An Semantic Rank for Web Crawler Based on Formal Concept Analysis

Yajun Du Xinchun Li

[1]School of Mathematical and Computers Science, Xihua University, Chengdu 610039, Sichuan, China

Abstract

Web Crawler is an important research in Search Engine. In this paper, a method for measuring the similarity of FCA concepts is proposed by using information content approach based on user Web log. In process of crawling Web pages for Web Crawler, in order to make choice of Web pages, the semantic rank of Web pages can be determined by using the similarity, other than relying on ontology with human domain expertise. The semantic rank can be made choice of Web pages for Web crawler.

Keywords: Formal concept analysis, Web crawler, Concept similarity, Ontology, Web log

1. Introduction

A crawler is a part program of search engine that it retrieves and downloads Web pages from internet and save these Web pages to local computer. Crawlers are used broadly in some search engines, e.g., AltaVista, DirectHit, Excite, Google, HotBot, Lycos and Yahoo, etc. The work flow of the crawler can be described roughly as follows [1, 2]:

1. Search engine assigns some URLs as an initial Web pages (Seed URLs) for every crawler. And then, the crawler pushes them into URL queue (QueueURLs) in which each one instructs the crawler to travel in Web.
2. The crawler starts works with these URLs.
3. When the crawler retrieves some Web pages, it extracts all URLs (CurrentURLs) in these Web pages.
4. The crawler choices some URLs and adds them to an URL queue.
5. Whereafter, to continue crawling, the crawler makes a choice of URLs from the QueueURLs, and deletes these crawled URLs.
6. The crawler repeats (2) to (5), until some URLs does not exist in URL queue.

Web Coverage, relevance and precision[3] are three indexes measuring the performance of various crawlers. Apparently, the performance of a crawler is decided in (3), (4) and (5). Some crawlers attempts to cover most Web pages, some attempt to crawl most professional Web pages, some attempt to crawl most accurate Web pages for user query. On the other hands, some crawlers desire to spend less time to crawl Web pages in the specialized domains. In order to implement these tasks, how to select URLs into the QueueURLs, and how to make a choice of URLs for next step are important challenges. To crawl efficiently and retrieve Web pages, there are two typical methods for above two challenges: linkage-based, content-based methods[4].

1.1. URL rank

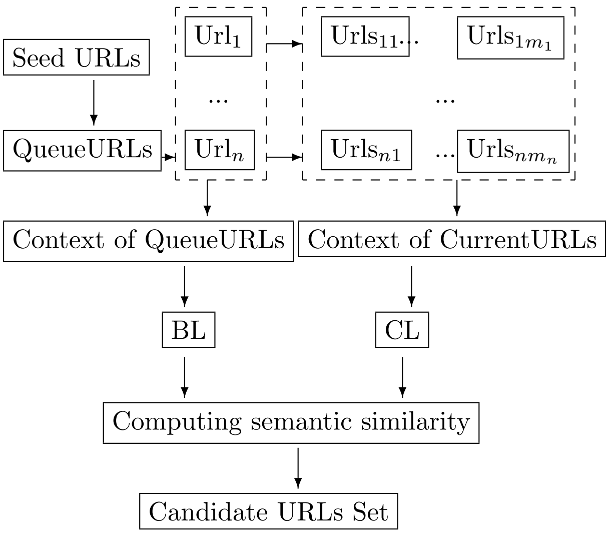
For the linkage-based method, (1). PageRank; (2).Hits; (3). back-link; (4).forward-link; (5). Location Metric[1]; For the content-based method: Similarity between Web page and user Query[1]:

In [5], a new focused Web crawler is proposed. During the crawling phase, the crawler associates a priority value with each URL, the URL with the higher priority value is being added to the queue, the priority of URL is computed by a combination of linkage and context of Web page. The visited Web pages are collected and clustered, whereas page sequences leading to target pages are extracted from the link structure among these pages from different clusters by using Hidden Markov Model, finally the priority of URLs to follow is a learned estimate value how likely the page is to lead to a target page. This Crawler performs better than Context-Graph crawler and Best-First crawler.

1.2. Our contribution

In order to retrieve web pages in which each web page is satisfactory for user, our crawler (Fig. 1) works as follows:

- Seed URLs of user query are selected by the famous search engine, such as Google, AltaVista, DirectHit, Excite, HotBot, Lycos and Yahoo,



BL: Concept lattice of QueueURLs, CL: Concept lattice of CurrentURLs

Fig. 1: Idea of the Semantical Crawler

etc. After submitting his query, the crawler expands the query, allocates some keywords to these famous search engine by their performance and forms the seed URLs by intersecting their search result.

- We consider these URLs as QueueURLs and Push them into queue, find their URLs (CurrentURLs) that they are link by these URLs of queue.
- Extract all concepts of QueueURLs and CurrentURLs respectively. then, constructs two concept lattices for these concepts, respectively.
- Compute the semantic similarity between each concept of QueueURLs and each concept of CurrentURLs.

2. Concept lattice of URLs

Concept Lattice is systematically built up by Wille R. in 1982. Concept lattice and corresponding Hasse diagram reflect a conceptual hierarchy[6, 7, 8]. It is constructed on formal context. In this section, we propose some notion such as formal context, formal concept, and concept lattice of Seed URLs and Current URLs.

Definition 1. A formal context of ULRs is a tripe $T = (ULRs, W, Q)$, where each url ∈ URLS is interpreted to object, each w ∈ W is interpreted to attribute, W is the common key word set of Web pages identified by url in URLS. URLS and W can not be empty sets. $Q \subseteq ULRs \times W$ is a binary relation, if $(url, w) \in Q$, then it means that url have the attribute w.

When a user submit his query to search engine, this query reflects a concept relating to user knowledge. In natural language, noun can express concept, however the essence of concept can be formalized into objects and attributes of objects. For example, student is a concept, student imply person who study in school and some characteristics which they own name and student-id, etc. In URLs, some Web pages also construct some concept relating to user query. a concept can be formalized into a duality (Objects, attributes), it reflects that these objects in duality take on the common attributes and these attributes only are shared by these objects. To introduce the definition of the formal concept of URLs, we rewrite two set-valued functions, $\uparrow$ and $\downarrow$[6], given by the expressions: $\uparrow$: $P(URLs) \rightarrow P(W), X^{\uparrow} = \{w|w \in W; \forall url \in X, (url, w) \in Q\}$, $\downarrow$: $P(W) \rightarrow P(URLs), Y^{\downarrow} = \{url|url \in G; \forall w \in Y, (url, w) \in Q\}$.

Definition 2. A concept of URLs is a duality $(X, Y) \in P(URLs) \times P(W)$ such that $X^{\uparrow} = Y$ and $Y^{\downarrow} = X$. The set X is called extent of the concept, the set Y intent of the concept.

The greatest concept I and smallest concept O of URLs can be described respectively as follows:

$$\bigvee_{i=1}^{n}(X_i, Y_i) = ((\bigcup_{i=1}^{n} X_i)^{\uparrow\downarrow}, \bigcap_{i=1}^{n} Y_i),$$

$$\bigwedge_{i=1}^{n}(X_i, Y_i) = (\bigcap_{i=1}^{n} X_i, (\bigcup_{i=1}^{n} Y_i)^{\downarrow\uparrow}).$$

Two concept of URLs $(X_1, Y_1) \leq (X_2, Y_2)$ if and only if $X_1 \subseteq X_2$ (or equivalently $Y_2 \supseteq X_1$).

Definition 3. Let C be all concepts of URLs, and $L(T)$ be $(C, O, I, \leq)$, we denote L(T) to a concept lattice of URLs.

Example1. A user query includes two keywords: "Web Page, Spider" be submitted to PISE. PISE selects 5 urls, represented into U_1, U_2, U_3, U_4, U_5, to form URLS of the user query. "Internet, Technology, Network, Web page, Information, Spider" are the key words of Web pages identified by urls. Table 1 is the formal context of this URLS. In this formal context, we extract some formal concepts as follows:
1.($\{U_1, U_2, U_3, U_5\}$, {W});
2.($\{U_2, U_3, U_4, U_5\}$, {T});
3.($\{U_1, U_4, U_5\}$, {Inf, S });
4.($\{U_1, U_2, U_5\}$, {W, S });
5.($\{U_2, U_3, U_4\}$, {T, N});
6.($\{U_2, U_3, U_5\}$, {T, W});

7.({U_1}, {W, Inf, S});
8.({U_2, U_4}, {T, N, S});
9.({U_2, U_5}, {T, W, S});
10.({U_2, U_3}, {Int, T, N, W});
11.({U_4}, {T, N, Inf, S});
12.({U_5}, {T, W, Inf, S});
13.({U_2}, {Int, T, N, W, S });
I.({U_1, U_2, U_3, U_4, U_5}, Φ);
O.(Φ, {Int, T, N, W, Inf, S});
The concept lattice of the URLs is showed in Fig 2.

Table 1: The key word frequency between 5 Web pages and 6 key words

	Int	T	N	W	Inf	S
U_1				√	√	√
U_2	√	√	√	√		√
U_3	√	√	√	√		
U_4		√	√		√	√
U_5		√		√	√	√

Int:Internet, T:Technology, N:Network , W: Web page, Inf: Information, S: Spider

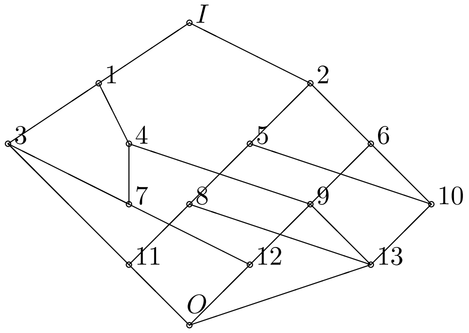


Fig. 2: Galois/concept lattice corresponding to Table 1.

Definition 4(The base concept lattice(BL) and the current concept lattice (CL)). Consider two formal context $T_1 = (QueueURLs, W, Q)$ and $T_2 = (CurrentULRs, W, Q)$. They generate two concept lattices BL and CL for T_1 and T_1, respectively.

3. Semantic rank of URL

Although domain Ontologies and Formal Concept Analysis (FCA) aim at different purposes, they offer a tool of modelling concepts [9, 10]. In real world application, a concept include the extensional and intensional aspects, the extension of the concept is all objects in which each object takes on all attributes of the concept, while the intension of the concept is all attributes in which each attribute is shared commonly by all objects of the concept. Given a formal context (domain), FCA support to formalize domain of interest by a formal pair (objects , attributes), its' concept give an exact description for a realistic concept. Ontologies emphasize on the intensional component to model the domain interest [14]. A domain ontology is a "formal, explicit specification of a shared conceptualization" [15]. A domain ontology contains a set of interrelated concepts, each associated with a formal definition providing an unambiguous meaning of the concept in the given domain [9, 10]. Therefore, a domain ontology should be explained as a set of concepts and their relations among them by a panel of experts in the given domain.

Some references [16] have introduced the approaches combining FCA and ontologies in many applications. The semantic similarity evaluation of two concepts (X_i, Y_i) and (X_j, Y_j), based on a notion of references [9, 10], includes not only the extension of two concepts, but also their intention. [9] considers a domain ontology ϑ and defines the concept similarity (Sim) of two concepts $C_1 = (X_1, Y_1)$ and $C_2 = (X_2, Y_2)$ of the different concept lattices. Let $n = |Y_1|$, $m = |Y_2|$, and suppose that $n \leq m$. The set $\mathrm{P}(Y_1, Y_2)$ is defined by all possible sets of n pairs of attributes $\mathrm{P}(Y_1, Y_2) = \{\{< a_1, b_1 >, \cdots, < a_n, b_n >\} | a_i \in Y_1, b_i \in Y_2, \forall i = 1, \cdots, n, and\, a_i \neq a_k, b_i \neq b_l, \forall k, l \neq i\}$. The $Sim(C_1, C_2)$ is defined as follows: $Sim(C_1, C_2) = \frac{|X_1 \cap X_2|}{r} * w + [\frac{1}{m} \max_{P \in \mathrm{P}(Y_1,Y_2)}(\sum_{<a,b> \in P} as(a, b))] * (1 - w)$.

3.1. Extension similarity degree of two concepts

In Web, A hyperlink, from Web page A to Web page B, is supposed as " Web page A and B might be on same topic" or " the author of Web page A recommend Web page B to the user "[12]. It implies an axiomatic semantic relations between Web page A and B. Page rank reflects the evaluations (out-degree and in-degree of Web page) of authors, these authors are the experts of the different domain. On the other hand, A clicking, from Web page A to Web page B, is supposed as "the user approve that Web page A and B might be on same topic". A clicking implies also an axiomatic semantic relations between Web page A and B. It indicates that the intension of concepts (topic) of Web page A is the similar as one of Web page B. This semantic relation denotes the ontology similarity relation. The user Web log of search engine recorded the abundant history data (click-data, browse-time, keywords, etc.) of user. In fact, the knowledge of the group users of the same interest is very outstanding knowledge of the special domain, the users are the most fine experts. In this paper, we defined this click-data of the same interest of the users as the user domain. To consider two concepts $(X_i, Y_i) \in BL$ and $(X_j, Y_j) \in CL$, the hyperlink of Url in X_i and Url in X_j reflects the semantic relation of their extents; the click-data of Url in X_i and Url in X_j reflects the semantic relation of their extents too. Because they can not share objects between concepts of BL and CL, however note that there exist some linkages and click-datas among URLs in BL and CL, they reflect the semantic distance of extensions of two concept in BL and CL respectively.

Definition 5(Extension Similarity Degree) Consider two concept lattice BL and CL, $(X_i, Y_i) \in BL$ and $(X_j, Y_j) \in CL$, let $Sim_{extension}$ be the extension similarity between (X_i, Y_i) and (X_j, Y_j). $Sim_{extension} = \frac{|X_i \rightarrow_L X_j| + |X_i \leftarrow_L X_j| + |X_i \rightarrow_C X_j| + |X_i \leftarrow_C X_j|}{4 * max(|X_i|, |X_j|)}$.

where $|X_i \rightarrow_L X_j|$ and $|X_i \rightarrow_C X_j|$ are the number of hyperlinks in which URL in BL link to URL in CL and clicks from url of CL to URL in BL, respectively, and $|X_i \leftarrow_L X_j|$ and $|X_i \leftarrow_C X_j|$ are the number of hyperlinks in which URL in CL link to URL in BL and clicks from url of BL to URL in CL, respectively.

3.2. Intension similarity degree of two concepts

In order to compute the semantic similarity of intensions of two concept, in a predefine domain ontology, the similarity degrees for any pair of concept descriptors should be contained. Reference[10] replaces axiomatically the similarity degrees with information content similarity scores that they can be automatically computed by any lexical database (such as Wordnet). The computing rely independently on domain expertise, it is convenient to program for a real application. In lexical database for the English nouns, the relationships among nouns such as ISA, PartOf, etc. are appointed by some linguists or other specialist on given domain. However, for the personal Web spiders, they retrieve timely Web pages from Internet by user queries. And so, the domain ontology determining the semantic relation of two concepts of CL and BL should rely on the history knowledge of the user. On the other hand, The user Web log of search engine offers the abundant history information, we make use of these information to measure the similarity of concept descriptors (attributes). To compute the information content similarity, we discuss the semantic relation among nouns as follows:

- If the different users of the same user group submit the different key words, they click the same Web pages, then we suppose that these key words own the semantic relationships ISA.
- If a user submits the different key words, he click the same Web pages, then we suppose that these key words own the semantic relationships ISA.
- If the different users of the same user group submit the different key words, they click these Web pages in which they exist hyperlinks, then we suppose that these key words own the semantic relationships Partof.
- If a user submits the different key words, he click these Web pages they exist hyperlinks, then we suppose that these key words own the semantic relationships Partof.

According to the four case above, if relation of two keywords belongs to ISA and Part, then we choice that their relation is ISA.

Definition 6(Lexical Nouns Database for the BL and CL). A lexical database (Ω) for the BL and CL is 4-tuple $(N_{BL}, N_{CL}, f(N), R)$, where N_{BL}, N_{CL} are the set of key words in which each one is attribute of formal context of BL, CL respectively, $f(N)$ is function from N_{BL} or N_{CL} to the positive integers which the value represents click-numbers of Web pages after submitting key word to Web spider in user Web log. R is a set of relationships between N_{BL} and N_{CL}(such as ISA and Partof).

For example, consider our Web log of PISE for user query: "internat spider", $N_{BL} = \{$ $internet$, $spider$ $\}$, $N_{CL} = \{$ $internet$, $technology$, $network$, $Webpage$, $information$, $spider$ $\}$. The part of Web log for User query:internat spider is listed in Table 2, $f(Internet)$=32510, ..., $f(Network)$=43891. $R = \{$ $ISA(Spider, Spider)$, $ISA(Internet, Internet)$, $ISA(Internet, Spider)$, $ISA(Technology, Network)$, $ISA(Webpage, Information)$, $Partof(Internet, Technology)$, $Partof(Internet, Network)$, $Partof(Spider, Information)$, $Partof(Spider, Webpage)$ $\}$.

The weighted hierarchy of the information content approach [18, 17] have not conceived for the

Table 2: A fragment of Web log of User query:internat spider

Key word	click-numbers	Click-URLs
Internet	3251	$\{U_1, U_3, ...\}$
Technology	2458	$\{U_3, U_4, ...\}$
Webpage	12983	$\{U_2, U_6, ...\}$
Information	67856	$\{U_2, U_6, ...\}$
Spider	565	$\{U_1, U_2, ...\}$
Network	4389	$\{U_3, U_5, ...\}$

Hyperlinks: $U_1 \rightarrow U_2$, $U_1 \rightarrow U_3$, ...

PartOf relationship, only for the ISA relationship. In order to compute the information content similarity, reference[10] focus on discussing the the ISA relationship. These notion of the weighted hierarchy is constructed on the probability of a concept noun n of every node. In this paper, we consider that the ISA and Partof relationships are then important and frequent semantic relationship, it allows us to express the notion of the Weighted ISA and Partof hierarchy.

Definition 7(Weighted ISA and Partof hierarchy). Given a lexical database (Ω) for the BL and CL, let ∂ be the ISA and Partof hierarchy, ∂ is a direct Graph. For these nodes(keywords) in the same layer, they reflect ISA relationships among these keywords. In the other hands, for these nodes(keywords) in the different layers, which they exists connecting paths, they reflect Partof relationships among these keywords. The probability of every keyword is computed as:

$$p(n) = \frac{f(N)}{\sum_{i=1}^{M} f(N)}$$

, where M is the total number of keywords in Lexical Nouns Database for the BL and CL, in our Web log of PISE, $\sum_{i=1}^{M} f(N) = 87482$. The connecting weight $W_{ISA}(n_1, n_2)$ of two keywords n_1, n_2 attaching to them for ISA relationships is 1.0, the connecting weight W_{Partof} of two adjacent keywords n_1, n_2 attaching to them for Partof relationships is defined as:

$$W_{Partof}(n_1, n_2) = \frac{1}{Max(path(n_1), path(n_2)) + \alpha}$$

, where path(n) is the node number of the maximum path of n from the root node to n, $\alpha \leq 0.5$ is a adjustment factor of the ISA and Partof hierarchy.

In user Web log, we only consider the Partof relationships from N_{BL} to N_{CL}, these Partof relationships from N_{CL} to N_{BL} are omitted. The weighted ISA and Partof hierarchy is a direct tree. The ISA and Partof hierarchy has a unique Top node-the keyword, the Top node is discussed as follows:

1. If the user submits a keyword, the keyword has not keywords that they keep ISA relation in the given Lexical Nouns Database for the BL and CL, then the Top node is the keyword.
2. If the user submits a keyword, the keyword has keywords that they keep ISA relation in the given Lexical Nouns Database for the BL and CL, we assume that the ISA and Partof hierarchy has the Top node (the most general keyword), it keeps the Partof relation with this keyword, let $p(Top) = max(p(N))$.
3. If the user submits some keywords, these keywords keep the ISA relations, then we assume that the ISA and Partof hierarchy has the Top node (the most general keyword), it keeps the Partof relation with these keywords, let $p(Top) = max(p(N))$.

For example, A fragment of the weighted ISA and partof hierarchy derived from Web log of User query: internat spider (Table 2) is shown in Fig. 3.

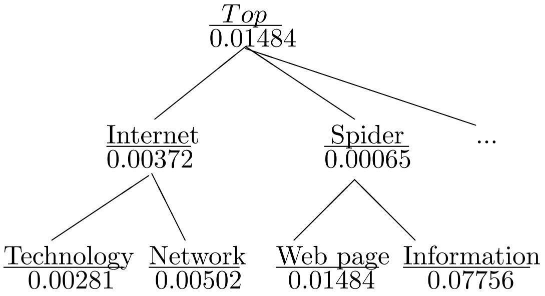


Fig. 3: A fragment of the weighted ISA and Partof hierarchy.

Definition 8(Information content similarity (ics)). Consider a lexical database (Ω) for the BL and CL, let ∂ be the weighted ISA and Partof hierarchy, two keywords $n_1 \in N_{BL}$, $n_2 \in N_{CL}$, in order to consider Partof relation n_1 and n_2 to their shared node, we revise The information content similarity $ics(n_1, n_2)$ of n_1, n_2 as follows:

$$ics(n_1, n_2) = \frac{2log^{p(n')}}{log^{p(n_1)} * w_1 + log^{p(n_2)} * w_2}$$

, where n' is a keyword providing the maximum information content shared by n_1, n_2. Let w_1 and

w_2 are the weight of 'n' to n_1' and 'n' to n_2', respectively, $w_1 = \sum_{n=n_1;m,n\in N_1} W_{Partof}(m,n)$, $w_2 = \sum_{n=n_2;m,n\in N_2} W_{Partof}(m,n)$, N_1 and N_2 are the set of keywords in which each key word belongs to the nodes of the maximum path from n' to n_1 and n_2, respectively. m is a keyword that it exists a Partof relationship with n.

For continuous example: (1). Consider $n_1 = network$ and $n_2 = network$, the maximum information content shared by n_1, n_2 is $network$,

$$ics(network, network) = 1.$$

(2). Consider $n_1 = Internet$ and $n_2 = Webpage$, the maximum information content shared by n_1, n_2 is Top, let $\alpha = 0.2$, $ics(Internet, Webpage) = \frac{2*log^{p(Top)}}{log^{p(Internet)}*w_1+log^{p(Webpage)}*w_2} = \frac{2*log^{0.01484}}{log^{0.00372}*\frac{1}{1+0.2}+log^{0.01484}*(\frac{1}{1+0.2}+\frac{1}{2+0.2})} = 0.8351.$

Definition 9(Intension Similarity Degree)Consider two concept lattice BL and CL, $(X_1, Y_1) \in BL$ and $(X_2, Y_2) \in CL$, let $Sim_{intension}$ be the intension similarity between (X_1, Y_1) and (X_2, Y_2). $Sim_{intension} = \frac{1}{n}\max_{P\in \mathrm{P}(Y_1,Y_2)}(\sum_{<a,b>\in P} ics(a,b))$.

3.3. The semantic Rank of Web Pages in Concept Lattice CL

Definition 10 Consider two concept lattice BL and CL, $(X_1, Y_1) \in BL$ and $(X_2, Y_2) \in CL$. The concept similarity (Sim) between (X_1, Y_1) and (X_2, Y_2) include two part: the extension and Intension Similarity. We define it as follows: $Sim((X_1, Y_1), (X_2, Y_2)) = Sim_{extension} * w + Sim_{intension} * (1 - w)$. Where $w \in [0, 1]$ is a weight which it is a proportion the extension in the whole concept. w is defined as $Sim_{extension}/(Sim_{extension} + Sim_{intension})$.

For a given Web page (URL $\in CL$), it's concept set Cp is all concept in which the objects of each concept contains the URL and each concept belong to CL. According to user query, Each URL in SEED URLs can stand for user requirement, these concepts of BL are derived from the SEED URLs, step by step. The semantic similarity degree of the URL $\in CL$ with BL reflects the semantic relation between the URL and user query Q. It allows us to define the semantic similarity degree of a concept in CL as follows:

Definition 11(The semantic similarity degree of the concept in CL). Consider two concept lattice BL and CL, $(X_1, Y_1) \in CL$. Let $Sim((X_1, Y_1), BL)$ be the sematic similarity degree of the concept (X_1, Y_1). $Sim((X_1, Y_1), BL) = \max_{(X,Y)\in BL} sim((X, Y), (X_1, Y_1))$.

In order to make conveniently choice Web pages for next steps, each Web pages in current URLs should be assigned a rank. The semantic rank of the Web page (URL) in concept lattice CL is computed by as follows formulary.

Definition 12 Consider two concept lattice BL and CL, $(X_1, Y_1) \in CL$. A URL is an object of X_1, let $SemRank(URL)$ be the semantic rank of the URL. $SemRank(URL) = \max_{(X_1,Y_1)\in Cp} Sim((X_1, Y_1), BL))$.

Acknowledgement

This work forms part of projects supported by the Cultivating Foundation of the Science and Technology Leader of Sichuan Province, the Application Foundation of Science& Technology Bureau Sichuan Province (2006J 13-056), the Science & Technology Foundation of Xihua University (R0622611), and the Nature Science Foundation of Education of Sichuan Province(2006A086).

References

[1] J. Cho, H. Garcia-Molina, L. Page, Efficient crawling through URL ordering. Computer Networks, 30(1-7): 161-172, 1998.

[2] Y. J. Du, H. M. Li, Z. Pei., H. Peng., Intelligent Spider¡¯s Algorithm of Search Engine Based on Keyword. *Ecti Transactions on Computer and Information Theory*, 01(01): 40-49, 2005.

[3] M.M. Sufyan Beg. A subjective measure of web search quality. *Information Sciences*, 169: 365¨C381, 2005.

[4] A. Rungsawang, N. Angkawattanawit. Learnable topic-specific web crawler. *Journal of Network and Computer Applications*, 28: 97-114, 2005.

[5] H.Y. Liu, J. Janssen, E. Milios. Using HMM to learn user browsing patterns for focused Web crawling. *Data & Knowledge Engineering*, 59: 270¨C291, 2006.

[6] R. Wille, *Restructuring the Lattice Theory: an Approach Based on Hierarchies of Concepts. Rival, I.(Ed.): Ordered Sets,* Reidel, Dordrecht, Boston, 445-470, 1982.

[7] R. Wille, *Lattices in Data Analysis: How to Draw Them with a Computer*, In: Algorithms and order, Kluwer Acad. Publ., Dordrecht, 1989.

[8] R. Wille. Concepe Lattices and Conceptual Knowledge Systems. *Comput. Math. Apll*, 23(6-9):493-515, 1992.

[9] A. Formica. Ontology-Based Concept Similarity in Formal Concept Analysis. *Information Sciences*, 176(18): 2624-2641, 2006.

[10] A. Formica. Concept similarity in Formal Concept Analysis: An information content approach. *Knowledge-Based Systems*, doi:10.1016/j.knosys, 2007.

[11] R. K. Rajapakse, M. Denham. Text retrieval with more realistic concept matching and reinforcement learning. *Information Processing and Management*, 42(5): 1260-1275, 2006.

[12] S. Brin, L. Page. The anatomy of a large-scale hypertextual Web search engine. *In: Thistlewaite P, et al. eds. Prof. the 7th ACM-WWW International Conference. Brisbane: ACM Press*, pp.107-117, 1998.

[13] M. Jones, H. Alani. Content-based Ontology Ranking. *Prof. the 9th International Protege Conference Ranking,* 2006.

[14] H. Uschold, M. Gruninger. Ontogies: Principles, Methods and Applications. *The Konwledge Engineering Review*, 11(2), 1996.

[15] Y. Ding, D. Fensel, M. Klein, B. Omelayenko. The Semantic Web: yet another hip . *Data and Knowledge Engineering*, 41(2-3):205-227, 2002.

[16] M. Bain, Inductive construction of ontologies from Formal Concept Analysis. *Australian Conference on Artificial Intelligence,* pp. 88-99, 2003.

[17] D. Lin, An Information-Theoretic Definition of Similarity, in: *Prof. the International Conference on Machine Learning, Morgan Kaufmann, Madison, Wisconsin, USA*, pp. 296-304, 1998.

[18] P. Resnik, Using Information Content to Evaluate Semantic Similarity in a Taxonomy, in: *Prof. the International Joint Conference on Artificial Intelligence, Montreal, Quebec, Canada, Morgan Kaufmann*, pp. 448-453, 1995.

[19] Y. J. Du, B. Yan, L. Song, Design of crawler¡¯s algorithm and Implement of crawler¡¯s program. *Journal of Computer Applications*, 24(1):33-35, 2004.