Jingfa LIU, Yongchuang WU, Zhaoxia LIU, 2025. A focused crawling strategy based on comprehensive priority evaluation of hyperlinks and improved Bayesian classifier. *Frontiers of Information Technology & Electronic Engineering*, 26(12):2569-2582. https://doi.org/10.1631/FITEE.2400939

A focused crawling strategy based on comprehensive priority evaluation of hyperlinks and improved Bayesian classifier

Key words: Focused crawler (FC); Bayesian classifier; Information retrieval; Priority evaluation

Corresponding author: Yongchuang WU
E-mail: wu112002@outlook.com
ORCID: Yongchuang WU, https://orcid.org/0009-0004-0467-9767; Jingfa LIU, https://orcid.org/0000-0002-0407-1522

Motivations

1. Avoidance of topic drift and enabling crossing tunnels are two main difficulties in focused crawling.

2. High accuracy and recall rates are notable gaps in designing efficient focused crawlers.

Main idea

1. Propose an improved Bayesian classifier with weights (BCW) that adds label weights to feature words, enhancing webpage classification accuracy.

2. Design a content block segmentation (CBS) technique based on backtracking to enable crossing tunnels for accessing relevant webpages from low-relevance ones.

3. Develop a comprehensive priority evaluation (CPE) method of unvisited hyperlinks by considering relevance of anchor text, context of hyperlink, and webpage text.

Bayesian Classifier with Weights

To improve webpage classification accuracy, label weights are incorporated into the Bayesian classifier, thereby reducing misclassification of webpages and preventing topic drift. The BCW method is as follows:

(1) Assign weights to HTML label groups based on their importance in representing the webpage's topic. For example, title tags (<title>, <h1>) are assigned a weight of 2.0, while body text tags (<p>, <li>) receive a weight of 1.0. See the table below for details：

Table 1 Division of labels and their weights

Group	Label	Meaning	Weight
Group 1	<title>, <description>, <keyword>, <h1>	Title, description, keyword, first-level headline	2.0
Group 2	<h2>, <h3>	Secondary-level headline, third-level headline	1.5
Group 3	<h4>, <h5>,	Fourth-level headline, fifth-level headline, bold text	1.2
Group 4	<p>, <td>,	Body information	1.0
Group 5	Other labels	Non-body information	0.2

(2) Calculate the weighted posterior probability for a webpage $X=\{U_1,U_2,\ldots,U_n\}$ belonging to class C_i using the formula:

$$P(C_i \mid X) \propto P(C_i) \prod_{k=1}^{n} P(U_k \mid C_i) L_j$$

L_j is the weight of the label group containing feature word U_k, $P(C_i)$ is the prior probability, and $P(U_k \mid C_i)$ is the conditional probability.

(3) Determine the classification results of webpage X:

$$C(X) = \{C_i | {}^{argmax}_{i} \{P(C_i) \times \prod_{k=1}^{n} (P(U_k|C_i) \times L_j) \mid i = 1, 2, \ldots, m;$$

$$U_k \; is \; located \; in \; Group_j, \; j \in \{1, 2, \ldots, J\}\}\}.$$

By applying the above method, BCW enhances the discrimination of topic-relevant webpages, significantly improving classification accuracy and reducing topic drift during focused crawling.

Content Block Segmentation

To enable crossing tunnels, every webpage is segmented into content blocks, and hyperlinks are extracted from relevant blocks, even if the overall page is irrelevant. The CBS method is as follows:

(1) Parse the webpage's DOM tree and identify innermost labels that contain no nested elements.

(2) For each content block B_i, compute its topic relevance $R(B_i)$ using TF-IDF and VSM (similar to Eq. (3) in the paper).

(3) If $R(B_i) > \lambda$ ($\lambda = 0.30$), proceed to extract all hyperlinks within B_i.

By applying the above method, you can obtain as many topic-related web pages as possible from unrelated ones.

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:

$$E(l) = \alpha R(G) + \beta R(\mathrm{AL}) + \gamma R(\mathrm{CL})$$

E(*l*) indicates the comprehensive priority value of the unvisited hyperlink *l*, *R*(*G*) represents the relevance of webpage text, *R*(AL) indicates the relevance of anchor text, and *R*(CL) denotes the relevance of link context. α, β, and γ are weight coefficients satisfying $\alpha + \beta + \gamma = 1$.

Weights Bayesian Classifier-based Focused crawler with CPE and CBS

Traditional crawlers, lacking tunnel traversal capability, may fail to access topic-relevant web pages hidden in irrelevant pages. Meanwhile, their classification accuracy is relatively low. Therefore, we propose a novel focused crawler termed BCW_CC that integrates Bayesian classifiers with weights, comprehensive priority evaluation (CPE), and content block segmentation (CBS) techniques.

Its most important idea is that BCW_CC combines machine learning classification with fine-grained content analysis, enabling the crawler to extract relevant hyperlinks from specific content blocks even when the overall webpage is classified as irrelevant. This crossing tunnels' mechanism significantly expands the coverage of topic-relevant webpage retrieval.

Experimental Results

Table 2 Comparison of metric indices of the eight crawling algorithms at DP=15 000 in the rainstorm disaster domain

Algorithm	RP	AC	AR	SD
BFS	3549	0.2366	0.2947	0.3096
OPS	9813	0.6542	0.6376	0.2599
FCSA	10 506	0.7004	0.6627	0.1953
FCITS_OH	12 384	0.8256	0.7412	0.1441
FCPSO	12 546	0.8364	0.8079	0.1424
FCCBS	9413	0.6275	0.6874	0.1937
FCBCW	12 680	0.8453	0.8246	0.2080
BCW_CC	13 522	0.9015	0.8412	0.1643

DP: number of downloaded webpages; RP: number of downloaded topic-relevant webpages; AC: accuracy; AR: average topic relevance; SD:standard deviation.

Experimental Results

Table 3 Comparison of five metric indices obtained by five crawling algorithms in the sports domain at DP=15 000

Algorithm	RP	AC	AR	SD	RC
BFS	4131	0.2754	0.3027	0.2952	0.0808
OPS	8828	0.5885	0.6184	0.2627	0.2067
FCCBS	9635	0.6423	0.6534	0.1937	0.2941
FCBCW	11 260	0.7507	0.8061	0.1987	0.2537
BCW_CC	11 828	0.7885	0.8586	0.1606	0.3539

DP: number of downloaded webpages; RP: number of downloaded topic-relevant webpages; AC: accuracy; AR: average topic relevance; SD:standard deviation; RC: recall rate.

Conclusions

1. Enhanced classification accuracy of pages through BCW. The BCW improves webpage topic identification by incorporating HTML label weights, effectively reducing topic drift.

2. Crossing tunnels via CBS. CBS enables extraction of relevant hyperlinks from specific content blocks, expanding coverage of topic-relevant content.

3. Superior performance of BCW_CC. The integrated BCW_CC algorithm demonstrates high-performance crawling, significantly outperforming traditional methods in both accuracy and recall.

Conclusions

The BCW_CC approach successfully addresses key challenges in focused crawling through synergistic integration of classification optimization and content analysis technologies, providing an effective solution for topic-specific information retrieval.

Jingfa LIU received the B.S. degree in mathematics from Hunan Normal University, Changsha, China, in 1995, and the M.S. degree in operational research and cybernetics from Shanghai Railway University, Shanghai, China, in 1999, and the Ph.D. degree in computer software and theory from Huazhong University of Science and Technology, Wuhan, 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 current research interests mainly include information retrieval, computational intelligence, and multi-objective constrained optimization.

Yongchuang WU is a graduate student of the School of Information Science and Technology from Guangdong University of Foreign Studies, Guangzhou, China. His main research interests include information retrieval and focused crawler.

Zhaoxia LIU is currently a senior engineer of the Network and Information Center from Guangdong University of Foreign Studies, Guangzhou, China. Her main research interests include information retrieval and multi-objective optimization.