International Journal of Geo-Information

MDPI

Article

Identifying the Relatedness between Tourism Attractions from Online Reviews with Heterogeneous Information Network Embedding

Peiyuan Qiu [1,2], Jialiang Gao [2,3] and Feng Lu [2,3,4,5,*]

1 School of Surveying and Geo-Informatics, Shandong Jianzhu University, Jinan 250101, China; qiupeiyuan20@sdjzu.edu.cn
2 State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; gaojl@lreis.ac.cn
3 University of Chinese Academy of Sciences, Beijing 100049, China
4 The Academy of Digital China, Fuzhou University, Fuzhou 350002, China
5 Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
* Correspondence: luf@lreis.ac.cn; Tel.: +86-10-6488-8966

Abstract: The relatedness between tourism attractions can be used in a variety of tourism applications, such as destination collaboration, commercial marketing, travel recommendations, and so on. Existing studies have identified the relatedness between attractions through measuring their co-occurrence—these attractions are mentioned in a text at the same time—extracted from online tourism reviews. However, the implicit semantic information in these reviews, which definitely contributes to modelling the relatedness from a more comprehensive perspective, is ignored due to the difficulty of quantifying the importance of different dimensions of information and fusing them. In this study, we considered both the co-occurrence and images of attractions and introduce a heterogeneous information network (HIN) to reorganize the online reviews representing this information, and then used HIN embedding to comprehensively identify the relatedness between attractions. First, an online review-oriented HIN was designed to form the different types of elements in the reviews. Second, a topic model was employed to extract the nodes of the HIN from the review texts. Third, an HIN embedding model was used to capture the semantics in the HIN, which comprehensively represents the attractions with low-dimensional vectors. Finally, the relatedness between attractions was identified by calculating the similarity of their vectors. The method was validated with mass tourism reviews from the popular online platform MaFengWo. It is argued that the proposed HIN effectively expresses the semantics of attraction co-occurrences and attraction images in reviews, and the HIN embedding captures the differences in these semantics, which facilitates the identification of the relatedness between attractions.

Keywords: relatedness between attractions; online tourism reviews; heterogeneous information network; embedding; attraction image; topic extraction

check for updates

Citation: Qiu, P.; Gao, J.; Lu, F. Identifying the Relatedness between Tourism Attractions from Online Reviews with Heterogeneous Information Network Embedding. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 797. https://doi.org/10.3390/ijgi10120797

Academic Editors: Andrea Marchetti, Angelica Lo Duca and Wolfgang Kainz

Received: 28 August 2021
Accepted: 26 November 2021
Published: 29 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The relatedness between geographic objects captures a broad relation between objects that can be close or far apart in location, can be linked by interaction, or may simply share a common property [1]. Identifying the relatedness between tourism attractions can be used in a variety of tourism applications, such as (1) destination collaboration, e.g., evaluating the connection between attractions and find the core attractions in a tourist destination [2]; (2) commercial marketing, e.g., testing how changes in links between destinations influence market equilibrium [3]; (3) travel recommendation, e.g., recognizing the popular tourist areas for tourism route recommendation based on the interactions between attractions [4].

In recent years, with the development of ICT (information and communications technology), big data, such as UGC (user-generated content) data, device data, and transaction data, has made great contributions to improving tourism research [5]. In particular, massive travel reviews of tourists are becoming easily accessible through social networks, such as Yelp, TripAdvisor, Booking, and so on. These reviews support the different types of information about visited attractions, visited times, travel notes and basic profiles of tourists, labels, ranks, review texts, and basic attributes of attractions. Intuitively, the relatedness between attractions can be identified by measuring the co-occurrence of attractions from the above information: the higher the frequency of co-occurrence of attractions (namely, the attractions are mentioned more in the information at the same time), the stronger the relatedness between them. On the one hand, the co-occurrence of attractions is reflected in the lists of tourists' visited attractions, which can be used to construct an attraction flow network. Then, the relatedness between attractions can be identified with network analytics. The results of identified relatedness are helpful to cognize the tourism movement patterns [6,7], evaluate the market position of different attractions [7,8], and reveal the factors affecting the network structure of the tourist flows [9,10]. On the other hand, the co-occurrence of attractions is expressed in review texts or travel note texts. For example, Haris et al. extracted the semantic relationships between tourist places from travel notes through the natural language processing (NLP) technique, then constructed a points of interest (POIs) graph to find the popular attractions and popular trip patterns which consist of the related attractions [11]. Yuan et al. implemented the frequent pattern mining method to identify the city's popular locations by their sequenced co-occurrences from travel blogs, then develop a max-confidence-based method to detect travel routes from the popular location network [12].

In addition to the co-occurrence of attractions, the implicit semantic information in tourism online reviews definitely contributes to modelling the relatedness from a more comprehensive perspective. The attraction image is one of these information types, which is the impression attractions on tourists, and it has different topics, such as the attractions to be seen (e.g., sand and beach), the environment to be perceived (e.g., weather, public hygiene), and experiences to remember (e.g., surfing, swimming) [13]. Thus, if two attractions have more similar images, they will have a stronger relatedness. Due to the attraction image being described in review texts and travel note texts, a topic model can be used to "understand" and extract the attraction image topics from these texts and divide the images into different semantic dimensions. The topic model is a probabilistic model for uncovering the underlying the semantic structure of a document collection based on a hierarchical Bayesian analysis of the original texts [14]. In tourism research, the topic model is used to discover the abstract "topics" in texts [15,16]. Then, the attraction images by tourists in different dimensions are obtained by fusing the topics related to this attraction, and the relatedness between attractions can be measured. The extracted attraction images facilitate the tourism destination analysis [13,17] or tourism personalized recommendation [18–21].

The key to using multi-dimensional semantic information to comprehensively identify relatedness is to quantify the importance of different dimensions of information and fuse them. That is, if two attractions have a higher frequency of co-occurrence, or more similar images, or both, they should have stronger relatedness. Determining the importance of these from massive online travel reviews manually is difficult. Thus, in this paper, we introduce a heterogeneous information network (HIN) to represent the tourism online reviews to characterise the co-occurrence and images of attractions, then comprehensively identify the relatedness between attractions through the HIN embedding technique automatically.

In the HIN, the type of nodes (or objects) or edges (or relations, links) is greater than one [22]. Therefore, the HIN can better model the real interacting system existing in multiple types of relationships. For example, a bibliographic information network can be organized as a HIN, which expresses many facts "one or more authors written a paper", "a paper has been published in a venue", and "a paper cited one or more papers" [23]. In this HIN, the types of nodes are "author", "paper", and "venue", and the types of edges are

"written" (links author and paper), "published in" (links paper and venue), and "citing" (links paper and paper). Then, the relationships between authors can be characterized with the semantics of research area topic from this HIN compared with the homogeneous information network. Moreover, the social network [24–26] and bioinformatic network [27–29] have been modelled as HINs. HINs have been applied to massive tasks as clustering, classification, link prediction, ranking, recommendation, information fusion, influence propagation, and so on [30]. The HIN embedding technique characterizes the nodes of HIN with low-dimensional vectors, i.e., embeddings [24]. Then, the semantic information is embedded in the low-dimensional vector space and the relationships between nodes can be calculated by vector operation.

Taking the HIN's advantage in expressing the different types of semantics between nodes, we utilize it to represent the tourism online reviews and use HIN embedding to comprehensively identify the relatedness between attractions. First, an online review-oriented HIN is designed to form the different types of elements in the reviews. Second, a topic model is employed to extract the nodes of the HIN from the review texts. Third, an HIN embedding model is used to capture the semantics in the HIN and comprehensively represent the attractions with low-dimensional vectors. Finally, we conduct several experiments to verify the effectiveness of the proposed method.

The remainder of this paper is structured as follows. Section 2 proposes the structure of an online review HIN, the construction method, and the embedding method of this HIN. Section 3 conducts a case study using online tourist review data. Section 4 is devoted to discussions, and Section 5 concludes this work.

2. Materials and Methods

The procedure of identifying the relatedness between tourism attractions from online reviews with HIN embedding is shown as Figure 1. Firstly, a structure of HIN is designed to represent the tourism online reviews. Next, the original online reviews are transformed into the form of the proposed HIN through direct extraction and image topic extraction. Then, the attractions in HIN are embedded into the n-dimensional vectors by HIN embedding technology. Finally, the relatedness between attractions is calculated based on the vector similarity.

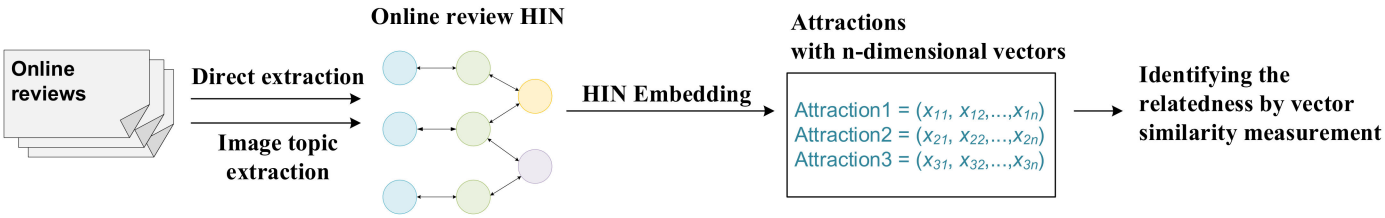


Figure 1. Flowchart for identifying the relatedness between tourism attractions with HIN embedding.

2.1. Online Review HIN Structure

In this research, we built an HIN for representing tourists' online reviews. Online reviews support which attractions are visited and the attraction images of tourists. Specifically, the attraction image in review is expressed around one or more topics, such as cost, dining, feature of attraction, traffic, and so on. So, the types of nodes in the proposed online review HIN are "attraction", "tourist", "review", and "topic". The types of edges between these nodes are "havingreview" (an attraction has a review), "reviewof" (a review of an attraction), "writing" (a tourist writes a review), "writtenby" (a review is written by a tourist), "hastopic" (a review has a topic) and "topicof" (an image topic of a review). Figure 2 illustrates an example of the online review HIN from four reviews about two tourists, three attractions and two topics.

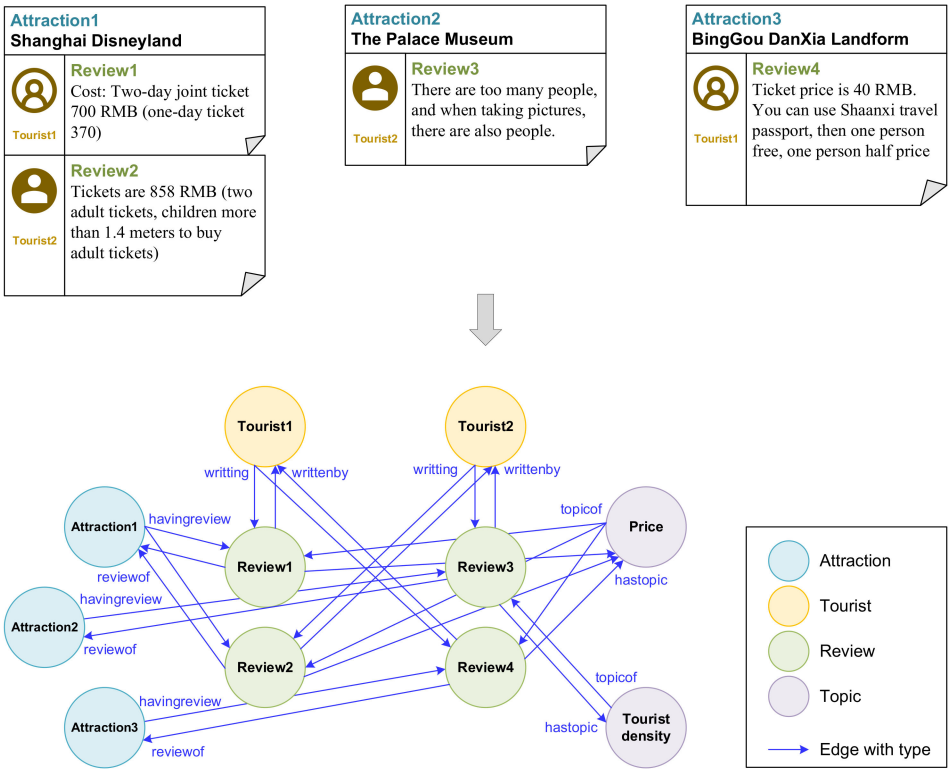


Figure 2. Example of an online review HIN.

In this online review HIN, the node path "attraction"→"review"→"tourist"→"review"→"attraction", that is, the edge path "havingreview"→"writtenby"→"writing"→"reviewof", holds the co-occurrence of attractions visited by the same tourists, and the node path "attraction"→"review"→"topic"→"review"→"attraction", that is, the edge path "havingreview"→"hastopic"→"topicof"→"reviewof", holds the relationship between attractions by the same topics of attraction images. Thus, this online review HIN expresses the co-occurrence of attractions and attraction images through the above long hop paths.

2.2. Topic Extraction and HIN Construction

The key task of constructing the presented online review HIN is extracting the nodes and edges from the reviews. The nodes "attraction" and "review" and their edge "havingreview"/"reviewof" can be directly extracted from the review list of the attraction. For the nodes "tourist" and "review", their edge "writing"/"writtenby" can be directly parsed from the basic information of the review, which contains tourist name, score given to an attraction, time of posting the review, etc. However, the image topic is not provided as basic information by the online review, so the node "topic" and the edge "hastopic"/"topicof" between "review" and "topic" are not directly extracted from the online review. Meanwhile, the image topic can be represented by certain words which make up the review text, so the image topic can be acquired from the review text through topic extraction.

Topic models are widely used for extracting abstract "topics" and hidden semantic structures from vast textual documents. Topic models as unsupervised machine learning models can automatically analyse the documents in the corpus and extract potential topics according to the co-occurrence of words in documents. For example, particular words such as "train", "subway" and "taxi" would co-occur more frequently in a document about the topic "traffic". In this study, we use the Latent Dirichlet Allocation (LDA) model [31], which is the most popular topic model, to extract the topics of review from the review text. Inputting several documents, the two main outputs of the LDA model are the probabilities that each document belongs to the different topics and the high-frequency keywords

of each topic. Then, the meaning of each topic can be summed up manually from its high-frequency keywords.

However, the original LDA model experiences large performance degradation over short texts due to the lack of word co-occurrence information in each short text [32]. Meanwhile, most of the tourism online review texts are short texts, and the word count in these texts is less than 100. Thus, we introduce the word embedding technique to extend the context of online tourism review texts to meet the word count requirement for the original LDA. For word embedding, the words in the corpus are encoded into a continuous low-dimensional semantic vector space, where each word is represented by a fixed dimensional real-valued vector [33,34]. For instance, the words "France" and "U.S.A" are represented by the 200-dimensional real-valued vectors, respectively, through word embedding; then, their distance can be calculated in the vector space. If the distance between two words is close, these words have similar semantics or related semantics [35]. For example, the distance between "France" and "U.S.A" (or "France" and "French") is less than the distance between "France" and "Mountain" in the vector space. Thus, words with similar semantics to the original words in a review text can be obtained through a similarity calculation.

The detailed procedure of acquiring the edges "hastopic"/"topicof" between "review" and "topic" from the online reviews through topic extraction is shown in Figure 3.

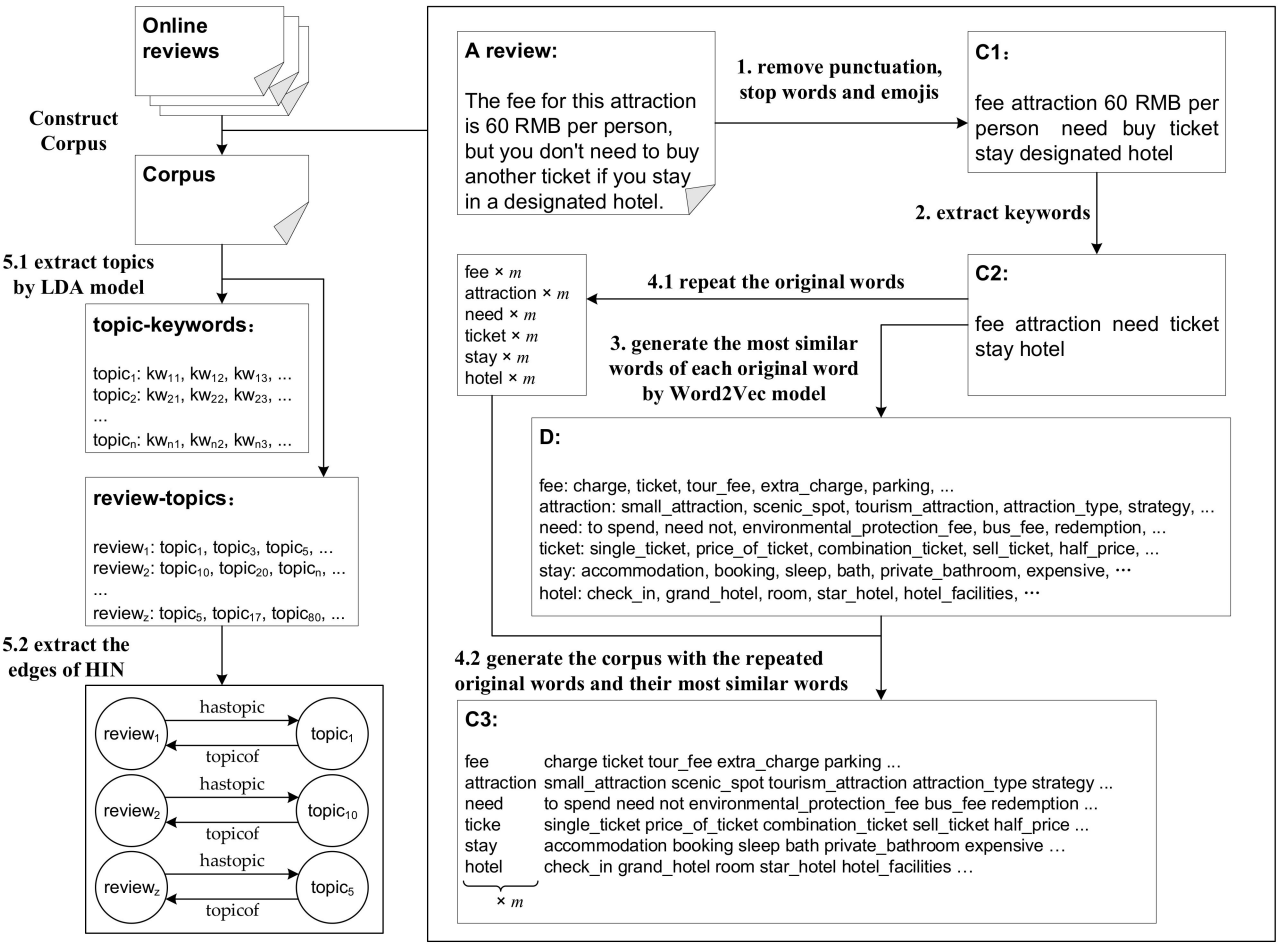


Figure 3. Flowchart for acquiring the edges "hastopic"/"topicof" between nodes "review" and "topic" through topic extraction.

Firstly, the punctuation, stop words, and emojis are removed from the original review text to reduce the interference of this meaningless information on the subsequent processing. The processed texts form a corpus "C1".

Secondly, a TextRank [36] algorithm is conducted to extract the keywords of each review in the corpus "C1" for highlighting the key information in review. The extracted k keywords of each review represent this review and form a new corpus "C2".

Thirdly, we use the word embedding model Word2Vec to obtain the low-dimensional semantic vectors of each word in the corpus "C2". Then, the semantic similarity between words can be measured by the cosine similarity as follows:

$$CosSim(x, y) = cos(\theta) = \frac{x \cdot y}{||x||||y||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \quad (1)$$

where x and y are the vectors of two words. x_i and y_i are components of vector x and y, respectively.

For one word, the semantic similarities between this word and each other word can be measured by Equation (1) and ranked in ascending order. Then, a dictionary records the top l most similar words of each word built. We can use this dictionary "D" to quickly obtain a similar word set of an input word.

Fourthly, each word of the review in the corpus "C2" has l semantic similar words as its extended words from the dictionary "D". The original words and their extended words in the corpus "C2" consist of a new corpus, e.g., extending the context of reviews. To avoid the importance of original words being diluted by their extended words, the original words can be repeated m times, respectively, in the new corpus. The final corpus is named "C3".

Fifthly, the number n topics, with their high-frequency keywords and the probabilities that each review belongs to the different topics, were obtained through using the corpus "C3" to train the LDA model. The meaning of each topic can be summed up manually from its high-frequency keywords. The topic with the highest probability is the image topic of a review. Then, the edges "hastopic"/"topicof" between reviews and topics are constructed from the reviews and their topics.

2.3. HIN Embedding and Identifying the Relatedness between Attractions

In order to achieve good performance in such tasks as clustering, classification, link prediction, recommendation, etc., the HIN embedding technique is proposed to embed the nodes of HIN into low-dimensional vectors, and then the embedded nodes can be input into the advanced machine learning models. In recent years, many HIN embedding models have been proposed, such as Metapath2Vec [37], HIN2Vec [38], HAN [39], HERec [24], and so on. While these models have been used to represent the nodes in HINs of a bibliography (e.g., from DBLP, AMiner), social media platforms (e.g., from an online blog, Flickr, Yelp, Douban), bioinformatics (e.g., from HMDD, aBiofilm), etc., they have not been applied to the HIN of tourism information before.

In this research, we select the HIN embedding model HIN2Vec to embed the online review HIN. The HIN2Vec model captures the semantic information contained in meta-paths (namely the node path or edge path mentioned in Section 2.1) and the whole network structure. Then, the relevant nodes which have semantic relationships are close to each other in the low-dimensional vector space. Compared with other HIN embedding models, the HIN2Vec model automatically constructs meta-paths with a given path length and captures the semantic information in these meta-paths instead of the limited short hop (one-hop or two-hop) meta-paths in other models. Thus, HIN2Vec can capture the semantic information in the long hop meta-paths of the online review HIN mentioned in Section 2.1: "havingreview"→"writtenby"→"writing"→"reviewof" and "havingreview"→"hastopic"→"topicof"→"reviewof".

Specially, the HIN2Vec model is a neural network model which learns the low-dimensional vectors of nodes and edges in HIN by a prediction task: input nodes x, nodes y and edges r to the model to predict whether r exists between x and y. The structure of the HIN2Vec model is shown in Figure 4. The input layer accepts the one-hot vectors $\vec{x}$, $\vec{y}$ and $\vec{r}$ of x, y and r. The latent layer transforms $\vec{x}$, $\vec{y}$ and $\vec{r}$ into latent vectors $W'_X \vec{x}$,

$W'_Y \vec{y}$ and $f_{01}\left(W'_R \vec{r}\right)$ in the d-dimensional vector space. Then, a Hadamard function is used to aggregate these latent vectors and an Identity function is applied for activation. Finally, the output layer uses the Summation as the input function and the Sigmoid function for activation to finish the prediction. The goal of the HIN2Vec model is to learn the optimal vectors $W'_X \vec{x}$, $W'_Y \vec{y}$ and $f_{01}\left(W'_R \vec{r}\right)$ of x, y and r to ensure that the predicting result is true if r exists between x and y in the real HIN, and false if r does not exist between x and y in the real HIN.

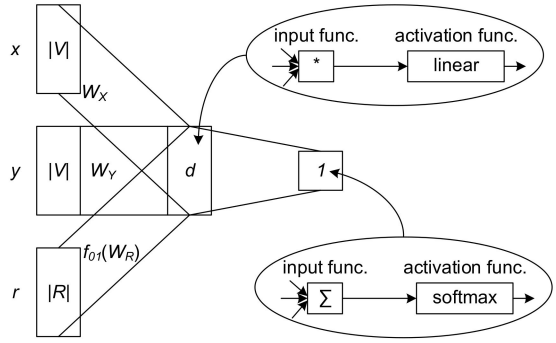


Figure 4. The structure of HIN2Vec model [38].

The process of identifying the relatedness between attractions through the HIN embedding model HIN2Vec is shown in Figure 5. Each edge in the online review HIN is re-represented by the tuple form $\langle node_i, node_j, edge_k \rangle$ for meeting the input of the model's prediction task and used to train a HIN2Vec model, where $node_i$ and $node_j$ are the head node and tail node in the edge $edge_k$. Then, the vectors of "attraction" nodes are extracted from the trained HIN2Vec model. Finally, the relatedness between two attractions can be identified by a variety of vector similarity measurements such as Euclidean distance, Manhattan distance, cosine similarity, and so on, according to applications.

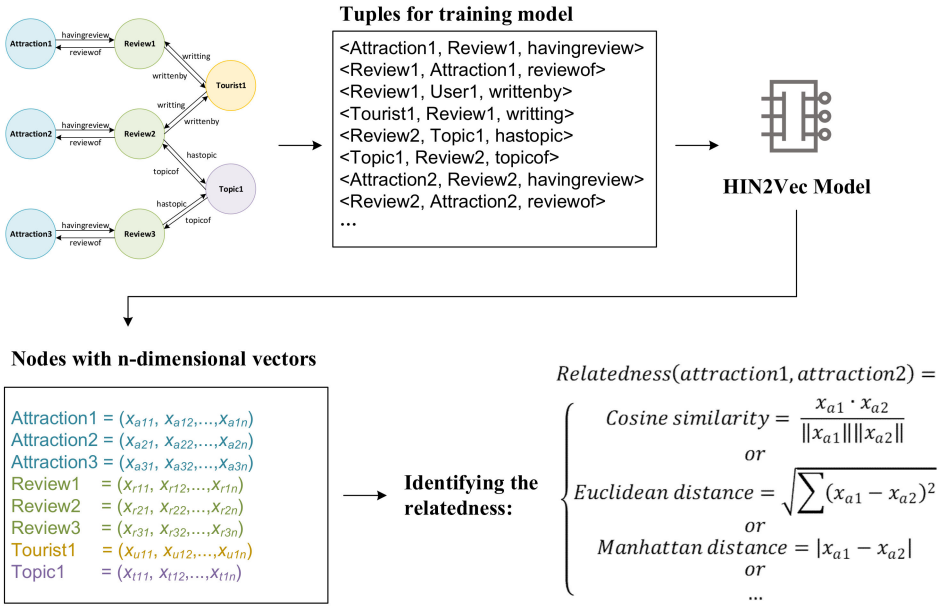


Figure 5. The process of identifying the relatedness between attractions through HIN2Vec model.

3. Case Study

In this section, we verify the performance of the proposed method with the mass tourism reviews. Firstly, the tourism review data and the constructed online review HIN are described. Then, three experiments are conducted: (1) visualization of the HIN embedding result, (2) top related attractions finding, and (3) attractions clustering.

3.1. Review Data

The tourism online review data were collected from the popular tourist-oriented information sharing platform MaFengWo (www.mafengwo.cn/). We selected attractions with reviews from within China (except Hong Kong, Macao, and Taiwan), and the period of reviews was from 2014 to 2018. Moreover, to ensure that each attraction had enough reviews for extracting tourists and topics to build paths to other attractions, attractions with fewer than 20 reviews were filtered out. The final review data to conduct the experiments contained 11,122 attractions, 202,777 tourists, and 1,087,438 reviews. The spatial distribution of attractions is shown in Figure 6.

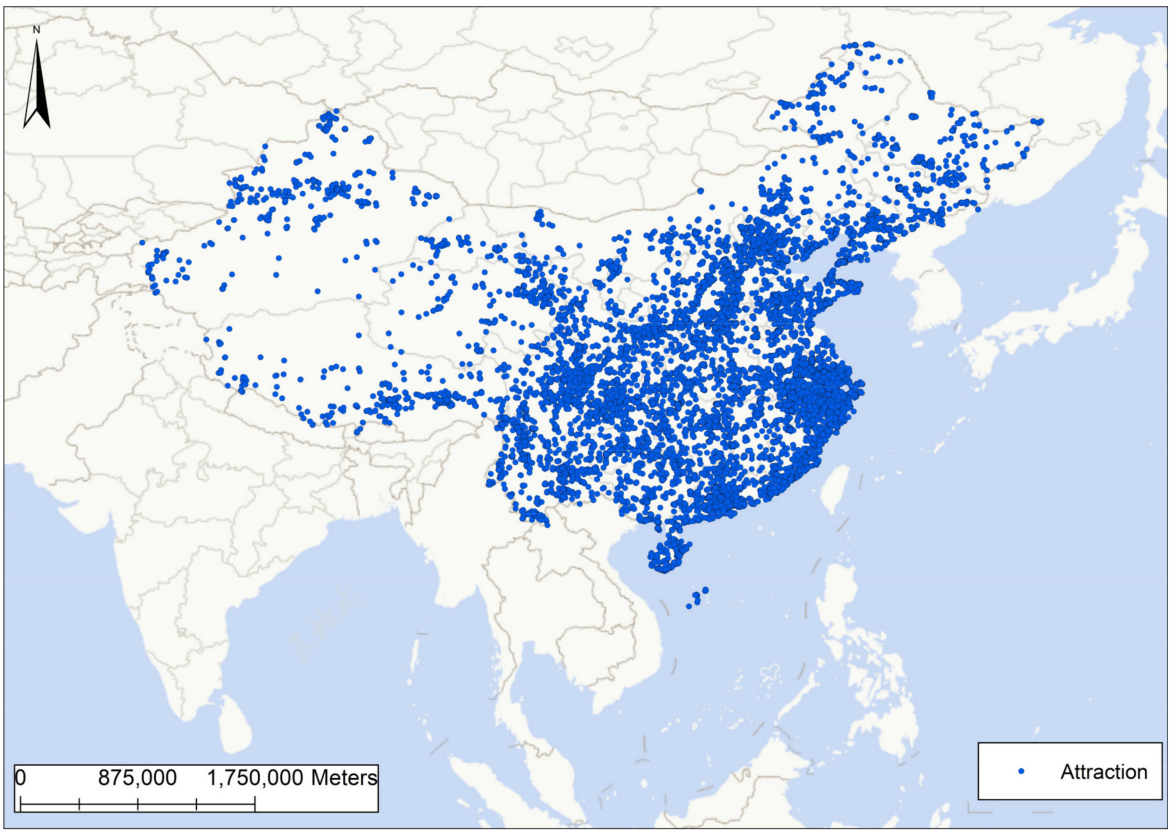


Figure 6. Spatial distribution of attractions in the review data. (Base map is obtained from Map World: http://lbs.tianditu.gov.cn/server/MapService.html).

3.2. Online Review HIN Construction

3.2.1. Image Topic Extraction

In the attraction image topic extraction, some model parameters are set considering the amount of data and efficiency. Primarily, the original review text was segmented into word sequences using the Chinese word segmentation tool because Chinese texts do not use space or another symbol to indicate different words. We used the HanLP2 tool (www.hanlp.com/) to segment the tourist reviews. Then, for TextRank which is also implemented in the HanLP2 tool, the maximum number k of keywords extracted was 50. Next, we used the gensim tool (radimrehurek.com/gensim/) to train the Word2Vec and

LDA models. For the Word2Vec model, the dimension of the vector is 250, the window is 5, the minimum word frequency is 5, and the skip-gram model [34] was selected. The number l of words used to extend the context of reviews is 25, and the repeat times m of original words to avoid the importance of original words being diluted by their extended words is 12. Then, the average word number of reviews in the corpuses "C1", "C2", and "C3" are 21.98, 12.21, and 440.78, respectively. So, the length of the reviews in corpus "C3" with extended context is suitable for the LDA model to extract image topics. The topic number n of the LDA model was set as 200 to fully distinguish the semantic differences between potential image topics. All training was conducted on a computer equipped with two 2.20 GHz Intel Xeon CPUs and 128 GB RAM.

The probabilities that each document belongs to the different topics and the high-frequency keywords of each topic are the two main outputs of LDA. The extracted 200 image topics can be further divided into 13 categories and 155 sub-categories through manually interpreting the high-frequency keywords of each image topic. The categories and sub-categories are shown in Table 1. Table 2 shows 20 image topics of the above 155 sub-categories and their top 10 high-frequency keywords. These results indicate that the proposed image topic extraction method can capture the semantic difference in the reviews, which will facilitate the relatedness identification between attractions. Finally, the topic with the highest probability is the image topic of a review. It needs to be emphasized that the interpreted categories were used for understanding the meanings of images and to verify the validity of the topic extraction results. So, we kept all the extracted 200 image topics in the next HIN construction instead of merging the topics that belonged to the same category for the HIN2Vec model, which captures the slight semantic differences between the topics.

Table 1. Image topic category and sub-categories (the number of extracted topics belong to the category is in the brackets).

Category	Sub-Categories
dining (5)	no sub-category (2); seafood (1); farmer meal (1); flavor (1)
cost (6)	no sub-category (1); ticket (4); discount (1)
environment (25)	no sub-category (1); quietness (1); safety (1); furnishings (1); panorama (1); scenery (7); air (1); facility (2); beauty (1); tourist density (2); color (2); hygiene (2); atmosphere (1); sunshine (1); vegetation (1)
advice (9)	no sub-category (6); book ticket (1); tour guide (1); scheduling (1)
traffic (11)	no sub-category (3); walk (2); subway (1); bus (1); riding (1); parking (1); road trip (1); rental car (1)
experience (26)	no sub-category (24); climbing mountain (2)
feature of attraction (58)	ice & snow (1); museum (2); urban core (1); Hmong village (1); village (1); panda (1); landmark (1); sculpture (1); animal (2); cave (1); high-rise building (1); college (1); park (1); historic site (1); piazza (1); beach (1); aquarium (1); waterfall (1); lake (1); flower (3); building (1); river (1); alley & street (1); attraction (2); Kaifeng (1); national custom (2); bridge (1); place to film (1); forest (1); mountain (2); business (1); stone carving (1); water town (1); temple (1); theme park (1); railroad (1); hot spring (1); protected historic site (2); cultural heritage (1); canyon (1); modern city (1); town (1); recreation (1); art (1); gingkgo (1); playground (1); garden (1); arboretum (1); natural landform (1); natural scenery (1)
introduction (22)	no sub-category (3); preservation condition (1); locals (1); allusion (1); style (1); custom (1); scale (1); building structure (1); history (3); area (1); entrance (2); picture (1); location (2); development (1); culture (1); religion (1)
evaluate (15)	no sub-category (2); negative (4); positive (9)
time (5)	peak and slack season (1); queuing time (1); opening time (2); duration (1)
activity (16)	no sub-category (1); boating (1); taking cable car (1); temple fair (1); rafting (1); family activity (1); shopping (2); photography (3); show (3); exhibition (1); worship (1)
accommodation (1)	no sub-category (1)
MaFengWo self (1)	no sub-category (1)

Table 2. Examples of image topics and top 10 high-frequency keywords.

Topic ID	Category	Top 10 High-Frequency Keywords
0	activity: show	role; music; image; wax; vivid (2); lifelike; dance; song and dance; wonderful
16	feature of attraction: landmark	different; landmark (3); seem; county; mark; road; landmark building; difference
21	time: opening time	aspect; open (2); crowds; crowd; close; museum close; off duty; closed on Monday
26	experience	pass by; relax; excellent; just pass by; highlight; destination; cannot miss; by pure chance; specially; mind
28	cost: ticket	free of charge; free to visit; watch; open for free; Admission with ID; get ticket; exchange; verify ID (2); Admission with ticket
34	evaluate: positive	be worth; tourism; be very worth; small and beautiful; nice nice nice; nice; with somebody; play; be worth to come
40	dining: seafood	seafood; characteristic; fresh; bathroom; mantis shrimp; seafood market; Musculus senhousei; taste; food stall; corner
41	traffic: parking	park; parking (2); parking fee; hold; park at; park at will; no bathroom; toll collector; park to the side of the road
54	introduction: history	a part; founded in (2); built in; divided into; AD; access; formal; period; old name
77	activity: photography	take photo with (3); take photo (3); Fairy Lake; be tempted to; population; area
85	environment: air	air; fun; breathe (2); interesting; natural oxygen bar; fresh air; facility; site; anion
97	feature of attraction: panda	lovely; panda; cute (2); panda kindergarten; very cute; charmingly nave; cute critter; so lazy; see panda
109	introduction: location	located (5); adjoin; to the north; take road as boundary; west of; to the west
126	environment: tourist density	tourist; protect; worshipper; foreign tourist; endless stream of tourists; surge; vociferously; popular; large party; too popular
140	evaluate: negative	problem; owner; manage; on file; warn; attitude; service attitude; very poor; chaos; servant
169	accommodation	hotel (2); romantic; see; big bedroom; check in; suite; starred hotel; booking; standard room
171	experience: climbing mountain	up; add; flower perfume; up to; climbing mountain; climb up; quite steep (2); wear knee; tired
183	feature of attraction: beach	sand beach; seawater; beach; comfortable; sand; white sand; soft; sandiness; swimming; fine sand beach
191	advice: booking ticket	in advance; put in; luggage; tourist center; information; book on; taobao; book; online shopping; buy ticket
195	traffic: subway	subway; scene; subway station; ordinary people; seek; Line 1; station; Xintiandi Station; Xincheng Station; Line 10

Chinese words translated into the same English word are merged and the number of these Chinese words is shown in brackets.

3.2.2. Online Review HIN

The final online review HIN was constructed based on the nodes and edges acquired from the above original review data and image topic extraction result. The online review HIN contains 1,301,537 nodes and 7,017,522 edges. The number of different types of nodes and edges is shown in Table 3.

Table 3. Statistics of nodes and edges in the online review HIN.

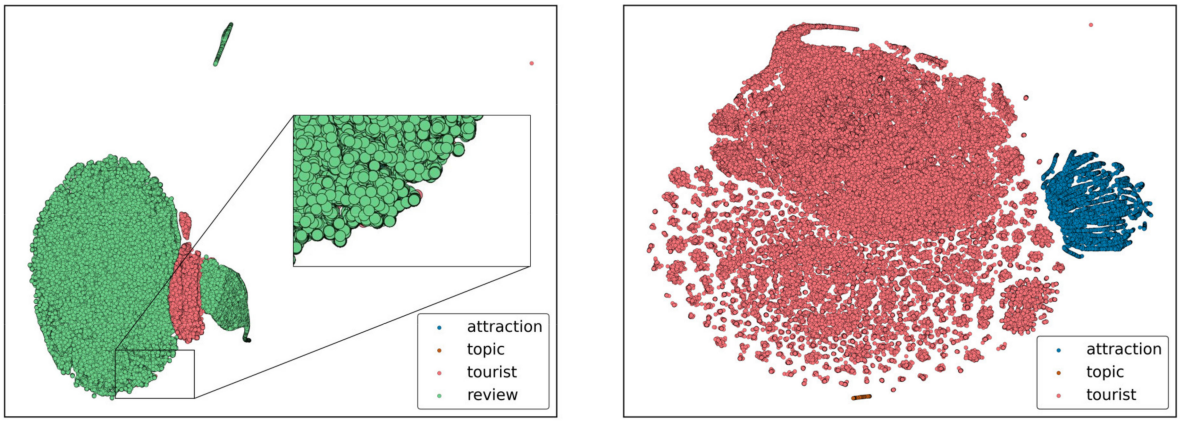
Node			
#Attraction	**#Tourist**	**#Review**	**#Topic**
11,122	202,777	1,087,438	200
Edge			
#Having Review/Review of	**#Writting/Written by**	**#Hastopic/Topic of**	
2,176,454	2,176,454	2,664,614	

3.3. Online Review HIN Embedding

The HIN2Vec model was implemented by the open code of the author (https://github.com/csiesheep/hin2vec). For training a HIN2Vec model, some important param-

eters were set considering the amount of data and efficiency: the dimension of vectors is 150, the number of negative samples is 5, and the length of random walks is 1000. The length of meta-paths was set to 4 to capture the semantics in the two edge paths "havingreview"→"writtenby"→"writing"→"reviewof" and "havingreview"→"hastopic" →"topicof"→"reviewof" presented in Section 2.1.

Inspired by the work of Liu et al. [40], we used t-SNE to reduce the HIN2Vec embedding result with 150 dimensions to two dimensions for visualization on a two-dimensional plane. The results are shown in Figure 7: (a) is the visualization of all nodes, and (b) is that of the nodes except for the "review" nodes. It illustrates that all nodes are mixed in the visualization result, but the nodes of "attraction", "topic", and "tourist" are grouped. The possible reason is the "review" nodes are connected with all the other kinds of nodes in the online review HIN, so the HIN2Vec model cannot discriminate the difference of semantics between "review" nodes and other kinds of nodes in the embedding process. Consequently, the HIN2Vec model captures the semantic differences between attractions, topics, and tourists, which ensures the effectiveness of the relatedness identification between attractions.



(a) all nodes **(b) nodes except reviews**

Figure 7. Visualization of the embedding result.

3.4. Top Related Attractions Finding

This experiment was conducted to find the top related attractions of a given attraction to verify the presented method. The relatedness $rel\_hin(a_i, a_j)$ between attraction a_i and a_j based on online review HIN embedding was identified through measuring the cosine similarity between the vectors of attractions, which is a common metric of measuring the similarity between high-dimensional vectors in machine learning.

3.4.1. Comparative Relatedness Identification Methods

We used two comparison relatedness identifying methods based on homogeneous co-occurrence attraction network embedding and image topic distribution as the contrasts of the proposed relatedness identification.

(1) Relatedness Identification Based on Homogeneous Network Embedding

We built a homogeneous co-occurrence attraction network from the assumption "a tourist written a review text to a tourism attraction" means "this tourist has visited this tourism attraction". Thus, if a tourist wrote different reviews of different tourism attractions, this tourist has visited all these tourism attractions. That is, these tourism attractions co-occur, which can be used to identify the relatedness between attractions, as mentioned in Section 1. Specifically, the node in the homogeneous co-occurrence attraction network is attraction. The edge represents that its two nodes (attractions) have been visited by the same tourists. Moreover, the edge has a weight to indicate the number of the same tourists.

A higher weight of the edge means that the nodes (attractions) of this edge are visited together more frequently.

Then, the homogeneous network embedding model LINE (Large-scale Information Network Embedding) was used to characterize the nodes with low-dimensional vectors. The LINE model is suitable for undirected, directed, and/or weighted networks containing millions of nodes [41]. This model (1) captures the first-order proximity between the nodes of the observed links in the network, and (2) explores the second-order proximity between the nodes, which is not measured through the observed links but through the shared neighborhood structures of the nodes. Thus, the LINE model can solve the problem of sparse edges in the large real homogeneous network, which leads to poor performance of node embedding.

The LINE model was implemented by the open code of the author (https://github.com/tangjianpku/LINE). For training a LINE model, some important parameters were set: the vector dimension was 128, the number of negative samples was 5, the total number of training samples is 10,000, the edge is undirected, and the *first-order* and second-order proximity were both used. Similarly, to the result of the HIN2Vec model, the relatedness $rel\_line(a_i, a_j)$ between a_i and a_j was also identified by calculating their cosine similarity in the vector space embedded by the LINE model.

(2) Relatedness Identification Based on Image Topic Distribution

An attraction has many different reviews, and a review has an image topic, so an attraction has different image topics, namely image topic distribution of this attraction. The image topic distribution of attraction can form a vector of this attraction: the vector dimension is the number of all image topics, and the dimension value is the reviews' number that belongs to the corresponding topic. Thus, the relatedness between two attractions was identified by these vectors: the high relatedness means these two attractions have similar image topic semantics. Specifically, the numbers of an attraction's reviews belonging to each image topic are counted from the result of topic extraction and as the dimension values. Therefore, the vector dimension was 200, consistent with the parameter of the LDA model. After normalizing each vector of the attraction, the relatedness $rel\_topic(a_i, a_j)$ between a_i and a_j was measured by the cosine similarity.

3.4.2. Results

Each attraction can obtain its top 1000 related attractions by identifying and sorting the *rel_line*, *rel_topic*, and *rel_hin*, which reflects the perspectives of attraction co-occurrence, image topic semantics, and HIN, respectively. Figure 8 shows the spatial distribution of the top 1000 related attractions of five attractions sorting by *rel_line*,*rel_topic*, and *rel_hin* (abbreviated as *SD_line*, *SD_topic*, and *SD_hin*, respectively, for brevity): the Palace Museum, Shanghai Disneyland, Qingdao Trestle, Mount Siguniang, and Potala Palace. To observe the difference in spatial distribution more clearly, the Kernel Density Estimation (KDE) surface generating from the top related attractions overlays each map. This figure illustrates that, compared with the attractions in *SD_hin* and *SD_topic*, the attractions in *SD_line* were closer to the given attractions. This phenomenon is consistent with the notion that frequent pairwise occurrences of points of interest (POIs) indicate their geographic proximity [11] because the *SD_line* is conducted from the co-occurrence attraction network. Meanwhile, compared with the attractions in *SD_hin* and *SD_line*, the attractions in *SD_topic* were more scattered in China (e.g., the high-density surfaces are greater). The reason is that the geographic proximity of attractions' image topics is not significant. For some image topics relating to certain types of natural terrain, the spatial distributions of these topics may present some rules. For instance, Qingdao Trestle is a wharf that stretches into the sea at Qingdao, so most of its attractions in *SD_topic* are located on the coastline of China. Overall, the *SD_hin* is situated between the *SD_line* and *SD_topic*, showing that the proposed method identifies the relatedness between attractions from the perspectives of attraction co-occurrence and attraction image topic comprehensively.

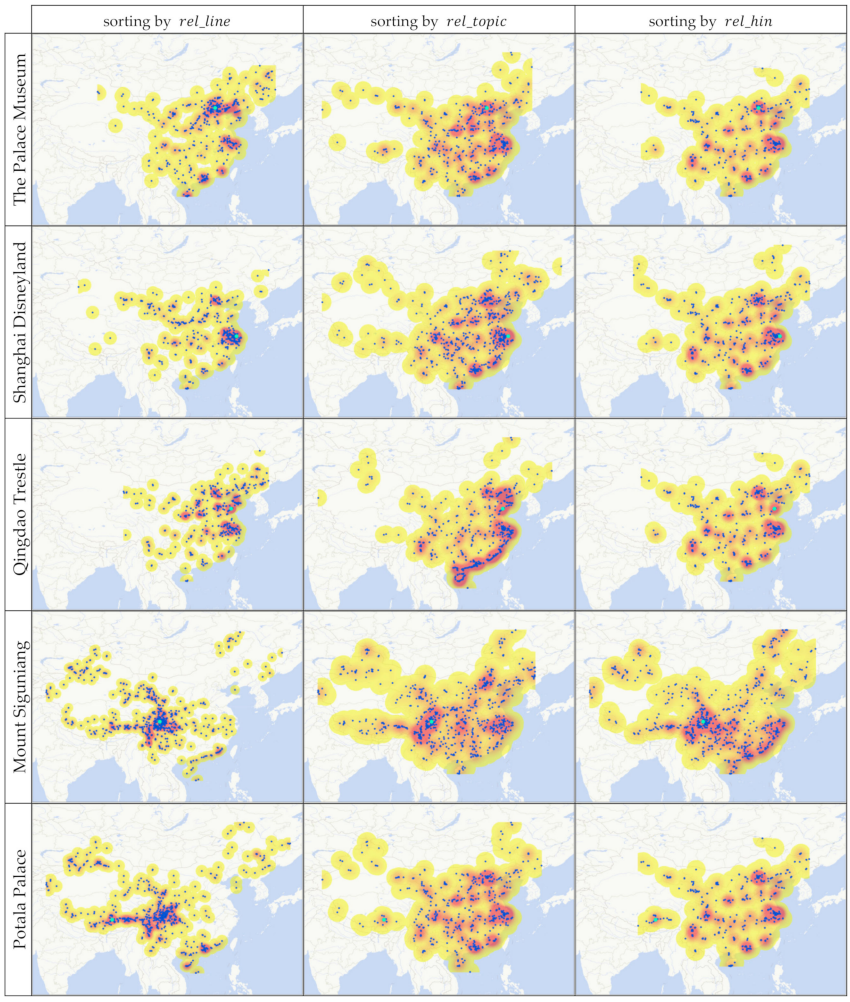


Figure 8. Spatial distribution of the top 1000 related attractions of the given attractions sorting by different relatedness identification (the cyan points are the given attractions; the blue points are the top related attractions; the "yellow-red" surfaces are the KDE surfaces generating from the top related attractions: "yellow" indicates a low density of attractions, and "red" indicates a high density of attractions).

3.4.3. Efficiency Analysis

We calculated the *rel_line* and *rel_topic* between each attraction and its top 1000 related attractions which were sorted by *rel_hin*. The statistical indicators' average, median, first quartile, and third quartile of the *rel_line*, *rel_topic*, and *rel_hin* on each sorting position are shown in Figure 9. This figure illustrates that the tendencies of *rel_line* and *rel_topic* both decreased when the *rel_hin* decreased. This result indicates that the HIN2Vec model is most efficient in terms of fusing information of attraction co-occurrence and the image topic semantics to comprehensively identify the relatedness between attractions.

Furthermore, we calculated the distances between each attraction and its top 1000 related attractions which were sorted by *rel_line*, *rel_topic*, and *rel_hin*, respectively. The statistical indicators' average, median, first quartile, and third quartile of the distances on each sorting position are shown in Figure 10. It can be seen that the distances between attractions and their top 1000 related attractions sorting by *rel_topic* are large, and the differences between the distances on different sorting positions are slight. It illustrates that the geographic proximity of attraction image topics is again not significant. Furthermore, the distances between attractions and their top 1000 related attractions sorting by *rel_line* and *rel_hin* increased as the relatedness decreased. Specifically, the distances based on *rel_hin* increased faster than the distances based on *rel_line*, e.g., the median distance

of each attraction and its 200th related attraction sorted by *rel_line* is 391.04 km, but the median distance of that sorted by *rel_hin* is 907.71 km. These show that the HIN2Vec model can capture the image topic similarity based on geographic proximity. That is, the HIN embedding listed not only the near co-occurrence attractions as the related attractions of an attraction, but also the attractions far away but with similar image topics.

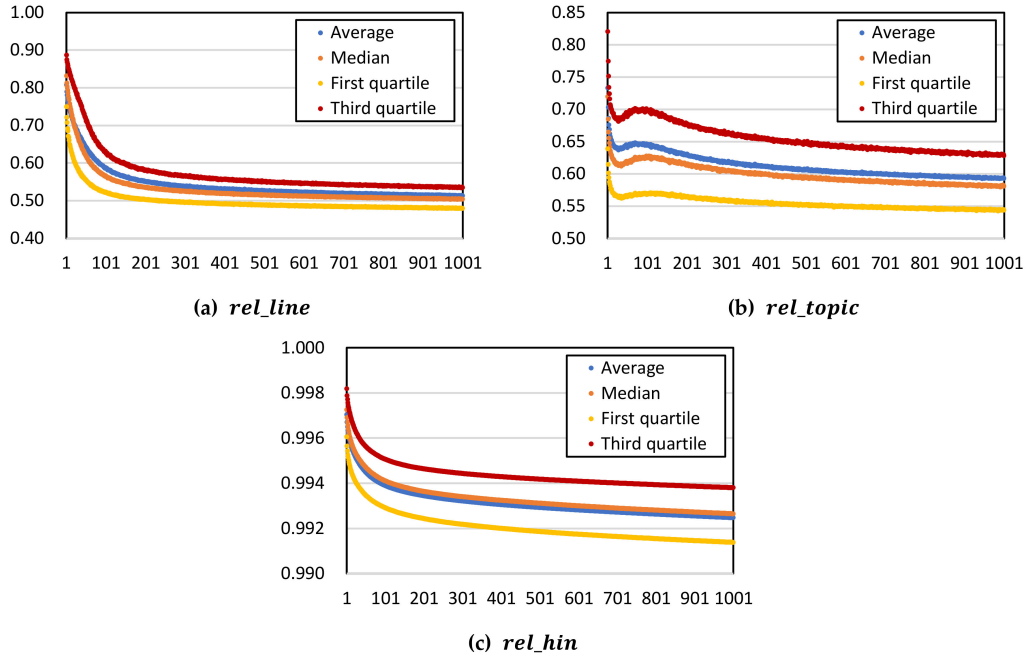


Figure 9. Statistical indicators of *rel_hin*, *rel_line*, and *rel_topic* between each attraction and its each top 1000 related attraction sorted by *rel_hin* (the *rel_hin* is calculated by cosine similarity).

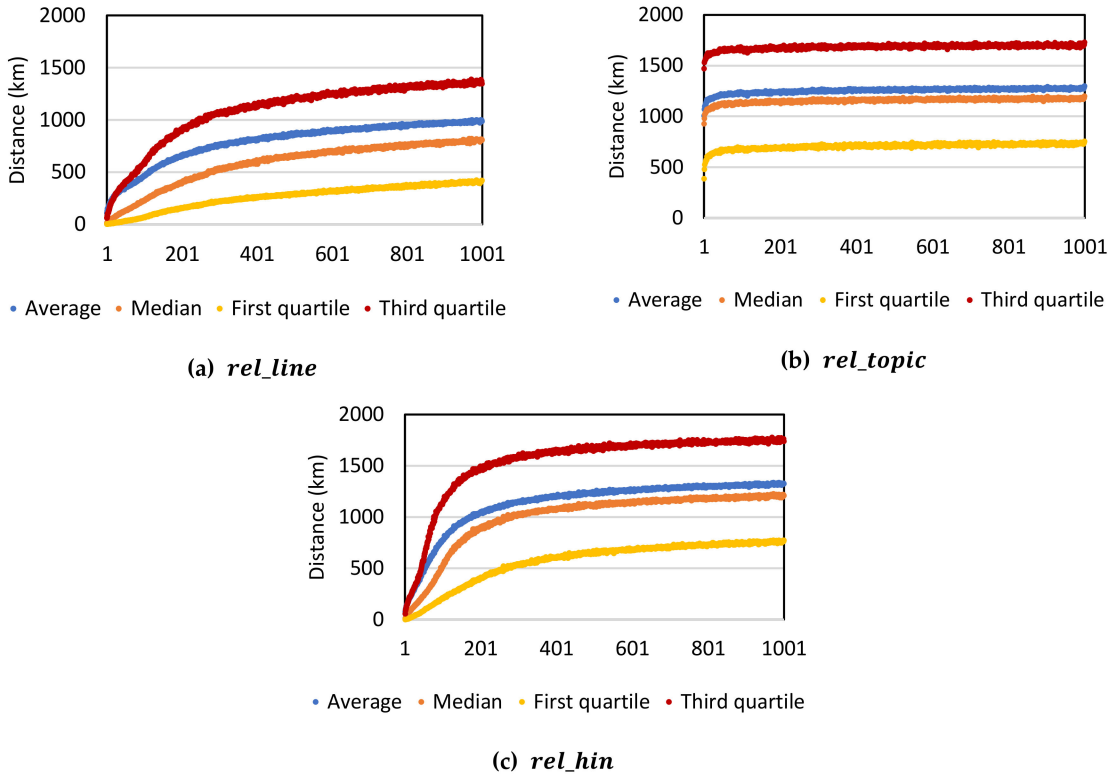


Figure 10. Statistical indicators of distances between each attraction and its top 1000 related attractions sorted by *rel_hin*, *rel_line*, and *rel_topic* (the *rel_hin* is calculated by cosine similarity).

3.5. Attractions Clustering

The attractions can be grouped using a clustering algorithm based on the vectors from HIN embedding. In this case study, the Affinity Propagation (AP) clustering algorithm was selected to group the attractions. AP clustering views each data point as a node in a network, then recursively transmits real-valued messages along the edges of the network until a good set of exemplars and corresponding clusters emerges [42]. Specially, the real-valued messages are divided into responsibility and availability. The former is the message sent from data point i to candidate clustering centre point j, reflecting the suitability that point j is the clustering centre of point i. The latter is the message sent from candidate clustering centre point j to data point i, reflecting the appropriateness that point i selects point j as its clustering centre. AP clustering determines the clustering centre of all data points by the iterative calculation of these two real-valued messages, then finishes the clustering. Thus, the number of clusters of the Affinity Propagation clustering algorithm was not prespecified, which is consistent with the lack of prior knowledge to determine the optimal number of clusters of attractions. Finally, 11,122 attractions were clustered into 467 clusters.

Then, we calculated the average of the relatedness based on the online review HIN (*ave_rel*), the average of distances (*ave_dis*) and the standard deviation of distances (*std_dis*) between all attractions in each cluster. The larger *ave_rel* of cluster indicated that the attractions in this cluster have stronger relatedness. The larger *ave_dis* of cluster indicated that the attractions in this cluster were distributed in a larger space range. The larger *std_dis* of cluster indicated that the attractions in this cluster were distributed more unevenly in space. Because the similarity between data points in AP clustering is measured by the negative Euclidean distance between vectors, we also used negative Euclidean distance to identify the relatedness between attractions in this experiment:

$$relatedness(x, y) = -dist(x, y) = -\sqrt{\sum_{i=1}^{n}(x_i - y_i)} \quad (2)$$

where x and y are the vectors of two attractions. x_i and y_i are components of vector x and y, respectively.

Figure 11 indicates the overall trend of *ave_dis* and *std_dis* decreasing with *ave_rel* increasing, while it is not strictly decreasing. It illustrates that the attractions which are spatially close and uniformly distributed have a higher probability of being clustered. That is, the HIN2vec model decides that attraction co-occurrence is a factor that may be more important than image topic in determining the semantic relationship between attractions from the proposed online review HIN. However, the HIN2vec model embeds the attractions from the structure of the online review HIN rather than simply combining co-occurrence relatedness and image topic relatedness between attractions directly. This process may take advantage of additional potential semantic relationships, so the trend is not strictly decreasing.

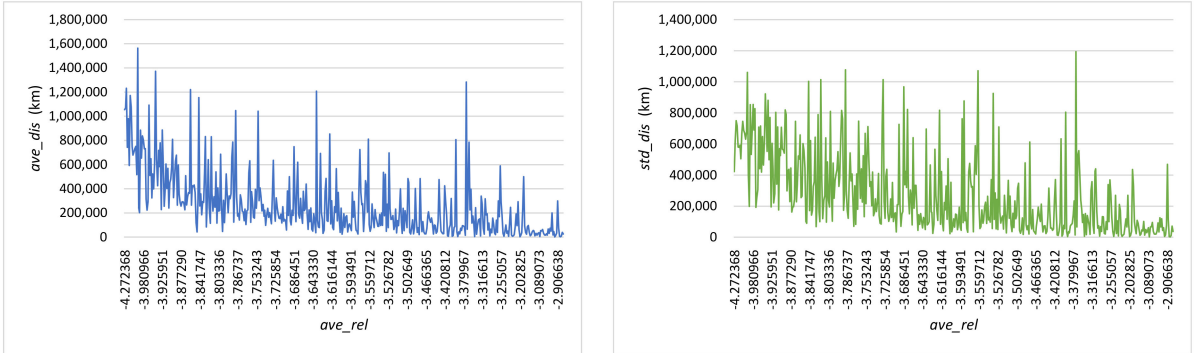


Figure 11. Trend of *ave_dis* and *std_dis* between attractions in each cluster as *ave_rel* increases (the *ave_rel* is calculated by negative Euclidean distance).

We used Jenks natural breaks classification method to further divide the above 467 clusters into five groups based on the *ave_rel* of these clusters. Then, one cluster for each group was chosen for exploring the validity of the clustering results. The spatial distributions of the five groups and the attractions in the five sample clusters are shown in Figure 12. It indicates that the attractions in each cluster were spatially concentrated as *ave_rel* increased. Most attractions in cluster #6 and all attractions in cluster #20 were concentrated in a city (Harbin and Wuhan). Besides, even the attractions of a cluster are distributed in a large space range, these attractions may have similar image topics, e.g., cluster #441 is about "museum", cluster #307 is about "beach", and cluster #6 is about "historic towns". Meanwhile, the attractions in cluster #20 and cluster #161 are clustered because if these attractions are distributed in a small space range, then they have a higher probability of being co-visited by tourists, resulting in a stronger co-occurrence relatedness between these attractions than image topic relatedness between them. Overall, the attractions in different clusters present co-occurrence relatedness or image topic relatedness, which demonstrates that the HIN embedding automatically adjusts the importance of attraction co-occurrence and attraction image in final relatedness from the characteristics of real data. The clustering result helps one to further discover the attraction communities, of which the attractions can establish close cooperation.

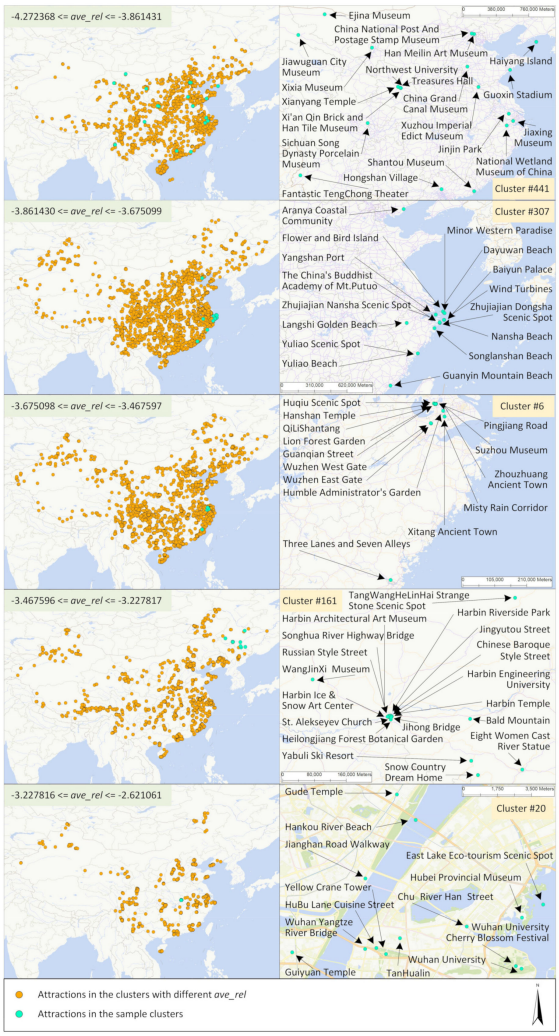


Figure 12. Spatial distributions of the attractions in the clusters with different *ave_rel*. The attractions in the left column are the attractions in all clusters with given range of *ave_rel*. The attractions in the right column are the attractions in the clusters sampled from the corresponding left clusters (the *ave_rel* is calculated by negative Euclidean distance).

4. Discussion

In this study, the HIN2Vec model was used to embed the online review HIN into low-dimensional vector space, whereas there are many other HIN embedding models, as mentioned in Section 2.3, such as Metapath2Vec, HAN, and HERec. These models also show a good performance in representing the nodes in HIN. The reasons we selected the HIN2Vec model in this research are: (1) as presented in Section 2.3, the HIN2Vec model can construct meta-paths automatically and avoid meta-path design. Although the two edge paths "havingreview"→"writtenby"→"writing"→"reviewof" and "havingreview"→"hastopic"→"topicof"→"reviewof" express the semantics of attraction co-occurrence and attraction image, as explained in Section 2.1, we think the other edge paths can still give clues for the HIN2Vec model to mine the semantic relationships between nodes, which may not have significant meanings for people to understand. (2) No model has demonstrated undisputed performance on HIN embedding, because the above models are verified in different tasks and evaluation metrics with different pre-processing [43]. Overall, the emphasis of this research illustrates that the HIN can retain the difference between different relationship semantics when the online reviews are reorganized into a network structure, and the HIN embedding model can capture and fuse these different relationship semantics, which facilitate identifying the relatedness between attractions from a comprehensive perspective.

The proposed relatedness identification between attractions based on online review HIN is a data-driven approach. The HIN2Vec model can automatically capture and fuse heterogeneous semantic information in the online review HIN and give the attraction vectors through fusing all information, without the need to manually set the weights of attraction co-occurrence and attraction image topic. Specifically, the strength of attraction co-occurrence is reflected by the heterogeneous network structure, rather than the weight of edges, as in the traditional network analytics. That is, if two attractions have more reviews written by the same tourists, the HIN2Vec model will ensure these attractions are closer to each other in the embedding vector space, i.e., these attractions have stronger relatedness. Moreover, the HIN2Vec model generates the training data from HIN based on random walk and negative sampling, which overcomes the data-sparsity problem and outputs the effective embedding vectors of attractions that have a few co-occurrences with other attractions or attraction image topics.

While the number of node types in the proposed online review HIN was four and the number of edge types was six, more information in the tourism online reviews should be introduced into the online review HIN in future to better identify the relatedness between attractions, such as the type of attraction, the level of attraction, the residence of the tourist, and so on. Nevertheless, the quality and reliability of the information needs to be noticed to avoid introducing noise into the HIN. For instance, the attraction level "National AAAAA level tourism attraction" is labelled as "National 5A level tourism attraction", "AAAAA attraction", "5A level attraction", etc. in Chinese on MaFengWo. The reason is that the information in social networks lacks strict inspection and revision. Thus, the model will determine these labels as having different semantics if these labels are not uniformed. Furthermore, the data size affects embedding efficiency. The training time of the HIN2Vec model exceeded 15 hours based on the constructed online review HIN. If the length of the meta-paths was set to 5, the HIN2Vec model would not have completed training for five days. Consequently, while HIN embedding showed good performance in identifying the relatedness between attractions, the HIN structure, data size, data quality, and HIN embedding model need to be carefully selected to ensure the training efficiency.

The related attractions of an attraction can be used as the recommendation information when a user browses this attraction online. Meanwhile, the attraction manager can regard the tourists who visited these related attractions as potential customers and take measures to attract these tourists. In addition, the HIN embedding model embeds not only the attractions, but also tourists and image topics in the online review HIN. Thus, the relatedness between tourists can also be identified, which helps to extract tourist profiles,

cluster tourists, and recommend related tourists based on fusing the multiple relationship semantics. Furthermore, the attractions that may be of interest to a tourist can be obtained based on the relatedness between tourists and attractions by the operation of their vectors.

5. Conclusions

Most studies identify the relatedness between attractions through measuring their co-occurrence extracted from online tourism reviews. However, the implicit semantic information in these reviews, which definitely contributes to modelling the relatedness from a more comprehensive perspective, is ignored due to the difficulty of quantifying the importance of different dimensions of information and fusing them. Thus, this paper introduces HIN to reorganize the tourism online reviews for representing the co-occurrence and images of attractions, and then uses HIN embedding to comprehensively identify the relatedness between attractions. First, an online review-oriented HIN was designed to form the different types of elements in the reviews. Second, a topic model was employed to extract the nodes of the HIN from the review texts. Third, an HIN embedding model was used to capture the semantics in the HIN and comprehensively represent the attractions with low-dimensional vectors. The effectiveness of the presented method was validated by three tasks based on the tourist review data from MaFengWo: (1) the visualization illustrates the HIN2Vec model accurately discriminates the attraction, topic, and attraction image types of elements in an online review HIN; (2) the top 1000 related attraction findings show that the presented method comprehensively identifies the relatedness between attractions from the perspectives of both attraction co-occurrence and attraction image; (3) the result of attraction clustering demonstrates the HIN embedding can automatically adjust the importance of attraction co-occurrence and attraction image in final relatedness based on the characteristics of real data. These results indicate that the online review HIN can correctly express the semantics of attraction co-occurrences and attraction images in reviews, and the HIN embedding can capture the differences in these semantics, which facilitates identifying the relatedness between attractions from a comprehensive perspective.

Limitations also exist in this study. Firstly, the structure of the proposed online review HIN only contained four node types and six edge types. Meanwhile, the tourism online reviews provided more types of information, such as the type of attraction, the level of attraction, the residence of the tourist, etc., which helped to identify the relatedness through integrating more semantics. Secondly, we only used the HIN2Vec model to verify the effectiveness of the proposed online review HIN, not to compare the effects of different HIN embedding models. Moreover, while the HIN2Vec model can capture the semantic information in the long hop edge paths, its training time increased significantly with the increase in data size. Therefore, in future work, we would like to (1) extend the online review HIN with more types of information; (2) improve the training efficiency in terms of model selection, model optimization, and HIN structure optimization; and (3) apply the proposed relatedness identification to tourism recommendation and tourism analytics.

Author Contributions: Conceptualization, Peiyuan Qiu and Feng Lu; methodology, Peiyuan Qiu; validation, Peiyuan Qiu and Jialiang Gao; formal analysis, Peiyuan Qiu; investigation, Peiyuan Qiu and Jialiang Gao; resources, Peiyuan Qiu and Jialiang Gao; data curation, Jialiang Gao and Peiyuan Qiu; writing—original draft preparation, Peiyuan Qiu; writing—review and editing, Feng Lu; supervision, Feng Lu; project administration, Feng Lu; funding acquisition, Feng Lu and Peiyuan Qiu. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Grant No. 41631177, Grant No. 42001341); Doctoral Research Fund of Shandong Jianzhu University, grant number X20084Z; and a grant from State Key Laboratory of Resources and Environmental Information System.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors would like to thank the four anonymous reviewers for their valuable suggestions, which significantly improve the quality of this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, Y.; Wang, F.; Kang, C.; Gao, Y.; Lu, Y. Analyzing Relatedness by Toponym Co-Occurrences on Web Pages. *Trans. GIS* **2014**, *18*, 89–107. [CrossRef]
2. Gu, Z.; Zhang, Y.; Chen, Y.; Chang, X. Analysis of Attraction Features of Tourism Destinations in a Mega-City Based on Check-in Data Mining—A Case Study of Shenzhen, China. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 210. [CrossRef]
3. Hong, T.; Ma, T.; Huan, T.-C. Network behavior as driving forces for tourism flows. *J. Bus. Res.* **2015**, *68*, 146–156. [CrossRef]
4. Du, S.; Zhang, H.; Xu, H.; Yang, J.; Tu, O. To make the travel healthier: A new tourism personalized route recommendation algorithm. *J. Ambient Intell. Humaniz. Comput.* **2019**, *10*, 3551–3562. [CrossRef]
5. Li, J.; Xu, L.; Tang, L.; Wang, S.; Li, L. Big data in tourism research: A literature review. *Tour. Manag.* **2018**, *68*, 301–323. [CrossRef]
6. Han, H.; Kim, S.; Otoo, F.E. Spatial movement patterns among intra-destinations using social network analysis. *Asia Pac. J. Tour. Res.* **2018**, *23*, 806–822. [CrossRef]
7. Kirilenko, A.P.; Stepchenkova, S.O.; Hernandez, J.M. Comparative clustering of destination attractions for different origin markets with network and spatial analyses of online reviews. *Tour. Manag.* **2019**, *72*, 400–410. [CrossRef]
8. Sugimoto, K.; Ota, K.; Suzuki, S. Visitor Mobility and Spatial Structure in a Local Urban Tourism Destination: GPS Tracking and Network analysis. *Sustainability* **2019**, *11*, 919. [CrossRef]
9. Liu, B.; Huang, S.; Fu, H. An application of network analysis on tourist attractions: The case of Xinjiang, China. *Tour. Manag.* **2017**, *58*, 132–141. [CrossRef]
10. Mou, N.; Zheng, Y.; Makkonen, T.; Yang, T.; Tang, J.; Song, Y. Tourists' digital footprint: The spatial patterns of tourist flows in Qingdao, China. *Tour. Manag.* **2020**, *81*, 104151. [CrossRef]
11. Haris, E.; Gan, K.H.; Tan, T.-P. Spatial information extraction from travel narratives: Analysing the notion of co-occurrence indicating closeness of tourist places. *J. Inf. Sci.* **2020**, *46*, 581–599. [CrossRef]
12. Yuan, H.; Xu, H.; Qian, Y.; Li, Y. Make your travel smarter: Summarizing urban tourism information from massive blog data. *Int. J. Inf. Manag.* **2016**, *36*, 1306–1319. [CrossRef]
13. Lin, M.S.; Liang, Y.; Xue, J.X.; Pan, B.; Schroeder, A. Destination image through social media analytics and survey method. *Int. J. Contemp. Hosp. Manag.* **2021**. Epub ahead of printing. [CrossRef]
14. Blei, D.M.; John, D.L. Topic models. In *Text Mining: Classification, Clustering, and Applications*; Taylor and Francis: London, UK, 2009.
15. Rossetti, M.; Stella, F.; Zanker, M. Analyzing user reviews in tourism with topic models. *Inf. Technol. Tour.* **2016**, *16*, 5–21. [CrossRef]
16. Guo, Y.; Barnes, S.J.; Jia, Q. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tour. Manag.* **2017**, *59*, 467–483. [CrossRef]
17. Wang, J.; Li, Y.; Wu, B.; Wang, Y. Tourism destination image based on tourism user generated content on internet. *Tour. Rev.* **2020**, *76*, 125–137. [CrossRef]
18. Kurashima, T.; Iwata, T.; Hoshide, T.; Takaya, N.; Fujimura, K. Geo topic model: Joint modeling of user's activity area and interests for location recommendation. In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM' 13), Association for Computing Machinery, Rome, Italy, 4–8 February 2013; pp. 375–384.
19. Zheng, X.; Luo, Y.; Sun, L.; Zhang, J.; Chen, F. A tourism destination recommender system using users' sentiment and temporal dynamics. *J. Intell. Inf. Syst.* **2018**, *51*, 557–578. [CrossRef]
20. An, H.; Moon, N. Design of recommendation system for tourist spot using sentiment analysis based on CNN-LSTM. *J. Ambient Intell. Humaniz. Comput.* **2019**, 1–11. [CrossRef]
21. Shafqat, W.; Byun, Y.-C. A Recommendation Mechanism for Under-Emphasized Tourist Spots Using Topic Modeling and Sentiment Analysis. *Sustainability* **2020**, *12*, 320. [CrossRef]
22. Shi, C.; Philip, S.Y. *Heterogeneous Information Network Analysis and Applications*; Springer: Berlin/Heidelberg, Germany, 2017.
23. Sun, Y.; Han, J. Mining heterogeneous information networks: A structural analysis approach. *ACM SIGKDD Explor. Newsl.* **2013**, *14*, 20–28. [CrossRef]
24. Shi, C.; Hu, B.; Zhao, W.X.; Yu, P.S. Heterogeneous Information Network Embedding for Recommendation. *IEEE Trans. Knowl. Data Eng.* **2019**, *31*, 357–370. [CrossRef]
25. Wang, C.; Raina, R.; Fong, D.; Zhou, D.; Han, J.; Badros, G. Learning relevance from heterogeneous social network and its application in online targeting. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR' 11), Association for Computing Machinery, Beijing, China, 24–28 July 2011; pp. 655–664.
26. Yu, J.; Gao, M.; Li, J.; Yin, H.; Liu, H. Adaptive Implicit Friends Identification over Heterogeneous Network for Social Recommendation. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM' 18), Torino, Italy, 22–26 October 2018; pp. 357–366.

27. Hosseini, A.; Chen, T.; Wu, W.; Sun, Y.; Sarrafzadeh, M. HeteroMed: Heterogeneous Information Network for Medical Diagnosis. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM' 18), Torino, Italy, 22–26 October 2018; pp. 763–772.
28. Chen, X.; Yin, J.; Qu, J.; Huang, L. MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction. *PLOS Comput. Biol.* **2018**, *14*, e1006418. [CrossRef]
29. Zhang, F.; Wang, M.; Xi, J.; Yang, J.; Li, A. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Sci. Rep.* **2018**, *8*, 3355. [CrossRef] [PubMed]
30. Shi, C.; Li, Y.; Zhang, J.; Sun, Y.; Yu, P.S. A Survey of Heterogeneous Information Network Analysis. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 17–37. [CrossRef]
31. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
32. Qiang, J.; Qian, Z.; Li, Y.; Yuan, Y.; Wu, X. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. *IEEE Trans. Knowl. Data Eng.* **2020**, 1–19. [CrossRef]
33. Bengio, Y.; Ducharme, R.; Vincent, P.; Janvin, C. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155.
34. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the Workshop at International Conference on Learning Representations 2013, Scottsdale, AZ, USA, 2–4 May 2013; pp. 1–12.
35. Liu, K.; Gao, S.; Qiu, P.; Liu, X.; Yan, B.; Lu, F. Road2Vec: Measuring traffic interactions in urban road system from massive travel routes. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 321. [CrossRef]
36. Mihalcea, R.; Tarau, P. TextRank: Bringing Order into Text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Association for Computational Linguistics, Barcelona, Spain, 25–26 July 2004; pp. 404–411.
37. Dong, Y.; Chawla, N.V.; Swami, A. Metapath2Vec: Scalable Representation Learning for Heterogeneous Networks. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 17), Halifax, NS, Canada, 13–17 August 2017; pp. 135–144.
38. Fu, T.; Lee, W.-C.; Lei, Z. HIN2Vec: Explore Meta-paths in Heterogeneous Information Networks for Representation Learning. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM' 17), Singapore, Singapore, 6–10 November 2017; pp. 1797–1806.
39. Wang, X.; Ji, H.; Shi, C.; Wang, B.; Ye, Y.; Cui, P.; Yu, P.S. Heterogeneous Graph Attention Network. In Proceedings of the 2019 World Wide Web Conference (WWW' 19), San Francisco, CA, USA, 13–17 May 2019; pp. 2022–2032.
40. Liu, N.; Huang, X.; Li, J.; Hu, X. On Interpretation of Network Embedding via Taxonomy Induction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery (KDD' 18), London, UK, 19–23 August 2018; pp. 1812–1820.
41. Tang, J.; Qu, M.; Wang, M.; Zhang, M.; Yan, J.; Mei, Q. LINE: Large-scale Information Network Embedding. In Proceedings of the 24th International Conference on World Wide Web (WWW' 15), International World Wide Web Conferences Steering Committee, Florence, Italy, 18–22 May 2015; pp. 1067–1077.
42. Frey, B.J.; Dueck, D. Clustering by Passing Messages between Data Points. *Science* **2007**, *315*, 972–976. [CrossRef] [PubMed]
43. Dong, Y.; Hu, Z.; Tang, J.; Sun, Y.; Wang, K. Heterogeneous Network Representation Learning. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20), Online, 7–15 January 2021; Volume 5, pp. 4861–4867.