Vertical Search Method of Tourism Information Based on Mixed Semantic Similarity

Honghong Chen(✉) and Hongshen Liu

Heilongjiang Polytechnic, Haerbin 150080, China
gyg210422@163.com

Abstract. With the rapid development of the tourism industry, the volume of tourism information has increased exponentially, making it difficult for tourists to obtain the tourism information they need, which has become the main factor restricting the development of the tourism industry. In order to solve the above problems, the research on vertical search method of tourism information based on mixed semantic similarity is proposed. The Heritrix web crawler is used to collect tourism information and de duplicate it. On this basis, the Nutch structure is used to process tourism information, calculate the mixed semantic similarity between tourism information and known topics, and determine the corresponding topics of tourism information based on this, and develop an adaptive vertical search algorithm for tourism information. The vertical search results of tourism information can be obtained by executing the formulation algorithm. The experimental data shows that after the application of the proposed method, the maximum recall rate of vertical search of tourism information is 96%, the maximum precision rate of vertical search of tourism information is 98%, and the minimum response time of vertical search of tourism information is 0.56s, which fully proves that the proposed method has better application performance.

Keywords: Mixed Semantic Similarity · Vertical Search · Similarity Calculation · Tourism Information

1 Introduction

The rapid development of the national economy, the continuous improvement of people's living standards and the government's policy support for the domestic tourism industry have led to the rapid development of the domestic tourism industry, which plays an increasingly important role in the national economy [1]. China is rich in tourism resources and has a large consumer group. Tourism has developed into a wide range of industrial groups, including people's basic consumption needs of clothing, food, housing and transportation, as well as higher level needs of play, entertainment and shopping. The development of Internet technology has fundamentally changed the way people obtain information. For the tourism industry and its related fields, there are a large number of websites in China, including tourist attraction information websites, professional

L. Yun et al. (Eds.): ADHIP 2023, LNICST 547, pp. 91–107, 2024.
https://doi.org/10.1007/978-3-031-50543-0_7

websites such as hotels and airlines, comprehensive vertical search websites, and tourism channels of large portal websites. However, these websites have more or less problems, which are mainly reflected in: less effective information, more garbage information, poor user experience, and lack of features in content. On the one hand, users hope to quickly and accurately understand tourism related information through the Internet [2], on the other hand, they are confused and at a loss when facing the huge amount of mixed information. The general search engine is usually used as the entrance of information retrieval, but it is ineffective in accurately querying the target information. Although a large number of results can be retrieved according to keywords, it is difficult to really meet the needs of users. Most of the time, users also need to manually analyze and filter these information accurate location of target information is a time-consuming and laborious task for ordinary users. In this context, tourism vertical search engine came into being.

The tourism vertical search engine is aimed at the tourism industry and its related fields, automatically collecting relevant data, sorting out, filtering and artificially optimizing the data, and providing users with accurate tourism information retrieval services. Compared with general search engines, tourism vertical search engines have structurally extracted and integrated information related to the tourism industry, such as hotels, scenic spots, air tickets, catering, transportation and other information, and provided special services according to users' personalized needs. The tourism vertical search engine automatically extracts the relevant data of the tourism industry, saving a lot of time in searching for information. At the same time, it carries out structural processing on these data and integrates the data in combination with the different needs and interests of users, which can not only meet the current needs of users, but also explore the potential needs of users [3]. The tourism vertical search engine has completed a lot of work for users, such as information search, sorting, filtering, etc., making it easier and faster for users to obtain the information they want, and making the process of obtaining information more intelligent and humane.

The main core function of tourism vertical search engine is to vertically search tourism information and provide users with relevant tourism information services. The emergence and development of search engines have greatly facilitated people's access to information and played a positive role in promoting social progress. However, with the advent of the era of big data, the amount of data has exploded, and the amount of data generated every day in the world has reached 1EB. Although the general search engine continues to innovate in technology and improve its processing speed, it is still unable to meet people's diverse and personalized information retrieval needs [4]. According to the existing research results, the information search methods that are frequently used are the information search method based on correlation coefficient and the tourism vertical search method based on MongoDB. There is a time delay (mainly because the search engine cannot respond to information updates in a timely manner, the time required for classification and indexing has increased significantly due to the dramatic increase in data volume, and many effective information cannot be added to the index in a timely manner, with a certain time delay) The search accuracy is low (users have increasingly high requirements for the efficiency and accuracy of information retrieval, while general search engines have no advantages in the efficiency and accuracy of search) and other

defects, which cannot meet the needs of tourists. Therefore, the research on vertical search method of tourism information based on mixed semantic similarity is proposed.

2 Research on Vertical Search Method of Tourism Information

2.1 Tourism Information Collection

Tourism information collection is the primary link of tourism information vertical search, which plays a vital role in the follow-up research. This research uses the Heritrix web crawler to collect tourism information, which lays a solid foundation for subsequent tourism information de duplication.

Heritrix is an open source web crawler tool developed by Java and used to crawl and archive web pages on the Internet. Herritrix provides interface management tools to control Herritrix, and also provides command line tools for users to select and call. Its structure is shown in Fig. 1.

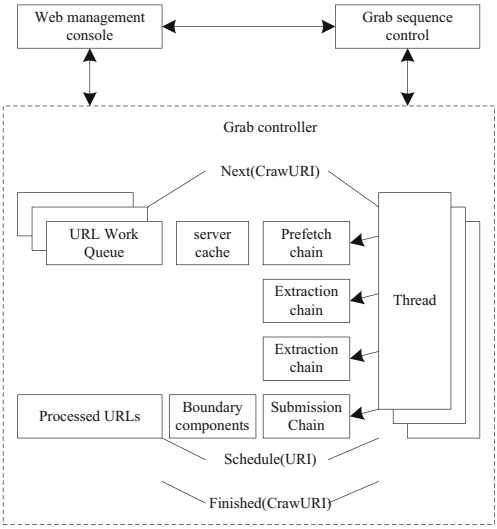

Fig. 1. Schematic Diagram of Heritrix Structure

As shown in Fig. 1, Heritrix is mainly divided into the following components: the Web management console, the crawl sequence controller, and the crawl controller [4]. The crawl controller is composed of Frontier, multithreaded processor, server cache and crawl range controller.

The functions of each module of Heritrix are as follows:

The web management console is implemented by jetty to provide users with a visual web-based management page. Through this page, you can set the modules used by the crawler at runtime. A good and convenient management interface promotes the wide application of Heritrix.In addition, Heritrix also provides a set of command line management tools to manage crawlers.

The crawl order controller is a configuration file based on XML format, which configures the crawl task. When the crawler runs, it calls the corresponding module according to the information in the configuration file to implement the crawl task [5]. In addition to the configuration of key modules, the configuration file also includes other important information, such as the definition of the captured URL range and the captured entry URL.

The grabbing controller is the core of the Heritrix web crawler, the general controller of the grabbing process, and the coordinator of each functional module to ensure that each module cooperates to complete the grabbing task together during the grabbing process. The crawl controller consists of three core modules, namely: Frontier, server cache and multithreaded processor.

Among them, Frontier is the URL scheduler, which is used to balance the fetching task, allocate the task to multiple different fetching threads to complete, and reduce the access pressure on a single Web server.

The server cache stores the persistent information of the captured server, including the server's IP address, DNS and other information, which can be called by various components of the crawler.

The multithreaded processor is the thread that actually executes the fetching work. In the running process, each thread will execute the URL processing chain once to complete the relevant operations of page fetching [6]. Herritrix generally runs in a multi-threaded way, that is, multiple URL processing chains are performing page fetching operations at the same time, and the fetching of each URL is independent of each other, and the fetching threads of different pages do not interfere with each other.

The advantages and disadvantages of Heritrix are shown in Table 1.

According to the content analysis in Table 1, Heritrix is suitable for vertical search of tourism information, so using it to collect tourism information on the Internet mainly includes two stages, namely web page collection and information extraction, as shown below:

(1) Web page collection.

Web page collection includes the collection of information on domestic and foreign tourism websites. Taking domestic tourism websites as an example, provide the specific collection process. Collect hotel data from Tongcheng, Ctrip and Xinxin websites, including HTML pages of hotel information and photos related to hotels, and store them locally in the form of mirror files. Collect scenic spot data from Tongcheng, Xinxin and Tuniu websites, including HTML pages of scenic spot information and photos related to scenic spots, and save them locally in the form of mirror files [7]. This function is mainly used to retrieve specific theme data from the target website using a web crawler and save it locally for subsequent operations.

(2) Information extraction.

Extract data in a specific format from the captured hotel data, generate Json format files, and store them in the collection. The root node of hotel data is hotel, and the specific data extraction format is shown in Table 2.

Table 1. Analysis of Advantages and Disadvantages of Heritrix

S/N	advantage	inferiority
1	Heritrix is fully functional.Heritrix itself only provides a framework for web crawlers, and there are many options in the same mode.Using the default provider module can record the captured pages well and avoid repeated work	Heritrix is only suitable for capturing Internet pages, that is, storing the image of the page. It cannot parse the content of the page itself. This part of parsing needs to be completed separately
2	Heritrix has excellent performance.Because Herritrix controls I/O and operations very well, it uses very little system resources when it runs. Even if it runs for a long time, it only generates very little system garbage, and its performance will not be reduced	Heritrix itself has many imperfections, which need to be expanded and improved to give full play to its potential.For example, Heritrix can't handle Chinese characters very well
3	Heritrix configurability.The Heritrix framework is flexible. Users can configure key modules to control the capture time, strategy, speed and scale of different websites. In addition, they can configure file archiving methods and capture target file formats	The configuration of Heritrix is very complicated, and there are many places that need customization.When changing a feature, you need to change many things. Some modules need to be changed to take effect
4	–	The fault tolerance and recovery mechanism of Heritrix are not perfect, and it needs to be expanded to achieve

Table 2. Hotel Information Extraction Format Table

content	Json	remarks
Original URL	OriginUrl	Original page link address
theme	Title	Original page subject information
Hotel name	HotelName	Name of the hotel
level	Level	Hotel star
Description	Description	Hotel related information
city	City	Hotel City
District/County	Zone	District/County where the hotel is located
address	Address	Address of the hotel
Picture Path	PicUrl	Hotel image storage path
floor price	LowestPrice	Hotel minimum price

Extract specific format data from the captured scenic spot data and integrate it into the sites collection. The root node of scenic spot data is site. Json format files are generated and stored in the specific data extraction format as shown in Table 3.

Table 3. Format table of scenic spot information extraction

content	Json	remarks
Original URL	OriginUrl	Original page link address
theme	Title	Original page subject information
Name of scenic spot	SiteName	Name of the scenic spot
Description	Description	Information about tourist attractions
city	City	City where the scenic spot is located
District/County	Zone	District/County where the scenic spot is located
Types of tourist attractions	SiteType	Types of tourist attractions
Picture Path	PicUrl	Storage path of scenic spot pictures

Collect tourism information through Heritrix according to the above process to provide basis for subsequent research.

2.2 De Duplication of Tourism Information

Most of the collected tourism information belongs to structured information, which has duplication and affects the effect of subsequent information search. Therefore, it is necessary to de duplicate the tourism information.

The de duplication processing of structured information in vertical search engines is one of the important structured information processing technologies in the improved vertical search engine model, which directly affects the accuracy of vertical search engines' search results for web information [8, 9]. The de duplication function of vertical search engines is mainly used in four parts, as shown in Table 4.

In the section of freeing storage space: By deduplicating data, duplicate data can be identified and deleted, thereby reducing the occupancy of storage space. This is particularly important in large-scale data storage and processing, as it can save costs and improve the overall performance of the system.

In the section of improving the efficiency of web page information collection: In web crawling and information collection tasks, data deduplication can effectively reduce the downloading and processing of duplicate web pages or data, and improve the efficiency of collection [10]. By removing duplicates, it is possible to avoid duplicate crawling of the same content, saving network bandwidth and computing resources.

Table 4. Application table of de duplication function

Application part	Application description
Free storage space	In the explosive development trend of the Internet, the repetition rate of information is constantly improving. De duplication of information with high repetition rate can significantly reduce the information storage space. Structured data is the main part of the data used by vertical search engines, and its storage requirements are high, so the problem of de duplication of structured information is crucial
Improve the efficiency of web information collection	In the process of crawling web information resources, search engines collect the content related to the search request sent by users. A large number of duplicate data resources in professional fields will lead to low efficiency of information collection of vertical search engines. After the de duplication of structured information of information resources, the collection efficiency of structured information resources can be better improved
Enhance user experience and improve user utilization	In the improved search engine model, the de reprocessing design of structured information effectively improves the sorting results of the index sorting module. Index sorting improves the accuracy of data retrieval results for the user interface. It can not only enhance users' experience of vertical search engines, but also improve users' utilization of vertical search engines
Improve the quality of retrieval data and the accuracy of index sorting	Vertical search engines have a preliminary problem of page duplication removal in the process of page crawling and page extraction. However, due to the diversity of the format of massive web information resources, the repetition rate of search results can only be optimized in a small range. Therefore, further data de reprocessing is needed in the vertical search engine to improve the quality of data retrieval results and the accuracy of index sorting

There is no distinction between enhancing user experience and improving user usage: for user generated content (such as social media, comments, etc.), data deduplication can avoid duplicate and redundant information being displayed to users, improving their reading and browsing experience. At the same time, reducing duplicate content can also increase user stickiness to the platform, improve user usage and satisfaction.

In improving the quality of retrieved data and the accuracy of index sorting, data deduplication can improve the quality of retrieved data in search engines and database systems. By removing duplicate data, the impact of redundant information on search results can be reduced, and search accuracy and efficiency can be improved. In addition, when building indexes and Sorting algorithm, data de duplication can avoid the interference of duplicate data on the sorting results, and improve the accuracy and effect of sorting.

The information processing technology carries out the pattern separation, data adjustment and relevant link analysis of the body content of the web page information stored in the database, and carries out a structural analysis process. Based on the structured data, the information is further processed such as de duplication and classification [11, 12]. An efficient de recalculation method should be used to process structured data, which improves the security performance of structured information. Through the study of the above several commonly used de duplication algorithms and in-depth analysis of their advantages and disadvantages, this paper proposes an improved algorithm with high efficiency and high security [13].

The basic idea of the improved de duplication method is that $T_i = \{t_1, t_2, \cdots, t_n\}$ represents the set of feature items of the top n tourism information with the highest weight, $W_i = \{w_1, w_2, \cdots, w_n\}$ represents the corresponding feature vector of the feature item set, $A(P_i)$ represents the webpage summary, $C(T_i)$ represents the string concatenated from the top n tourism information, $C[S(T_i)]$ represents the string concatenated from the top n tourism information after alphabetical sorting, $MD5(X)$ represents the N hash value of string X, $M\left(P_i, P_j\right)$ represents P_i and P_j are mutually repeated webpage. Use $A \Rightarrow B$ to represent 'A holds, then B holds'. Then the de duplication algorithm expression is.

$$\left.\begin{array}{c}(MD5(C(T_i))) = \left(MD5\left(C\left(T_j\right)\right)\right) \\ \left(\frac{|w_i - w_j|^2}{|w_i|^2 - |w_j|^2}\right) < \alpha\end{array}\right\} \Rightarrow M\left(P_i, P_j\right) \tag{1}$$

In formula (1), α represents an auxiliary parameter with a value range of [0, 1]; $MD5$ processes input information in 512 bit packets. Each packet is divided into 16 32-bit sub packets. After a series of processing, the output of the algorithm consists of four 32-bit packets. Cascading the four 32-bit packets will generate a 128 bit hash value. According to formula (1), the tourism information is de duplicated to obtain more indirect tourism information, which facilitates the follow-up research.

2.3 Theme Identification of Tourism Information

The reason for choosing the Nutch framework for tourism information segmentation is its open source, configurability, distributed architecture, Data cleansing and de duplication functions, as well as scalability and active community support. These advantages

can quickly establish a tourism information segmentation system and achieve efficient and accurate data collection and processing. Among them, (1) Open source framework: Nutch is an open source web crawler framework that can be freely accessed and used. (2) High configurability: The Nutch framework provides rich configuration options that can be flexibly configured and adjusted according to specific needs. (3) Distributed architecture: The Nutch framework supports distributed crawling and processing, which can execute tasks in parallel on multiple machines, improving crawling speed and efficiency. (4) Data cleansing and de duplication: The Data cleansing and de duplication functions are built into the Nutch framework, which can clean and de duplicate crawled data through configuration and plug-ins. (5) Scalability and flexibility: The Nutch framework adopts a modular design, allowing you to expand functionality and customize development based on requirements and scenarios. (6) Community Support and Activity: Nutch is an active open source project with a large user and developer community. Therefore, based on the above de duplicated tourism information, the Nutch architecture is used to segment tourism information. On this basis, the mixed semantic similarity between tourism information and known topics is calculated, and the corresponding topics of tourism information are determined, providing support for the subsequent launch of adaptive vertical search algorithm.

The processing of Chinese information is mostly based on word processing. The Chinese information stored in the computer does not have obvious segmentation marks between words. Therefore, we must use the segmentation specification of Chinese words to convert Chinese information into words, which is the so-called Chinese word segmentation problem [14]. Chinese word segmentation module is a part of the preprocessing module, which is mainly used by the indexer. The indexer transmits the original text information to be processed to the word segmentation module for processing, while the Chinese word segmentation module returns the processed word segmentation results to the indexer with the corresponding data structure for subsequent processing, so that the indexer can use the word segmentation results to index documents.

Nutch is developed for English environment, so its Chinese word segmentation needs to be improved. The NutchAnalyzer class is an extension point for extending analysis text in Nutch. Writing your own Chinese plug-in must extend from this extension point. The specific word segmentation method is determined by the user according to the application. GB2312–80 divides the Chinese character code table into 94 areas, corresponding to the first byte; Each area has 94 bits, corresponding to the second byte. The value of two bytes is the area code value and the tag number value plus 32 (20H) respectively. The collected Chinese characters are placed in the 16–87 area, and the 0 bit of each area does not store Chinese characters. The offset calculation formula of Chinese characters in the code table is.

$$O = (C_1 - 0xB0) * 94 + (C_2 - 0xA1) \tag{2}$$

In formula (2), O represents the position of a Chinese character in the code table; C_1 and C_2 represents the internal code of Chinese characters.

The logical structure of the word segmentation dictionary is the data structure form after the dictionary is added to the memory. The logical structure of the dictionary used in this paper is shown in Fig. 2.

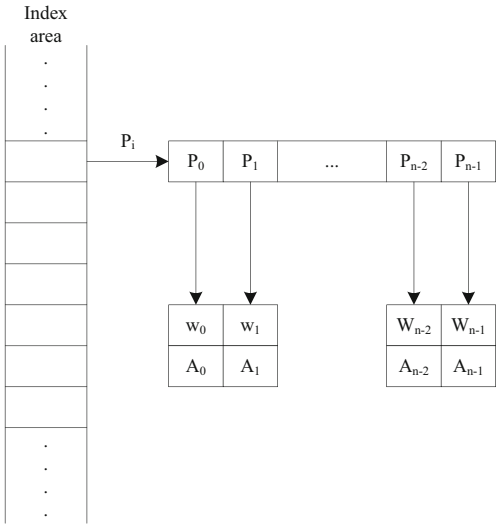

Fig. 2. Structure diagram of Chinese word list in memory

Among them, P_i is a pointer to all entries with the first character being the i-th Chinese character C_i; w_i is the k-th word with the initial character C_i (entries are arranged in internal code order from smallest to largest), excluding the initial character; A_i is the attribute of w_i (including value, ambiguity, position, part of speech, etc.); I_i is the index entry of the word C_i, occupying 5 bytes.

The tourism information is segmented according to the rules shown in Fig. 2. Based on this, the mixed semantic similarity between the tourism information and the known topics is calculated [15]. The semantic computing method proposed in this section emphasizes the importance of syntactic structure in computing sentence semantic similarity. The similarity between two sentences can be obtained by weighted summation of similarity between terms of syntactic components shared between sentences [16]. Unlike SyMSS, which only selects core words in dependency pairs and compares them, this paper calculates all words in the same dependency relationship. The syntactic structure and semantic hybrid algorithm in this paper uses the following formula to measure the semantic similarity between sentences. The expression is.

$$sim_1(S_1, S_2) = \frac{1}{n}\sum_{i=1}^{n} sim(E_{1i}, E_{2i}) - (N_1 + N_2 - N) \cdot PF \tag{3}$$

In Eq. (3), S_1 and S_2 are composed of N_1 and N_2 syntactic components, respectively, with N common syntactic components. Each syntactic component in S_1 is represented by E_1. Sentence S_2 is the same, where E_{1i} and E_{2i} have the same syntactic function; $(N_1 + N_2 - N)$ represents indicates the number of redundant syntactic components between two sentences, because if there are different grammatical components between two sentences, this method will ignore the syntactic information of these different components, and the syntactic information of these different components may also cause semantic differences. Due to different syntactic components, the similarity

between these sentences cannot be calculated directly. Here, a balance parameter is used to indicate that there is additional information between these sentences; PF represents auxiliary calculation parameters.

Semantic similarity recognition will give different weight values to different syntactic components, so this paper also proposes a weighted syntactic combination method, which can measure the similarity between sentences through formula (4), which is.

$$sim_2(S_1, S_2) = \frac{w_0 A_0 + \cdots + w_n A_n + \sum_{i=1}^{n} w^o \cdot sim(E_{1i}, E_{2i})}{w_0 + \cdots + w_n} - (N_1 + N_2 - N) \cdot PF \tag{4}$$

In Eq. (4), $w_n A_n$ represents the product of the weight assignment of the subject and the semantic similarity value at the lexical level between the subjects. The product of the direct object and the verb is omitted. E_i is no defined common syntactic component. According to the formula, the calculation result of the above example is 0.7261. Since the subject and verb of the two sentences are consistent, their similarity increases after weighting.

The syntactic information is used to improve the semantic similarity at the sentence level, which largely avoids the neglect of sentence meaning in previous sentence level semantic similarity calculation, and improves the effect of sentence level semantic similarity. However, depending on the syntactic information of the sentence, there are also shortcomings. For example, in the two sentences "the train from Nanjing to Beijing leaves at ten o'clock" and "the train from Beijing to Nanjing leaves at ten o'clock", "the train from Nanjing to Beijing" and "the train from Beijing to Nanjing" are both attributes of the subject "train", so they are used as a dependency unit to calculate the lexical similarity between them. However, the algorithm at the lexical level is insensitive to the order of words, so the final similarity result of the two sentences is 1, which means they are identical. However, it is obvious that the two sentences are not identical semantically. Therefore, the order of words in the sentence is also related to semantics. This paper uses formula (5) to introduce editing distance as a feature to correct the effect of similarity calculation. The expression is.

$$sim^*(S_1, S_2) = sim_1(S_1, S_2) + \beta \frac{\min dist(w_{i1}, w_{i2})}{|S_1| + |S_2|} \tag{5}$$

In formula (5), $\min dist(w_{i1}, w_{i2})$ refers to the editing distance of the new sentence formed by extracting the same syntactic information vocabulary in order between two sentences. For example, the editing distance of the new sentence extracted in the above example is 1. $|S_1| + |S_2|$ refers to the sum of the number of words in the sentence, so that the influence of the editing distance of long and short sentences is consistent, and the value of this item is less than 1; β is a feature parameter configuration, and its value range is $(-\infty, 0]$.

According to the calculation result of formula (5), the tourism information theme recognition rules are formulated, as shown in the following formula:

$$\begin{cases} sim^*(S_1, S_2) \geq \delta & \text{Subject matching} \\ sim^*(S_1, S_2) < \delta & \text{Contradictory themes} \end{cases} \tag{6}$$

In formula (6), δ refers to the recognition threshold of tourism information subject, which needs to be set according to the actual situation.

Through the above process, the identification of tourism information topics is completed, which facilitates the subsequent adaptive vertical search.

2.4 Introduction of Adaptive Vertical Search Algorithm

Based on the above recognition results of tourism information topics, an adaptive vertical search algorithm for tourism information is developed, and the vertical search results of tourism information can be obtained by implementing the developed algorithm, so as to provide better services for tourists.

In this paper, the vector space model is used A_P and H_P. Set an initial threshold value to judge the relevance between the content of the web page corresponding to the network node and the query subject, and only the web pages with high relevance can be included in the root set or the corresponding base set. The algorithm is implemented as table 5:

Table 5. Implementation process of adaptive vertical search algorithm

Step	Content
1	Construct the root collection
2	Expand the base set
3	Loop through steps 1 and 2 and determine if the required number of extension pages that meet the pre-set threshold has been reached
4	Calculate the *Hub* and *Authority* values and normalize them. Until A_P and H_P converge, otherwise return to step 3
5	Output vertical search results for tourism information

Step 1: Construct the root collection [17]. The construction of the root set is very critical. It first requires users to focus on keywords in the field or industry; Secondly, set the weight according to the role of these keywords; Finally, we began to use these keywords to construct the root set. When constructing the root set, the first thing to be calculated is the relationship between query q and page t in the root set, expressed as.

$$\begin{cases} \zeta(q,t) < \gamma & \text{delete} \\ \zeta(q,t) \geq \gamma & \text{reserve} \end{cases} \tag{7}$$

In Eq. (7), $\zeta(q,t)$ represents the calculated value of the relationship between query q and page t;γ represents the limit value.

According to formula (7), if their relationship is not greater than the predetermined limit, t will be removed from the root set; If the calculated value exceeds the pre-set limit, then t becomes a node in the root set.

The number of root sets can be set as an upper limit according to the actual situation. When the upper limit is reached, the expansion of the number of root sets will stop. Then the root set expansion rules are as follows:

$$\begin{cases} K < \psi \text{ continue} \\ K \geq \psi \quad \text{cease} \end{cases} \tag{8}$$

In Eq. (8),K represents the number of root set nodes;ψ represents the upper limit of the root set node.

Step 2: Expand the base set. The root set is used to extend the base set. The extension method is the same as that of constructing the root set;

Step 3: Step 1 and Step 2 can operate circularly until the number of expansion pages that meet the preset threshold reaches the required number;

Step 4: Calculation *Hub* and *Authority* Value and normalize it. Until A_P and H_P convergence, otherwise return to step 3;

Step 5: Select the top n with the highest A_P and H_P values as the return result, which is the vertical search result for tourism information.

The above process completes the vertical search of tourism information, provides more abundant and accurate tourism information support for tourists, and promotes the development of the tourism industry.

3 Experiment and Result Analysis

3.1 Experiment Preparation Stage

The main task of the experiment preparation stage is to select evaluation indicators, and the selection results are as follows:

First, recall: refers to the calculation formula of the results returned most by the search engine according to the user's query criteria:

$$L = \frac{Q_1}{Q_{total}} \times 100\% \tag{9}$$

In Eq. (9),L represents the recall rate;Q_1 represents the number of pages related to the subject in the search results;Q_{total} refers to the number of pages related to all topics.

The second is the precision ratio: refers to the ratio of the number of theme related pages to the total number of returned pages in the results returned to users by the search engine. The calculation formula is.

$$G = \frac{Q_1}{Q_2} \times 100\% \tag{10}$$

In Eq. (10), G is the precision; Q_2 indicates the number of all pages in the search results.

Third, response time: refers to the time spent by the search engine from the user submitting the query criteria to returning the results, which directly reflects the efficiency of the search engine query mechanism. Due to the limitation of research space, its calculation formula will not be repeated.

3.2 Analysis of Experimental Results

Based on the above selected evaluation indicators, the information search method based on correlation coefficient and the tourism vertical search method based on MongoDB are used as comparison methods 1 and 2 to carry out the comparative experiment of tourism information vertical search. The specific analysis process of the experimental results is as follows:

3.2.1 Analysis of Recall Rate

The recall rate obtained through experiments is shown in Fig. 3.

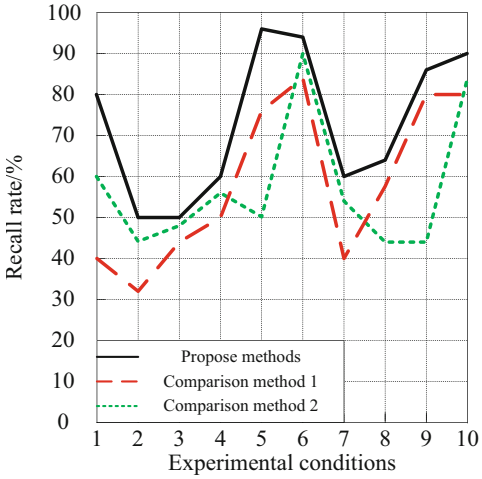

Fig. 3. Schematic diagram of recall rate

As shown in the data in Fig. 3, under the background conditions of different experimental conditions, the vertical search recall rate of tourism information obtained after the application of the proposed method is far higher than that of comparison method 1 and comparison method 2, and the maximum vertical search recall rate of tourism information obtained under the background of the fifth experimental condition is 96%.

3.2.2 Precision Analysis

The precision obtained through experiments is shown in Fig. 4.

As shown in the data in Fig. 4, under the background conditions of different experimental conditions, the vertical search precision of tourism information obtained after the application of the proposed method is far higher than that of comparison method 1 and comparison method 2, and the maximum vertical search precision of tourism information obtained under the background of the fifth experimental condition is 98%.

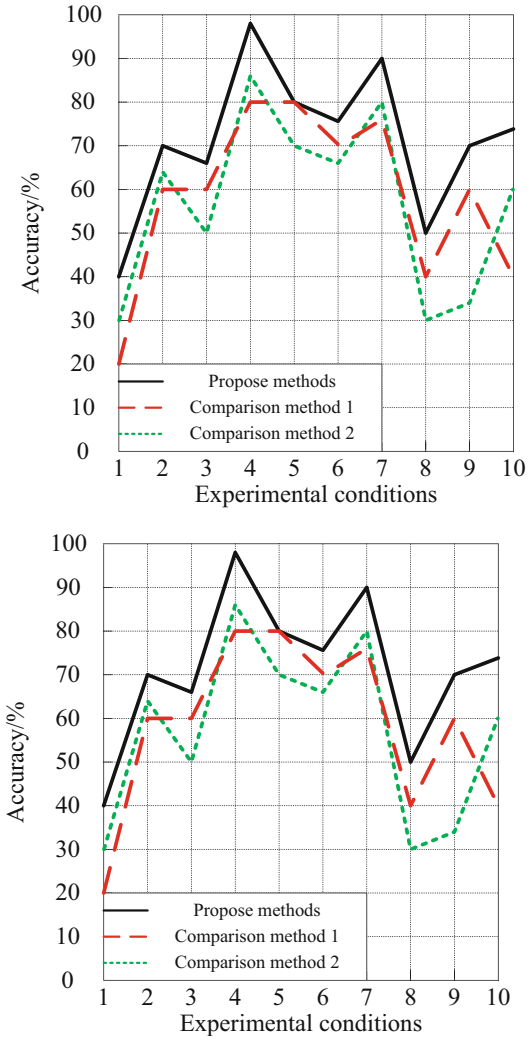

Fig. 4. Schematic diagram of precision

3.2.3 Response Time Analysis

The response time obtained through experiments is shown in Table 6.

As shown in the data in Table 6, under the background conditions of different experimental conditions, the vertical search response time of tourism information obtained after the application of the proposed method is far lower than that of comparison method 1 and comparison method 2, and the minimum vertical search response time of tourism information obtained under the background of the first experimental condition is 0.56s.

Table 6. Response Time Data Table/s

Test conditions	Propose method	Comparison method 1	Comparison method 2
1	0.56	2.56	4.26
2	1.23	3.02	5.78
3	0.80	4.56	6.35
4	0.75	5.01	4.15
5	1.11	5.89	7.25
6	1.25	6.12	6.59
7	1.04	4.12	8.01
8	0.89	4.52	6.89
9	1.01	6.35	7.45
10	1.20	7.01	8.10

4 Conclusion

With the development of the tourism industry, self-help travel has also developed rapidly. Tourism strategy information search has become a new demand. The development model of tourism strategy information search engine website, which takes tourism strategy information search as a development opportunity, and integrates a series of profit points such as hotels, air tickets, tourism products, is quietly emerging. Under this development trend, this paper proposes a new vertical search method for tourism information, introduces vertical search engine technology, provides users with more efficient and faster search tools, and creatively adds mobile elements. Users can download tourism information to mobile clients in the form of maps, upload photos anytime and anywhere, and combine GPS functions. Realize mobile phone search for surrounding living facilities and send scenic spots introduction to users. Experimental data show that the proposed method greatly improves the recall and precision of vertical search of tourism information, shortens the response time of vertical search of tourism information, and can provide users with more effective tourism information services.

Acknowledgement. 2022 Project of Heilongjiang Vocational College: "Preliminary Exploration of the Training Strategies for Higher Vocational Tourism Talents in the Background of Digital Economy" (No. XJYB2022097).

References

1. Jabalameli, M., Nematbakhsh, M., Ramezani, R.: Denoising distant supervision for ontology lexicalization using semantic similarity measures. Expert Syst. Appl. **177**(2), 114922 (2021)
2. Li, X., Li, H., Pan, B., et al.: Machine learning in internet search query selection for tourism forecasting. J. Travel Res. **60**(6), 1213–1231 (2021)

3. Chandrasekaran, D., Mago, V.: Evolution of semantic similarity—a survey. ACM Comput. Surv.Comput. Surv. **54**(2), 1–37 (2021)
4. Jiang, S., Cao, L.: Research on the secret homomorphism retrieval method of multiple keywords in privacy database. Comput. Simul. **39**(4), 408–412 (2022)
5. Liu, S., Xiyu, X., Zhang, Y., Muhammad, K., Weina, F.: A reliable sample selection strategy for weakly-supervised visual tracking. IEEE Trans. Reliab.Reliab. **72**(1), 15–26 (2023)
6. Kamran, A.B., Naveed, H.: GOntoSim: a semantic similarity measure based on LCA and common descendants. Sci. Rep. **12**(1), 1–10 (2022)
7. Lou, B., Zhao, W., Liu, X., Li, L., Ma, H.: Teaching design of tourism management based on information-based teaching method: a case study of selection of hotel construction site. Asian Agric. Res. **13**(12), 55–61 (2021)
8. Zhao, F., Zhu, Z., Han, P.: A novel model for semantic similarity measurement based on wordnet and word embedding. J. Intell. Fuzzy Syst. **40**(5), 1–12 (2021)
9. Yang, Z., Yang, L., Huang, W., et al.: Enhanced deep discrete hashing with semantic-visual similarity for image retrieval. Inf. Process. Manage. **58**(5), 102648 (2021)
10. Orlando, L., Ortega, L., Defeo, O.: Perspectives for sandy beach management in the Anthropocene: satellite information, tourism seasonality, and expert recommendations. Estuarine Coastal Shelf Sci. **262**, 107597 (2021). https://doi.org/10.1016/j.ecss.2021.107597
11. van der Vegt, A., Zuccon, G., Koopman, B.: Do better search engines really equate to better clinical decisions? If not, why not? J. Assoc. Inf. Sci. Technol. **72**(2), 141–155 (2021). https://doi.org/10.1002/asi.24398
12. Dias, L., Aldana, I., Pereira, L., et al.: A measure of tourist responsibility. Sustainability **13**(6), 3351 (2021)
13. Sánchez-Cervantes, J.L., Alor-Hernández, G., Paredes-Valverde, M.A., Rodríguez-Mazahua, L., Valencia-García, R.: NaLa-Search: a multimodal, interaction-based architecture for faceted search on linked open data. J. Inf. Sci. **47**(6), 753–769 (2021). https://doi.org/10.1177/0165551520930918
14. Varthis, E., Poulos, M., Giarenis, I., Papavlasopoulos, S.: A novel framework for delivering static search capabilities to large textual corpora directly on the Web domain: an implementation for Migne's Patrologia Graeca. Int. J. Web Inf. Syst. **17**(3), 153–186 (2021). https://doi.org/10.1108/IJWIS-10-2020-0062
15. Wu, C., Zhuo, L., Chen, Z., et al.: Spatial spillover effect and influencing factors of information flow in urban agglomerations—case study of china based on Baidu search index. Sustainability **13**(14), 8032 (2021)
16. Ting, X.: Chemistry course network teaching based on key information search and big data cloud platform. J. Intell. Fuzzy Syst. Appl. Eng. Technol. **40**(4), 7347–7358 (2021)
17. Wicaksono, A.F., Moffat, A.: Modeling search and session effectiveness. Inf. Process. Manage. **58**(4), 102601 (2021)