected in Table 2.

From Table 2, we can tell that some of the keywords have similar meaning but were considered separately rather than as semantic unit, thus resulting in very different ranking results. Some of the keywords only appeared once, but they could not be ignored as they also stand for very closely related research directions. In addition, some keywords like “geograph_inform_system” can also be written as “geograph_inform_system_gis”. Both of the two expressions have a relatively high frequency. In this case, merging these two keywords with similar meanings will make this semantic unit rank in much higher place.

To analyze the results of TF-KAI more intuitively, we adopted word embedding to generate the heat map depicted in Fig. 4. The points scattered on the map are the 59 unique keywords of TF-KAI that are different from TF and TF-IDF results, each of the points representing a keyword. The closer the data points are, the more similar the semantics of corresponding words. As the default word embedding setting is a 1 * 100 dimensioned vector, dimensional reduction applied in the t-SNE (Der Maaten and Hinton 2008) method generates two-dimensional data vectors. 1868 word vectors are used to generate the heat map using the method kernel density estimation (Rosenblatt 1956). As the keywords do not distribute evenly in the semantic space, the clustering degree of keyword is varied. Deeper

Table 2 Exemplar synonyms in TF-KAI results

Keyword by TF-KAI	TF	Rank	Synonyms of the keywords	TF	Rank
geograph_inform_system	47	3	geograph_inform_system_gis	27	6
flash_flood	28	6	flood	22	9
risk_assess	20	12	multi_risk_assess	1	1175
risk_manag	9	35	optim_risk_manag	1	1274
wenchuan_earthquak	2	291	haiti_earthquak	1	866
geomorpholog_map	2	187	geotechn_microzon_map	1	816
northern_china	1	1250	northeast_china	1	1249

All the keywords have been processed with the stemming methods to keep the basic forms

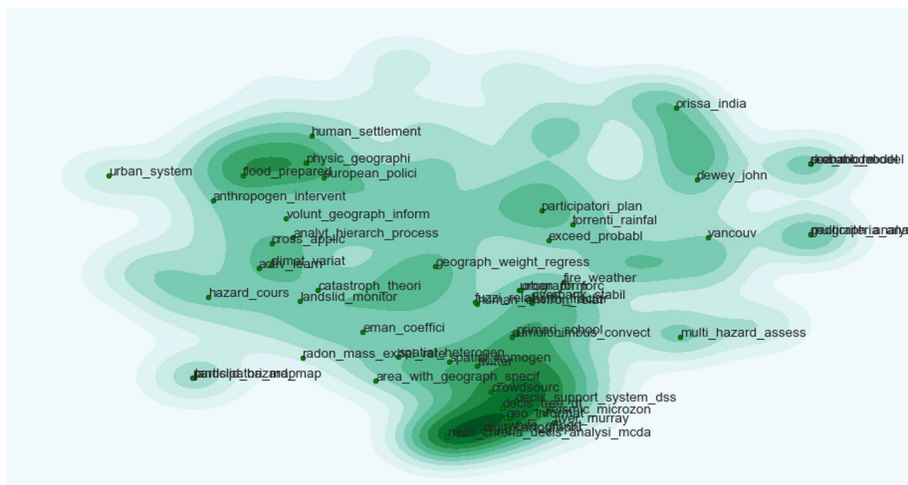


Fig. 4 Semantic density map of 59 TF-KAI unique keywords in the all domain keywords

color means that there is a higher possibility to find a similar keyword in that area. Overlapping of different keywords indicates semantic high similarity among those words, suggesting the inefficiency of the TF-KAI keyword extraction process. Semantic overlapping keywords must be merged into independent semantic units. Therefore, it is necessary to take semantic meanings of keywords into consideration.

Semantic based results (SF, SF-SIDF, and SF-SAI)

Semantics are significant and unneglectable in bibliometric and scientometric analysis. Here we facilitate the proposed SF, SF-SIDF and SF-SAI method to extract domain-specific keywords.

Generating SF results

It is important to set the granularity of the abstraction level of the semantic meanings. Too large or too small granularity creates interpretation difficulties. Many domain specific keywords are distributed in low-density areas. This is also in line with that the assumption that domain-specific keywords are often very unique. In this paper, we use the similarity threshold to control the clustering granularity of the semantic units. In Table 3, the most similar keywords for five exemplar keywords are collected.

From the table, we can tell that the similarity computation can reveal the semantic meaning to some extent. However, because of the semantic characteristics, keywords do not distribute evenly in the semantic space. Some keywords tend to have more similar keywords and others do not. Therefore, setting different thresholds will lead to different numbers of similar keywords, result in semantic units with different size. Table 4 is the table for semantic units under different similarity thresholds (ranged from 0.95 to 0.99).

We found by experimentation with different similarity thresholds that a similarity threshold of 0.97 better represents word-level semantic meanings although word-level meaning is a relatively vague concept that is hard to quantitatively measure.

Table 3 Examples of the most similar keywords for five exemplar keywords

Keyword	Top five similar keywords in the background corpus
fuzzi_logic	(fuzzi_logic, 0.999), (fuzzi_hazop, 0.989), (fuzzi, 0.987), (fuzzi_ahp, 0.985), (fuzzi_arithmet, 0.983)
electr_power_outag	(electr_power_outag, 1.0), (electr_power, 0.997), (power_outag, 0.973), (solar_power, 0.961), (power, 0.961)
standard_precipit_index_spi	(standard_precipit_index_spi, 1.0), (standard_precipit_index, 0.997), (standardis_precipit_index_spi, 0.976), (standard_runoff_index_sri, 0.948)
geograph_perspect	(geograph_perspect, 0.999), (geograph_represent, 0.927), (geograph_healthcar, 0.913), (geograph_inform_system_gis, 0.908), (geograph_technolog, 0.906)
european_starl	(european_starl, 1.00), (european_union, 0.985), (european_windstorm, 0.977), (european_alp, 0.952), (east_european_craton, 0.932)

All the keywords have been processed with the stemming methods to keep the basic linguistic forms

Table 4 The top five semantic units with different similarity thresholds by SF ranking

Rank	ST = 0.95	ST = 0.97	ST = 0.99
1	gis; participatori_gis	gis; participatori_gis	gis
2	geograph_inform_system; geograph_inform_system_gis; geograph_inform; volunt_geograph_inform	geograph_inform_system; geograph_inform_system_gis	natur_hazard
3	natur_hazard	nature_hazard	geograph_inform_system
4	landslid_inventori; landslid; landslid_inventori_map; shallow_landslid; landslid_suscept_Is; landslid_suscept; landslid_dam; landslid_map	socioeconom_vulner; socio_demograph_vulner; differenti_vulner; vulner; social_vulner; port_vulner	landslid
5	socio_demograph_vulner; socioeconom_vulner; differenti_vulner; vulner; social_vulner; port_vulner; vulner_matrix; wildfir_vulner; physic_vulner; vulner_aware	landslid; landslid_inventori	vulner

ST stands for the similarity threshold. All the keywords have been processed with the stemming methods to keep the basic forms

Generating SF-SIDF and SF-SAI results

Background information must be discriminated prior to generating SF-SIDF and SF-SAI results. The discrimination factor in TF-IDF method is the IDF value, counted by a function of $\log(n(\text{all})/n(i, \text{all}))$. In semantic based methods, the discrimination factor; SIDF values, are counted by the function, $\log(n(\text{all})/n(i_{\text{sim}}, \text{all}))$. The $n(i_{\text{sim}}, \text{all})$ value is not exactly the same semantic unit as the SF function $n(i_{\text{cluster}}, j)$, but keywords in the background corpus similar or belonging to the semantic unit. Because the background

corpus has a larger amount of keywords, and many of them are different from the keywords in the semantic units generated by domain corpus. We can tell from the Table 2 that keywords in the semantic units have various similarity values with keywords from the background corpus. Finding similar words in a background corpus relies on word-level meaning; therefore, we set the similarity threshold to a rigid range. We obtained similar keywords by setting the similarity threshold to 0.97. SF-SIDF and SF-SAI results are obtained based on the setting, as illustrated in the “Appendix 2”. Based on semantic units, we made a ranking list for total 1355 semantic units. To see functionality of the discrimination factor, we also generated the correlations of between SF and SF-SIDF, SF and SF-SAI results.

As Fig. 5 illustrates, SF-SIDF also has a very high correlation to SF results ($R^2 = 0.5057$). Thus, SIDF does not discriminate the representative keywords in the current corpus from the background corpus. SF-SAI results, on the other hand, showed a low correlation with SF ($R^2 = 0.1972$). Therefore, SF-SAI methods based on semantic meanings, produces more effective results than SF-SIDF, when calculating representative keywords.

Qualitative analysis of the TF-KAI and SF-SAI results

To examine the performance of our proposed methods, we have collected and compared the keyword results produced by TF-KAI and the proposed SF-SAI. Because, the TF-KAI method is regarded as effective in domain analysis, thus we use the TF-KAI results as the baseline. We regard the keywords belonging to one semantic units represent the same meaning. We selected a random keyword of each “semantic unit” to compare with the keywords found in the list generated by TF-KAI. Examining the results, we find that 66 keywords in the TF-KAI list are included in SF-SAI list and 61 semantic units contain TF-KAI keywords. We regard the 61 semantic units as overlapping with the TF-KAI. Because, some of the TF-KAI words are clustered into the same semantic units in the SF-SAI results, 66 keywords are regarded as overlapping with SF-SAI. With this relatively high overlapping rate, we can conclude that SF-SAI achieves relatively complete results that TF-KAI provides.

Whether SF-SAI results can better stand for the domain knowledge depends on whether the unique part of SF-SAI results are better than the unique part of TF-KAI results. We collected all 137 keywords, including 38 unique semantic units, 33 unique keywords, and 66 overlapping keywords, as shown in “Appendix 3”. In line with the tradition in keyword analysis, we explored the structure of domain knowledge using keyword clusters extracted through the hierarchical clustering method, as shown in Fig. 6.

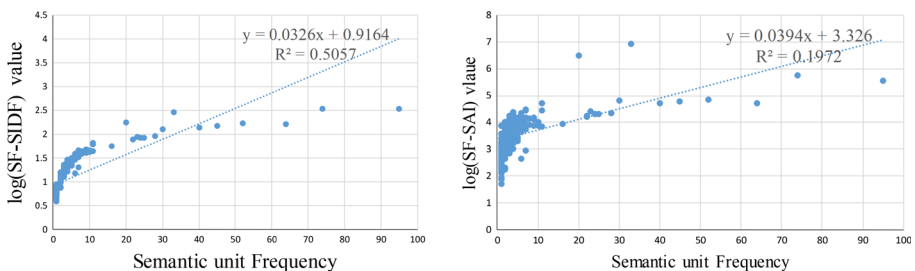


Fig. 5 Semantic unit frequency correlated to the SF-SIDF (left) and SF-SAI (right)

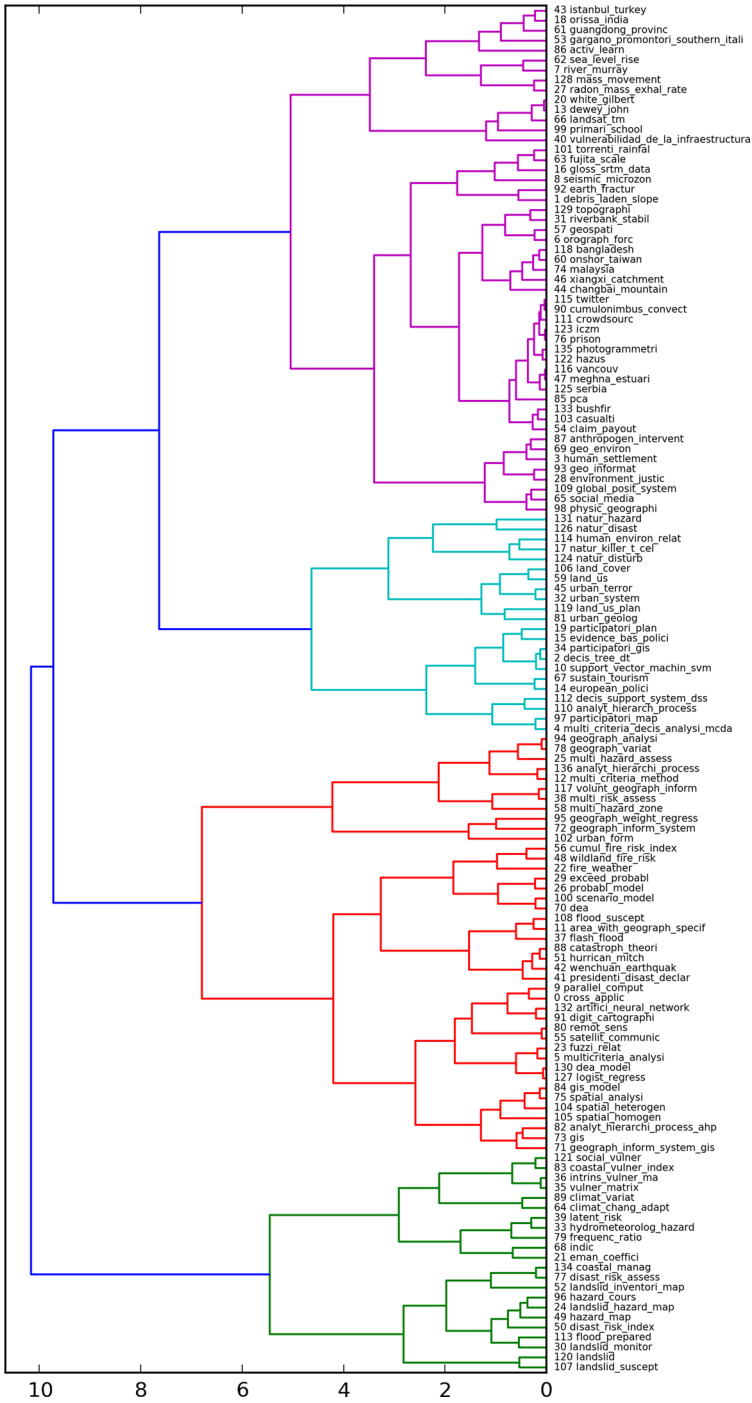


Fig. 6 the hierarchical clustering results for all 137 keywords. The 33 TF-KAI keywords, 38 SF-SAI keywords, and 66 overlapping keywords can be found in the “[Appendix 3](#)”

Table 5 lists the corresponding clustering results obtained hierarchical clustering as visualized in in Fig. 6. The keyword clusters are listed from top to down as in the denotogram, from cluster 1 to cluster 7.

We analyzed the corresponding seven clusters, demonstrating how SF-SAI more effectively illustrates the knowledge structure of a specific domain:

Cluster 1 contains the keywords related to the topic of natural hazards and primary school education. TF-KAI unique keywords includes author names like “dewey_john” and “white_gilbert”. They are indeed important scholars in the geographic research field and natural hazard research. However, TF-KAI methods only consider the string-level uniqueness. The author names are rarely used as keywords but TF-KAI selected them rather than the research topics behind the author names, given their uniqueness. In contrast, SF-SAI results yields keywords related to geolocations where geographical disasters happened, like “istanbul_turkey”, “gargano_promontori_southern_itali”, and “guangdong_province”.

Cluster 2 includes topics related to human issues caused by geographical natural hazards. The TF-KAI results contain human related keywords like “human_settlement” and “environment_justice”; but these terms are not directly specified to geographic natural hazards. The SF-SAI results however, bring in keywords that are more related to disasters. The term “social_media” as found in SF-SAI results reflects the popularity of these tools for understanding human reactions towards disasters. The term “claim_payout” seen in the SF-SAI results is a frequently mentioned keyword used in disaster management scenarios.

Cluster 3 is related to urban planning issues associated with disasters. TF-KAI results contain words such as “european_policci”, and “participatori_plan” which are related to the urban planning and management but hardly connected to natural hazard scenarios. Moreover, other TF-KAI keywords are also general terms like “decis_tree_dt”, “multi_criteria_decis_analisi_mcdm”, and “support_vector_machin_svm”. They were most likely selected because of the string-level uniqueness rather than the semantic-level uniqueness. SF-SAI produced more relevant results including topics like “urban_terror”, “land_us”, “sustain_tourism”, and “participatori_gis”. Specially, “participatori_gis” is a public participation approach in GIS that enables the planner to collect information and suggestions from the local citizens, that can help disaster management and relief.

Cluster 4 contains keywords related to risk assessment. Cluster 4 is a small cluster, each of TF-KAI and SF-SAI contains two unique keywords and these keywords are similar. But there are some differences. Expressions such as “multi_criteria_method” produced by TF-KAI are very general phrases and may not sufficiently reflect the geographic natural hazard research focus. SF-SAI keywords place more emphases on the disasters, like “multi_risk_assess” and “multi_hazard_zone”.

Cluster 5 is related to the hazard evaluation and prediction approaches. Unique TF-KAI keywords contain the words “cross_applic”, “multicriteria_analisi”, “parallel_comput”, “fuzzi_relat”, and “probabl_model”. These reflect a limited range of geographical characteristics. They are more general terms that can be universally applied in all natural hazard related research, rather than specially geographic natural hazards. The expressions “flash_flood”, “wenchuan_earthquak”, “wildland_fire_risk”, “hurrican_mitc”, and “DEA” however, generated by SF-SAI are more specific to the hazard related topics and maybe more interesting to the domain experts.

Cluster 6 reveals the keywords on the vulnerability topic in multiple disaster scenarios. “eman_coeffici” is relevant to the human health hazard exposed to the radiation, not very relevant to geographical disasters. They are selected because they have special linguistic form, but they have limited semantic relevance to the geographical natural hazard topics. In

Table 5 keyword clusters generated by TF-KAI and SF-SAI methods

Clusters	Unique TF-KAI	Unique SF-SAI	Overlapping keywords
Cluster 1	river_murray dewey_john orissa_india white_gilbert radon_mass_exhal_rate	vulnerabilidad_de_la_infraestructura istanbul_turkey sea_level_rise gargano_promontori_southern_itali guangdong_province landsat_tm	activ_learn primari_school mass_movement
Cluster 2	debris_laden_slope human_settlement orograph_forc seismic_microzon gloss_sfrm_data environment_justic riverbank_stabil	changbai_mountain xiangxi_catchment meghna_estuari claim_payout geospati onshor_taiwan fujita_scale social_media geo_environ	malaysia prison pea anthropogen_intervent cumulonimbus_convect earth_fractur_geo_informat physic_geographi torrenti_rainfal casualti_global_posit_system crowdsourc twitter_vancouv bangladesh hazus teczm serbia topographi bushfir photogrammetri urban_geolog participatori_map land_cover
Cluster 3	decis_tree_dt urban_system multi_criteria_decis_analysi_meda support_vector_machin_svm european_polici evidence_bas_polici natur_killer_lcel participatori_plan	participatori_gis urban_terror_land_us sustain_tourism	analyt_hierarch_process decis_support_system_dss human_environ_relat_land_us_plan natur_disturb_natur_disast natur_hazzard
Cluster 4	multi_criteria_method multi_hazard_assess	multi_risk_assess multi_hazard_zone	geograph_inform_system geograph_variat geograph_analysi geograph_weight_regress urban_form volumt_geograph_inform analyt_hierarchi_process
Cluster 5	cross_applic multicriteria_analysi parallel_comput area_with_geograph_specif_fire_weather_fuzzi_relat probabl_model exceed_probabl	flash_flood presidenti_disast_declar wenchuan_earthquak_wildland_fire_risk hurrican_mitch satellit_communic cumul_fire_risk_index_dea	geograph_inform_system_gis spatial_analysi remot_sens analyt_hierarchi_process_ahp_gis_model catastroph_theori digit_cartographii scenario_model spatial_heterogen spatial_homogen_flood_suscept_logist_regress_dea_model artifici_neural_network
Cluster 6	eman_coeffici	hydrometeorolog_hazard_vulner_matrix intrins_vulner_ma latent_risk climat_chang_adapt_indic	frequene_ratio_coastal_vulner_index climatat_variat_social_vulner
Cluster 7	landslid_hazard_map landslid_monitor	hazard_map_disast_risk_index landslid_inventori_map	disast_risk_assess hazard_cours landslid_suscept_flood_prepared landslid_coastal_manag

All the keywords have been processed with the stemming methods to keep the basic forms

contrast, SF-SAI keywords offer more details for the vulnerability index, which are more representative in the field of geographical natural hazard research.

Cluster 7 contains keywords related to disaster risk evaluation. In this cluster the, TF-KAI and SF-KAI methods yielded similar keywords. But, the expression “landslid_monitor” found in the TF-KAI results is a board concept, and may not be sufficient to capture the research focuses of domain experts. The term “disast_risk_index” seen in the SF-SAI results however, is a more concrete index used to quantitatively evaluate the possibility of the disaster events, and would likely be of interest to many domain experts.

To more intuitively understand the results produced by TF-KAI and SF-SAI, we also describe the results in semantic density maps, as illustrated in Figs. 7 and 8. In the “land_use” area of the SF-SAI map, TF-KAI map shows a keyword, “cross_application”, which is a general term. Where the “participate_GIS” area appears in the SF-SAI map, a corresponding area of TF-KAI map shows “parallel_computing”. In this regard, participate GIS refers to GIS applications that incorporate public involvement, a set of information collection methods that support rapid information updates in disaster areas. The term “parallel_computing” is useful in geographic natural hazard research, but is a much more general term referring to data-processing applicable to all natural hazard datasets. In the most dense semantic area in the lower part of Fig. 8, SF-SAI offers keywords “claim_payout” and “vulnerabilidad_de_la_infraestructura”. Correspondingly, TF-KAI methods produce “decis_tree_dt” and “multi_criteria_decis_analysi_mcda” in the results, terms that have limited domain-specific representative capability.

Quantitative experiments

Through the analysis above, we can only understand the advantages of SF-SAI subjectively. To make this advantages more quantitatively measured, we conducted an blind testing similar to that reported in reference (Chen and Xiao 2016). By asking experts to evaluate whether the extracted keywords are representative for the domain knowledge, the estimation results can be obtained. The possibility of a unique keyword to be identified by experts as representative keywords can be described as the percentage of the identified

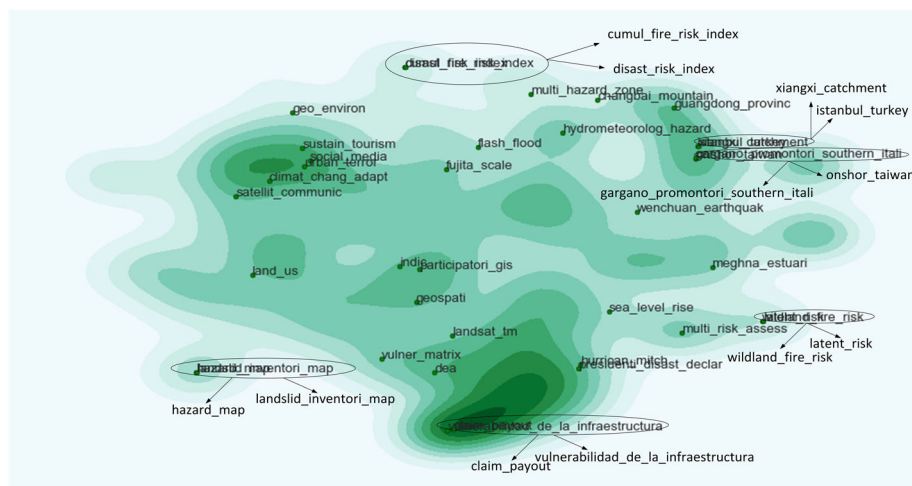


Fig. 7 The SF-SAI unique keywords on a semantic density map (comparing with TF-KAI)

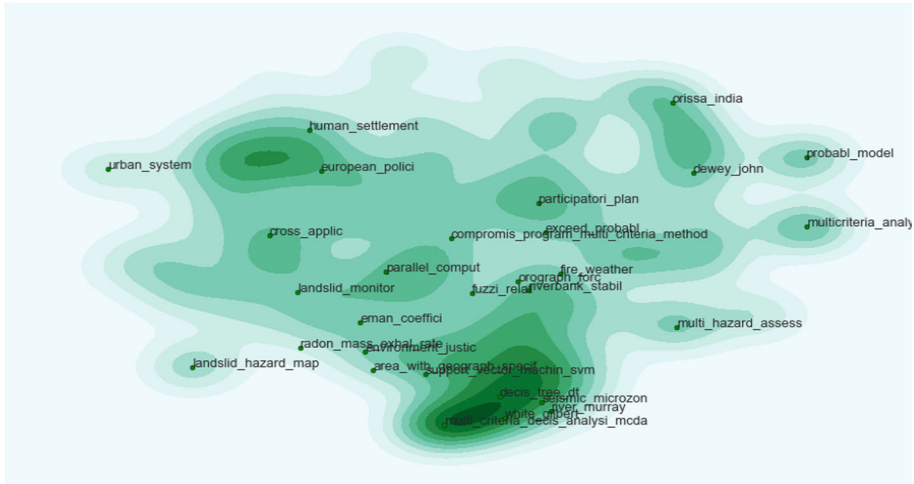


Fig. 8 The TF-KAI unique keywords on a semantic density map (comparing with SF-SAI)

keywords in the total TF-KAI or SF-SAI unique keywords. As shown in Table 6, each semantic unit of SF-SAI provides a keyword. Adding up with the 33 unique TF-KAI keywords, 61 keywords in total were selected as the testing materials, as seen in “Appendix 4”.

From the result table, we can tell that most of the unique keywords are more representatvie in SF-SAI values than the TF-KAI methods. Therefore, we can conclude that the efficiency of selecting domain-specific keywords is improved.

The number of the selected keywords also affect the estimation result of the proposed method. For example, selecting top 100 and top ten semantic units will surely results in different estimation results of SF-SAI. Therefore, we set up an experiment to see how the number of selected keywords can affect the estimation results. As in previous experiment, we use the TF-KAI results as the baseline. By undertaking ten groups comparisons, from top ten to top 99 units of TF-KAI and SF-SAI list, the overlapping counts and proportions are illustrated in Fig. 9.

Table 6 Results of the blind-testing by experts

Expert ID	Identified keyword number	TF-KAI unqie keywords (33)		SF-SAI unqie keywords (38)	
		Count	Percentage	Count	Percentage
1	29	13	0.393	16	0.421
2	17	5	0.1512	12	0.316
3	24	10	0.303	14	0.368
4	26	11	0.333	15	0.395
5	20	7	0.212	13	0.342
6	20	12	0.363	20	0.526
Average	22.666	9.666	0.292	15	0.395

As in the domain keyword analysis, no more than 100 keywords were selected. When we compare with the TF-KAI methods and SF-SAI methods, we find that when less than 50 units are selected, the overlapping percentage of TF-KAI and SF-SAI keywords are relatively low. From 50 to 99 units, the overlapping percentage increased at a steady rate. In the situation of selecting 99 units, we can tell that 61 units are overlapped and the overlapping percentage achieved a value of 0.616. SF-SAI and TF-KAI are highly correlated.

To eliminate the possible bias incurred by selecting only one sample of top 99 keywords, we conducted additional blind tests with different sample sizes, the top 20, top 40, top 60, top 80 keywords as identified by SF-SAI and TF-KAI methods, to see if SF-SAI is also more effective than TF-KAI with different sample sizes. Table 7 shows the experimental results with different sample sizes.

From Table 7, we can tell that the sampling size affects the evaluation of the TF-KAI and SF-SAI but in a slight way. When the sample size is smaller, the average ratio of SF-SAI keywords considered representative by our experts is lower. TF-KAI results however, are affected by the sampling size differently. When the sample size was changed from 40 to 60 keywords, because the overlapping keywords increased, the identification ratio decreased to 0.038. Overall however, we can still conclude that SF-SAI gives more complete results.

Discussion

The advantages of the semantic over frequency based methods

The advantage of semantic based methods can be concluded as improvement of accuracy, which means that more keywords identified by the semantic keywords are regarded as more representative for domain knowledge. The words of same or similar meaning but in different forms (synonyms) are merged into same semantic units. TF-KAI are regarded as efficient in finding domain keywords. 61 semantic units in SF-SAI results have 66 overlapping keywords out of total 99 TF-KAI keywords and the rest unique 39 semantic units

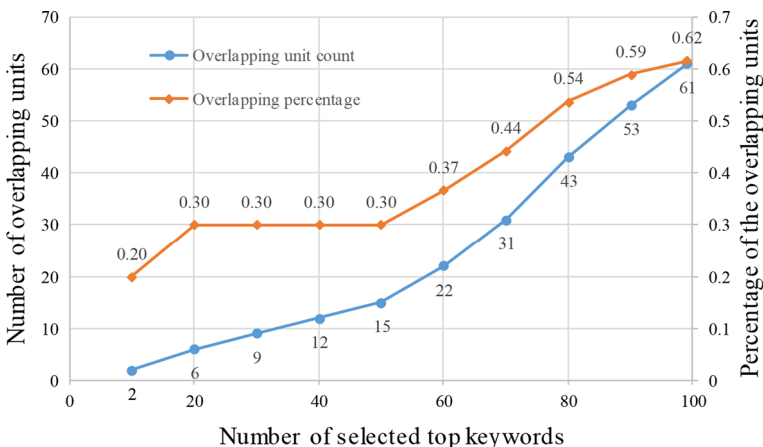


Fig. 9 The overlapping units of SF-SAI comparing with TF-SAI ranged from top ten to top 99

Table 7 Results of blind-test with different sample sizes

ID	Sample size	Count of unique keywords		Average identification percentage	
		TF-KAI	SF-SAI	TF-KAI	SF-SAI
1	20	12	14	0	0.214
2	40	26	28	0.122	0.22
3	60	35	38	0.038	0.289
4	80	33	37	0.227	0.338
5	99	33	38	0.292	0.395

of SF-SAI results are demonstrated to be more efficient to represent the domain knowledge, which indicates that SF-SAI can achieve the similar or better performance of the TF-KAI methods.

The reason of better performance in the SF-SAI is that we take the word context into consideration by introducing the Word2Vec model and the properly set the similarity threshold. Synonyms in the keywords were merged into semantic units by relatedness and similarity. Note that the results of the SF-SAI methods are sensitive to the similarity threshold. This can be decided by the experts of different analysis purpose. In our paper, we obtained the semantic units by setting the similarity threshold and merging conditions to a rigid range of “> 0.97”. In extreme case, the SF-SAI can equal to the TF-KAI methods when the similar threshold setting as “= 1”. If the experts wish to understand a more general domain concepts, they can alter the similarity threshold to generate the semantic units at different leveled details.

The effectiveness of the proposed methods is also related to the special organizing way of the data source. Traditional data source processed in the TF-IDF is the full-text dataset. In full-text datasets, different documents contain same high frequency words like “the” or “a”. Therefore, the TF-IDF method can efficiently remove the meaningless words “the” or “a”. In our paper and in reference (Chen and Xiao 2016), the corpus however, is built up by keyword collections. The keywords stand for domain knowledge that are often rarely used in other scientific domains and are often in the multi-word forms, both of which decreases the overlapping rate of the word in the domain and background corpus. Thus, it is not easy to find the same or similar words in the background documents and the discriminating factor of IDF is ineffective. Therefore, the SAI methods can improve the situation by enhancing the discriminating domain and background.

In addition to the effectiveness of keyword selection, SF-SAI also provide additional similar words in the semantic units, which can be helpful for domain analysis. Keywords are the abstraction of the corresponding knowledge. Single keywords can be hard to interpret, because the related information is not clear. The final purpose of domain keywords extraction is the domain knowledge analysis. If the relationships between the keywords are not clear, interpreting the domain knowledge can be not effective enough. For example, the keywords “spatial_heterogen” and “spatial_homogen” are not serious synonyms but are much related to each other and both are important concepts of the spatial statistical analysis. The information of relatedness cannot be provided in the frequency based methods like TF-KAI. Therefore, the semantic based methods SF-SAI can help more effective interpretation than sole frequency based method.

The limitation of the proposed semantic based methods

The semantic based methods have the limitation because of the vagueness of the semantics and corpus size. Some semantic units in our semantic based methods contain keywords that might not have exactly same meanings. As the semantic units are generated by the similarity threshold, therefore, the granularities of the semantic units are hard to be fixed. Though, many of the semantic units are not serious synonyms, they do have relatively high relatedness. As the experiment results tell, the efficiency can still be assured in the domain knowledge analysis. Therefore, one of the limitations is stemmed from the vagueness of semantic and the similarity threshold methods.

In addition, the number of selected keywords can also affect the analysis results. In our paper, we can find that different number of selected keyword will result in different overlapping rates in SF-SAI and TF-KAI results. Thus, the ranking is important factor in the domain keyword analysis, words with ranking lower than threshold (like top 100, top 40, or top 20 keywords) will be ignored in the analysis. Therefore, the future work can move to combine the ranking into evaluating the performance of the extraction methods.

Conclusions

In this paper, we propose a new method that introduces a word-embedding model for domain keyword extraction that extends existing TF, TF-IDF and TF-KAI domain-specific keyword methods by adding semantic measurements, thus becoming SF, SF-SIDF and SF-SAI. A case study using a dataset derived from geographic natural hazard literature, demonstrates that the proposed methods improves quality of the keyword extraction results. We compared the TF-KAI and SF-SAI results, finding that SF-SAI results better represent the domain knowledge and can extract domain-specific keywords more effectively.

Domain-specific knowledge is a relative concept and is present in every domain. In our experiments, the extracted domain keywords from the geographic natural hazard field are the mixture of terms from many different fields including computer science, geographical information sciences, geophysics, cartography, and so on. These keywords are also often associated with governmental tasks, including urban planning, disaster response, and other public interest issues, which require varied expert knowledge. Therefore, these keywords relate to diverse sets of background information, especially when experts from different backgrounds evaluate them. Understanding background and domain specific characteristics increases understanding of domain development and can help researchers identify potential focuses and directions in their work. Therefore, the proposed SF-SAI method can help a wide range of disciplinary or topical analyses.

The improvements we introduce to existing methods tackle the semantic vagueness in the natural language of the literature. The limitations are also resulted partially from the vagueness of semantic meanings. In this paper, a similarity threshold is set up by experience, and thus lacks quantitative criterion. Therefore, our future work will consider ways to automatically set the similarity threshold and build up an evaluation approach to determine whether the semantic unit granularity is suitable for various analytical demands.

Acknowledgements This work was supported by the National Natural Science Foundation of China (No. 41371372). Thanks Mr. Stephen C. McClure for helping us with the English revisions.

Appendix 1: Top 99 selected keywords from the TF, TF-IDF and TF-KAI methods (the unique keywords are in bold)

Rank	TF	Freq.	TF-IDF	Freq.	TF-KAI	Freq.
1	gis	94	gis	306	geograph_inform_system_gis	5,934,569
2	natur_hazard	64	geograph_inform_system	216	geograph_inform_system	4,634,698
3	geograph_inform_system	47	natur_hazard	158	gis	2,444,443
4	landslid	43	geograph_inform_system_gis	150	malaysia	1,139,113
5	vulner	37	landslid	130	spatial_analysi	999,469.3
6	geograph_inform_system_gis	27	remot_sens	109	prison	979,072
7	Hazard	26	vulner	107	disast_risk_assess	925,529
8	remot_sens	26	natur_disast	83	geograph_variat	925,529
9	flood	22	risk_assess	67	frequenc_ratio	920,429.7
10	natur_disast	22	flood	60	remot_sens	790,241.5
11	risk	22	earthquak	51	urban_geolog	697,015.1
12	risk_assess	20	landslid_suscept	51	analyt_hierarachi_process_ahp	674,641.8
13	disast	18	social_vulner	49	coastal_vulner_index	674,641.8
14	earthquak	16	disast	46	gis_model	674,641.8
15	climat_chang	13	climat_chang	45	Pca	674,641.8
16	resili	11	malaysia	44	activ_learn	573,675
17	landslid_suscept	10	resili	40	anthropogen_intervent	573,675
18	social_vulner	10	bangladesh	39	catastroph_theori	573,675
19	drought	9	risk	39	climat_variat	573,675
20	risk_manag	9	drought	38	cross_applic	573,675
21	bangladesh	7	frequenc_ratio	38	cumulonimbus_convect	573,675
22	environ	7	logist_regress	37	debris_laden_slope	573,675
23	hurrican	7	hazard	35	decis_tree_dt	573,675
24	logist_regress	7	risk_manag	35	digit_cartographi	573,675
25	malaysia	7	hazard_map	32	earth_fractur	573,675

Rank	TF	Freq.	TF-IDF	Freq.	TF-KAI	Freq.
26	adapt	6	hurricane	31	geo_informat	573,675
27	flash_flood	6	flash_flood	30	geograph_analysi	573,675
28	freque_ratio	6	land_us_plan	30	geograph_weight_regress	573,675
29	geomorpholog	6	mass_movement	29	hazard_cours	573,675
30	hazard_map	6	topographi	29	human_settlement	573,675
31	itali	6	ahp	28	multi_criteria_decis_analysi_mcda	573,675
32	model	6	itali	28	multicriteria_analysi	573,675
33	monitor	6	seismic_hazard	28	orograph_forc	573,675
34	seismic_hazard	6	spatial_analysi	28	participatori_map	573,675
35	sustain	6	sustain_develop	28	physic_geographi	573,675
36	ahp	5	sea_level_rise	27	primari_school	573,675
37	debris_flow	5	urban_geolog	27	river_murray	573,675
38	disast_manag	5	disast_manag	26	scenario_model	573,675
39	environment_hazard	5	environment_hazard	26	seismic_microzon	573,675
40	eros	5	land_cover	26	torrenti_rainfal	573,675
41	fire	5	sustain	26	urban_form	573,675
42	indic	5	geomorpholog	25	casualti	562,201.5
43	insur	5	monitor	25	parallel_comput	562,201.5
44	land_us_plan	5	analyt_hierarchy_process	24	support_vector_machin_svm	562,201.5
45	mass_movement	5	artifici_neural_network	24	area_with_geograph_specif	489,536
46	risk_percept	5	bushfir	24	multi_criteria_method	489,536
47	sea_level_rise	5	coastal_manag	24	dewey_john	489,536
48	suscept	5	flood_disast	24	european_polici	489,536
49	sustain_develop	5	land_use_plan	23	evidence_bas_polici	489,536
50	topographi	5	debris_flow	22	gloss_srtm_data	489,536
51	tsunami	5	disast_risk_assess	22	natur_killert_cel	489,536
52	urban	5	geograph_variat	22	orissa_india	489,536

Rank	TF	Freq.	TF-IDF	Freq.	TF-KAI	Freq.
53	analyt_hierarchi_process	4	insur	22	participatori_plan	489,536
54	artifici_neural_network	4	slope_stabil	22	spatial_heterogen	489,536
55	bushfir	4	adapt	21	spatial_homogen	489,536
56	china	4	analyt_hierarchi_process_ahp	21	white_gilbert	489,536
57	coastal_manag	4	casualti	21	land_cover	470,065.8
58	flood_disast	4	coastal_vulner_index	21	landslid_suscept	452,160.2
59	flood_hazard	4	environ	21	flood_suscept	437,085.7
60	groundwat	4	gis_model	21	global_posit_system	437,085.7
61	himalaya	4	himalaya	21	analyt_hierarch_process	430,256.3
62	infrastructur	4	iran	21	crowdsourc	430,256.3
63	iran	4	parallel_comput	21	decis_support_system_dss	430,256.3
64	land_cover	4	pca	21	eman_coeffici	430,256.3
65	land_use_plan	4	risk_percept	21	fire_weather	430,256.3
66	rainfal	4	support_vector_machin_svm	21	flood_prepared	430,256.3
67	rockfal	4	suscept	21	fuzzi_relat	430,256.3
68	slope_stabil	4	tsunami	21	human_environ_relat	430,256.3
69	spatial_analysi	4	environment_justic	20	landslid_hazard_map	430,256.3
70	turkey	4	flood_hazard	20	multi_hazard_assess	430,256.3
71	uncertainti	4	flood_suscept	20	probabl_model	430,256.3
72	urban_geolog	4	global_posit_system	20	radon_mass_exhal_rate	430,256.3
73	wildfir	4	hazus	20	twitter	430,256.3
74	analyt_hierarchi_process_ahp	3	iczm	20	vancouver	430,256.3
75	caribbean	3	indic	20	volunt_geograph_inform	430,256.3
76	casualti	3	natur_disturb	20	bangladesh	401,176.9
77	climat_chang_adapt	3	rockfal	20	land_us_plan	382,450
78	coastal_eros	3	serbia	20	landslid	353,191.5
79	coastal_vulner_index	3	turkey	20	social_vulner	346,514.1

Rank	TF	Freq.	TF-IDF	Freq.	TF-KAI	Freq.
80	damag	3	environment_chang	19	environment_justic	339,955.6
81	databas	3	estuari	19	hazus	339,955.6
82	dea	3	fire	19	lczm	339,955.6
83	decis_support_system	3	infrastructur	19	natur_disturb	339,955.6
84	disast_mitig	3	ndvi	19	serbia	339,955.6
85	disast_risk_assess	3	photogrammetri	19	natur_disast	311,798.6
86	ecosystem_servic	3	rainfal	19	logist_regress	307,984.7
87	emerg_manag	3	urban	19	mass_movement	306,324.2
88	emerg_respons	3	china	18	topographi	306,324.2
89	environment_chang	3	eros	18	dea_model	299,840.8
90	environment_justic	3	groundwat	18	exceed_probabl	299,840.8
91	epidemiolog	3	peru	18	landslid_monitor	299,840.8
92	estuari	3	uncertainti	18	riverbank_stabil	299,840.8
93	exposur	3	wildfir	18	urban_system	299,840.8
94	flood_suscept	3	caribbean	17	natur_hazard	297,429.3
95	gender	3	coastal_eros	17	artifici_neural_network	293,721.6
96	geograph_variat	3	decis_support_system	17	bushfir	293,721.6
97	gis_model	3	disast_mitig	17	coastal_manag	293,721.6
98	global_posit_system	3	ecosystem_servic	17	photogrammetri	276,128.9
99	hazus	3	liquefact	17	analyt_hierarchi_process	259,166.1

All the keywords have been processed with the stemming methods to keep the basic linguistic forms

Appendix 2: Top 99 semantic units in the semantic based results (SF, SF-SIDF, SF-SAI)

Rank	SF	Freq.	SF-SIDF	Freq.	SF-SAI	Freq.
1	gis; participatori_gis	95	gis; participatori_gis	350,5554	fluvial_hazard; hydrometeorolog_hazard; hazard_geographi; geoenvironment_hazard; multi_hazard; hazard_ontolog; hazard_informat; hazard;	8,329,761
2	geograph_inform_system; geograph_inform_system_gis	74	geograph_inform_system; geograph_inform_system_gis	341,2582	car_model; multiscal_model; dea_model; loglinear_model; model; ensembl_model; nois_model; traffic_model; model_chain; hidden_markov_model; mathemat_model; geochem_model; bayesian_hierarch_model; probabilist_model	3,059,600
3	natur_hazard	64	natur_hazard	163,5489	geograph_inform_system; geograph_inform_system_gis	551,130.6
4	socioeconom_vulner; socio_demograph_vulner; different_vulner; vulner; social_vulner; port_vulner	52	socioeconom_vulner; socio_demograph_vulner; different_vulner; vulner; social_vulner; port_vulner	171,622	gis; participatori_gis;	361,425.3
5	landslid; landslid_inventori	45	landslid; landslid_inventori	151,6579	socioeconom_vulner; socio_demograph_vulner; different_vulner; vulner; social_vulner; port_vulner	73,343.6
6	vulner_matrix; vulner; port_vulner; different_vulner	40	vulner_matrix; vulner; port_vulner; different_vulner	138,9708	hyperspect_remot_sens; remot_sens_rs; remot_sens; satellit_remot_sens	65,562.86
7	fluvial_hazard; hydrometeorolog_hazard; hazard_geographi; geoenvironment_hazard; multi_hazard; hazard_ontolog; hazard_informat; hazard	33	fluvial_hazard; hydrometeorolog_hazard; hazard_geographi; geoenvironment_hazard; multi_hazard; hazard_ontolog; hazard_informat; hazard	295,0969	landslid; landslid_inventori	58,894.39
8	hyperspect_remot_sens; remot_sens_rs; remot_sens; satellit_remot_sens	30	hyperspect_remot_sens; remot_sens_rs; remot_sens; satellit_remot_sens	128,6511	natur_hazard	52,744.62
9	flash_flood; flood	28	flash_flood; flood	93,62943	vulner_matrix; vulner; port_vulner; different_vulner;	51,638.82

Rank	SF	Freq.	SF-SIDF	Freq.	SF-SAI	Freq.
10	latent_risk; risk; predat_risk; risk_zonat	25	latent_risk; risk; predat_risk; risk_zonat	86.54228	analt_hierarch_process; analt_hierarch_process_ahp; analt_hierarch_process_ahp; fuzzi_analt_hierarchi_process_fahp	51,418.28
11	multi_risk_assess; risk_assess; participatori_risk_assess; ecolog_risk_assess; individu_risk_assess	24	multi_risk_assess; risk_assess; participatori_risk_assess; ecolog_risk_assess; individu_risk_assess	85.94491	frequenc_ratio; likelihood_frequenc_ratio	31,233.42
12	natur_disast; natur_disast_prepared	23	natur_disast; natur_disast_prepared	90.27616	landslid_suscept; landslid_suscept_ls	28,922.78
13	chamoli_earthquake; earthquake; haiti_earthquake; chi_chi_earthquake; wenchuan_earthquake; devast_earthquake	22	chamoli_earthquake; earthquake; haiti_earthquake; chi_chi_earthquake; wenchuan_earthquake; devast_earthquake	77.67505	malaysia	28,830.85
14	presidenti_disast_declar; manmad_disast; disast; post_disast_reconstruct; disast_prepared	22	presidenti_disast_declar; manmad_disast; disast; post_disast_reconstruct; disast_prepared	78.47514	natur_disast; natur_disast_prepared	26,796.83
15	car_model; multiscal_model; dea_model; loglinear_model; model; ensembl_model; nois_model; traffic_model; model_chain; hidden_markov_model; matemat_model; geochem_model; bayesian_hierarch_model; probabilist_model	20	car_model; multiscal_model; dea_model; loglinear_model; model; ensembl_model; nois_model; traffic_model; model_chain; hidden_markov_model; matemat_model; geochem_model; bayesian_hierarch_model; probabilist_model	178.8466	geograph_inform; volunt_geograph_inform	24,476.8
16	climat_chang; climat_chang_adapt	16	climat_chang; climat_chang_adapt	56.49095	intrins_vulner_map; map_vulner; vulner_map; specif_vulner_map; spatial_analysi; spatial_multi_criteria_analysi; tempor_analysi	24,476.8
17	analt_hierarch_process; analt_hierarchi_process; analt_hierarch_process_ahp; analt_hierarchi_process_ahp; fuzzi_analt_hierarchi_process_fahp	11	analt_hierarch_process; analt_hierarchi_process; analt_hierarch_process_ahp; analt_hierarchi_process_ahp; fuzzi_analt_hierarchi_process_fahp	66.57154		22,947
18	landslid_suscept; landslid_suscept_ls	11	landslid_suscept; landslid_suscept_ls	60.24254	flash_flood; flood;	22,210.43
19	resili	11	resili	44.24584	multi_risk_assess; risk_assess; participatori_risk_assess; ecolog_risk_assess; individu_risk_assess;	20,684.62

Rank	SF	Freq.	SF-SIDF	Freq.	SF-SAI	Freq.
20	gargano_promontori_southern_itali; southern_itali; itali; western_jilim_province; central_liaon_province	10	gargano_promontori_southern_itali; southern_itali; itali; western_jilim_province; central_liaon_province	46.51871	land_cover	20,397.33
21	hurricane; hurrican_mitch; hurrican_katrina	10	hurricane; hurrican_mitch; hurrican_katrina	46.65664	latent_risk; risk; predat_risk; risk_zonat;	19,919.27
22	optim_risk_manag; risk_manag	10	optim_risk_manag; risk_manag	42.97939	nord_pas_de_calai; vulnerabilidad_de_las_instalacion_critica; santiago_de_chile; vulnerabilidad_de_la_infraestructura;	17,483.43
23	ayvalik_turkey; turkey; findik_turkey; egirdir_turkey; yenisehir_turkey; istanbul_turkey	9	ayvalik_turkey; turkey; findik_turkey; egirdir_turkey; yenisehir_turkey; istanbul_turkey	47.05882	casualti	17,210.25
24	drought	9	drought	41.04273	disast_risk_assess	17,210.25
25	geotechn_microzon_map; neotecton_map; krige_map; participatori_map; map; geomorpholog_map	9	geotechn_microzon_map; neotecton_map; krige_map; participatori_map; map; geomorpholog_map	45.64016	geograph_variat	17,210.25
26	environ; geo_environ	8	environ; geo_environ	38.65165	presidenti_disast_declar; manmad_disast; disast; post_disast_reconstruct; disast_prepared;	17,139.43
27	urban; urban_geographi; urban_terror; urban_abandon	8	urban; urban_geographi; urban_terror; urban_abandon	43.32776	chamoli_earthquake; earthquake; haiti_earthquake; chi_chi_earthquake; wenchuan_earthquake; devast_earthquake;	16,527.3
28	alborz_mountain; fagara_mountain; apusen_mountain; changbai_mountain; mountain	7	alborz_mountain; fagara_mountain; apusen_mountain; changbai_mountain; mountain	40.06418	land_plan; land_us_plan	15,298
29	bangladesh	7	bangladesh	39.52545	hazus; hazus_mh	15,298
30	bayesian_analysis; multicriteria_analysis; spatiotemporal_analysis; discours_analysis; meta_analysis; 3d_visual_analysis	7	bayesian_analysis; multicriteria_analysis; spatiotemporal_analysis; discours_analysis; meta_analysis; 3d_visual_analysis	20.23169	urban_geolog	15,298
31	binghamton_geomorpholog_symposium; geomorpholog	7	binghamton_geomorpholog_symposium; geomorpholog	35.07353	ayvalik_turkey; turkey; findik_turkey; egirdir_turkey; yenisehir_turkey; istanbul_turkey;	15,111.44

Rank	SF	Freq.	SF-SIDF	Freq.	SF-SAI	Freq.
32	freqenc_ratio; likelihood_freqenc_ratio	7	freqenc_ratio; likelihood_freqenc_ratio	45.20196	alborz_mountain; fagara_mountain; apuseni_mountain; changbai_mountain; mountain;	14,992.04
33	logist_regress	7	logist_regress	35.4979	topographi	14,709.62
34	malaysia	7	malaysia	44.64167	urban; urban_geographi; urban_terror; urban_abandon;	14,398.12
35	northeast_china; fujian_china; china; northern_china	7	northeast_china; fujian_china; china; northern_china	33.05976	bangladesh	13,881.52
36	semi_natur_habitat; natur_disturb; natur_geohazard; natur_calam; natur_conserv	7	semi_natur_habitat; natur_disturb; natur_geohazard; natur_calam; natur_conserv	34.80427	rainfal; torrenti_rainfal	13,768.2
37	sustain; sustain_tourism	7	sustain; sustain_tourism	36.10698	gis_model	13,768.2
38	artifici_neural_network; neural_network	6	artifici_neural_network; neural_network	33.45021	pca	13,768.2
39	asset_manag; habitiat_manag; stock_manag; wildlif_manag; self_manag; manag	6	asset_manag; habitiat_manag; stock_manag; wildlif_manag; self_manag; manag	29.60998	catchment; saratel_catchment; xiangxi_catchment; loess_catchment;	13,598.22
40	big_data; data_fusion; multisensor_data_fusion; strtm_data; lidar_data; administr_data	6	big_data; data_fusion; multisensor_data_fusion; strtm_data; lidar_data; administr_data	31.5207	natur_disturb_regim; natur_disturb	13,598.22
41	cardiovascular_diseas; coronari_diseas; diarrheal_diseas; hiv_diseas_progress; diseas_progress; coronari_heart_diseas	6	cardiovascular_diseas; coronari_diseas; diarrheal_diseas; hiv_diseas_progress; diseas_progress; coronari_heart_diseas	15.01788	geotechn_microzon_map; neotecton_map; krige_map; participatori_map; map; geomorpholog_map	12,907.69
42	fire; grassland_fire	6	fire; grassland_fire	28.70068	estuari; megna_estuari;	12,238.4
43	hazard_map	6	hazard_map	34.84102	wildland_fire_risk; fire_risk; forest_fire_risk;	12,238.4
44	land_plan; land_us_plan	6	land_plan; land_us_plan	36.31175	hazard_map;	11,972.35
45	monitor	6	monitor	31.5207	coastal_vulner_index	11,473.5
46	rainfal; torrenti_rainfal	6	rainfal; torrenti_rainfal	35.67959	dea;	11,473.5
47	sea_level; sea_level_rise	6	sea_level; sea_level_rise	33.45021	disast_risk_index; ecolog_disast_risk_index; grassland_snow_disast_risk_index;	11,473.5
48	seismic_hazard	6	seismic_hazard	30.30306	coastal_manag; urban_manag	11,248.53

Rank	SF	Freq.	SF-SIDF	Freq.	SF-SAI	Freq.
49	slope; slope_movement; slope_stabil	6	slope; slope_movement; slope_stabil	30.81401	hurricane; hurricane_mitch; hurricane_katrina;	10,623.61
50	social_media; social_geographi; social_disadvantag; social; social_contact	6	social_media; social_geographi; social_disadvantag; social; social_contact	32.85957	landslid_inventori_map; landslid_map; landslid_suscept_map;	10,623.61
51	spatial_analysi; spatial_multi_criteria_analysi; tempor_analysi	6	spatial_analysi; spatial_multi_criteria_analysi; tempor_analysi	38.74454	gargano_promontori_southern_itali; southern_itali; itali; western_jilin_province; central_haon_province;	10,478.08
52	arctic_region; region; himalayan_region; arid_region; mediterranean_region	5	arctic_region; region; himalayan_region; arid_region; mediterranean_region	27.70566	spatial; spatial_smooth; spatial_heterogen; spatial_homogen	10,198.67
53	coastal_manag; urban_manag	5	coastal_manag; urban_manag	30.54558	activ_learn	10,198.67
54	debril_flow	5	debril_flow	22.99262	anthropogen_intervent	10,198.67
55	decis_support_system; decis_support_system_dss	5	decis_support_system; decis_support_system_dss	29.25644	catastroph_theori	10,198.67
56	disast_manag	5	disast_manag	27.22911	claim; claim_payout;	10,198.67
57	eros	5	eros	25.67834	climat_variat	10,198.67
58	flash_flood_hazard; flood_hazard	5	flash_flood_hazard; flood_hazard	27.70566	communic_satellit; satellit_communic;	10,198.67
59	guangdong_province; shandong_province; hebei_province; jilin_province; western_jilin_province	5	guangdong_province; shandong_province; hebei_province; jilin_province; western_jilin_province	29.73299	crowdsourc	10,198.67
60	himalaya; kashmir_himalaya	5	himalaya; kashmir_himalaya	27.70566	cumul_fire_risk_index; fire_risk_index;	10,198.67
61	indic	5	indic	29.03418	cumulonimbus_convect	10,198.67
62	insur	5	insur	25.56844	earth_fractur	10,198.67
63	landslid_inventori_map; landslid_map; landslid_suscept_map	5	landslid_inventori_map; landslid_map; landslid_suscept_map	30.25979	flood_prepared	10,198.67
64	mass_movement	5	mass_movement	29.98946	geo_informat	10,198.67
65	po_river; river_murray; river; yangtz_river_delta	5	po_river; river_murray; river; yangtz_river_delta	23.54112	geograph_analysi	10,198.67
66	risk_percept	5	risk_percept	21.88991	geograph_weight_regress	10,198.67
67	sustain_develop	5	sustain_develop	28.61727	geospati;	10,198.67
68	topographi	5	topographi	31.8869	hazard_cours	10,198.67

Rank	SF	Freq.	SF-SIDF	Freq.	SF-SAI	Freq.
69	tsunami	5	tsunami	21.94227	hazard_zone; multi_hazard_zone;	10,198.67
70	alp; apuan_alp; swiss_alp	4	alp; apuan_alp; swiss_alp	23.05711	human_environ_relat	10,198.67
71	bushfir	4	bushfir	25.21309	land_us;	10,198.67
72	catchment; saratel_catchment; xiangxi_catchment; loess_catchment	4	catchment; saratel_catchment; xiangxi_catchment; loess_catchment	26.98042	offshor_taiwan; onshor_taiwan;	10,198.67
73	central_america; central_greec; central_haon_province	4	central_america; central_greec; central_haon_province	21.54793	physic_geographi	10,198.67
74	coastal_area; coastal_urban_area; urban_area	4	coastal_area; coastal_urban_area; urban_area	24.20783	primari_school	10,198.67
75	cultur; cultur_geographi; cultur_heritag	4	cultur; cultur_geographi; cultur_heritag	23.78639	prison	10,198.67
76	eco_friend_method; interdisciplinari_method; geostatist_method; semi_qualit_method	4	eco_friend_method; interdisciplinari_method; geostatist_method; semi_qualit_method	21.0138	scenario_model	10,198.67
77	environment_anthropolog; environment_justic	4	environment_anthropolog; environment_justic	22.30014	twitter	10,198.67
78	estuari; meghna_estuari	4	estuari; meghna_estuari	26.55898	urban_form	10,198.67
79	flood_disast	4	flood_disast	24.43647	vancouv	10,198.67
80	forest; minudash_forest; oak_forest	4	forest; minudash_forest; oak_forest	22.73693	mass_movement	10,064.47
81	fujita_scale; scale	4	fujita_scale; scale	25.50952	digit; digit_cartographi	9834.429
82	geograph_inform; volunt_geograph_inform	4	geograph_inform; volunt_geograph_inform	29.33157	flood_suscept	9834.429
83	geolog_hazard; geomorpholog_hazard	4	geolog_hazard; geomorpholog_hazard	22.4405	global_posit_system	9834.429
84	groundwat	4	groundwat	20.12123	iczm	9834.429
85	hazus; hazus_mh	4	hazus; hazus_mh	27.45155	photogrammetri	9834.429
86	human_capit; human_settlement; human_biomonitor	4	human_capit; human_settlement; human_biomonitor	23.22734	guangdong_province; shandong_province; hebei_province; jilin_province; western_jilin_province;	9561.25
87	infrastructur	4	infrastructur	24.20783	artifici_neural_network; neural_network	9495.31
88	integr_vulner_assess; vulner_assess	4	integr_vulner_assess; vulner_assess	22.73693	sea_level; sea_level_rise;	9495.31
89	intrinsic_vulner_map; map_vulner; vulner_map; specif_vulner_map	4	intrinsic_vulner_map; map_vulner; vulner_map; specif_vulner_map	29.33157	fujita_scale; scale;	9414.154

Rank	SF	Freq.	SF-SIDF	Freq.	SF-SAI	Freq.
90	iran	4	iran	21.43525	climat_chang; climat_chang_adapt;	8741.714
91	land_cover	4	land_cover	28.60228	bushfir	8741.714
92	land_use_plan	4	land_use_plan	23.05711	decis_support_system; decis_support_system_dss	8692.045
93	natur_break; natur_geohazard; natur_calam; socio_natur	4	natur_break; natur_geohazard; natur_calam; socio_natur	18.44639	social_media; social_geographi; social_disadvantag; social; social_contact;	8605.125
94	natur_disturb_regim; natur_disturb	4	natur_disturb_regim; natur_disturb	26.98042	landsat; landsat_fm;	8605.125
95	nival_process; poisson_process; torrenti_process; geochem_process	4	nival_process; poisson_process; torrenti_process; geochem_process	20.81864	serbia	8605.125
96	nord_pas_de_calai; vulnerabilidad_de_las_instalacion_critica; santiago_de_chile; vulnerabilidad_de_la_infraestructura	4	nord_pas_de_calai; vulnerabilidad_de_las_instalacion_critica; santiago_de_chile; vulnerabilidad_de_la_infraestructura	27.98568	sustain; sustain_tourism;	8518.205
97	rockfal	4	rockfal	20.63256	indic;	8314.13
98	satellit_imag; satellit_imageri; satellit_imag_classif	4	satellit_imag; satellit_imageri; satellit_imag_classif	22.58597	environ; geo_environ;	8025.18
99	soil; soil_termiticid; soil_suction	4	soil; soil_termiticid; soil_suction	16.68658	logist_regress	7808.354

All the keywords have been processed with the stemming methods to keep the basic linguistic forms

Appendix 3: All 137 keywords including 33 unique TF-KAI keywords, 38 unique SF-SAI keywords, and 66 overlapping keywords

Id	Keyword	Id	Keyword	Id	Keyword	Id	Keyword	Id	Keyword		
0	cross_applic	23	fuzzi_relat	46	xiangxi_catchment	69	geo_environ	92	earth_fractur	115	twitter
1	debris_laden_slope	24	landslid_hazard_map	47	meghna_estuari	70	dea	93	geo_informat	116	vancouv
2	decis_tree_dt	25	multi_hazard_assess	48	wildland_fire_risk	71	geograph_inform_system_gis	94	geograph_analysi	117	volunt_geograph_inform
3	human_settlement	26	probabl_model	49	hazard_map	72	geograph_inform_system	95	geograph_weight_regress	118	bangladesh
4	multi_criteria_decis_analysi_mcdm	27	radon_mass_exhal_rate	50	disast_risk_index	73	gis	96	hazard_cours	119	land_us_plan
5	multicriteria_analysi	28	environment_justic	51	hurrican_mitch	74	malaysia	97	participatori_map	120	landslid
6	orograph_forc	29	exceed_probabl	52	landslid_inventori_map	75	spatial_analysi	98	physic_geographi	121	social_vulner
7	river_murray	30	landslid_monitor	53	gargano_promontori_southern_itali	76	prison	99	primari_school	122	hazus
8	seismic_microzon	31	riverbank_stabil	54	claim_payout	77	disast_risk_assess	100	scenario_model	123	iczm
9	parallel_comput	32	urban_system	55	satellit_communic	78	geograph_variat	101	torrenti_rainfal	124	natur_disturb
10	support_vector_machin_svm	33	hydrometeorolog_hazard	56	cumul_fire_risk_index	79	frequenc_ratio	102	urban_form	125	serbia
11	area_with_geograph_specif	34	participatori_gis	57	geospati	80	remot_sens	103	casualti	126	natur_disast
12	multi_criteria_method	35	vulner_matrix	58	multi_hazard_zone	81	urban_geolog	104	spatial_heterogen	127	logist_regress
13	dewey_john	36	intrinsic_vulner_ma	59	land_us	82	analyt_hierarchi_process_abp	105	spatial_homogen	128	mass_movement
14	european_polici	37	flash_flood	60	onshor_taiwan	83	coastal_vulner_index	106	land_cover	129	topographi
15	evidence_bas_polici	38	multi_risk_assess	61	guangdong_province	84	gis_model	107	landslid_suscept	130	dea_model
16	gloss_srtm_data	39	latent_risk	62	sea_level_rise	85	pca	108	flood_suscept	131	natur_hazard
17	natur_killer_t_cel	40	vulnerabilidad_de_la_infraestructura	63	fujita_scale	86	activ_learn	109	global_posit_system	132	artifici_neural_network

Id	Keyword	Id	Keyword	Id	Keyword	Id	Keyword	Id	Keyword		
18	orissa_india	41	presidenti_disast_ declar	64	climat_chang_adapt	87	anthropogen_ intervent	110	analyt_hierarch_ process	133	bushfir
19	participatori_ plan	42	wenchuan_earthquak	65	social_media	88	catastroph_theori	111	crowdsourc	134	coastal_manag
20	white_gilbert	43	istanbul_turkey	66	landsat_tm	89	climat_variat	112	decis_support_ system_dss	135	photogrammetri
21	eman_coeffici	44	changbai_mountain	67	sustain_tourism	90	cumulonimbus_ convect	113	flood_prepared	136	analyt_ hierarchi_ process
22	fire_weather	45	urban_terror	68	indic	91	digit_cartographi	114	human_environ_ relat		

Appendix 4: Experimental materials for testing the efficiency of the SF-SAI and TF-KAI methods

Rank	SF-SAI semantic units	SF-SAI keywords	TF-KAI unique keywords
1	fluvial_hazard; hydrometeorolog_hazard; hazard_geographi; geoenvironment_hazard; multi_hazard; hazard_ontolog; hazard_informat; hazard	hydrometeorolog_hazard	cross_applic
2	gis; participatori_gis	participatori_gis	debris_laden_slope
3	vulner_matrix; vulner; port_vulner; differenti_vulner	vulner_matrix	decis_tree_dt
4	intrins_vulner_map; map_vulner; vulner_map; specif_vulner_map	intrins_vulner_ma	human_settlement
5	flash_flood; flood	flash_flood	multi_criteria_decis_analisi_mcda
6	multi_risk_assess; risk_assess; participatori_risk_assess; ecolog_risk_assess; individu_risk_assess	multi_risk_assess	multicriteria_analisi
7	latent_risk; risk; predat_risk; risk_zonat	latent_risk	orograph_forc
8	nord_pas_de_calai; vulnerabilidad_de_las_instalacion_critica; santiago_de_chile; vulnerabilidad_de_la_infraestructura	vulnerabilidad_de_la_infraestructura	river_murray
9	presidenti_disast_declar; manmad_disast; disast; post_disast_reconstruct; disast_prepared	presidenti_disast_declar	seismic_microzon
10	chamoli_earthquake; earthquake; haiti_earthquake; chi_chi_earthquake; wenchuan_earthquake; devast_earthquake	wenchuan_earthquake	parallel_comput
11	ayvalik_turkey; turkey; findik_turkey; egirdir_turkey; yemisehir_turkey; istanbul_turkey	istanbul_turkey	support_vector_machin_svm
12	alborz_mountain; fagara_mountain; apuseni_mountain; changbai_mountain; mountain	changbai_mountain	area_with_geograph_specif
13	urban; urban_geographi; urban_terror; urban_abandon	urban_terror	multi_criteria_method
14	catchment; sarate_catchment; xiangxi_catchment; loess_catchment	xiangxi_catchment	dewey_john
15	estuari; meghna_estuari	meghna_estuari	european_polic
16	wildland_fire_risk; fire_risk; forest_fire_risk	wildland_fire_risk	evidence_bas_polic
17	hazard_map	hazard_map	gloss_srtm_data
18	disast_risk_index; ecolog_disast_risk_index; grassland_snow_disast_risk_index	disast_risk_index	natur_killer_t_cel
19	hurricane; hurrican_mitch; hurrican_katrina	hurrican_mitch	orissa_india
20	landslid_inventori_map; landslid_map; landslid_suscept_map	landslid_inventori_map	participatori_plan

Rank	SF-SAI semantic units	SF-SAI keywords	TF-KAI unique keywords
21	gargano_promontori_southern_itali; southern_itali; itali; western_jilin_province; central_liaon_province	gargano_promontori_southern_itali	white_gilbert
22	claim; claim_payout	claim_payout	eman_coeffici
23	communic_satellit; satellit_communic	satellit_communic	fire_weather
24	cumul_fire_risk_index; fire_risk_index	cumul_fire_risk_index	fuzzi_relat
25	geospati	geospati	landslid_hazard_map
26	hazard_zone; multi_hazard_zone	multi_hazard_zone	multi_hazard_assess
27	land_us	land_us	probabl_model
28	offshor_taiwan; onshor_taiwan	onshor_taiwan	radon_mass_exhal_rate
29	guangdong_province; shandong_province; hebei_province; jilin_province; western_jilin_province	guangdong_province	environment_justic
30	sea_level; sea_level_rise	sea_level_rise	exceed_probabl
31	fujita_scale; scale	fujita_scale	landslid_monitor
32	climat_chang; climat_chang_adapt	climat_chang_adapt	riverbank_stabil
33	social_media; social_geographi; social_disadvantag; social; social_contact	social_media	urban_system
34	landsat; landsat_tm	landsat_tm	
35	sustain; sustain_tourism	sustain_tourism	
36	indic	indic	
37	environ; geo_environ	geo_environ	
38	dea	dea	

All the keywords have been processed with the stemming methods to keep the basic linguistic forms

References

- Bird, S. (2006). NLTK: The natural language toolkit. In *Proceedings of the COLING/ACL on interactive presentation sessions, 2006* (pp. 69–72). Association for Computational Linguistics.
- Borgatti, S. P. (2005). Centrality and network flow. *Social networks*, 27(1), 55–71. <https://doi.org/10.1016/j.socnet.2004.11.008>.
- 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.
- Chen, G., & Xiao, L. (2016). Selecting publication keywords for domain analysis in bibliometrics: A comparison of three methods. *Journal of Informetrics*, 10(1), 212–223.
- Chen, G., Xiao, L., Hu, C.-P., & Zhao, X.-Q. (2015). Identifying the research focus of Library and Information Science institutions in China with institution-specific keywords. *Scientometrics*, 103(2), 707–724.
- Der Maaten, L. V., & Hinton, G. E. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.
- Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing and Management*, 37(6), 817–842.
- Feng, J., Zhang, Y. Q., & Zhang, H. (2017). Improving the co-word analysis method based on semantic distance. *Scientometrics*, 111(3), 1521–1531.
- Handler, A. (2014). *An empirical study of semantic similarity in WordNet and Word2Vec*. Citeseer.
- Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers—Volume 1, 2012* (pp. 873–882): Association for Computational Linguistics.
- Meng, L., Huang, R., & Gu, J. (2013). A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology*, 6(1), 1–12.
- Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI, 2006* (Vol. 6, pp. 775–780).
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). *Efficient estimation of word representations in vector space*. arXiv preprint [arXiv:1301.3781](https://arxiv.org/abs/1301.3781).
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Neural information processing systems* (pp. 3111–3119).
- Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11), 39–41.
- Newman, M. E. (2008). The mathematics of networks. *The New Palgrave Encyclopedia of Economics*, 2(2008), 1–12.
- Quoniam, L., Balme, F., Rostaing, H., Giraud, E., & Dou, J. M. (1998). Bibliometric law used for information retrieval. [journal article]. *Scientometrics*, 41(1), 83–91. <https://doi.org/10.1007/bf02457969>.
- Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27(3), 832–837.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5), 513–523.
- Su, H.-N., & Lee, P.-C. (2010). Mapping knowledge structure by keyword co-occurrence: A first look at journal papers in Technology Foresight. *Scientometrics*, 85(1), 65–79. <https://doi.org/10.1007/s11192-010-0259-8>.
- Wang, Z.-Y., Li, G., Li, C.-Y., & Li, A. (2012). Research on the semantic-based co-word analysis. *Scientometrics*, 90(3), 855–875.
- 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.
- Zhao, R., & Wang, J. (2010). Visualizing the research on pervasive and ubiquitous computing. *Scientometrics*, 86(3), 593–612.