Jingfa LIU, Fan LI, Ruoyao DING, Zi’ang LIU, 2022. Focused crawling strategies based on ontologies and simulated annealing methods for rainstorm disaster domain knowledge. *Frontiers of Information Technology & Electronic Engineering*, 23(8):1189-1204. https://doi.org/10.1631/FITEE.2100360

Focused crawling strategies based on ontologies and simulated annealing methods for rainstorm disaster domain knowledge

Key words: Focused crawler; Ontology; Priority evaluation; Simulated annealing; Rainstorm disaster

Corresponding author: Fan LI
E-mail: bj2014_lifan@163.com
ORCID: https://orcid.org/0000-0001-7836-0522

Motivation

1. The information about webpages related to rainstorm disasters is sparse, showing the characteristics of big data. In the field of information retrieval (IR), traditional focused crawlers face great challenges in improving their accuracy. The main difficulties are the establishment of topic benchmark models, the assessment of topic relevance (including hyperlinks and texts), and the design of crawler strategies.

2. Domain ontology is a formal description of the background knowledge in a specific field.

3. The simulated annealing (SA) algorithm has a strong global search capability and can accept the sub-optimal links based on Metropolis sampling and avoid the focused crawling falling into local search.

Main idea

1. A novel multiple-filtering strategy based on local ontology and global ontology (MFSLG) is proposed to find more topic-relevant hyperlinks.

2. A comprehensive priority evaluation method (CPEM) considering four indicators (topic relevance of the webpage containing the unvisited hyperlink, topic relevance of anchor text, the PageRank value, and topic relevance of the webpage to which the unvisited hyperlink points) is used to evaluate the unvisited hyperlinks.

3. An annealing strategy based on Metropolis sampling is applied to avoid the focused crawler falling into a local optimal search.

4. A new focused crawler combining domain ontology and the SA algorithm has been used to obtain the effective domain knowledge of the rainstorm disaster for the first time.

Method

1. By incorporating SA into the focused crawler with MFSLG and CPEM for the first time, two novel focused crawler strategies based on ontology and SA (FCOSAs) are proposed to obtain topic-relevant webpages about rainstorm disasters from the network.

2. One is a focused crawler strategy based on only global ontology (FCOSA_G), and the other is a focused crawler strategy based on both the global ontology and local ontology (FCOSA_LG).

Method (Cont'd)

Algorithm 3 FCOSA_LG

Input: seed URLs

Output: downloaded webpages

1: Add seed URLs to Q_w. Set σ, φ, and η. Let DP=0 and LP=0

2: Select the first link ordered in Q_w, and mark it as **Header-link**. The webpage to which **Header-link** points is marked as the **Current-page**

3: Remove **Header-link** from Q_w and download the **Current-page**

4: Let DP=DP+1

5: Remove the noise and extract tag information (Table 2) from the **Current-page**. Use IK for word segmentation and gain the feature vector **DK** of the **Current-page**

6: Calculate the topic relevance R(**Current-page**) of the **Current-page** text according to Eq. (6)

7: **If** R(**Current-page**)$>\sigma$ **then**

  Download the **Current-page** and let LP=LP+1

  End if

8: Extract all the **child-links** and the corresponding anchor texts from the **Current-page**, and remove repeated links

  // Local ontology is used to implement the first filtering

  // of **child-links**

9: **For** i=1 to k_1 **do**

  // k_1 is the size of the **child-links**

    For j=1 to k_2 **do**

    // k_2 is the number of local ontologies, and k_2=3 in this

    // study

      Calculate the topic relevance $R_{i,j}$ of **child-link**$_i$ based on the j^{th} local ontology according to Eq. (11):

      $R_{i,j}$= Sim(LTK$_j$, UK)

      If $R_{i,j} \geqslant \varphi$ **then**

      // φ is a positive parameter

        Save **child-link**$_i$

      break

      Else if $R_{ij}<\varphi$ and $j=k_2$

        Discard **child-link**$_i$

      End if

    End for

  End for

  // Global ontology is used to implement the second fil-

  // tering of the saved **child-links**

10: **For** j=1 to k_3 **do**

  // k_3 is the number of the saved **child-links**

    Calculate the comprehensive priority of the **child-link**$_j$ according to Eq. (12), where TK is replaced by **GTK**= (gtk$_1$, gtk$_2$, ..., gtk$_r$)

    If Priority(**child-link**$_j$)$>\eta$ **then**

    // η is a positive parameter

      Insert **child-link**$_j$ into Q_w

    Else give up **child-link**$_j$

    End if

  End for

11: Recalculate PR values of all the downloaded webpages and update the comprehensive priority values of all links in Q_w

12: **If** Q_w is not empty **then**

    Let l=ISA (Q_w)

    // Return link l

    Insert link l into the head of Q_w

  Else the algorithm ends

  End if

13: **If** DP<15 000 **then**

    Go to step 2

  Else the algorithm ends

  End if

The FCOSA_G algorithm is obtained by deleting step 9 in the FCOSA_LG algorithm.

Major results

Table 4 Experimental results of different algorithms about evaluation indices of Accuracy, LP, AR_{LP}, SD_{LP}, AR_{DP}, SD_{DP}, and retrieval time and when DP reaches 1000, 5000, 10 000, and 15 000

DP	Algorithm	Accuracy	LP	AR_{LP}	SD_{LP}	AR_{DP}	SD_{DP}	Time (h)
1000	BFS	0.1840	184	0.7760	0.0538	0.4122	0.2662	
	OPS	**0.7440**	**744**	**0.7769**	**0.0342**	0.6007	0.2258	
	WSE	0.4020	402			0.6500	0.1620	
	ITS	0.6960	696			**0.7027**	0.1830	
	On-ITS	0.7020	702			0.6982	0.1624	
	FCSA	0.6010	601	0.7498	0.0367	0.5819	0.2956	
	FCOSA_G	0.7140	714	0.7663	0.0651	0.6909	0.1359	
	FCOSA_LG	0.7100	710	0.7378	0.0644	0.6779	**0.1129**	
5000	BFS	0.1438	719	0.7723	0.0491	0.2856	0.2563	
	OPS	**0.7900**	**3950**	0.7782	**0.0274**	0.6736	0.1494	
	WSE	0.6130	3065			0.7000	0.1620	
	ITS	0.6580	3290			0.6577	0.1556	
	On-ITS	0.7000	3500			0.7076	0.1629	
	FCSA	0.6264	3132	0.7616	0.0449	0.5952	0.2365	
	FCOSA_G	0.7314	3657	0.7871	0.0633	0.7106	0.1478	
	FCOSA_LG	0.7620	3810	**0.7954**	0.0688	**0.7498**	**0.1199**	

To be continued

Major results (Cont’d)

Table 4

10 000	BFS	0.0965	965	0.7776	0.0425	0.2927	0.2726	
	OPS	0.5376	5376	0.7784	**0.0321**	0.5716	0.2139	
	WSE	0.7000	7006			0.7250	0.1620	
	ITS	0.6600	6600			0.6436	0.2013	
	On-ITS	0.7010	7010			0.7266	0.1622	
	FCSA	0.6043	6043	0.7798	0.0472	0.6228	0.2424	
	FCOSA_G	0.7123	7123	0.7808	0.0643	0.6913	0.1693	
	FCOSA_LG	**0.7882**	**7882**	**0.8023**	0.0604	**0.7562**	**0.1287**	
15 000	BFS	0.0657	985	0.7788	0.0447	0.2262	0.2552	**8.54**
	OPS	0.4426	6639	0.7785	**0.0375**	0.5631	0.2020	9.12
	WSE	0.7330	11 002			0.7290	0.1600	12.23
	ITS	0.6364	9546			0.6627	0.1953	11.48
	On-ITS	0.7340	11 010			0.7295	0.1619	13.24
	FCSA	0.5817	8726	0.7895	0.0462	0.6463	0.2475	11.16
	FCOSA_G	0.6693	10 040	0.7906	0.0644	0.6871	0.1677	12.55
	FCOSA_LG	**0.7653**	**11 479**	**0.8095**	0.0581	**0.7511**	**0.1462**	13.12

Best results are in bold

Major results (Cont'd)

Table 5 Friedman test ranks of eight algorithms for the three representative evaluation indices of Accuracy, AR_{DP}, and SD_{DP} when DP reaches 15 000

Index	Rank							
	BFS	OPS	WSE	ITS	On-ITS	FCSA	FCOSA_G	FCOSA_LG
Accuracy	8	7	3	5	2	6	4	1
AR_{DP}	8	7	3	5	2	6	4	1
SD_{DP}	8	6	2	5	3	7	4	1
Average	8	6.67	2.67	5	2.33	6.33	4	**1**

Table 6 Computational results obtained by FCOSA_LG algorithm with different threshold sizes of σ, φ, and η when DP= 15 000

φ	(Accuracy, LP)								
	σ=0.5			0.6			0.7		
	η=0.10	0.15	0.20	0.10	0.15	0.20	0.10	0.15	0.20
0.05	(0.4845, 7268)	(0.9050, 13 575)	–	(0.4085, 6127)	(0.8055, 12 082)	–	(0.2850, 4275)	(0.6798, 10 197)	–
0.10	(0.5305, 7958)	(0.9117, 13 675)	–	(0.4469, 6704)	(0.8490, 12 735)	–	(0.3058, 4587)	(0.7297, 10 945)	–
0.15	(0.6055, 9083)	**(0.9280, 13 920)**	–	(0.4605, 6907)	**(0.8555, 12 832)**	–	(0.3283, 4924)	**(0.7653, 11 479)**	–
0.20	(0.7354, 11 031)	–	–	(0.6358, 9537)	–	–	(0.5543, 8314)	–	–
0.25	(0.8518, 12 777)	–	–	(0.7426, 11 139)	–	–	(0.6615, 9923)	–	–

"–" means that the algorithm has ended prematurely when DP has not reached 15 000

Major results (Cont’d)

Table 7 Experimental results of FCOSA_LG algorithm over five independent times when σ=0.7, φ=0.15, η=0.15, and DP=15 000

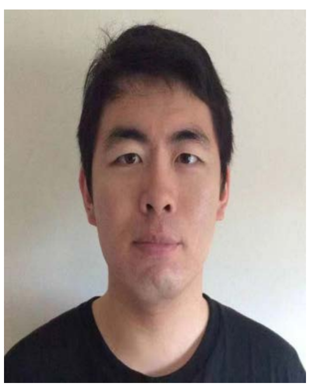
No.	Accuracy	LP	AR_{LP}	SD_{LP}	AR_{DP}	SD_{DP}
1	0.7498	11 247	0.8029	0.0311	0.7125	0.1515
2	0.7516	11 274	0.8183	0.0389	0.7187	0.1489
3	0.7630	11 445	0.7915	0.0286	0.7110	0.1476
4	0.7528	11 292	0.8060	0.0301	0.7244	0.1517
5	0.7653	11 479	0.8095	0.0581	0.7511	0.1462
Average	0.7565	11 347	0.8056	0.0374	0.7235	0.1492

Conclusions

1. The FCOSA_LG algorithm achieved state-of-the-art performance and was capable of finding more topic-relevant webpages. The crawler algorithms proposed here can effectively obtain relevant knowledge about rainstorm disasters from the network, and provide a reference plan for disaster warning and preventive measures. In addition, crawlers can promote the construction of ontology knowledge in the domain of rainstorm disasters.

2. It was proved that the combination of the SA algorithm and the MFSLG strategy to guide crawlers to filter hyperlinks can improve the stability of focused crawlers.

Jingfa LIU received his BS degree in mathematics from Hunan Normal University, Changsha, China, in 1995, and MS degree in operational research and cybernetics from Shanghai Railway University, Shanghai, China, in 1999, and PhD degree in computer software and theory from Huazhong University of Science and Technology, Wuhan, China, in 2007. He is currently a professor of the School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China. His current research interests include mainly information retrieval, computational intelligence, and multi-objective constrained optimization.



Ruoyao DING received his BS and MS degrees in electronics and information engineering from Beijing Jiaotong University, Beijing, China, in 2009 and 2011, respectively, and also received his MS and PhD degrees in computer science from University of Delaware, USA, in 2013 and 2017, respectively. He is currently an associate professor in the School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China. His research interests include text mining, machine learning, and natural language processing.