Knowledge-Based Systems 243 (2022) 108495

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A novel focused crawler combining Web space evolution and domain ontology

Jingfa Liu [a,b], Xin Li [c,*], Qiansheng Zhang [d], Guo Zhong [a,b]

[a] Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangdong University of Foreign Studies, Guangzhou 510006, China
[b] School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou 510006, China
[c] School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044, China
[d] School of Mathematics and Statistics, Guangdong University of Foreign Studies, Guangzhou 510006, China

ARTICLE INFO

Article history:
Received 20 June 2021
Received in revised form 19 February 2022
Accepted 22 February 2022
Available online 28 February 2022

Keywords:
Focused crawler
Web space evolution
Multi-objective optimization
Pareto optimal
Ontology

ABSTRACT

In many fields, how to catch the related-topic Web resources is crucial. As a vertical search method, focused crawler has received great attention in recent years. Currently, most focused crawlers consider multiple evaluating factors of the hyperlinks and use the weighted sum approach to compute the priorities of unvisited hyperlinks. However, the proper weighted coefficients are hard to determine, and their unsuitable values may even cause the direction of crawlers to deviate seriously from the topic. To overcome this issue, this article builds a multi-objective optimization model based on Web text and link structure and designs a crawler framework called the Web space evolution (WSE), where a hyperlink bank whose radius is gradually increased is introduced to extend the search scape of crawlers in Web space. To improve the uniformity and diversity of hyperlinks, a nearest and farthest candidate solution method is combined with the fast non-dominated sorting to choose Pareto-optimal solutions (hyperlinks). A domain ontology based on the formal concept analysis is applied to establish the topic model. By incorporating the WSE and the domain ontology into the focused crawling, a novel focused crawler called FCWSEO is proposed to collect topic-relevant webpages. The experimental results on the rainstorm disaster domain show that the FCWSEO outperforms other focused crawler strategies in terms of the quantity and quality of retrieved relevant webpages.

1. Introduction

The crawler, which is an important part of search engines for information retrieval (IR), is a technology for automatically obtaining webpages. However, for acquiring domain-specific knowledge, traditional crawlers have difficulty in implementing semantic analysis. The focused crawler is a Web crawler that selectively crawls webpages related to a predefined topic. Therefore, the focused crawler technologies characterized with topic preference have received great attention in recent years [1,2]. Focused crawlers have been widely used to build domain-specific Web corpora, national Web archives [3], domain-specific search engines [4], and so on. This article addresses the focused crawler about the topic of rainstorm disaster, which is one of the most frequent meteorological disasters. It is extremely important to obtain early warnings, preventive measures, and emergency response information about rainstorm disaster to reduce and avoid the losses caused by the rainstorm disaster and ensure the safety of people's lives and properties.

The target of focused crawlers is to retrieve topic-relevant webpages with larger quantity and higher quality in a short time. At present, the main difficulties of focused crawlers focus on three aspects: establishment of the topic benchmark model, assessment of unvisited hyperlinks, and design of the crawling strategies.

The topic benchmark model provides a judgment basis for topic identification of webpages and can be established in two ways: keywords specified by domain experts and feature-words described by a semantic structure [5]. Keywords express a topic by using a static or dynamic keywords list. However, this method has the disadvantage of polysemy and ignores the intrinsic link of keywords, which may have a negative impact on the accuracy of the topic description. Differently, feature-words are extracted from the corpus and focus on the semantic relationship. The most promising ways based on feature-words are context graph (CG) [6–8] and domain ontology [9–11]. Currently, the method of determining the central concept of CG has never been unified, and the construction of CG relies on the user's query history, which may cause topic deviation due to insufficient user knowledge. Therefore, most focused crawlers use ontology to describe the concepts and relations among domain knowledge and carry out topic crawling through a semantic way.

[*] Corresponding author.
E-mail address: lixin19931020@163.com (X. Li).

https://doi.org/10.1016/j.knosys.2022.108495
0950-7051/© 2022 Elsevier B.V. All rights reserved.

The primary methods of assessing unvisited hyperlinks include two categories: hyperlink structure-based method and Web text analysis-based method. At present, most of the related researches considers the combination of these two types of methods.

In the design of crawling strategies, the common ones are the breadth-first search (BFS) [12] and the optimal priority search (OPS) [13]. Because the BFS ignores the assessment of the priority of hyperlinks, the performance of the BFS is generally inferior to that of the OPS. Most scholars now use the OPS-based crawling strategy, but the OPS strategy is a greedy algorithm that is easy to make the search trap into the local optima. To avoid inherent flaws of the greedy algorithm, researchers have recently proposed some intelligent crawler methods based on meta-heuristic strategies, such as the genetic algorithm (GA) [14, 15], the improved tabu search (ITS) algorithm [16], the simulated annealing (SA) algorithm [17], the particle swarm optimization (PSO) algorithm [18] and the ant colony optimization (ACO) algorithm [19,20]. Among these intelligent crawlers, most researchers use the weighted sum approach which linearly integrates multiple indicators of evaluating unvisited hyperlinks and the corresponding weighted coefficients to compute priorities of hyperlinks, and then regard the selection of unvisited hyperlinks as a single-objective optimization problem (SOOP). However, the proper weighted coefficients are difficult to determine, and their unsuitable values may even cause the direction of crawlers to seriously deviate from the topic and fetch some irrelevant webpages.

This article builds a multi-objective optimization model based on Web text and link structure for evaluating the unvisited hyperlinks and designs a novel framework called the Web space evolution (WSE) to guide the search of the focused crawler. The experimental results of the focused crawler on the rainstorm disaster domain verify the effectiveness of the proposed methods. The main contributions of this article are as follows. (1) A domain ontology based on the formal concept analysis (FCA) is applied to construct the topic benchmark model of rainstorm disaster. (2) A novel focused crawler called FCWSEO that incorporates the WSE and the ontology into focused crawling is proposed to collect relevant webpages. (3) A combination of the nearest and farthest candidate solution (NFCS) method and the fast non-dominated sorting is proposed to select the Pareto-optimal solutions (hyperlinks) from the unvisited hyperlink bank. (4) A comprehensive priority evaluation method considering the topic relevance of webpages where the hyperlink to be visited is located, webpage of out-hyperlink, anchor text, and PageRank (PR) value of webpage is used to assess the unvisited hyperlinks.

The structure of this article is as follows: Section 2 reviews related works. In Section 3, the construction process of ontology for the rainstorm disaster is introduced. In Section 4, the ontology-based semantic similarity calculation method is proposed. Section 5 presents the comprehensive priority assessment of hyperlinks. In Section 6, a novel focused crawler combining the WSE and ontology is proposed, and the experimental results are analyzed in Section 7. Section 8 presents the conclusions and outlines future research.

2. Related works

In this section, we survey the related works of the focused crawling techniques: classic heuristic focused crawlers, conceptual semantic analysis-based focused crawlers, and intelligent optimization algorithm-based focused crawlers.

2.1. Classic heuristic focused crawlers

Crawling methods based on hyperlink structure or webpage text are two major categories of classic heuristic focused crawlers. In hyperlink structure-based focused crawlers, the PageRank (PR) and the hyperlink-induced topic search (HITS) are two popular algorithms. The PR algorithm first proposed by Brin and Page [21] iteratively calculates the degree of importance of all webpages on the Web until stable importance of each page is achieved. Here, the degree of importance can be depicted by the similarity between a webpage and the given topic. To search for webpages with greater importance, Wang and Ji [22] proposed an improved PR algorithm based on user interest and topic (ITPR), which utilized user feedback (staying time in webpages and the number of clicks on hyperlinks) to modify the PR value appropriately, and the experimental results showed that the ITPR improved the quality of webpage ranking, accuracy, and user satisfaction. The HITS algorithm proposed by Kleinberg [23] is another typical method based on hyperlink structure. This algorithm calculates hub score and authority score for each webpage, and eventually webpages with the highest authority scores are selected as the output. Asano et al. [24] designed several types of improved HITS algorithms, which mainly consisted of three phases. First, the HITS algorithm used query terms to collect a set of start pages as the root set, and then started an iterative estimate to calculate the authority and hub weights, and finally output a list of pages with authorities and hubs. Although the HITS algorithm performs well in various crawlers, it still has limitations compared to the PR algorithm, including lower computational efficiency, easy cheater manipulation, and unstable structure. The above methods tend to search webpages through hyperlink structure, but ignore the influence of webpage text on the importance of webpages, which easily leads to the phenomenon of "topic drift".

As classic heuristic focused crawler methods based on webpage text, the fish-search algorithm [25] and shark-search algorithm [26] are typical ones. The core idea of fish-search stems from the behavior of schools of fish. When a school of fish finds food (topic-relevant webpages), it reproduces and continues to search for more food. Of course, if no food is found, the school of fish will gradually die. The fish-search algorithm has the advantage of the dynamic search but at the same time, it also has the disadvantage that discrete values cannot delineate the variation of topic relevance. Compared with the fish-search algorithm, the shark-search algorithm divided the topic relevance of anchor text, link context, and webpage text into finer granularity to calculate the priorities of hyperlinks. Chen et al. [27] proposed an improved shark-search algorithm, which derived topics from the open directory project and predicted the relevance of webpages to the specific topic based on multi-information. Focusing on the problem that the relevance judgment is not comprehensive enough in the shark-search algorithm, Cheng et al. [28] treated the content of links by the method of word embedding clustering, and effectively improved the efficiency of focused crawlers. Regarding the topic relevance of webpage text, Liu and Du [29] summarized two types of focused crawler technology: vector space model (VSM) crawler and semantic similarity retrieve model (SSRM) crawler. The VSM considers the cosine similarity between the text and the topic vector as the topic relevance of the text, and the SSRM acquires the topic relevance of the text by associating term frequency and term semantic similarity. Taking advantage of VSM and SSRM, Du et al. [30] researched the calculation method of topic similarity and proposed a semantic similarity vector space model (SSVSM) crawler. Obviously, the above researches show that priorities of unvisited hyperlinks can be obtained by analyzing the webpage text, but the influence of the hyperlink structure on the accuracy is ignored, which may cause the crawlers to only find relevant webpages within a local scope and cannot guarantee the coverage of Web crawling.

To solve the problems of topic drift and local search, some researchers designed ranking methods based on a combination of hyperlink structure and webpage text instead of using them individually. To overcome the shortcoming that the shark-search algorithm cannot consider global search, Qiu et al. [31] designed a novel shark-PageRank algorithm that utilized the authority value of the PR algorithm to compensate for this deficiency. Seyfi et al. [32] combined both link-based and content-based methods to crawl and index relevant webpages. They used specific HTML elements of webpages to predict the topic focus of the unvisited webpage and adopted the T-Graph hierarchical structure to assign an appropriate priority score to each unvisited hyperlink. To harvest high-quality topical Web resources, Zhao et al. [33] proposed an on-line topical quality estimation (OTQE) crawler, which intelligently evaluated the topical quality of unvisited webpages to prioritize their corresponding URLs by combining link-based and content-based methods. The experimental results showed that the OTQE significantly outperformed the other six different frontier prioritizing algorithms.

In conclusion, the hyperlink structure-based and webpage text-based methods all play important roles in evaluating the topic-relevance of webpages or hyperlinks, and the focused crawlers should absorb the advantages of these two methods. In this article, we propose a comprehensive scheme, which integrates link structure and Web text, to evaluate the priority of unvisited hyperlinks.

2.2. Conceptual semantic analysis-based focused crawlers

With the development of Web technology, the data of the deep Web [34], the semantic Web, extensible markup language (XML), and domain-specific markup language (DSML) are growing. How to obtain useful information depends on the method of knowledge discovery [34]. It is important to apply conceptual semantic analysis to establish a topic benchmark model in focused crawling. Recently, researches have shown that context graph (CG) and ontology can describe the topic and compute topic relevance in the semantic layer. Some scholars did a lot of researches on the CG and built the concept context graph (CCG) [8], the relevancy context graph (RCG) [7], the concept similarity context graph (CSCG) [6] the path trust knowledge graph (PTKG) [35], and knowledge graphs (KG) [36] and so forth. However, a limitation of the CG method is that the performance of CG relies on the user's query history information, which is difficult to determine its relevance to the crawling topic. On the other hand, domain ontology clarifies conceptual semantic hierarchies and relationships between concepts. For example, the YAGO [37], containing more than 1.7 million entities and 15 million facts, is a large ontology with high coverage and precision. YAGO was automatically derived from Wikipedia and WordNet. Currently, the basic methods of constructing domain ontology mainly include the latent Dirichlet allocation (LDA) and the formal concept analysis (FCA). LDA is a typical data mining model, which identifies hidden semantic topic information in corpus and defines a topic with good coverage and granularity. Based on the probability distribution relationships defined by the LDA model, Wang et al. [9] constructed a set of rules that identified the domain ontology concepts and the basic semantic relationships between related terms. Rani et al. [10] explored the construction method of domain ontology based on the MapReduce LDA and applied it to the semantic retrieval. LDA trains topic documents to mine the potential semantic relationships, but it is obvious that the training process is too complex to improve the efficiency of crawlers. Additionally, FCA is a semi-automated method of constructing domain ontology, which can represent the formal background of domain knowledge in an effective way. Zhu et al. [11] introduced the construction steps of the FCA-based domain ontology, which involved the extraction of domain concepts, the establishment of the document-term matrix, the construction of concept lattice, the analysis of concept hierarchy relationship, and the visualization of domain ontology. In the ontology hierarchy, Daoui et al. [38] considered semantic impact factors such as the concept distance and the concept density to obtain the semantic similarity between two concepts. In fact, the establishment of a high-quality topic benchmark model completes the query expansion of topic semantic description and provides a good start for focused crawling. Ontology has widely been used for focused crawling and information retrieval. Ehrig and Maedche [39] first proposed an ontology-based focused crawler in 2003 and constructed a comprehensive crawler framework with a complex ontology and associated instance elements. Bedi et al. [40] proposed a multi-threaded semantic focused crawling strategy for the topic of educational learning, which extended the topic terminology by using manually constructed domain ontology. Zhang et al. [41] proposed an ontology-based semantic retrieval scheme and applied ontology to extend semantic queries. In addition, some studies combined ontology and machine learning methods to design the learning focused crawler [42], which can learn new rules by the previous running results in their training set. Saleh et al. [43] introduced a domain distiller that combined Naïve Bayes (NB) and support vector machines (SVM) to decide whether the retrieved webpage was related to the given topic. Hassan et al. [44] combined the unsupervised ontology-learning method and web-reasoning to achieve scalable and adaptive focused crawling. Although the machine learning method can study the characteristics of webpages and compute their topic relevance, its time complexity is in quadratic time, owing to the massive training data. Thus, it is not suitable for gathering webpages in time-constraint focused crawlers. In addition, in order to make the intelligent crawler recognize the commonsense knowledge (CSK) in the corpus more accurately, Tandon et al. [45] proposed the CSK acquisition method and the use of repositories such as WebChild, CSK in natural language for addressing collocation issues based on linguistic classification as well as detecting and correcting collocation errors. Razniewski et al. [46] presented text-extraction-based, multi-modal, and transformer-based techniques to compile and consolidate the CSK. Sheng et al. [47] proposed a novel multi-document semantic extraction system, called *MuReX*, to aid users in quickly discerning salient and meaningful facts and connections in a collection of relevant documents, via graph-based visualizations of relationships between concepts even across documents.

In summary, with the maturity of semi-automatically building ontology, ontology has been successfully applied to semantic annotation, semantic query, topic classification, and so on. This article builds the domain ontology of rainstorm disaster based on FCA and realizes an ontology-based focused crawler for rainstorm disaster domain knowledge.

2.3. Intelligent optimization algorithm-based focused crawlers

The essence of the search strategy is to guide crawlers to traverse the Web. The BFS algorithm [12] and the OPS algorithm [13] are the most common heuristic search strategies for focused crawling. The BFS is a traditional graph algorithm that uses the first-in-first-out (FIFO) strategy to sequentially crawl webpages. Compared with the BFS, the OPS is a greedy algorithm that gives priorities to webpages with high importance. Regretfully, the local search of greedy algorithms is difficult to guarantee the coverage of the crawler, which may result in the

Table 1
Formal context constituted by 4 documents and 4 terms.

Documents	*Terms*			
	T1	T2	T3	T4
D1	×	×		×
D2		×	×	×
D3	×	×	×	
D4	×		×	

loss of massive relevant webpages. To further improve the global search abilities of focused crawlers, researchers introduced some intelligent optimization algorithms, such as the GA, the ITS, the SA, the PSO, and the ACO. Jing et al. [14] proposed a GA-based focused crawling search strategy, in which dynamic fitness function and genetic operator enabled crawlers to be adaptive. Yan and Pan [15] also used the GA to optimize focused crawling and improved the fitness function by using user browsing behavior. Liu et al. [16] provided a focused crawler strategy combining ontology and the ITS algorithm, which improved the crawling performance by designing a different acceptance principle. He et al. [17] presented an SA-based crawling strategy, which allowed crawlers to access sub-optimal hyperlinks for searching webpages with high importance more broadly. Tong [18] considered two hyperlink evaluation values (immediate value and future value) and proposed a heuristic focused crawler search algorithm based on the adaptive dynamical evolutional PSO to predict the importance of webpage more accurately. Aiming at the shortcoming of "near-sighted" for heuristic-based focused crawlers, Chen et al. [19] proposed a focused crawling strategy based on the ACO to improve the global search ability of crawlers. Zheng [20] proposed a focused crawler based on the GA and ACO (GAAA), which made full use of the fast, random, and global convergence of the GA and the parallelism and positive feedback of the ACO to improve the performance of focused crawlers. Liu et al. [29] proposed a crawling strategy based on the cell-like membrane computing optimization algorithm to obtain the weighted coefficients of various impact factors of topic relevance, and the optimization process was regarded as the single-objective optimization problem (SOOP).

In fact, most of the above intelligent optimization algorithm-based focused crawlers are regarded as a SOOP when evaluating priorities of unvisited hyperlinks, which have the common disadvantages: optimal weighted coefficients are difficult to determine and the crawling process is easy to fall into local optima. To avoid this issue, this article proposes a novel WSE strategy based on a multi-objective optimization model, which can optimize the selection of unvisited hyperlinks and expand the search scope of the focused crawler.

3. Construction of domain ontology

In this section, we give the construction process of domain ontology based on the formal concept analysis (FCA) for rainstorm disaster. The FCA-based ontology construction is roughly divided into three steps: formal context creation, concept lattice construction, and domain ontology construction.

(1) Formal context creation. We give the definition of the formal context and an example of creating the formal context.

Definition 1. Let a triple $F=(O, M, R)$ be a formal context, where O represents a set of objects, M represents a set of attributes, and R is a binary relation between O and M. A relation $(o, m)\in R$ is regarded as "object o has attribute m".

Example 1. In this article, the formal context is represented as $F=(Documents, Terms, I)$, where *Documents* represents a set of documents, *Terms* represents a set of terms, and I is a binary relation between *Documents* and *Terms*. If $(document, term)\in I$, then the term appears in the document. Firstly, we obtain the topic document corpus from three databases Elsevier, Springer, and CNKI (China National Knowledge Infrastructure) as the set of documents, and extract the titles, abstracts, and keywords from the corpus as a candidate set of domain terms. Then, we count the frequency of each term in the candidate set and select the top several operative terms with the highest frequency as the set of terms. In this process, the knowledge of domain experts is applied to remove some irrelevant terms. Table 1 shows the formal context constituted by 4 documents and 4 terms, where the elements on the left side are documents, the elements at the top are terms, and "×" represents the term that appears in the document.

(2) Concept lattice construction. The concept lattice contains all formal concepts of a formal context and their relationships between sub-concepts and super-concepts.

Definition 2. Let a pair (X, Y) be a formal concept of formal context $F=(O, M, R)$, where $X\subseteq O$, $Y\subseteq M$. The extension of concept (X, Y) is X and the intention is Y.

Definition 3. Let (X_1, Y_1) and (X_2, Y_2) be two formal concepts of a formal context $F=(O, M, R)$. If $X_1\subseteq X_2$(or $Y_2\subseteq Y_1$), (X_1, Y_1) is the sub-concept of (X_2, Y_2) and (X_2, Y_2) is the super-concept of (X_1, Y_1).

Example 2. Fig. 1 is the concept lattice and Hasse graph that correspond to the formal context in Table 1, where a node represents a concept described by its extension and intention, and edges represent the relationships between sub-concepts and super-concepts. The concept lattice can be visualized by the corresponding Hasse graph, which is constructed by the formal concept analysis tool called ConExp.[1]

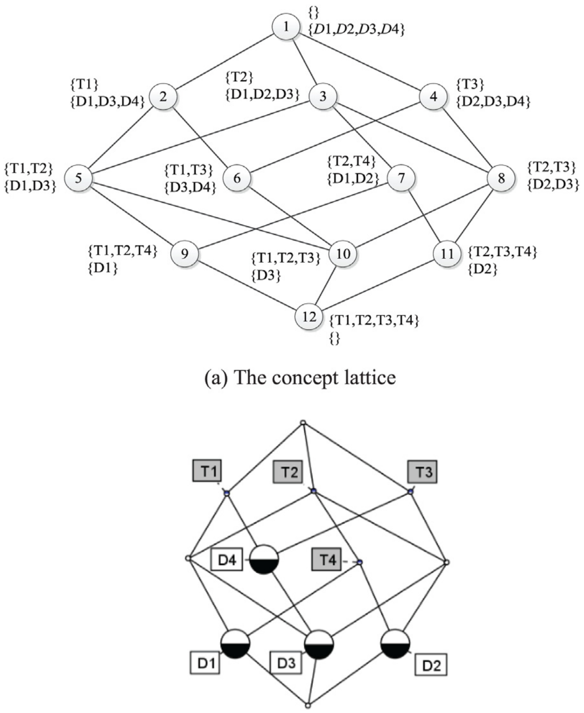
(3) Domain ontology construction. On the basis of the construction of concept lattice, we use ontology Web languages (OWL[2]) to describe the concept hierarchy. In the OWL, a term is defined as a class, and the relationship between sub-concepts and super-concepts is defined as the hypernym/hyponym relation. In general, the development tool Protégé[3] is used to visualize the OWL, and the domain ontology is obtained finally. Ontology is a formal expression of the domain concepts and their relations, and each of its nodes is a feature word. The hypernym/hyponym relationship is necessary for building a concept hierarchy in ontology, which is usually defined as three basic semantic relationships, including "*is-a*", "*part-of*", and "*attribute-of*". The "*is-a*" represents the inheritance relationship between concepts, the "*part-of*" denotes the whole and part relationship between concepts, and the "*attribute-of*" represents that a concept is an attribute of another concept.

Example 3. This article builds domain ontology (see Fig. 2(a)–(i)) about the topic of rainstorm disaster, including all related concepts, attributes, and base knowledge hierarchy. The ontology makes full use of WordNet resources [48] to modify the ontology and expand the relationships of "*synonym*", "*induce*". The "*synonym*" represents the synonymous relationship between

[1] https://sourceforge.net/projects/conexp/.

[2] https://www.w3.org/TR/owl-features/.

[3] https://protege.stanford.edu/.



(a) The concept lattice

(b) Hasse graph

Fig. 1. The concept lattice and Hasse graph that correspond to the formal context in Table 1.

concepts and the "*induce*" represents that a concept is triggered by another concept. Fig. 2 shows that the constructed domain ontology includes 71 concepts and a 7-level hierarchical structure. In order to display the structure of this ontology plainly, the main part of the ontology is displayed in Fig. 2(a), and the extendible parts (which are framed by rectangles in the figure) of the ontology are displayed in different sub-graphs in Fig. 2(b)–(i). The ontology of rainstorm disaster provides a complete semantic model that supports the exchange and sharing of information in the domain and is also of great significance for the semantic similarity calculation between feature words.

4. Ontology-based semantic similarity calculation

In this section, we give the ontology-based semantic similarity calculation (OSSC) method. The OSSC considers five impact factors (semantic distance (IF_{Dis}), concept density (IF_{Den}), concept depth (IF_{Dep}), concept coincidence degree (IF_{Coi}), and concept semantic relationship (IF_{Rel})) to compute the semantic similarity between any two concepts in the domain ontology. The concrete calculation of five impact factors except for the concept semantics relation can refer to the Ref. [49].

In the domain ontology, we consider five concept semantic relationships: "*is-a*", "*part-of*", "*attribute-of*", "*synonym*", and "*induce*". Different from the Ref. [49], the impact factor of concept semantic relationship on the semantic similarity is computed by Eq. (1).

$$IF_{Rel} = \frac{\sum_{i=1}^{L} Value\left(C_i, F_i\right)}{L}$$

$$Value\left(C_i, F_i\right) = \begin{cases} 1 & Synonym\left(C_i, F_i\right) \\ 1/2 & Induce\left(C_i, F_i\right) \\ 1/3 & Is\text{-}a\left(C_i, F_i\right) \\ 1/3 & Part\text{-}of\left(C_i, F_i\right) \\ 1/3 & Attribute\text{-}of\left(C_i, F_i\right) \end{cases} \qquad (1)$$

Table 2
Weights of different labels.

Groups	*Labels*	*Meanings*	W_j
group 1	⟨title⟩, ⟨keyword⟩, ⟨description⟩ ,⟨h1⟩	title, keyword, description, first-level title	2
group 2	⟨h2⟩ ,⟨h3⟩	second-level title, third-level title	1.5
group 3	⟨h4⟩ ,⟨h5⟩ ,⟨strong⟩	fourth-level title, fifth-level title, bold text	1.2
group 4	⟨p⟩ , ⟨td⟩ ,⟨li⟩	body information	1.0
group 5	other labels	non-body information	0.2

Here, L is the number of connected edges between two concepts C_i and F_i, where F_i is the parent-concept of concept C_i, and $Value(C_i, F_i)$ is the weight of the edge between concept C_i and its parent-concept F_i.

The semantic similarity calculation between two concepts C_1 and C_2 based on the OSSC is given as follows.

$$Sem(C_1, C_2) = \alpha \times IF_{Dis} + \beta \times IF_{Den} + \gamma \times IF_{Dep} + \delta \times IF_{Coi} + \varepsilon \times IF_{Rel} \qquad (2)$$

Here, the adjustment factors $\alpha,\ \beta,\ \gamma,\ \delta,\ \varepsilon\ \geq\ 0$, and satisfy $\alpha + \beta + \gamma + \delta + \varepsilon = 1$.

We define $FK = \{fk_1, fk_2, \ldots, fk_i, \ldots, fk_n\}$ as the vector of feature words in the ontology and $W_{FK} = \{w_{fk_1}, w_{fk_2}, \ldots, w_{fk_i}, \ldots, w_{fk_n}\}$ as the semantic weight vector of feature words, where n is the number of feature words in the domain ontology, fk_i is the ith feature word, and w_{fk_i} is the semantic weight of the ith feature word. If the topic word of the domain ontology is t (t is rainstorm disaster in this article), the calculation expression of w_{fk_i} is as follows:

$$w_{fk_i} = (Sem(t, fk_1), Sem(t, fk_2), \ldots, Sem(t, fk_i), \ldots, Sem(t, fk_n)) \qquad (3)$$

5. Comprehensive relevance assessment of hyperlinks

In this section, we introduce the calculation method of topic relevance of hyperlinks. First, we use the vector space model (VSM) to calculate the topic relevance of webpages and anchor texts. Then, an improved PageRank algorithm is introduced to compute the PR value of webpages. Finally, we design a comprehensive priority assessment method to score an unvisited hyperlink.

5.1. Topic relevance of webpages

Webpages are generally expressed as HTML (Hyper Text Markup Language) files. All Web content is written inside HTML labels. Common labels that express the topic of webpages include <title>, <keyword>, <description>, <h1>, <strong>, and so on. Feature words that appear in different labels have different influences on the topic relevance of webpages. In this article, we choose the main labels from HTML and divide them into five groups, as shown in Table 2.

To describe the vectorization of the webpage text, we first remove noise from the downloaded webpage and then calculate

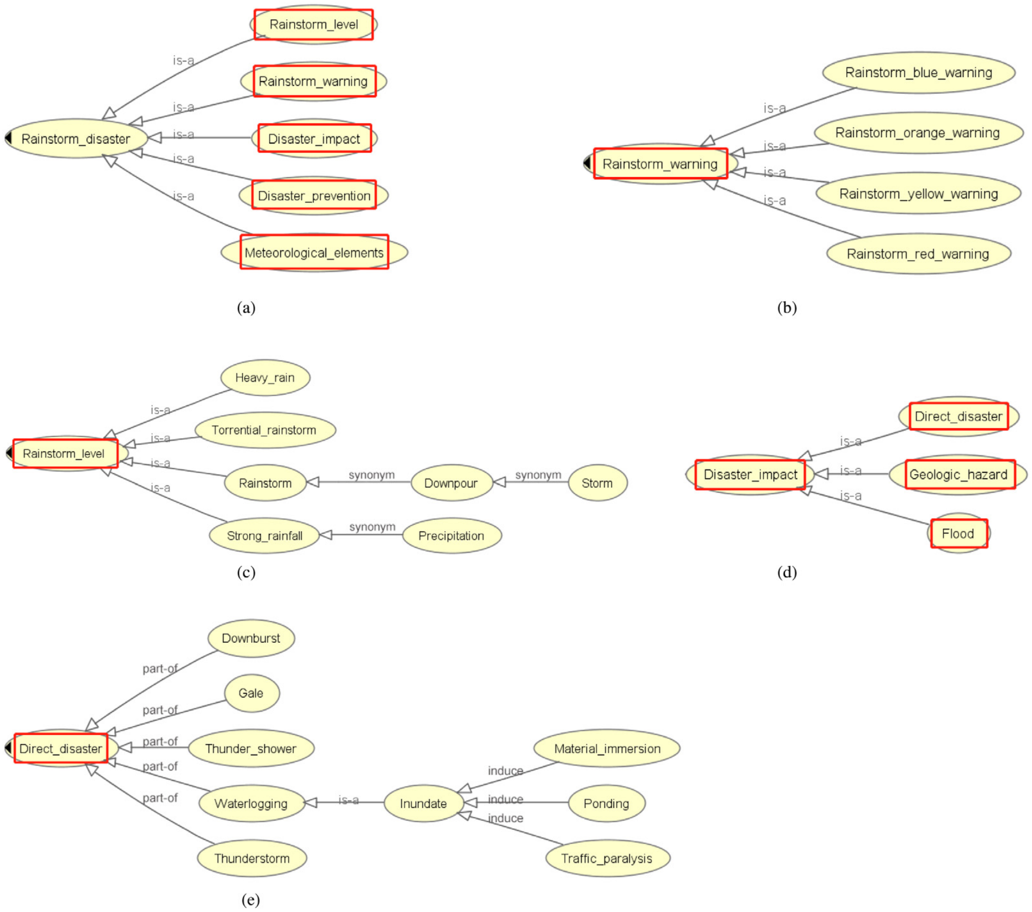


Fig. 2. A domain ontology structure about the topic of rainstorm disaster: (a) Main part of the rainstorm disaster ontology; (b) The rainstorm warning part of the rainstorm disaster ontology; (c) The rainstorm level part of the rainstorm disaster ontology; (d) The disaster impact part of the rainstorm disaster ontology; (e) The direct disaster part of the rainstorm disaster ontology; (f) The geologic hazard part of the rainstorm disaster ontology; (g) The flood part of the rainstorm disaster ontology; (h) The disaster prevention part of the rainstorm disaster ontology; (i) The meteorological elements part of the rainstorm disaster ontology.

the term frequency of feature words. Suppose $DK = \{dk_1, dk_2, \ldots, dk_i, \ldots, dk_n\}$ is the term frequency vector of feature words in the webpage text, where dk_i is the term frequency of the ith feature word and n is the number of feature words. Considering the different label weights, DK can be expressed as $((dk_{1,1}, dk_{1,2}, \ldots, dk_{1,\mathrm{l}}), (dk_{2,1}, dk_{2,2}, \ldots, dk_{2,\mathrm{l}}), \ldots, (dk_{i,1}, dk_{i,2}, \ldots, dk_{i,j}, \ldots, dk_{i,l}), \ldots, (dk_{n,1}, dk_{n,2}, \ldots, dk_{n,l}))$, where l is the number of label groups in the webpage, $dk_{i,j}$ is the term frequency of the ith feature word in the jth group, $i \in \{1, 2, \ldots, n\}$ and $j \in \{1, 2, \ldots, l\}$. The weight w_{dk_i} of the ith feature word in the webpage text is computed as follows:

$$w_{dk_i} = \sum_{j=1}^{l} \frac{dk_{i,j}}{\max dk_{i,j}} \times W_j \tag{4}$$

Here, *max dk* $_{i,j}$ is the maximum term frequency of the ith feature word in all label groups, and W_j is the weight of the jth label group.

In this article, we compute the topic relevance of various texts (webpage and anchor text) by the VSM. The topic relevance of texts is acquired by calculating the inner product between text vector and topic vector. The topic relevance $R(p)$ of the webpage p is calculated by Eq. (5).

$$R(p) = Sim(DK, FK) = \frac{W_{DK} \times W_{FK}}{\|W_{DK}\| \times \|W_{FK}\|} = \frac{\sum_{i=1}^{n} \left(w_{dk_i} \times w_{fk_i} \right)}{\sqrt{\sum_{i=1}^{n} w_{dk_i}^2} \times \sqrt{\sum_{i=1}^{n} w_{fk_i}^2}} \tag{5}$$

Here, $W_{DK} = \{w_{dk_1}, w_{dk_2}, \ldots, w_{dk_i}, \ldots, w_{dk_n}\}$ is the feature weight vector of webpage p, and $W_{FK} = \{w_{fk_1}, w_{fk_2}, \ldots, w_{fk_i}, \ldots, w_{fk_n}\}$ is the semantic weight vector of feature words in the domain ontology.

5.2. Topic relevance of anchor texts

The anchor text usually contains few words or phrases, but it is an important basis for predicting the topic relevance of the webpage to which the hyperlink points. According to the idea of term frequency-inverse document frequency (TF-IDF), if a feature word appears frequently in an anchor text and rarely appears in other texts, the feature word should be important and has a good classification ability. The weight w_{ak_i} of the ith feature word in the anchor text is computed as follows:

$$w_{ak_i} = \frac{g_i}{\sum_{j=1}^{n} g_j} \times \log_c \left(\frac{N}{N_i} + 0.01 \right) \tag{6}$$

Here, g_i is the term frequency of the ith feature word in the anchor text; N is the number of retrieved webpages; N_i is the number of retrieved webpages that contain the ith feature word, $i \in \{1, 2, \ldots, n\}$, and c is a positive constant. Suppose $AK = \{ak_1, ak_2, \ldots, ak_i, \ldots, ak_n\}$ is the feature vector of the anchor text of a hyperlink. By considering the inner product between FK and AK, the topic relevance $R(a_l)$ of anchor text a_l is computed by:

$$R(a_l) = Sim(AK, FK) = \frac{W_{AK} \times W_{FK}}{\|W_{AK}\| \times \|W_{FK}\|}$$

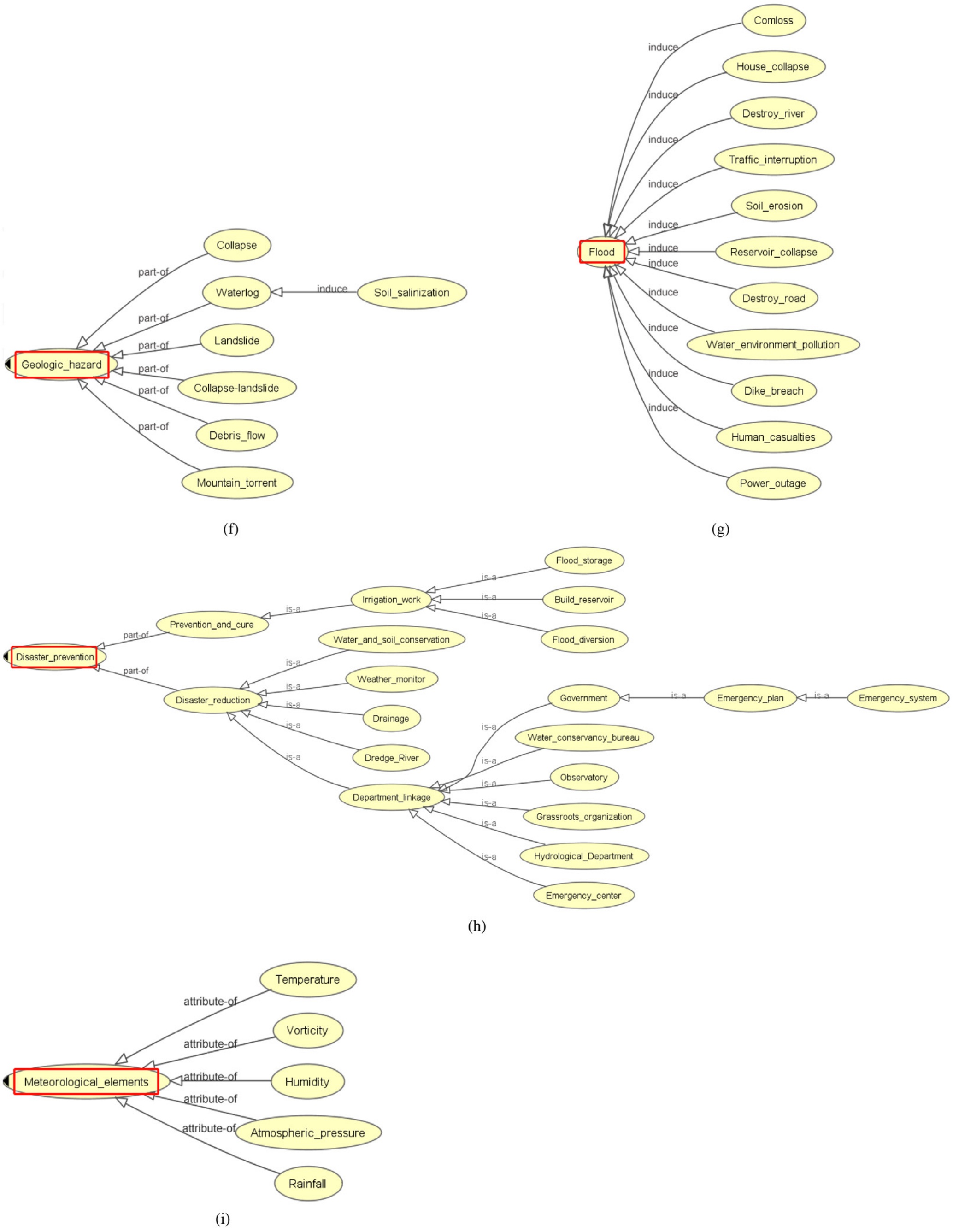


Fig. 2. (*continued*).

$$= \frac{\sum_{i=1}^{n} \left(w_{ak_i} \times w_{fk_i} \right)}{\sqrt{\sum_{i=1}^{n} w_{ak_i}^2} \times \sqrt{\sum_{i=1}^{n} w_{fk_i}^2}} \tag{7}$$

Here, $W_{AK} = \{w_{ak_1}, w_{ak_2}, \ldots, w_{ak_i}, \ldots, w_{ak_n}\}$ and $W_{FK} = \{w_{fk_1}, w_{fk_2}, \ldots, w_{fk_i}, \ldots, w_{fk_n}\}$ are the feature weight vector of anchor text a_l and the semantic weight vector of feature words in the domain ontology, respectively.

5.3. Improved PageRank value computation

For a webpage p, the traditional PageRank (PR) value is computed by Eq. (8).

$$PR(p) = (1-d) + d \times \sum_{i=1}^{h} \frac{PR(p_i)}{C(p_i)} \tag{8}$$

Here, h is the number of in-hyperlinks of the webpage p in the retrieved webpage set, p_i is the ith in-hyperlink webpage of the webpage p, $C(p_i)$ is the number of out-hyperlinks of the webpage p_i, and $d = 0.85$ is the damping coefficient.

In Eq. (8), the traditional PR value computation only analyzes the hyperlink structure of the webpage and does not judge whether the hyperlink in the webpage is topic-relevant, which may lead to the crawler catching a large number of irrelevant webpages, i.e. cause topic drift. To overcome this flaw, the topic relevance of anchor text is introduced into the calculation of the PR value [49]. The improved PR value computation (see Eq. (9)) can increase the importance (PR value) of topic-relevant webpages.

$$PR(p) = (1-d) + d \times \sum_{i=1}^{h} \left[\frac{PR(p_i)}{C(p_i)} \times (1 + \omega \times R(a_l)) \right] \tag{9}$$

Here, ω is an adjustment factor, l is the hyperlink of the webpage p_i and points to webpage p, a_l is the anchor text corresponding to hyperlink l, and $R(a_l)$ is the topic relevance of anchor text of hyperlink l.

5.4. Topic relevance evaluation of hyperlinks

In order to evaluate the topic relevance of unvisited hyperlink l, a comprehensive priority assessment method is given as follows:

$$E(l) = \lambda_1 \times R(p_l) + \lambda_2 \times \left(\frac{1}{h}\sum_{i=1}^{h} R(p_i)\right) + \lambda_3 \times R(a_l) + \lambda_4 \times PR(p_l) \quad (10)$$

Here, $E(l)$ is the comprehensive priority of the unvisited hyperlink l. $R(p_l)$ is the topic relevance of webpage p_l, where p_l is the webpage to which hyperlink l points. $R(p_i)$ is the topic relevance of webpage p_i, where p_i is the ith webpage that contains hyperlink l, and h is the number of webpages that contain hyperlink l. $R(a_l)$ is the topic relevance of anchor text of hyperlink l. $PR(p_l)$ is the PR value of webpage p_l. λ_1, λ_2, λ_3, and λ_4 are four weight factors and satisfy $\lambda_1+\lambda_2+\lambda_3+\lambda_4 = 1$. In this article, the weights of these four evaluating factors are the results of many experiments based on the grid search method. Their values are $\lambda_1 = 0.3$, $\lambda_2 = 0.15$, $\lambda_3 = 0.35$, and $\lambda_4 = 0.2$, respectively.

6. Focused crawling strategy combining WSE and ontology

Because it is hard for the traditional weighted sum approach (see Eq. (10)) to determine the proper weighted coefficients, we introduce four objective functions to evaluate the unvisited hyperlinks and design the nearest and farthest candidate solution (NFCS) method combining with the fast non-dominated sorting method to pick out optimal hyperlinks. For the constructed multi-objective optimization model based on Web text and link structure, we design a framework called WSE to guide the search of focused crawlers. Finally, the specific process of focused crawler FCWSEO combining WSE and ontology is introduced.

6.1. Objective functions of evaluating hyperlinks

The topic relevance of a hyperlink l mainly depends on the content of the webpages related to the hyperlink and the link structure. Based on the above analysis, in this article, we use the topic relevance of the webpage to which hyperlink l points, the average topic relevance of the webpages where hyperlink l is located, the topic relevance of the anchor text, and the PageRank (PR) value of the webpage to which the hyperlink points as four objective functions, and establish a multi-objective optimization model for evaluating unvisited hyperlinks, as shown in Eqs. (11)–(14). When the evaluation of the hyperlinks involves more than one objective, which is conflicting, the selection of the hyperlinks is treated as a multi-objective optimization problem. The formal descriptions of the four objective functions are as follows:

$$\max F_1(l) = R(p_l) \quad (11)$$

$$\max F_2(l) = \frac{1}{h}\sum_{i=1}^{h} R(p_i) \quad (12)$$

$$\max F_3(l) = R(a_l) \quad (13)$$

$$\max F_4(l) = PR(p_l) \quad (14)$$

Here, $F_1(l)$ represents the topic relevance of the webpage p_l to which hyperlink l points, $F_2(l)$ the average topic relevance of webpages that contain hyperlink l, $F_3(l)$ the topic relevance of anchor text a_l of hyperlink l, and $F_4(l)$ the PR value of webpage P_l to which hyperlink l points. In Eq. (12), h is the number of webpages containing hyperlink l.

6.2. Selection of optimal hyperlinks

It is desired that an algorithm that solves a multi-objective optimization problem (MOP) maintains a good spread of the non-dominated solutions. In the fast non-dominated sorting multi-objective genetic algorithm (NSGA-II) [50], the crowded degree comparison (CDC) method is used to keep the spread of solutions. Thereafter, the CDC method has been widely used for diversity maintenance of solutions in many multi-objective evolutionary algorithms [51,52]. However, it has an obvious flaw that the solutions in the high-density space have less chance to be selected so the spread of solutions is not good enough. In this article, we replace the CDC method with a novel method, called the nearest and farthest candidate solution (NFCS) method, which can improve the uniformity of the Pareto front and maintain the diversity of the obtained solutions.

In the NFCS, for two solutions (hyperlinks) l_s and l_t, we define a distance calculation method based on objective functions. The distance between two solutions l_s and l_t is calculated as follows:

$$Dis\,(l_s,\ l_t) = \sqrt{\sum_{i=1}^{m} \left(F_i\,(l_s) - F_i\,(l_t)\right)^2} \quad (15)$$

Here, m is the number of objective functions, $F_i(l_s)$ and

$F_i(l_t)$ are the ith objective function value of l_s and l_t, respectively. Suppose that we need to select q optimal solutions OS from the candidate solution set CS which contains r ($\geq$q) non-dominated solutions. The process is as follows:

Let $OS = \varnothing$. Calculate all objective function values $F_i(l_j)$ ($i = 1, 2, \ldots, m$) of each solution l_j ($j = 1, 2, \ldots, r$) in the candidate solution set CS. Find solutions with the largest objective function values based on the preferences of different objectives and put them into the optimal solution set MS. Considering the following two cases:

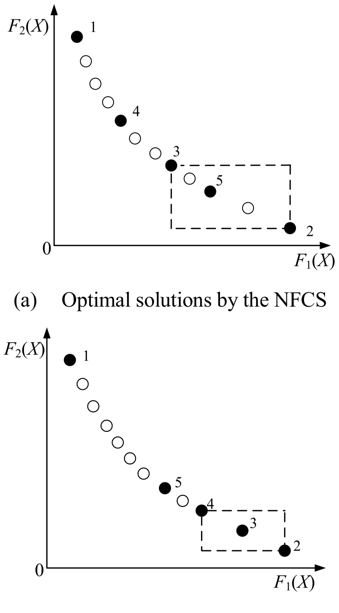
(1) If $q \leq m$, select q solutions randomly from the set MS and put them into the optimal solution set OS.

(2) If $q > m$, put all solutions of MS into the optimal solution set OS, and delete them from the candidate solution set CS. The rest of the solutions (the number of which is q-m) can be selected as follows: for each solution l_c in the updated candidate solution set CS, we calculate the nearest distance (see Eq. (15)) between l_c and all optimal solutions in OS. Select the solution l_f with the farthest distance from CS and put it into the optimal solution set OS, and delete it from candidate solution set CS. Repeat the above operations, until the number of solutions in set OS reaches q.

Fig. 3 shows an optimal solution set consisting of five Pareto-optimal solutions selected from twelve Pareto-optimal solutions, where two objectives F_1 and F_2 are considered. The solid circles in Fig. 3(a) are the optimal solution set obtained by the NFCS method, and the solid circles in Fig. 3(b) are the optimal solution set obtained by the CDC in NSGA-II [50]. The numbers in the figures indicate the order of selecting solutions. It is not hard to see that the NFCS method can obtain a more uniform Pareto front than the traditional CDC in NSGA-II, and the Pareto-optimal solutions obtained are also more diverse.

6.3. Focused crawling strategy combining web space evolution and ontology

To measure the distance between any two hyperlinks obtained in the process of focused crawling, we suppose there is a virtual URL (Uniform Resource Locator) which is linked to all seed URLs for focused crawling. As is shown in Fig. 4, a connected graph of hyperlinks is constructed, where the virtual URL is used to connect all seed URLs which combines all their child URLs (webpages) by hyperlinks. Let $G = (V, E)$ be a connected graph, where



(a) Optimal solutions by the NFCS

(b) Optimal solutions by the CDC

Fig. 3. Five optimal solutions selected by different methods for 12 Pareto-optimal solutions.

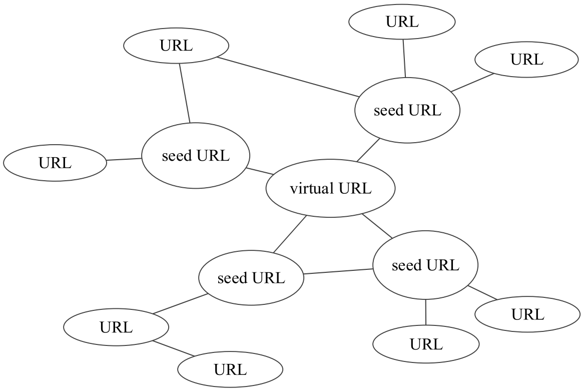


Fig. 4. A connected graph of hyperlinks with 13 vertices.

the set V of vertices is composed of a set of URLs (webpages) and the set E of edges is composed of a set of hyperlinks between webpages.

Definition 4. The distance between two URLs in the connected graph of hyperlinks is measured by computing the number of edges of the shortest path between these two URLs, and the distance between two URLs corresponding to two hyperlinks is called the distance between these two hyperlinks.

At the beginning of the Web space evolution (WSE) framework, k topic-relevant initial hyperlinks (called the seed hyperlink bank S_{seed}) are produced. Taking every hyperlink in S_{seed} as the center, we construct a circular region with the radius $d_{space} = d_{ave}/2$, where d_{ave} equals the average distance among k hyperlinks. Based on the comprehensive priority assessment method, we select all unvisited child-links with scores above preset the threshold, located in webpages that seed hyperlinks point to, and put them into a bank (called the retrieved hyperlink bank $S_{retrieve}$). Next, we use the fast non-dominated sorting method [50] and the NFCS method (see Section 6.2) to pick out 2×k optimal hyperlinks (called the evolutionary hyperlink bank S_{evolve}) from the retrieved hyperlink bank $S_{retrieve}$. Subsequently, we select k hyperlinks (called the test hyperlink bank S_{test}) with the larger topic relevance of the anchor text from S_{evolve}. Finally, we apply these k hyperlinks to update the seed hyperlink bank S_{seed} (see below for details). Once the seed hyperlink bank S_{seed} completes the update, we let $d_{space} = d_{space}\times 2$ expand the search space of the crawler. The updating mechanism of the seed hyperlink database S_{seed} is described as follows (see Fig. 5).

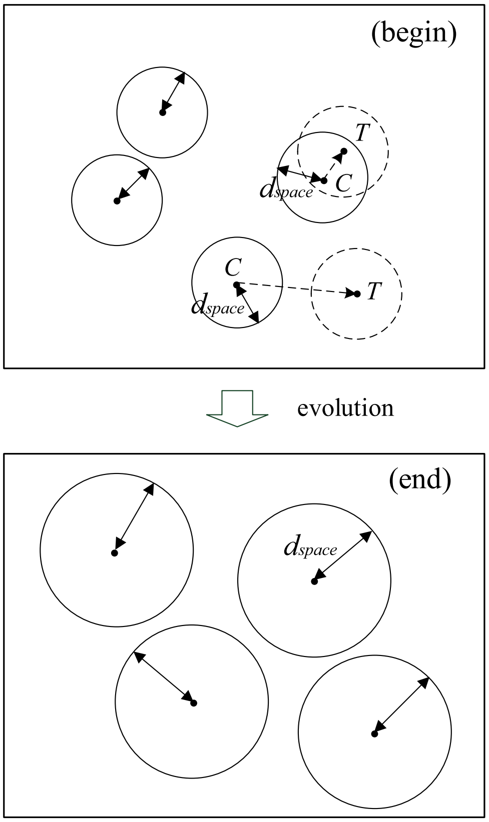


Fig. 5. Updating mechanism of the seed hyperlink database and procedure of Web space evolution.

For each hyperlink T in the test hyperlink bank S_{test}, we calculate the distance between it and every seed hyperlink in S_{seed}. Pick out the hyperlink with the nearest distance from S_{seed}, named by C. The distance between T and C is denoted by $D(T, C)$. Two cases are considered:

(1) If $D(T, C) > d_{space}$, test hyperlink T is far away from all hyperlinks in S_{seed}. We perform the fast non-dominated sorting [50] to all hyperlinks in S_{seed} and remove hyperlink Y that is located in the lowest layer. Meanwhile, we add T into seed hyperlink bank S_{seed}, and transfer the center of the circle from Y to T.

(2) If $D(T, C) \leq d_{space}$, test hyperlink T is contained in the circular region of C, which means these two hyperlinks are similar. Three cases are considered: if T dominates C, we add T into S_{seed}, delete C from S_{seed}, and transfer the center of the circle from C to T; if C dominates T, we abandon C with probability P_t and once C is deleted, then transfer the center of the circle from C to T; if T and C do not dominate each other, we randomly select one from T and C to update S_{seed}, and once C is deleted, we transfer the center of the circle from C to T.

Obviously, case (1) happens when the circle is small and case (2) is more likely to happen when the circle is large. In the early stage of focused crawling, a smaller value of d_{space} is easy to produce more diversity of hyperlinks. With the expansion of value of d_{space}, the crawler will gradually expand the search scope in Web space and shift the focus of crawling to fetch sub-optimal hyperlinks under different hosts. Therefore, to search efficiently for the webpages, it is necessary to maintain the diversity of hyperlinks in the early stages and then gradually shift the crawling emphasis toward sub-optimal hyperlinks, which is realized in the WSE by slowly increasing the value of d_{space}. Obviously, in the WSE, the number of hyperlinks in S_{seed} remains unchanged, which is k always.

By incorporating the WSE and the domain ontology into the focused crawling, a novel focused crawler strategy called FCWSEO is proposed. In the FCWSEO, we set two thresholds τ and η ($> \tau$). In the process of crawling, for each extracted child-link, if its comprehensive priority (see Eq. (10)) exceeds the threshold τ ($0<\tau<1$), this child-link will be downloaded and put into the retrieved hyperlink database $S_{retrieve}$. If the topic relevance of webpage to which the child-link points is bigger than threshold η, the webpage is considered to be topic relevant and is saved. The specific process of the FCWSEO is shown in Algorithm1.

Algorithm 1. FCWSEO

```
    Input: Seed hyperlinks
    Output: Downloaded webpages
1:  Determine the topic and construct domain ontology (see Section 3);
2:  Compute the semantic weight vector $W_{FK}$ of topic feature words based on the constructed domain ontology (see Section 4);
3:  Produce $k$ topic-relevant initial hyperlinks and add them into seed hyperlink bank $S_{seed}$. Set thresholds $\tau$ and $\eta$. Let $M = 0$ and $N = 0$;
    //N is the number of downloaded webpages, and M is the number of downloaded topic-relevant webpages
4:  Take every hyperlink in $S_{seed}$ as the center and construct a circular region with the radius $d_{space}$;
5:  Extract all the child-links which are located in webpages to which seed hyperlinks point and remove duplicate links;
6:  For $i = 1$ to $q$ do      // q is the number of child-links
        Calculate the comprehensive priority of child-link_i based on Eq.(10);
        If $E(child\text{-}link_i) > \tau$ then
            Download child-link_i and put it into retrieved hyperlink bank $S_{retrieve}$. Let $N=N+1$;
            Calculate $R(current\text{-}page)$ according to Eq. (5);
            //current-page is the webpage to which the child-link_i points
            If $R(current\text{-}page) > \eta$ then
                Save the current-page and let $M=M+1$;
            End if
            If $N \geq 15{,}000$ then
                The algorithm ends;
            Else
                continue;
            End if
        End if
    End for
7:  Pick out 2*k optimal unmarked hyperlinks from $S_{retrieve}$ by using the fast non-dominated sorting method and the NFCS method (see Section 6.2) and put them into evolutionary hyperlink bank $S_{evolve}$, where the marking operation is used to avoid repeated operations of hyperlinks;
8:  Pick out $k$ hyperlinks with the larger topic relevance of the anchor text (see Eq.(13)) from $S_{evolve}$ and put them into test hyperlink bank $S_{test}$. Mark chosen hyperlinks in $S_{retrieve}$ and empty $S_{evolve}$;
9:  Apply these $k$ hyperlinks in $S_{test}$ to update the seed hyperlink bank $S_{seed}$ based on the above-mentioned updating mechanism, and empty $S_{test}$;
10: Let $d_{space} = d_{space} \times 2$ and go to step 5.
```

7. Experimental results and discussion

To test the performance of the proposed focused crawler FCWSEO, we develop the crawling system for rainstorm disaster domain knowledge and compare the crawling results of the FCWSEO with those of other advanced algorithms, including the breadth-first search (BFS) [12], the optimal priority search (OPS) [13], the focused crawler based on simulated annealing (FCSA) [53], the improved tabu search algorithm combined ontology (On-ITS) [16] and the multi-objective ant colony optimization algorithm combined ontology (MOACO) [54]. All crawler algorithms are compiled in Java language and run on a PC with 3.2 GHz CPU and 8.0 GB RAM.

7.1. Experimental setup and evaluation indices

The required information involved in running the experiments includes the initial seed hyperlinks and topic document corpus. The initial seed hyperlinks are acquired from Baidu, which is the most powerful Chinese search engine. We search a certain number of webpages through this site using the rainstorm disaster as a keyword. We select 30 top-ranked webpages from the resultant webpages as the initial seed hyperlinks of all six crawler strategies. In addition, the corpus is drawn from three databases, including Elsevier, Springer, and CNKI. The corpus is the source of the terms in the domain ontology, which determines the accuracy of the ontology describing the topic. The topic documents are collected from the 50 top-ranked documents of each database by submitting the query word (i.e. rainstorm disaster), covering Journals, Doctoral Thesis, Masters' Thesis, Conferences and Books, and a total of 150 documents are collected finally.

The performance of the focused crawlers can be generally evaluated by the recall rate (RC) and harvest rate (HR). RC is equal to the ratio of the number of retrieved relevant webpages over the total number of all relevant webpages on the Web, which measures how well the crawler doing at finding all relevant webpages. HR is equal to the ratio of the number of retrieved relevant webpages over the total number of retrieved webpages, which measures how well the crawler doing at rejecting irrelevant webpages. Since it is difficult to count the total number of relevant webpages on the Web, this article does not use RC to evaluate the performance of crawler algorithms and choose HR as an evaluation index. We also use the average topic relevance (AR) and the standard deviation (SD) of retrieved webpages to analyze the results of different crawler algorithms. The three evaluation indices in this article are expressed as follows:

$$HR = \frac{M}{N} \tag{16}$$

$$AR = \frac{1}{N}\sum_{i=1}^{N} R(p_i) \tag{17}$$

$$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(R(p_i) - AR\right)^2} \tag{18}$$

Here, HR is the harvest rate, M is the number of retrieved relevant webpages, and N is the total number of retrieved webpages. AR is the average topic relevance of all retrieved webpages, and $R(p_i)$ is the topic relevance of the webpage p_i. SD is the standard deviation of all retrieved webpages compared to AR and is used to measure the spread of the topic relevance of all retrieved webpages. The value of SD is in [0, 1]. The lower the value of SD is, the better the stability of the algorithm is.

7.2. Experimental results for different crawlers

The experimental results are analyzed to evaluate the validity, superiority, and adaptability of each focused crawler algorithm. In this article, the same FK is used to evaluate the relevance of webpages, and the predefined topic relevance threshold η is set to 0.7 [29], which measures whether the retrieved webpage is a relevant page. Table 3 shows the results of M, HR, AR, SD, and $Time$ (h) of the six crawlers in a complete experiment on rainstorm disaster theme when the retrieved webpages gradually increase to 15,000.

It is not hard to see from Table 3 that the FCSA and the OPS in the early stages, the On-ITS in the middle stages, and the FCWSEO in the late stages of crawling obtain the optimal values of M and HR, respectively. Although the MOACO obtains the optimal result of AR in the early stage of crawling, it is exceeded by the FCWSEO in the later stage. The FCWSEO has the optimal SD value in the early stage of crawling and is slightly inferior to the MOACO in the late stage of crawling. In fact, we find that the ability of the FCWSEO to retrieve relevant webpages is not obvious in the early stage of crawling. However, the FCWSEO can find more relevant webpages as the number of retrieved webpages increases. As a result, when the number of retrieved webpages reaches 15,000, the results by the FCWSEO have been further improved. This is because the crawler uses the WSE framework to gradually expand the search scope, which is realized by slowly increasing the value of d_{space}. In addition, its stability (SD) which is second only to the MOACO is also competitive. About the retrieval time, we can find that the BFS has the shortest retrieval time, while the running time of the FCWSEO is slightly longer than other several algorithms. This is because the BFS directly crawls child hyperlinks of seed webpages and does not sort them according to their relevance, while the FCWSEO has the increased cost of ontology construction and relevance calculation. From Table 3, it is not hard to find that on the whole, the FCWSEO crawler overmatches the other five crawler algorithms.

Results by six different crawling algorithms for M, HR, AR, and SD are shown graphically in Figs. 6–9. It is not hard to see from these figures that when the number of downloaded webpages exceeds 9000, the performance indicators of each crawler tend to be stable. In order to better reflect the trend of each crawler algorithm, this article sets the number of retrieved webpages as 15,000.

Fig. 6 shows a comparison of the number of retrieved relevant webpages (M) for six tested crawler algorithms. Obviously, the number of retrieved relevant webpages grows rapidly as the number of retrieved webpages increases for the OPS, the FCSA, the On-ITS, the MOACO, and the FCWSEO, but the number of retrieved relevant webpages slowly increases for the BFS. When the number of retrieved webpages is greater than 10,000, the number of the topic-relevant webpages obtained by FCWSEO is more than that of the other five focused crawler methods. When the total number of downloaded webpages is 15,000, the FCWSEO outperforms the other five algorithms, and obtain 12,162 topic-relevant webpages finally. Fig. 6 indicates that the FCWSEO has the ability to collect more absolute quantities of relevant webpages than the other five crawlers.

Fig. 7 shows a comparison of the harvest rate (HR) for the six crawlers. When the number of downloaded webpages reaches 10,000, the HR value of FCWSEO tends to be stable, and its harvest rate becomes higher than the other five crawlers. When the total number of downloaded webpages is 15,000, the harvest rate of the BFS, the OPS, the FCSA, the On-ITS, the MOACO, and the FCWSEO are 6.6%, 44.3%, 70.6%, 78%, 74.2%, and 81.1%, respectively. These values show that the harvest rate of the FCWSEO is 12.3, 1.83, 1.15, 1.04, and 1.09 times as large as that of the other five algorithms, respectively.

From Figs. 6 and 7, it is not hard to find that the harvest rate of BFS is always low, this is because it does not consider the

Table 3
Comparison of experimental results by the six crawlers for M, HR, AR, SD, and Time (h) on rainstorm disaster theme.

N	*Crawlers*	*M*	*HR*	*AR*	*SD*	*Time/h*
1000	BFS	184	0.184	0.412	0.266	–
	OPS	744	0.744	0.601	0.226	–
	FCSA	**796**	**0.796**	0.710	0.183	–
	On-ITS	673	0.673	0.688	0.509	–
	MOACO	619	0.619	**0.717**	0.282	–
	FCWSEO	434	0.434	0.715	**0.166**	–
3000	BFS	666	0.222	0.379	0.286	–
	OPS	**2349**	**0.783**	0.641	0.186	–
	FCSA	2256	0.752	0.682	0.171	–
	On-ITS	2182	0.727	0.729	0.536	–
	MOACO	2033	0.678	**0.752**	0.216	–
	FCWSEO	1644	0.548	0.730	**0.161**	–
5000	BFS	720	0.144	0.286	0.256	–
	OPS	3950	0.790	0.674	**0.149**	–
	FCSA	**4130**	**0.826**	0.707	0.156	–
	On-ITS	3897	0.779	0.731	0.441	–
	MOACO	3473	0.695	**0.770**	0.185	–
	FCWSEO	3137	0.627	0.749	0.179	–
7000	BFS	721	0.103	0.270	0.254	–
	OPS	4718	0.674	0.606	0.210	–
	FCSA	5738	0.820	0.699	0.172	–
	On-ITS	**5815**	**0.831**	0.714	0.361	–
	MOACO	5039	0.720	**0.780**	**0.170**	–
	FCWSEO	4642	0.667	0.765	0.172	–
9000	BFS	855	0.095	0.309	0.279	–
	OPS	5157	0.573	0.583	0.207	–
	FCSA	6241	0.693	0.633	0.203	–
	On-ITS	**7403**	**0.823**	0.725	0.315	–
	MOACO	6528	0.725	0.775	**0.160**	–
	FCWSEO	6870	0.763	**0.784**	0.165	–
11000	BFS	946	0.086	0.278	0.265	–
	OPS	5662	0.515	0.576	0.210	–
	FCSA	7715	0.701	0.648	0.199	–
	On-ITS	8753	0.796	0.717	0.320	–
	MOACO	7880	0.716	0.774	**0.150**	–
13000	FCWSEO	**8841**	**0.804**	**0.791**	0.160	–
	BFS	988	0.076	0.255	0.253	–
	OPS	6240	0.480	0.563	0.204	–
	FCSA	8840	0.680	0.646	0.197	–
	On-ITS	9989	0.768	0.723	0.359	–
	MOACO	9514	0.732	0.777	**0.143**	–
	FCWSEO	**10473**	**0.806**	**0.815**	0.159	–
15000	BFS	990	0.066	0.226	0.255	**7.56**
	OPS	6645	0.443	0.563	0.202	9.17
	FCSA	10596	0.706	0.663	0.195	10.95
	On-ITS	11705	0.780	0.721	0.385	9.21
	MOACO	11126	0.742	0.778	**0.137**	11.43
	FCWSEO	**12162**	**0.811**	**0.822**	0.157	11.64

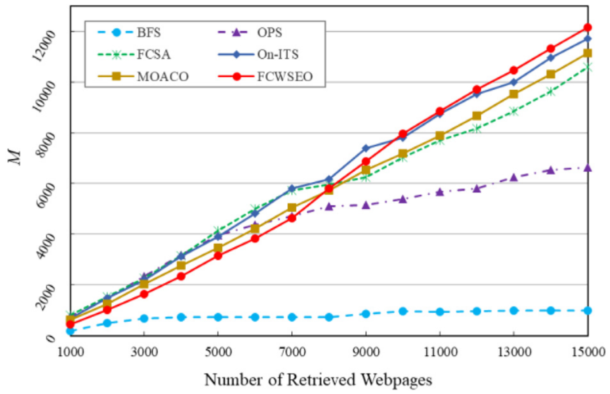


Fig. 6. Comparison of the number (M) of retrieved relevant webpages for the six crawlers on rainstorm disaster theme.

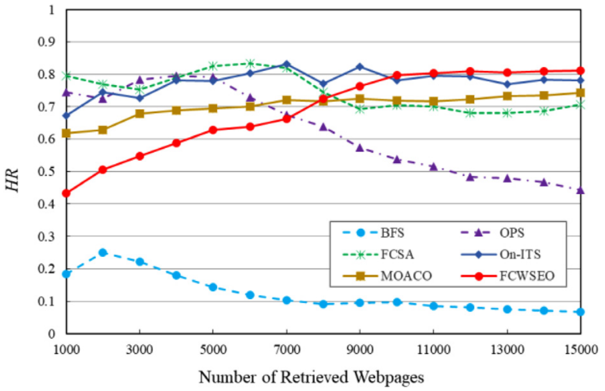


Fig. 7. Comparison of the harvest rate (HR) for the six crawlers on rainstorm disaster theme.

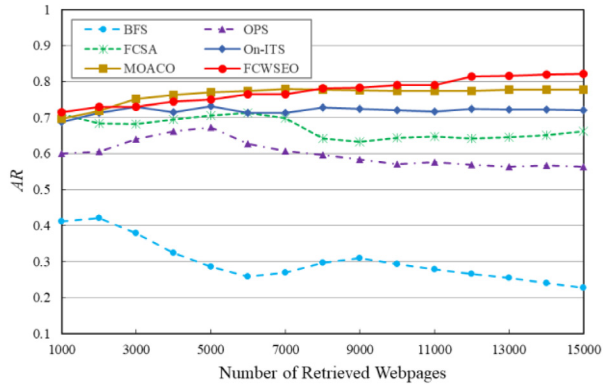


Fig. 8. Comparison of the average topic relevance (AR) for the six crawlers on rainstorm disaster theme.

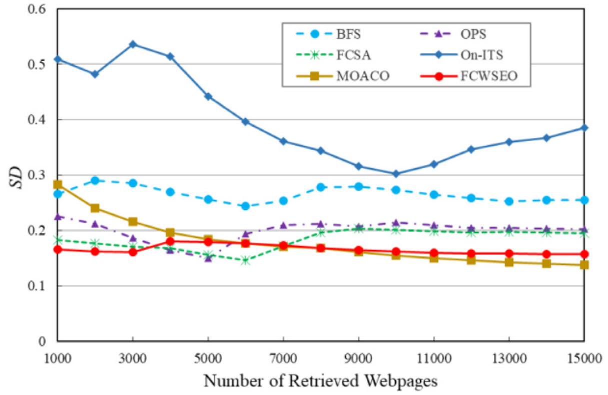


Fig. 9. Comparison of the standard deviation (SD) for the six crawlers on rainstorm disaster theme.

topic relevance of webpages. In addition, we can find that the OPS method has a good effect in the initial stage of webpage search on rainstorm disaster, but starts to diverge when the number of downloaded webpages increases. This is due to the OPS always gives the priority to the hyperlinks with the highest topic relevance in the process of crawling. The greedy strategy of the OPS makes it obtain larger M and HR in the early stage but gradually falls into local optima of the search as the crawler continues. Compared with the OPS, the FCSA can search more topic-relevant webpages in the later stage of crawler, this is because it accepts some sub-optimal hyperlinks with a certain probability during the search process, which can make it jump out of the local optima of the search. However, its main flaw is that parameters such as the initial temperature and the annealing speed are difficult to control. The HR value of the FCWSEO shows an upward trend in the process of searching webpages, and surpasses the On-ITS algorithm and the MOACO algorithm to obtain the highest harvest rate when the number of downloaded webpages reaches 8000 and 10,000, respectively. The On-ITS maintains a high harvest rate in the early stage, because it sets taboo object and aspiration criteria to avoid crawling visited hyperlinks, which avoid the detour search to a certain extent. However, due to the acceptance strategy does not give the hyperlink with the highest priority enough opportunity to grab its child hyperlinks, the global search cannot be performed in the later stage.

Fig. 8 shows a comparison of the average topic relevance (AR) for the six crawlers. In the whole process of Web search, the AR value curve of the FCWSEO shows a slow upward trend. When the number of downloaded webpages reaches 8000, the average topic relevance of the FCWSEO is higher than the other five crawlers. When the total number of downloaded webpages is 15,000, the average topic relevance of the BFS, the OPS, the FCSA, the On-ITS, the MOACO, and the FCWSEO are 0.226, 0.563, 0.663, 0.721, 0.778, and 0.822, respectively. These values show that the average topic relevance of the FCWSEO is increased by 46%, 24%, 14%, 5.7% compared with that of the OPS, the FCSA, the On-ITS, and the MOACO, respectively. Notably, the average topic relevance of the FCWSEO is 3.64 times as large as that of the BFS. As a whole, Fig. 8 indicates that the FCWSEO can collect higher quality webpages compared with the other five crawlers.

Fig. 9 shows a comparison of the standard deviation (SD) for the six crawlers. When the total number of downloaded webpages is 15,000, the SD of the BFS, the OPS, the FCSA, the On-ITS, the MOACO, and the FCWSEO are 0.255, 0.202, 0.195, 0.385, 0.137, 0.157, respectively. We can find that the SD of the On-ITS is about 0.5 in the early stage. As the number of crawled webpages increases, it decreases to 0.385 in the later stage, but it is higher than the other five algorithms as a whole. Throughout the whole process of crawling search, the standard deviation curve of the FCWSEO is relatively flat. After the number of downloaded webpages is greater than 4,000, the standard deviation of the FCWSEO gradually decreases and is finally stabilized at about 0.157. The standard deviation of the MOACO keeps a downward trend and is finally stabilized at 0.137. According to the criterion of lower standard deviation corresponding to better algorithm stability, the FCWSEO has the best stability in the early stage of Web search and is surpassed by the MOACO in the later stage. This is because, in the beginning, the FCWSEO is easy to generate more diverse non-dominated hyperlinks based on the smaller circular regions. As the radii of the circular regions are enlarged gradually,

Table 4
Comparison of performance of the FCWSE and FCWSEO crawlers on rainstorm disaster when retrieving 15,000 webpages.

Algorithms	*M*	*HR*	*AR*	*SD*	Time/h
FCWSE	11747	0.783	0.714	0.164	11.35
FCWSEO	12162	0.811	0.822	0.157	11.64

it is easy to grab the suboptimal hyperlinks by comparing with the nearby hyperlinks. However, in the MOACO, as the crawler continues, ants will accumulate more pheromones and are easier to find the optimal crawling path and fetch more topic-relevant hyperlinks.

Among the above six focused crawler methods, the FCWSEO, the On-ITS, and the MOACO adopt domain ontology as the topic benchmark model and analyze the content of the search webpages from a semantic perspective to guide the crawling direction of the crawlers. From the above experimental results, we can find that the performance of the FCWSEO, the On-ITS and the MOACO based on domain ontology is far superior to that of the other three focused crawlers without ontology. At the same time, from the results of the above focused crawler experiments, it can also be found that the FCWSEO has better performance and a stronger ability to search topic-relevant webpages than the other five methods.

In addition, in order to further investigate the ontology approach on the effectiveness and stability of the proposed FCWSEO crawler, we run the FCWSEO crawler without the ontology (abbreviated as FCWSE below). Table 4 shows the *M*, *HR*, *AR*, *SD,* and running time of the FCWSE and FCWSEO for the topic of rainstorm disaster when the crawlers retrieves 15,000 webpages. We can find that experimental results of the FCWSEO for all performance evaluation indices except the running time are better than those of the FCWSE. This further confirms the significance of the semantic approach to focused crawlers. In addition, we find that the running time of FCWSE is close to that of the FCWSEO, which proves that the topic relevance calculation based on ontology does not consume a lot of time. In fact, most of the crawling time is spent on parsing the webpage content. Because the FCWSEO searches topic-relevant webpages as much as possible, it will spend more time parsing a larger number of webpages. In short, we can conclude that the proposed FCWSEO crawler is an effective semantic retrieval method.

7.3. Numerical experiments for the FCWSEO under different parameters

In this section, we design the contrast experiments for testing and verifying the effects of parameters in the FCWSEO crawler. The important parameters in the FCWSEO include the hyperlink evaluation threshold τ and the topic relevance threshold η. In order to test the effects of τ and η on the results of the experiments, we select some representative values within a reasonable range. The FCWSEO is run 10 times under each group of parameters. The harvest rate of the FCWSEO with different threshold sizes of τ and η are listed in Table 5.

The value of threshold τ is set to 0.1, 0.15, 0.2, 0.25, respectively, and the value of threshold τ is set to 0.6, 0.65, 0.7, respectively. In Table 5, "—" means the FCWSEO cannot download 15,000 webpages under the thresholds of τ and η. From Table 5, it is clear that when the value of τ increases from 0.1 to 0.2, the harvest rates increase gradually, and when the threshold τ is set to 0.2, the harvest rate reaches the optimal value. However, when the value of τ is 0.25, the FCWSEO cannot find 15,000 webpages because the filter capacity of 0.25 is excessive. In addition, we can see from Table 5 that when the value of η increases from 0.6 to 0.7, the harvest rates decrease gradually. However, the high harvest rate under low threshold η does not mean the acquirement of more topic-relevant webpages, and the low harvest rate under high threshold η does not mean the acquirement of fewer topic-relevant webpages. This is because the FCWSEO with a high harvest rate may retrieve some irrelevant webpages under the low value of η, and the FCWSEO with a low harvest rate may miss some topic-relevant webpages under the high value of η. According to the predefined threshold in Ref. [29] and the analysis of webpages fetched by the FCWSEO with different parameters, we set the values of τ and η to 0.2 and 0.7, respectively.

Table 5
Harvest rates of the FCWSEO with different thresholds of τ and η under rainstorm disaster theme when retrieving 15,000 webpages.

η	τ			
	0.1	0.15	0.2	0.25
0.6	0.643	0.779	0.871	–
0.65	0.536	0.724	0.835	–
0.7	0.468	0.695	0.811	–

8. Conclusion and future work

Classical focused crawlers generally focus on the calculation of topic relevance, which linearly integrate multiple impact factors by the weighted sum approach to predict the topic priorities of unvisited hyperlinks. However, the optimal weighted coefficients are difficult to obtain and the subsequent single-objective optimization-based crawlers are easy to fall into local optima of the search. In contrast, this research focuses on a multi-objective optimization model for evaluating unvisited hyperlinks based on Web text and link structure. The main contributions of this article include: (1) A novel focused crawler strategy FCWSEO which combines the domain ontology and the so-called the WSE framework based on multi-objective optimization model is proposed to retrieve the topic-relevant webpages, where the ontology is applied to establish the topic model and the WSE is used to guide the search direction of the focused crawler; (2) In the WSE, a combination of the NFCS method and the fast non-dominated sorting is proposed to select the Pareto-optimal hyperlinks from the unvisited hyperlink bank. The results of comparative experiments show that the proposed FCWSEO is an effective focused crawler. It outperforms the other six crawlers, including the BFS, OPS, FCSA, On-ITS, MOACO, and FCWSE, and is capable of finding more quantity and higher quality topic-relevant webpages.

According to the methods presented in this article and the experimental results, the following conclusions could be drawn: (1) The performance of focused crawlers based on domain ontology is far superior to that of the other focused crawlers without ontology, which confirms the significance of semantic approach to focused crawlers. (2) The WSE framework based on multi-objective optimization model effectively extend the search scape of crawlers in Web space, characterizing with strong global search ability, while avoiding conflicts of multiple targets. (3) A combination of the NFCS method and the fast non-dominated sorting is good for improving the uniformity of the Pareto front and maintaining the diversity of the obtained hyperlinks.

From the constructing process of the FCWSEO, it is not hard to see that the method of this article can be easily extended to other fields. It is necessary to change the keywords and construct ontologies of the other domains and therefrom compute the semantic weight vector of the topic words based on the constructed ontologies, appropriately adjust the seed hyperlinks corresponding to the chosen domains, and set the thresholds τ, η by using the grid search method. It is necessary to build a novel ontology for the focused crawling of another keyword. This is because this study uses the VSM by computing the cosine similarity of the semantic weight vector of feature words in the ontology and the feature weight vector of a webpage to obtain the similarity of a webpage related to the given keyword (topic word). Furthermore, the proposed method based on WSE with the combination of the NFCS and the fast non-dominated sorting can be extended to other multi-objective optimization problems. In the future, there are some works to study deeply and widely. Firstly, the domain ontology should be constructed based on a smarter method. Moreover, we should focus on the running time of crawlers and consider more efficient crawler strategies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the Special Foundation of Guangzhou Key Laboratory of Multilingual Intelligent Processing, China(No. 20195010008), Guangdong Basic and Applied Basic Research Foundation of China (No. 2021A1515011974), and the Program of Science and Technology of Guangzhou, China (Grand No. 202002030238).

References

[1] S. Chakrabarti, M.V.D. Berg, B. Domc, Focused crawling: a new approach to topic-specific web resource discovery, Comput. Netw. 31 (11–16) (1999) 1623–1640.
[2] P. Hegade, N. Lingadhal, S. Jain, U. Khan, K.L. Vijeth, Crawler by contextual inference, SN Comput. Sci. 2 (216) (2021).
[3] T. Tamura, K. Somboonviwat, M. Kitsuregawa, A method for language-specific web crawling and its evaluation, Syst. Comput. Japan 38 (2) (2007) 10–20.
[4] K. Pavani, G.P. Sajeev, A novel web crawling method for vertical search engines, in: The Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI, Udupi, India, 2017, pp. 1488–1493.
[5] C.J. Fei, B.S. Liu, Focused crawler based on LDA extended topic terms, Comput. Appl. Softw. 35 (4) (2018) 49–54.
[6] Y.K. Yang, Y.J. Du, J.Y. Sun, Y.F. Hai, A topic-specific web crawler with concept similarity context graph based on FCA, in: The Proceedings of the 4th International Conference on Intelligent Computing: Advanced Intelligent Computing Theories & Applications-with Aspects of Artificial Intelligence, Shanghai, China, 2008, pp. 840–847.
[7] C.C. Hsu, F. Wu, Topic-specific crawling on the web with the measurements of the relevancy context graph, Inf. Syst. 31 (4–5) (2006) 232–246.
[8] W.G. Guan, Y.G. Luo, Design and implementation of focused crawler based on concept context graph, Comput. Eng. Des. 37 (2016) 2679–2684.
[9] H. Wang, H. Zhang, J.C. Shi, Research on domain ontology concept acquisition method based on LDA and application, Comput. Eng. Appl. (2017) 1–7.
[10] M. Rani, A.K. Dhar, O.P. Vyas, Semi-automatic terminology ontology learning based on topic modeling, Eng. Appl. Artif. Intell. 63 (2017) 108–125.
[11] G. Zhu, J.Y. Yang, X.H. Wu, M.N. Feng, Research on construction of hierarchy relationship and ontology of meteorological disaster based on FCA, J. Mod. Inf. 37 (5) (2017) 79–88.
[12] Y. Wang, Design and Implementation of Focused Crawler Based on Breadth-First, Fudan University, Shanghai, 2011.
[13] S. Rawat, D.R. Patil, Efficient focused crawling based on best first search, in: The Proceedings of the 2013 IEEE International Advance Computing Conference, Ghaziabad, India, 2013, pp. 908–911.
[14] W.P. Jing, Y.J. Wang, W.W. Dong, Research on adaptive genetic algorithm in application of focused crawler search strategy, Comput. Sci. 43 (8) (2016) 254–257.
[15] W. Yan, L. Pan, Designing focused crawler based on improved genetic algorithm, in: The Proceedings of the 2018 Tenth International Conference on Advanced Computational Intelligence, ICACI, Xiamen, China, 2018, pp. 319–323.
[16] J.F. Liu, P.Y. Gu, J.W. Liu, Focused crawler method combining ontology and improved Tabu search for meteorological disaster, J. Comput. Appl. 40 (8) (2020) 2255–2261.
[17] S. He, J.X. Cheng, X.B. Cai, Focused crawler based on simulated anneal algorithm, Comput. Technol. Dev. 19 (12) (2009) 55–58.
[18] Y.L. Tong, Application of focused crawler using adaptive dynamical evolutional particle swarm optimization, Geomat. Inf. Sci. Wuhan Univ. 33 (12) (2008) 1296–1299.
[19] Y.B. Chen, Z. Zhang, T. Zhang, A searching strategy in topic crawler using ant colony algorithm, Microcomput.Appl. 30 (1) (2011) 53–56.
[20] S. Zheng, Genetic and ant algorithms based focused crawler design, in: The Proceedings of the 2011 2nd International Conference on Innovations in Bio-Inspired Computing & Applications, Shenzhen, Guangdong, 2011, pp. 374–378.
[21] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN Syst. 30 (1–7) (1998) 107–117.
[22] C. Wang, X.H. Ji, Improved pagerank algorithm based on user interest and topic, Comput. Sci. 43 (3) (2016) 275–278.
[23] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, J. ACM 46 (5) (1999) 604–632.
[24] Y. Asano, Y. Tezuka, T. Nishizeki, Improvements of HITS algorithms for spam links, in: The Proceedings of the Joint 9th Asia-Pacific Web and 8th International Conference on Web-Age Information Management Conference on Advances in Data and Web Management, Huang Shan, China, 2007, pp. 479–490.
[25] P.D. Bra, G.J. Houben, Y. Kornatzky, R. Post, Improvements of HITS algorithms for spam links, in: The Proceedings of the Joint 9th Asia-Pacific Web and 8th International Conference on Web-Age Information Management Conference on Advances in Data and Web Management, Huang Shan, China, 2007, pp. 479–490.
[26] M. Hersovici, M. Jacovi, Y.S. Maarek, D. Pelleg, M. Shtalhaima, S. Ur, The shark-search algorithm-an application: tailored web site mapping, Comput. Netw. ISDN Syst. 30 (1–7) (1998) 317–326.
[27] Z.M. Chen, J. Ma, J.S. Lei, B. Yuan, L. Lian, An improved shark-search algorithm based on multi-information, in: The Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD, Haikou, China, 2007.
[28] Y.K. Cheng, W.J. Liao, G. Cheng, Strategy of focused crawler with word embedding clustering weighted in shark-search algorithm, Comput. Digit. Eng. 46 (1) (2018) 144–148.
[29] W.J. Liu, Y.J. Du, A novel focused crawler based on cell-like membrane computing optimization algorithm, Neurocomputing 123 (2014) 266–280.
[30] Y.J. Du, W.J. Liu, X.J. Lv, G.L. Peng, An improved focused crawler based on semantic similarity vector space model, Appl. Soft Comput. 36 (2015) 392–407.
[31] L. Qiu, Y.S. Lou, M. Chang, An improved shark-search algorithm for theme crawler, Microcomput. Appl. 33 (2) (2017) 19–21.
[32] A. Seyfi, A. Patel, J.C. Júnior, Empirical evaluation of the link and content-based focused Treasure-Crawler, Comput. Stand. Interf. 44 (2016) 54–62.
[33] W. Zhao, Z.Y. Guan, Z.W. Cao, Z. Liu, Mining and harvesting high quality topical resources from the web, Chin. J. Electron. 25 (1) (2016) 48–57.
[34] R. Nayak, P. Senellart, F.M. Suchanek, A.S. Varde, Discovering interesting information with advances in web technology, ACM SIGKDD Explor. Newsl. 14 (2) (2012) 63–81.
[35] Y.J. Du, C.X. Li, Q. Hu, X.L. Li, X.L. Chen, Ranking web page with path trust knowledge graph, Neurocomputing 269 (2017) 58–72.
[36] Z. Jia, S. Pramanik, R.S. Roy, G. Weikum, Complex temporal question answering on knowledge graphs, in: The Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Queensland, Australia, 2021, pp. 792–802.
[37] F.M. Suchanek, G. Kasneci, G. Weikum, YAGO: A large ontology from wikipedia and WordNet, J. Web Semant. 6 (3) (2008) 203–217.
[38] A. Daoui, N. Gherabi, A. Marzouk, An enhanced method to compute the similarity between concepts of ontology, in: The Proceedings of the 2017 International Conference on Information Technology and Communication Systems, Sydney, Australia, 2017, pp. 95–107.
[39] M. Ehrig, A. Maedche, Ontology-focused crawling of web documents, in: The Proceedings of the 2003 ACM Symposium on Applied Computing, Melbourne, Florida, 2003, pp. 1174–1178.
[40] P. Bedi, A. Thukral, H. Banati, A. Behi, V. Mendiratta, A multi-threaded semantic focused crawler, J. Comput. Sci. Technol. 27 (6) (2012) 1233–1242.
[41] X.T. Zhang, X. Hou, X.F. Chen, T. Zhuang, Ontology-based semantic retrieval for engineering domain knowledge, Neurocomputing 116 (2013) 382–391.
[42] A. Capuano, A.M. Rinaldi, C. Russo, An ontology-driven multimedia focused crawler based on linked open data and deep learning techniques, Multimedia Tools Appl. 79 (2020) 7577–7598.
[43] A.I. Saleh, A.E. Abulwafa, M.F.A. Rahmawy, A web page distillation strategy for efficient focused crawling based on optimized naïve bayes (ONB) classifier, Appl. Soft Comput. 53 (2017) 181–204.

[44] T. Hassan, C. Cruz, A. Bertaux, Ontology-based approach for unsupervised and adaptive focused crawling, in: The Proceedings of the International Workshop on Semantic Big Data, Chicago, Illinois, 2017, p. 2.

[45] N. Tandon, A.S. Varde, G. d. Melo, Commonsense knowledge in machine intelligence, ACM SIGMOD Rec. 46 (4) (2017) 49–52.

[46] S. Razniewski, N. Tandon, A.S. Varde, Information to wisdom: Commonsense knowledge extraction and compilation, in: The Proceedings of the 14th ACM International Conference on Web Search and Data Mining, Israel, 2021, pp. 1143–1146.

[47] Y.P. Sheng, Z.L. Xu, Y.F. Wang, G.d. Melo, Multi-document semantic relation extraction for news analytics, World Wide Web 23 (2020) 2043–2077.

[48] A.B. Rios-Alvarado, I. Lopez-Arevalo, V.J. Sosa-Sosa, Learning concept hierarchies from textual resources for ontologies construction, Expert Syst. Appl. 40 (2013) 5907–5915.

[49] L.L. Ma, H.W. Li, S.W. Lian, R.P. Liang, H. Chen, A strategy of disaster focused crawler based on ontology semantics, Comput. Eng. 42 (11) (2016) 50–56.

[50] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Trans. Evolut. Comput. 6 (2) (2002) 182–197.

[51] S. Kukkonen, K. Deb, Improved pruning of non-dominated solutions based on crowding distance for bi-objective optimization problems, in: The Proceedings of the 2006 IEEE International Conference on Evolutionary Computation, Vancouver, Canada, 2006, pp. 1179–1186.

[52] N. Hallam, P. Blanchfield, G. Kendall, Handling diversity in evolutionary multi-objective optimization, in: The Proceedings of the 2005 IEEE Congress on Evolutionary Computation, Edinburgh, Scotland, UK, 2005, pp. 2233–2240.

[53] J.F. Liu, F. Li, S.Y. Jiang, Focused annealing crawler algorithm for rainstorm disasters based on comprehensive priority and host information, Comput. Sci. 46 (2) (2019) 215–222.

[54] J.F. Liu, Y. Dong, J.W. Liu, Focused crawler strategy based on multi-objective ant colony algorithm, Comput. Eng. 46 (9) (2020) 274–282.