Jingfa LIU, Zhen WANG, Guo ZHONG, Zhihe YANG, 2023. A new focused crawler using an improved tabu search algorithm incorporating ontology and host information. *Frontiers of Information Technology & Electronic Engineering,* 24(6):859-875. https://doi.org/10.1631/FITEE.2200315

A new focused crawler using an improved tabu search algorithm incorporating ontology and host information

Key words: Focused crawler; Tabu search algorithm; Ontology; Host information; Priority evaluation

Corresponding author: Zhen WANG
E-mail: 1007427607@qq.com
ORCID: https://orcid.org/0000-0003-4940-2812

Motivation

1. The problems of incomplete topic description and repetitive crawling of visited hyperlinks exist in traditional focused crawling methods.

2. The current strategies used to guide crawling direction, such as breadth-first search (BFS), optimal priority search (OPS), and simulated annealing (SA), are prone to falling into local optima of the search.

3. It is possible that a crawler recursively crawls under a few hosts, resulting in premature convergence of the crawler and limitation to retrieve more topic-relevant webpages.

Main idea

1. To solve the problems of incomplete topic description, we propose the strategy of building ontology based on formal concept analysis (FCA) to describe topics at the semantic and knowledge levels.

2. A comprehensive priority evaluation method based on Web text and link structure is designed to improve the assessment of topic relevance for unvisited hyperlinks.

3. To guide the direction of the crawler, an improved tabu search (ITS) strategy with host information is presented to select the next hyperlink.

Ontology construction

The detailed steps of ontology construction based on FCA are as follows:

(1) select five keywords for the determined domain and search for keywords through search engines such as Baidu and Google to obtain the top 50 webpages of each search engine;

(2) use the tool IK-Analyzer to perform word segmentation;

(3) extract document sets and term sets that describe the topic;

(4) build a document–term matrix, which is input into the tool ConExp (https://sourceforge.net/projects/conexp/) to generate a concept lattice and obtain a Hasse diagram;

(5) describe the hierarchical relations among concepts by ontology Web language (OWL) (https://www.w3.org/TR/owl-features/);

(6) visualize the ontology by Protégé (https://protege.stanford.edu/).

Applying the above method, we construct a tourism ontology and a rainstorm disaster ontology. The tourism ontology includes 61 concepts and a seven-level hierarchical structure. The rainstorm disaster ontology includes 50 concepts and a six-level hierarchical structure.

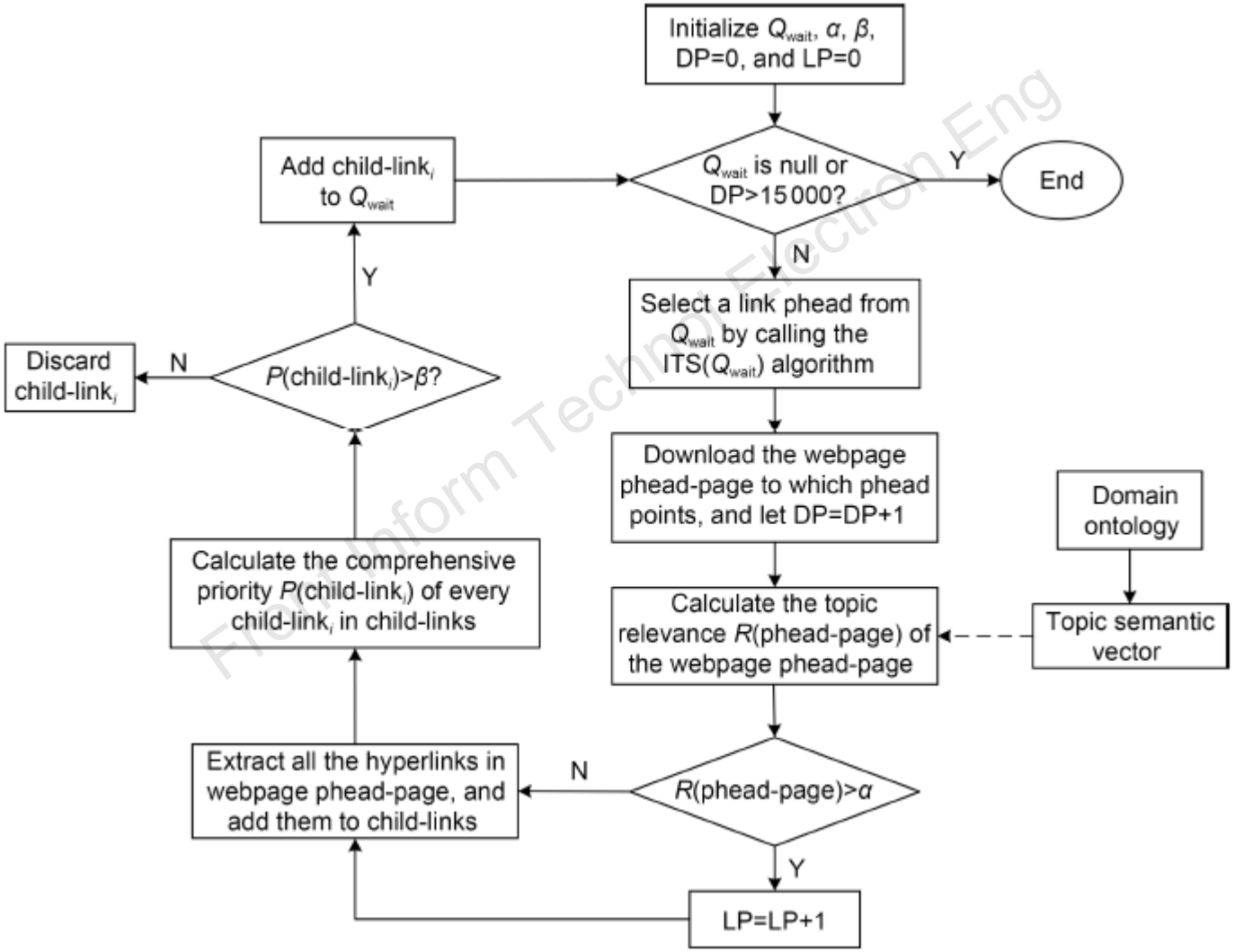
Comprehensive priority evaluation

A comprehensive priority evaluation method is given to evaluate the topic relevance of unvisited hyperlink *l*. Its expression is shown as follows:

$$P(l) = r_1 \cdot \mathrm{PR}(p_l) + r_2 \cdot \frac{1}{m}\sum^{m} R(p_i) + r_3 \cdot R(A_l).$$

$P(l)$ represents the comprehensive priority value of the unvisited hyperlink l, $R(A_l)$ represents the topic relevance of anchor text A_l of hyperlink l, $R(p_i)$ represents the topic relevance of webpage p_i that contains hyperlink l, and $\mathrm{PR}(p_l)$ is the PageRank (PR) value of webpage p_l containing hyperlink l.

Flowchart of the proposed FCOITS algorithm



Focused crawler combining the FCOITS algorithm and host information

It is possible that a crawler recursively crawls under a few hosts, resulting in premature convergence of the crawler and limitation to retrieve more topic-relevant webpages. We propose a new focused crawler that integrates host information into FCOITS, called FCITS_OH.

Its most important idea is that at the beginning of FCITS_OH, the hyperlinks that are located under different hosts and have higher comprehensive priorities are selected as seed hyperlinks to avoid premature convergence of the algorithm.

Experimental results

Table 2 Comparison of results obtained by nine crawling algorithms in the tourism and rainstorm disaster domains when DP reaches 15 000

Algorithm	Tourism domain					Rainstorm disaster domain				
	LP	AC	AR	SD	Time (h)	LP	AC	AR	SD	Time (h)
BFS (2015)	5610	37.40	0.4247	0.2848	**7.78**	3549	23.66	0.2947	0.3096	**8.24**
OPS (2013)	10 230	68.20	0.6966	0.2317	8.56	9813	65.42	0.6376	0.2599	8.93
FCSA (2019)	11 255	75.03	0.7292	0.1769	10.72	10 506	70.04	0.6627	0.1953	11.08
FCWSEO (2022)	12 129	80.86	0.7553	**0.1293**	13.94	12 162	81.03	**0.8200**	0.1570	11.64
OLMOACO (2022)	–	–	–	–	–	11 126	74.17	0.7781	**0.1375**	16.00
FCTS	10 822	72.15	0.7066	0.1726	10.11	10 254	68.36	0.6523	0.1962	9.98
FCITS	11 581	77.21	0.7534	0.1446	11.26	11 054	73.69	0.7002	0.1589	10.29
FCOITS	12 679	84.53	0.7806	0.1413	11.27	11 954	79.69	0.7306	0.1495	11.24
FCITS_OH	**13 082**	**87.21**	**0.7912**	0.1340	11.97	**12 393**	**82.62**	0.7421	0.1444	11.51

DP: number of downloaded webpages; LP: number of downloaded topic-relevant webpages; AC: accuracy; AR: average topic relevance; SD: standard deviation. The optimal value of every metric is marked in bold

Experimental results (Cont'd)

Table 3 Friedman ranks of nine crawling algorithms for the four representative evaluation metrics in the tourism and rainstorm disaster domains when DP reaches 15 000

Algorithm	Friedman rank									
	Tourism domain					Rainstorm disaster domain				
	LP	AC	AR	SD	Average	LP	AC	AR	SD	Average
BFS (2015)	8	8	8	8	8.00	9	9	9	9	9.00
OPS (2013)	7	7	7	7	7.00	8	8	8	8	8.00
FCSA (2019)	5	5	5	6	5.25	6	6	6	6	6.00
FCWSEO (2022)	3	3	3	1	2.50	2	2	1	4	2.25
OLMOACO (2022)	–	–	–	–	–	4	4	2	1	2.75
FCTS	6	6	6	5	5.75	7	7	7	7	7.00
FCITS	4	4	4	4	4.00	5	5	5	5	5.00
FCOITS	2	2	2	3	2.25	3	3	4	3	3.25
FCITS_OH	1	1	1	2	**1.25**	1	1	3	2	**1.75**

DP: number of downloaded webpages; LP: number of downloaded topic-relevant webpages; AC: accuracy; AR: average topic relevance; SD: standard deviation. The optimal value is marked in bold

Conclusions

1. The experimental results indicate that the use of domain ontology to describe topics proposed in this study can improve the level of description of crawling topics to a certain extent.

2. This study proposes a comprehensive priority evaluation method of hyperlinks based on Web text and link structure, and experimental results show that this is an effective method for hyperlink evaluation.

3. In order to address the shortcomings of traditional crawling methods, this study proposes the FCOITS and FCITS_OH crawling methods, and the results show that these two methods are high-performance crawling strategies.

Jingfa LIU received the B.S. degree in mathematics from Hunan Normal University, China in 1995, the M.S. degree in operational research and cybernetics from Shanghai Railway University, China in 1999, and the Ph.D. degree in computer software and theory from Huazhong University of Science and Technology, China in 2007. He is currently a professor and a member of the School of Information Science and Technology, Guangdong University of Foreign Studies. His research interests include information retrieval, computational intelligence, and multi-objective constrained optimization.

Zhen WANG received the B.S. degree in communication engineering from Xiangnan University, China in 2020, and the M.S. degree in electronic information from Guangdong University of Foreign Studies, China in 2022. He joined China Unicom Central South Research Institute as a technical developer in August 2022. His research interests include information retrieval and heuristic algorithm.

Guo ZHONG received the Ph.D. degree in computer science from University of Macau, Macao, China. He is currently an associate professor with the School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China. His research interests are machine learning and its applications, such as pattern recognition, data mining, and information retrieval.

Zhihe YANG received the B. S. degree in computer science and technology from Guangdong University of Foreign Studies, China in 2020. He is currently a master student in Guangdong University of Foreign Studies, China. His research interests include information retrieval and heuristic algorithm.