CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. (2010)
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/cpe.1624

Automatically constructing semantic link network on documents

Hai Zhuge[1,2,*,†] and Junsheng Zhang[1]

[1]*Knowledge Grid Research Group, Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, 100190 Beijing, People's Republic of China*
[2]*Southwest University, People's Republic of China*

SUMMARY

Knowing semantic links among resources is the basis of realizing machine intelligence over large-scale resources. Discovering semantic links among resources with limited human interference is a challenge issue. This paper proposes an approach to automatically discovering and predicting semantic links in a document set based on a model of document semantic link network (SLN). The approach has the following advantages: it supports probabilistic relational reasoning; SLNs and the relevant rules automatically evolve; and, it can adapt to the update of the adopted techniques. The approach can support cyber space applications, such as documentation recommendation and relational queries, on large documents. Copyright © 2010 John Wiley & Sons, Ltd.

Received 6 May 2010; Accepted 16 May 2010

KEY WORDS: semantic link network; probability; rules; relational reasoning; inference

1. INTRODUCTION

1.1. Motivation

Knowing relations between resources is important for realizing intelligent applications on large-scale resources. Manually establishing relations is time-consuming; hence, it is necessary to study the approach to automatically discover relations among resources.

*Correspondence to: Hai Zhuge, China Knowledge Grid Research Group, Key Lab of Intelligent Information Processing, Institute of Computing Technology, P. O. Box 2704-28, 100190 Beijing, People's Republic of China.
†E-mail: zhuge@ict.ac.cn

Contract/grant sponsor: National Basic Research and Development Program; contract/grant number: 2003CB317001
Contract/grant sponsor: International Cooperation Project of Ministry of Science and Technology of China; contract/grant number: 2006DFA11970
Contract/grant sponsor: National High Technology Research and Development Program of China; contract/grant number: 2007AA12Z220
Contract/grant sponsor: National Science Foundation of China; contract/grant numbers: 60773057, 60703018

Rethinking the success of the World Wide Web indicates the right way to the future semantic Web: inheriting the features of the Web—the simple hyperlink mechanism and the easy utility mode.

The World Wide Web is a network of references implemented by hyperlink. The content of a web page is explained by its text and the reference pages. The hyperlink represents the relevant topics of the page, whereas it does not represent the relation between pages explicitly. Therefore, the hyperlinked Web cannot recommend useful web pages.

The semantic link network (SLN) model extends the Web by attaching semantic indicators to hyperlinks [1]. A typical SLN consists of semantic nodes, semantic links and relational reasoning rules. A semantic node can be any type of resource, event, thought or even an SLN [2]. Potential semantic links can be derived from an existing SLN according to a set of reasoning rules. Adding a semantic link to the network could derive out new semantic links. The major advantages of the SLN are its simplicity, the ability of relational reasoning, and the nature of semantic self-organization: *any node can link to any semantically relevant node*.

The SLN model has the characteristics of autonomy and evolution. An easy construction approach can promote the adoptability of the SLN and facilitate its applications. This paper focuses on automatic discovery of semantic links among documents, which is useful in many applications. We use a probabilistic SLN model to represent, discover and predict probabilistic semantic links between documents. The approach integrates existing technologies and needs limited user interference. The model has the following characteristics: *loosely coupled*, *automatic*, *autonomous*, *self-evolutionary* and *extensible*. The SLN of documents supports applications such as document recommendation, document space reconstruction and relation-based retrieval.

1.2. Related works

Relations between two documents are related with their metadata, contents and references. The metadata describes the attributes of a document such as author, publishing time and venue; the contents of document include title, author(s), institution(s), abstract, body text and appendices; and the references show the relations to the existing documents.

Automatic discovery of semantic links on document contents needs technologies such as text analysis, document clustering and classification. To cluster and classify documents, document similarity needs to be measured by considering body text, anchor text and hyperlinks [3]. An approach to measuring document pair similarity in high-dimensional spaces was proposed [4]. Human-created metadata for Web directories was used to estimate semantic similarity and semantic maps to visualize relations between content and cues and what these cues suggest about page meaning [5].

Resource attributes and related classifications form a hierarchical structured network. Similarity between resources can be calculated according to their hierarchical structured networks, and links among resources can be used to cluster resources [6].

An approach to automatic categorization of documents was proposed for exploiting contextual information extracted according to the analysis of HTML structure of Web documents as well as the topology of the Web [7]. Hyperlinks between web pages have been used to classify web pages [8].

Prediction of semantic links in a network of documents has drawn many attentions. An approach to predicting links in graph was proposed based on the length of cycles in networks [9]. Exact inference is generally intractable and approximate inference techniques are necessary [8]. A joint probabilistic model was proposed to predict semantic links among documents from contents and link structure of documents [10].

DOI: 10.1002/cpe

SLN of documents can support relational query and query expansion. Graph model was used to expand the query for answering relational queries on entities [11]. To discover the relations between two entities on the Web, document pairs containing potential semantic links are ranked based on the connecting words between documents [12].

The SLN has emerging characteristics [13]. When a new semantic link is added, more semantic links may be derived out by reasoning rules and inference rules on the whole network. Besides, if semantic links are considered, communities in the SLN can be discovered to support advanced applications on the SLN. SLN is adopted as the self-organized semantic model to implement the interactive semantics [14]. The SLN model is systematically introduced in the 2nd edition of the book Knowledge Grid [15].

Semantic associations between elements of RDF graphs were studied and the approaches to searching and ranking semantic path in RDF graph were proposed [16, 17]. Existing data models, such as RDF (www.w3.org/RDF/) and OWL (www.w3.org/2004/OWL/), focus on the approach to representing relations. Reasoning link between the linked data is based on Description Logic, RuleML or SWRL (www.w3.org/Submission/SWRL/). As a knowledge representation model, semantic network represents relations, such as *IS-A* and *Part-Of*, among entities and attributes [18].

2. DOCUMENT SLN

A document semantic link network (SLN-D) consists of the following components:

(1) Document set $D=\{d_1, \ldots, d_x\}$.
(2) Document cluster set $C=\{c_1, \ldots, c_n\}$, where $c_i \in C (1 \leq i \leq n)$ is a cluster of similar documents. Clustering on clusters form a hierarchical network of clusters. During the initialization stage of its formation, documents are clustered into document clusters; then, clusters are clustered into higher clusters. Document clusters may not be independent, and a document could belong to a document cluster with a certain probability; hence, a document can belong to different clusters with different probabilities.
(3) Semantic link set $SL=\{s_1[l_1, u_1], s_2[l_2, u_2], \ldots, s_m[l_m, u_m]\}$ where $s_i (1 \leq i \leq m)$ are semantic indicators between documents, between keywords, between document and cluster or between clusters; $l_i \in (0, 1]$ is the lower-bounded probability of s_i, $u_i \in (0, 1]$ is the upper-bounded probability of s_i. Semantic links can be assigned by users, predicted by inference rules or derived by reasoning rules. $[l_i, u_i]$ represents the minimum probability and the maximum probability that a semantic link can be derived from different semantic paths.
(4) Keyword set of D, $T=\{t_1, t_2, \ldots, t_q\}$, where $1 \leq i \leq q$. Each document or document cluster has its corresponding keyword set.
(5) Rule set RULES $=\{LR, IR, CR, AR\}$,
   - *LR* is a set of relational reasoning rules, each of which takes the following form: $\alpha[l_\alpha, u_\alpha] \times \beta[l_\beta, u_\beta] \xrightarrow{cd} \gamma[l_\alpha l_\beta, u_\alpha u_\beta]$, where *cd* is the *certainty degree* of the rule. The SLN model is equipped with a basic set of rules. Users can define more domain-specific rules.
   - Inference rule set *IR* consists of rules of two types: statistical inference rules and assertion rules of semantic links according to metadata on documents and citations among documents.

DOI: 10.1002/cpe

Table I. Comparison between typical SLN and SLN-D.

Components	Typical SLN	SLN-D	Explanation
Semantic nodes	A, B, C	A, B, C	A node can be any type of resource, a concept, a document or even an SLN
Semantic links	$A—\alpha\rightarrow B$	$A—\alpha[l_\alpha, u_\alpha]\rightarrow B$	l_α and u_α are lower-bounded probability and upper-bounded probability of semantic link α
Reasoning rules	$\alpha\times\beta\rightarrow\gamma$	$\alpha[l_\alpha, u_\alpha]\times\beta[l_\beta, u_\beta] \xrightarrow{cd} \gamma[l_\alpha l_\beta, u_\alpha u_\beta]$	α and β are semantic links; l_α and l_β are the lower-bounded probability values of α and β; u_α and u_β are the upper-bounded probability values of α and β; *cd* is the *certainty degree* of the rule
Classification rules		$p(c\|t_i)$	The probability of a keyword t_i in a document belongs to a cluster c
Inference rules		$src—r[l_r, u_r]\rightarrow tgt$	l_r and u_r are the lower-bounded probability and the upper-bounded probability of semantic link r between clusters *srt* and *tgt*

- Classification rule set *CR* consists of rules for classifying documents according to the probable relations between keywords, documents and document clusters.
- Attribute rule set *AR* consists of rules for discovering semantic links between documents according to the attributes of documents. The attributes of documents are regarded as the metadata of documents.

Document clustering technologies support the formulation of document clusters. During the evolutionary stage, documents are classified by the classification rules. The SLN-D extends the typical SLN by adding the probability ranges for the uncertain semantic links.

Table I shows the difference between the typical SLN [15] and the SLN-D. The probabilistic values of semantic links take part in semantic link reasoning. The document classification rules are for classifying documents and discovering semantic links. The semantic link inference rules are for predicting the semantic links.

An SLN-D consists of the following parts:

- *Cluster–document network* consists of document clusters and document entities. The *instanceOf* link exists between a document and a cluster. The *equal*, *similar* and *subCluster/partOf* links exist between clusters and between documents.
- *Citation network* mainly consists of *refer* link, which can be refined as *cocite*, *cocited* and *sequential* links.
- *Metadata network* consists of such relations as *sequential* and *equal* links between document attributes. The *equal* link can be specialized as *sameAuthor*, *sameJournal* and *sameConference*. Document metadata and citation relations between documents can be extracted from relevant digital libraries.
- *Keyword–document network* consists of the *occur* links between keywords and documents. The probability of *occur* link is the weight of the keyword in the document. The *co-occur* link exists between a pair of keywords if they appeared in the same document. Counting the co-occurrence of keywords in the documents can help find the co-occurrence associations between keywords.

DOI: 10.1002/cpe

- *Document classification rules* classify new documents and insert them into the SLN-D. This type of rules can be obtained from the cluster networks and the keyword–document networks according to the *Bayesian* formula.
- *Semantic link inference rules* predict semantic links between documents according to the semantic links between the existing document clusters. This type of rules is acquired by statistical method from the existing SLN.
- *Attribute rules* reflect the relationship between two kinds of semantic links: one kind is between attributes of documents, the other kind is between documents.
- *Relational reasoning rules* reflect the relationship between semantic links. A general reasoning rule can be specialized in different domains. For example, the *equal* link can be explained as the *sameTopic* link; the *partOf* link can be explained as the *subSite* link. SLN supports relational reasoning in the following form: R_1—$\alpha[l_\alpha, u_\alpha] \rightarrow R_2$, R_2—$\beta[l_\beta, u_\beta] \rightarrow R_3 \Rightarrow_{cd}$ R_1—$\gamma[l_\gamma, u_\gamma] \rightarrow R_3$, where R_1, R_2 and R_3 are resources, α, β and γ are semantic indicators, and $[l_\gamma, u_\gamma] = f([l_\alpha, u_\alpha], [l_\beta, u_\beta], cd)$, e.g. $f([l_\alpha, u_\alpha], [l_\beta, u_\beta], p) = [cd \cdot l_\alpha \cdot l_\beta,\ cd \cdot u_\alpha \cdot u_\beta]$.

Document cluster network, keyword–document, citation network and metadata network are the sub-networks of SLN-D. Two or more sub-networks can be merged into a larger one. The sub-networks contain the semantic links between resources in the SLN-D. Semantic links between documents are determined by document contents or document metadata.

Figure 1 shows the SLN-D model and its construction and evolution mechanism. Besides the embedded semantic link reasoning rules, the SLN-D model requires neither any background knowledge nor ontology. When new documents are inserted into the SLN-D, semantic links between the new document and the existing nodes are inferred by the inference rules and the reasoning rules. Then, the statistical inference rules of semantic links are updated, the keywords of document clusters change and the document classification rules are also recalculated.

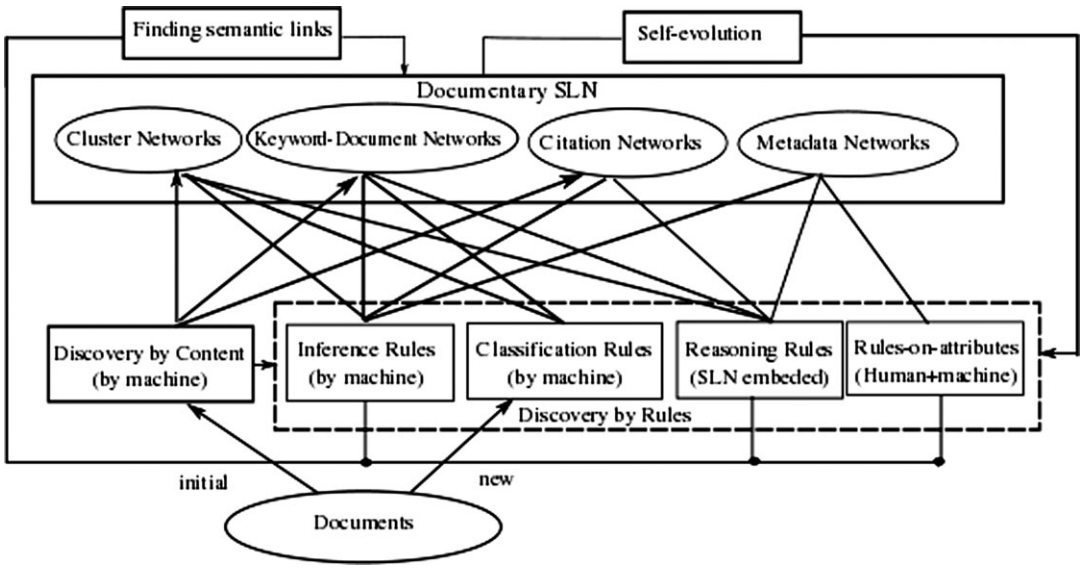

Figure 1. The SLN-D discovery and evolution mechanism.

DOI: 10.1002/cpe

The SLN-D has the following characteristics:

(1) *Loosely coupled*. Each node can link to any other semantically related node via a semantic link; hence, different SLN-Ds can be merged by the shared nodes.
(2) *Self-evolutionary*. It automatically adjusts document classification rules, keyword sets and inference rules.
(3) *Autonomous.* It derives more semantic links by reasoning rules, and predicts the semantic links among documents according to the statistical inference rules.
(4) *Automatic*. It initiates from a document set without any predefined ontology, and only needs a little user interference such as assigning reasoning rules and assertion rules.
(5) *Extensible*. Relationship types and reasoning rules can be extended.
(6) *Relational reasoning and relation prediction*. SLN-D has the ability of relational reasoning based on the embedded reasoning rules and the relational prediction ability based on the statistics on the semantic links in SLN-D.

3. CONTENT-BASED SEMANTIC LINK DISCOVERY

This stage includes finding keywords of documents, clustering documents hierarchically, building keyword co-occurrence network and discovering semantic links between documents, clusters and keywords.

3.1. Document analysis

This step needs to determine the document source, input documents, carry out text analysis and build keyword–document networks and keyword co-occurrences network. The open source software package *wvtool* (sourceforge.net/projects/wvtool) is used to find keywords. The *wvtool* (word vector tool) is a simple but flexible Java library to create word vector representations of text documents.

The co-occurrence networks of keywords can be built as follows. If two keywords co-occur in a document, then add a *co-occur* link between the two keywords. The number of *co-occur* link is recorded, and all the link weights are normalized by the largest *co-occur* link weight. The co-occurrence network of keywords shows semantic associations between keywords. By setting the threshold of link weights, the co-occurrence network of keywords can be partitioned. Keywords in the same connected component are relevant and can be used to expand query keywords.

3.2. Pairwise document similarity calculation

Document similarity calculation is the basis of document clustering and classification. Document similarity can be calculated by text analysis on documents. Figure 2 shows two documents including their keywords and connecting words.

According to the comparison of different similarity measures [19], *cosine* and *Jaccard* similarities can be adopted. The cosine similarity considers both the common keywords and their weights, whereas the Jaccard similarity only considers the overlap of two keyword sets. During the document pair similarity calculation, it is necessary to combine the two similarity calculation approaches. According to TF-IDF [20, 21], a document can be represented as a vector $\{(t_1, \omega_1), (t_2, \omega_2), \ldots, (t_m, \omega_m)\}$, where $t_i (1 \leq i \leq m)$ is a keyword and $\omega_i (1 \leq i \leq m)$ is its weight. The similarity between documents d_1 and d_2 can be measured as follows:

DOI: 10.1002/cpe

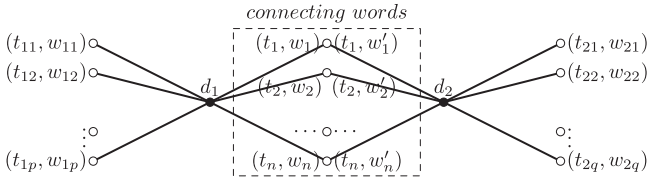

Figure 2. Two documents and their connecting words. The *occur* link exists between keywords and documents. The *co-occur* link also exists between keywords of the same document.

- The cosine similarity *sim_vec* is measured by

$$sim\_vec = \frac{\sum_{i=1}^{n} \omega_i \omega'_i}{\sqrt{\sum_{i=1}^{n} (\omega_i)^2 \sum_{i=1}^{n} (\omega'_1)^2}},$$

  where n is the number of common keywords between d_1 and d_2.
- The Jaccard similarity *sim_content* is measured by the keyword sets of documents. Suppose keyword sets of d_1 and d_2 are $S(d_1)$ and $S(d_2)$, respectively, $S(d_1) = (t_1, t_2, \ldots, t_m)$, and $S(d_2) = (t_1, t_2, \ldots, t_n)$, where $m \geq 0$ and $n \geq 0$, then

$$sim\_content = \frac{|S(d_1) \cap S(d_2)|}{|S(d_1) \cup S(d_2)|}.$$

- The similarity between d_1 and d_2 can be measured by combining *sim_vec* and *sim_content* as follows for clustering documents:

$$sim(d_1, d_2) = \alpha \times sim\_vec + (1 - \alpha) \times sim\_content, \quad \text{where } \alpha \geq 0.$$

  During the pairwise document similarity calculation, *occur* links are built between keywords and documents, and *co-occur* links are built between keywords.

3.3. Hierarchical document clustering

The single-pass clustering algorithm that processes documents sequentially and compares each document to all existing clusters can be used to cluster documents. Usually the method for determining the similarity between a document and a cluster is done by computing the average similarity of the document to all documents in that cluster.

A document *cluster* can be represented by the union of the keyword sets of documents sharing keywords. The mean of document vectors can be used as the *cluster vector*.

Document clusters form a hierarchical SLN. *InstanceOf* and *subCluster* are two important semantic links in the hierarchical structure of documental networks. Each pair of clusters at the same level are *similar*, and *subCluster* links may exist between clusters of different levels. Documents

DOI: 10.1002/cpe

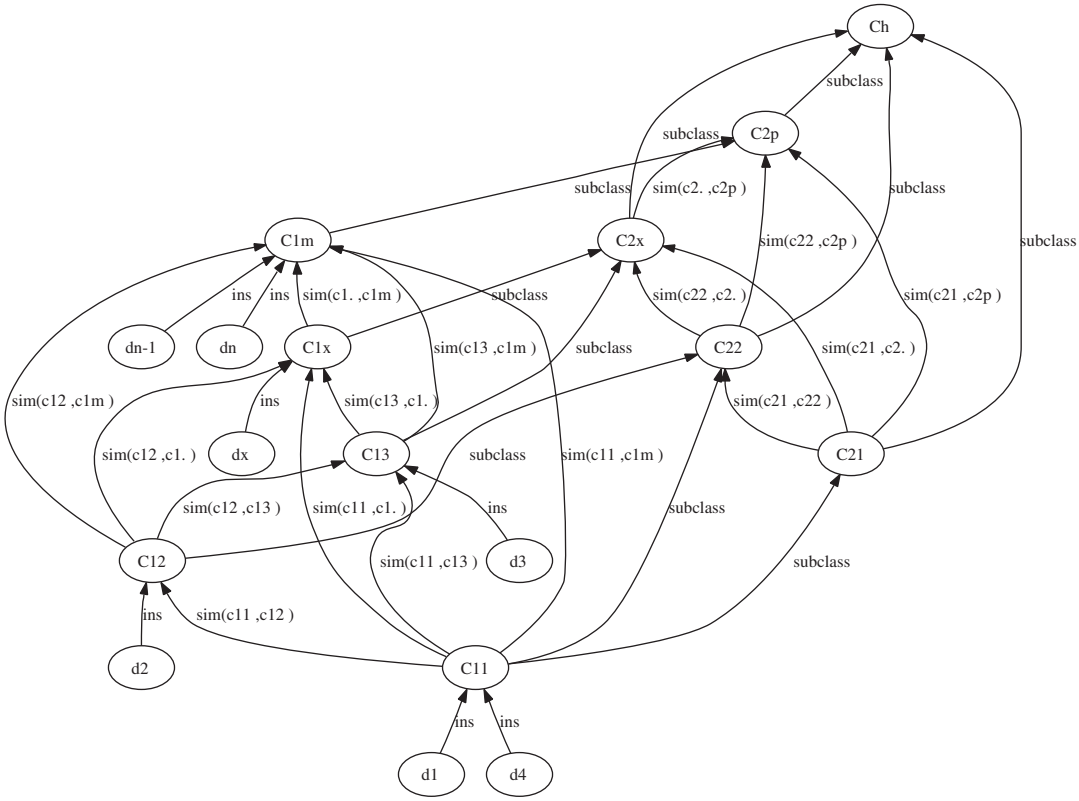

Figure 3. The hierarchical document cluster network consists of the *instanceOf* link between documents and clusters, the *association* link between clusters at the same level and the *subCluster* link between the clusters at different levels.

are clustered iteratively until one cluster contains all of the documents or the iteration time exceeds the maximum iteration time predefined.

Figure 3 shows an SLN segment of document clusters, where d_x represents document, c_{1x} and c_{2x} represent the clusters at levels 1 and 2, respectively, and c_h represents the cluster networks that are iteratively clustered h times into one cluster. The *subCluster* links between clusters exist at different levels, whereas the *ins* (*instanceOf*) links are between documents and clusters. Clusters at the same level have semantic associations, and the association values are the similarity values between clusters such as $sim(c_{1x}, c_{1y})$ and $sim(c_{2x}, c_{2y})$.

Each document cluster has a keyword set. Similar to the global keyword co-occurrence network, each document cluster has a local keyword co-occurrence network recording the co-occurrence of keywords in the documents in the same cluster. Global keyword co-occurrence networks can expand query keywords globally in the whole SLN-D, whereas local keyword co-occurrence networks can expand query keywords in the specified document clusters. The search results are different because of different query keywords expanded by the global and local keyword co-occurrence networks.

DOI: 10.1002/cpe

Table II. Discovering semantic links according to the keyword sets.

Semantic link	Characteristics
irrelevant	$\|T_1 \cap T_2\| = 0$
similar	$0 < \|T_1 \cap T_2\| < min(\|T_1\|, \|T_2\|)$, the similarity can be defined as $\|T_1 \cap T_2\| / \|T_1 \cup T_2\|$
partOf	$\|T_1 \cap T_2\| = min(\|T_1\|, \|T_2\|) < max(\|T_1\|, \|T_2\|)$
equal	$T_1 = T_2$

After documents are clustered, some semantic links between documents, semantic associations between document clusters, semantic associations between keywords, and semantic links between documents and clusters can be discovered and inserted into the SLN-D.

3.4. Discovering semantic links from text contents

If T_1 and T_2 are keyword sets of documents d_1 and d_2 in the same cluster, semantic links between them such as *similar*, *partOf* and *equal* can be discovered according to the rules in Table II.

Let $T(c_1)$ and $T(c_2)$ be the keyword sets of cluster c_1 and c_2, respectively. The following semantic links between clusters can be discovered:

(1) c_1 is the *subCluster* of c_2 (denoted as c_1—*subCluster*$\rightarrow c_2$) if $T(c_1) \subset T(c_2)$.
(2) c_1 is *equivalent* to c_2 (denoted as c_1—*equal*$\rightarrow c_2$) if $T(c_1) = T(c_2)$.
(3) c_1 is *similar* to c_2 (denoted as c_1—$similar \rightarrow c_2$) if $T(c_1) \cap T(c_2) \neq \emptyset$, $T(c_1) \cap T(c_2) \neq T(c_1)$ and $T(c_1) \cap T(c_2) \neq T(c_2)$.

If there are semantic links between two documents, the two documents should share some *connecting words* that imply the semantic links. The larger the number and weights of the *connecting words*, the higher the probability that semantic links exist between them. The initial inference rules depend on the initial document set.

An experiment on constructing SLN-D based on the contents of documents is given in the appendix.

4. DISCOVERING SEMANTIC LINKS ACCORDING TO ATTRIBUTE RULES

Humans have the ability to find various relations in the world because they have semantic worldview [14]. It is hard for machines to automatically discover semantic links among general objects except content analysis and learning through samples. The attributes of resources and rules on attributes are the minimum semantic worldview needed for machines to automatically discover semantic links. The attributes and rules should be separated from algorithm to make the mechanism suitable for any applications. Figure 4 shows the general framework for automatically discovering semantic links between resources. To discover semantic links between resources according to attributes is the first step.

Some semantic links between documents are implied by metadata such as topic, author, creation or publication time, document length and language. For example, scientific document metadata includes *authors*, *editors*, *publish time*, *journal/conference*, *page number*, *author address*, etc. Comparing author attributes can find the *sameAuthor* link, and comparing the *publishTime* can find *sequential* link. Optional information contains page number, project, institution, language, etc.

DOI: 10.1002/cpe

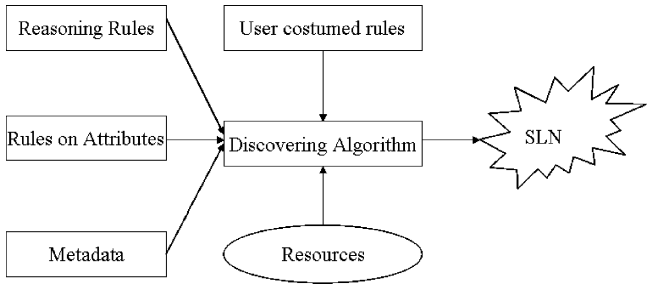

Figure 4. Discovering semantic links according to rules.

According to document metadata, semantic links, such as *sameAuthor*, *sameEditor*, *sameJournal*, *sameConf*, *sameYear*, *sameProject*, *sameAffiliation* and *sameLanguage*, can be found.

The following is the general form of rule on attribute for discovering semantic links between attributes, where A_k $(k \geq 1)$ is the common attribute of documents d_1 and d_2:

$$\text{Rule-on-Attribute: } A_k(d_1) — \alpha \rightarrow A_k(d_2) \Rightarrow d_1 — \beta \rightarrow d_2.$$

The following are two instances: $time(d_1) > time(d_2) \Rightarrow d_1 — laterThan \rightarrow d_2$ and $author(d_1) — equal \rightarrow author(d_2) \Rightarrow d_1 — sameAuthor \rightarrow d_2$. More semantic links can be derived out according to the discovered semantic links, existing semantic links and reasoning rules.

5. REASONING

More semantic links can be derived by using the reasoning rules. If some semantic links exist between two nodes in the SLN, there would be one or more semantic link paths between them. The probability values of the derived semantic links can be calculated by the production of all probability values of the semantic links in the path; in this way, the probability of the derived semantic links reduces with the increase of the length of semantic link path. If there are no reasoning rules for two neighboring semantic links, the reasoning result of the semantic link path can be regarded as *null;*, that is, the semantic links between the source and the target are *unknown*.

A document d can be regarded as an instance of c (denoted as $d — instanceOf \rightarrow c$) if a document d belongs to cluster c. Table III lists some reasoning rules of semantic links discovered from document content. These rules enable the SLN to carry out relational reasoning.

Relations between two resources are mutual, and each semantic link type can have an inverse semantic link type. Suppose N_1, N_2 and N_3 are semantic nodes; N_1 and N_3 connect to N_2; $s_1, s_2 \ldots$, and s_m $(m \geq 1)$ are semantic link types. To derive the semantic links between N_1 and N_3, the semantic links from N_1 to N_2 are denoted as $s_i'[l_i, u_i]$ $(0 \leq l_i \leq u_i \leq 1, 1 \leq i \leq m)$. The semantic links from N_2 to N_3 are denoted as $s_j'[l_j, u_j]$ $(0 \leq l_j \leq u_j \leq 1, 1 \leq j \leq m)$. Algorithm 2 shows the reasoning process on the semantic links. Not only the relations but also the probabilistic values of semantic links change during semantic link reasoning. The probabilistic interval of the existing semantic links may change.

DOI: 10.1002/cpe

Table III. Relational reasoning rules for semantic link network of documents.

Relational reasoning rules	Characteristics	Nodes
$subCluster \times subCluster \rightarrow subCluster$	$T(c_1) \subset T(c_2), T(c_2) \subset T(c_3) \Rightarrow T(c_1) \subset T(c_3)$	Clusters
$partOf \times irrelevant \rightarrow irrelevant$	$T(D_1) \subset T(D_2), T(D_2) \cap T(D_3) = \emptyset \Rightarrow T(D_1) \cap T(D_3) = \emptyset$	Documents
$partOf \times partOf \rightarrow partOf$	$T(D_1) \subset T(D_2), T(D_2) \subset T(D_3) \Rightarrow T(D_1) \subset T(D_3)$	Documents
$instanceOf \times subCluster \rightarrow instanceOf$	$D_1 \in c_1, c_1 \subset c_2 \Rightarrow D_1 \in c_2$	Document and cluster
$partOf \times instanceOf \rightarrow instanceOf$	$T(D_1) \subset T(D_2), D_2 \in c \Rightarrow D_1 \in c$	Document and cluster

Algorithm 2. Semantic link reasoning algorithm between N_1 and N_3 via N_2.

```
Input:  S(N1,N2)=(s'1[l11,u11],s'2[l12,u12],...,s'm[l1m,u1m]);
        S(N2,N3)=(s'1[l21,u21],s'2[l22,u22],...,s'm[l2m,u2m]);
        S(N2,N3)=(s'1[l21,u21],s'2[l22,u22],...,s'm[l2m,u2m]);
Output: S'(N1,N3)=(s'1[l'1,u'1],s'2[l'2,u'2],...,s'm[l'mu'm])
Procedure reasoning (S(N1,N2),S(N2,N3))
    for i=1 to m do [l'i,u'i]=[li,ui];
    for i=1 to m do
        for j=1 to m do {
        if ((s'i × s'j --cd--> s'k) and ((cd·l1i·l2j < l'k) or (l'k==0))) then
             l'k=cd·l1i l2j;
        if (( s'i × s'j --cd--> s'k) and (cd·u1i·u2j > u'k)) then u'k=cd·u1i·u2j;
     }
return S'(N1,N3);
```

6. EVOLUTION

SLN-D evolves with the changes of semantic nodes and semantic links. New documents may lead to the occurrence of new clusters. The vectors and keywords of new documents activate the changes of cluster vectors and keyword sets.

Figure 5 shows the evolution process of the SLN-D. When the size of document set is not too large, document clustering approaches are preferred. With the increase of documents, document clustering is time consuming [22]. When a new document comes, the document vector and the keyword set are calculated, and then new documents can be classified with the k-NN algorithm. A document may belong to several clusters; hence, the keyword–cluster association rules can be used to infer document classifications. Then, new documents can be inserted into the document cluster networks. New documents will change the cluster networks and the keyword networks, and the document–cluster–keyword networks will influence the document classification rules according to the Bayesian formula.

DOI: 10.1002/cpe

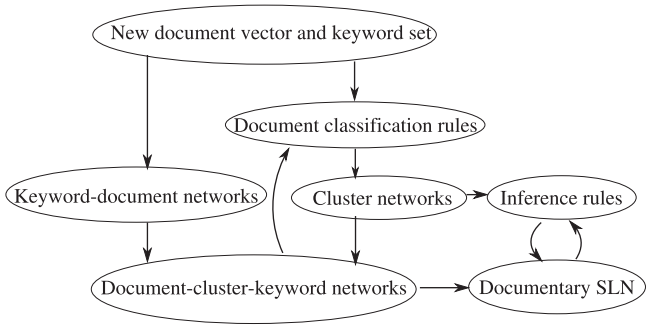

Figure 5. Evolution of semantic link network of documents.

The evolution of cluster networks carries out with:

(1) *New document insertion*. Semantic links from the newly inserted documents to their clusters and other documents are discovered. The cluster vectors evolve with the changes of the document vectors. New documents may cause the change of the keyword set.
(2) *New document clusters occurrence*. Semantic association degrees between the new cluster and old clusters are calculated. If new clusters are clustered into higher-level clusters, then the depth of the cluster networks will increase.

An inference rule is influenced by the following factors:

(1) Change of the source cluster or target cluster of the semantic links. New documents will lead to the change of the clusters or the occurrence of more clusters.
(2) Occurrence of new semantic link types. When new documents are inserted, semantic link types may be increased. The change of the number of semantic links leads to the change of inference rules.
(3) Change of classification rules. The association rules between keywords, documents and clusters will evolve with the changes of the cluster vectors and keyword sets. The classification rules change with the probability values of the semantic links between documents and clusters.

Because inference rules evolve with the changes of the SLN-D, the semantic link types and the probability values are related with the insertion order. Even the same document is inserted into the SLN-D in different orders, semantic links and probability values may be different. When duplicate documents are inserted into the SLN-D, inconsistency may occur; however, this reflects the uncertainty of the SLN-D. Different inference results are caused by different initial document sets. Documents inserted into the SLN-D will act as the initial documents that will influence the classification and semantic link inference of the new documents. With the evolution of the SLN-D, more semantic link types are generated, and the probabilistic intervals of semantic links keep evolving.

6.1. Creating classification rules

Documents are clustered according to keywords and their weights. Figure 6 shows the relations between keywords, documents and classes. An arrow between document d_1 and class c_1 means

DOI: 10.1002/cpe

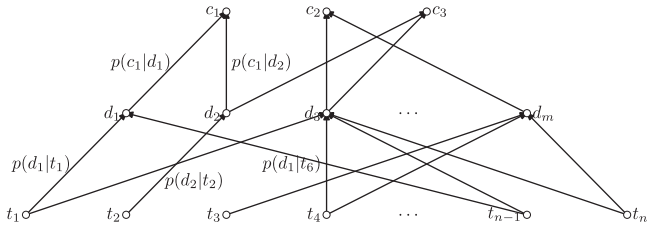

Figure 6. Classifying documents according to the relations between keywords, documents and clusters.

that d_1 belongs to c_1, and $p(c_1|d_1)$ means the probability of d_1 belonging to c_1. An arrow between keyword t_1 and document d_1 means that t_1 occurs in d_1, and $p(d_1|t_1)$ means the probability that t_1 occurs in d_1.

After the initial documents are clustered, each document cluster is assigned by an automatically generated name. Each cluster has the representative keywords that are chosen from the intersection of document keyword sets or according to the weights of keywords in the documents of the cluster.

The inference rules between keywords and classes are acquired by statistic method. If a keyword t occurs in documents d_1, d_2, …, and d_n of cluster c_1, then the association between keyword t and cluster c_1 can be calculated by

$$P(c_1|t)=\sum_{i=1}^{n} P(c_1|d_i)P(d_i|t),$$

where $P(c_1|d_i)$ is the probability that document d_i belongs to class c_1, and $P(d_i|t)$ is the probability that word $t$occurs in document d_i. $P(c_1|d_i)$ is calculated by the cluster algorithms, whereas $P(d_i|t)$ is calculated by the Bayes formula as follows:

$$P(d_i|t)=\frac{P(t|d_i)p(d_i)}{\sum_{t\in d} P(t|d)p(d)},$$

where $P(t|d_i)$ means the probability that word t occurs in document d_i.

6.2. Building semantic link inference rules

Each document may belong to several clusters at the same time. Let l and u be the lower-bounded probability and the upper-bounded probability of a semantic link s, and *src* and *tgt* be the source and the target clusters. The probability of s is calculated as follows:

$$\text{Min\_Pr}(s,src,tgt)=\frac{\sum Pr(l(s),src,tgt)}{\sum Pr(u(S),src,tgt)}, \tag{1}$$

$$\text{Max\_Pr}(s,src,tgt)=\min\left(\frac{\sum Pr(u(s),src,tgt)}{\sum Pr(l(S),src,tgt)},1\right), \tag{2}$$

$$src\text{—}s[\text{Min\_Pr}(s,src,tgt),\text{Max\_Pr}(s,src,tgt)]\rightarrow tgt, \tag{3}$$

where *Pr* (*l*(*s*), *src*, *tgt*) is the lower-bounded probability of a semantic link *s* between the source cluster *src* and the target cluster *tgt*; *Pr*(*u*(*s*), *src*, *tgt*): the upper-bounded probability of a semantic

DOI: 10.1002/cpe

link s between the source cluster *src* and the target cluster *tgt*. S is any semantic link between cluster *src* and cluster *tgt*.

Equation (3) is the semantic link inference rule. If two documents d_1 and d_2 are given, their classifications can be found by using the classification rules, and the probability of semantic link r between d_1 and d_2 can be inferred if d_1 belongs to class *src* and d_2 belongs to class *tgt*.

6.3. Discovering semantic links at evolution stage

The semantic links between the newly added documents and the existing documents can be inferred. For two documents d_1 and d_2 with keyword sets T_1 and T_2, semantic links between them are inferred as follows:

(1) Calculate the document vectors of d_1 and d_2.
(2) Find document clusters for documents d_1 and d_2 by one of the following ways:

- using document vector (one way is k-NN algorithm, the other is to calculate the similarity between the document vectors and the cluster vectors);
- using the document keyword set (comparing the similarity of the document keyword sets or the cluster keyword sets by using the *Jaccard* similarity and the most similar clusters are chosen as the document classification) or
- using the keyword–cluster association rules among documents, keywords and clusters.

The first way calculates document similarity in vector space, the second way calculates the document similarity with the similarity of two sets and the third way has the hypothesis that each pair of keywords are independent.

(3) If d_1 and d_2 belong to the same cluster, find the semantic links by using document keyword sets according to the rules in Table II. As the consequence, semantic links, such as *irrelevant*, *similar*, *partOf* or *equal*, can be discovered.

7. CONCLUSION

It is very important for document-based applications to know semantic links among documents. The approach to automatically discovering semantic links relies on the discovery of the rules that imply semantic links. The proposed approach has the following advantages: (1) new semantic links can be derived from the evolving SLN; (2) the document classification rules, inference rules and cluster association networks automatically evolve with the change of the network; (3) the approach can adapt to the update of the adopted techniques. The approach can be used to automatically construct a semantic overlay on a document set to support advanced applications, such as recommendation and relational query, on scientific documents or web pages. The approach can be used for reference in discovering other types of SLNs like the SLN of events SLN-E. SLN-D provides semantic middle layers for advanced applications.

APPENDIX A: EXPERIMENT ON CONSTRUCTING SLN-D BY CONTENT ANALYSIS

The first step is to collect papers and related data such as metadata and citations for experiment. At the initiation stage of SLN-D construction, 39 papers formulate the initial document set. The approach is also suitable for larger-scale document set.

DOI: 10.1002/cpe

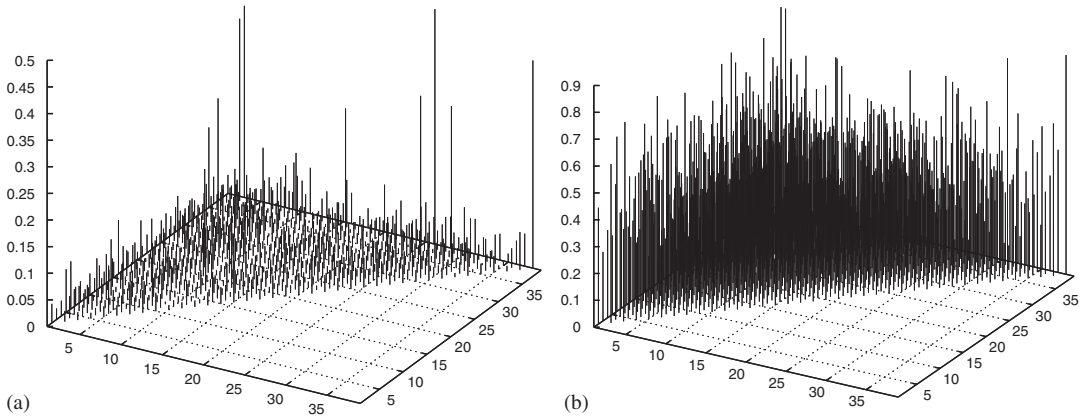

Figure A1. Document pair similarity calculated by classic and new cosine vector: (a) classic-cos and (b) new-cos.

Table AI. Clustering with different edge numbers.

Edge	Similarity	Cluster	Edge	Similarity	Cluster
15	0.689902	25	30	0.623188	13
50	0.582651	7	100	0.508107	1

When computing the pairwise document similarity, the classical vector cosine similarity formula considers all keywords. However, the document pairs may not contain all the keywords. It leads to the parse and low similarity between documents. Then, we calculate document pair similarity based on common keywords and the corresponding weights. Experiment results show that the document similarity distribute evenly than the classical calculation method. Especially, if two documents have no common keywords, then their similarity is 0.

Figure A1 shows the difference of document pair similarity calculation. Since the similarity values of document pairs are symmetric, only half of the result is plotted.

To combine cosine similarity and the Jaccard coefficient similarity, the similarity intervals are mapped onto $[0, 1]$. Experiments show that the cosine similarity and the Jaccard coefficient similarity are similar to each other except for the difference in similarity intervals. Then, the document pairs are sorted in descending order according to the similarity values. The similarity values are within $[0.0774412, 0.817713]$. By controlling the number of top similar document pairs, the document clustering results from cosine similarity are listed in Table AI. Table AII shows the document clusters and the numbers of keywords, and the pairwise document similarities are calculated by $0.5\cos + 0.5\text{jaccard}$. Other documents not listed in Table AII are isolated nodes in the SLN-D.

With document keyword intersection and union, we use the following two indicators for clustering evaluation:

- *Semantic abstract degree*. The more abstract document cluster owns more documents, and it is closer to the root of taxonomy tree. Semantic abstract degree *sem_abstract* is

DOI: 10.1002/cpe

Table AII. Document clusters and their union and intersection of keyword sets.

Cluster	Documents	Union	Common	Total
c_1	(d_1, d_{12}, d_{17})	2334	216	12 363
c_2	(d_3, d_{38}, d_{39})	2579	153	12 363
c_3	(d_5, d_{27})	2157	390	12 363
c_4	$(d_6, d_{23}, 35)$	838	45	12 363
c_5	(d_8, d_{22})	1979	351	12 363
c_6	(d_{11}, d_{36})	2199	450	12 363
c_7	(d_{15}, d_{26})	2261	468	12 363
c_8	$(d_{19}, d_{29}, d_{31}, d_{34})$	2986	161	12 363

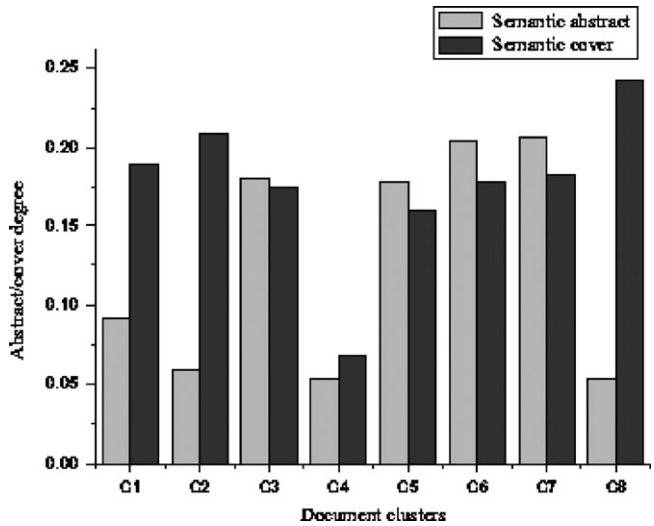

Figure A2. Document clustering evaluation.

calculated by

$$sem\_abstract(c) = \frac{\bigcap_{d_i \in c} T_i}{\bigcup_{d_i \in c} T_i},$$

where d_i is a document, T_i is the keyword set of d_i and c is a document cluster. The less the *sem_abstract*, the more abstract the document cluster.

- *Semantic cover degree*. It reflects the ratio between the document cluster and all document set. Semantic cover degree *sem_cover* is calculated by

$$sem\_cover(c) = \frac{\bigcup_{d_i \in c} T_i}{\bigcup_{d_i \in \{c_1, c_2, \ldots, c_n\}} T_i},$$

where d_i is a document, T_i is keyword set of d_i, c is a document cluster and $\{c_1, c_2, \ldots, c_n\}$ are the existing document clusters.

DOI: 10.1002/cpe

Table AIII. Paper clusters.

Cluster	Papers in the cluster
c_{11}	$\{d_2, d_6, d_{13}, d_{20}, d_{23}, d_{29}, d_{31}\}$
c_{12}	$\{d_3, d_{38}, d_{39}\}$
c_{13}	$\{d_{10}, d_{21}, d_{22}, d_{25}, d_{28}, d_{33}, d_{35}\}$
c_{21}	$\{d_1, d_2, d_3, d_6, d_9, d_{10}, d_{11}, d_{12}, d_{13}, d_{19}, d_{20}, d_{21}, d_{22}, d_{23}, d_{25}, d_{28}, d_{29}, d_{31}, d_{32}, d_{33}, d_{34}, d_{35}, d_{36}, d_{37}, d_{38}, d_{39}\}$
c_{22}	$\{d_5, d_{27}\}$
c_{31}	$\{d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}, d_{11}, d_{12}, d_{13}, d_{17}, d_{19}, d_{20}, d_{21}, d_{22}, d_{23}, d_{25}, d_{26}, d_{27}, d_{28}, d_{29}, d_{31}, d_{32}, d_{33}, d_{34}, d_{35}, d_{36}, d_{37}, d_{38}, d_{39}\}$
c_{41}	$\{d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}, d_{11}, d_{12}, d_{13}, d_{14}, d_{15}, d_{16}, d_{17}, d_{18}, d_{19}, d_{20}, d_{21}, d_{22}, d_{23}, d_{24}, d_{25}, d_{26}, d_{27}, d_{28}, d_{29}, d_{30}, d_{31}, d_{32}, d_{33}, d_{34}, d_{35}, d_{36}, d_{37}, d_{38}, d_{39}\}$

Table AIV. Semantic links between clusters.

Semantic links	Cluster pairs
similar	c_{11}—*similar*$\rightarrow c_{12}$, c_{11}—*similar*$\rightarrow c_{13}, \ldots$
partOf	c_{11}—$partOf \rightarrow c_{21}$, c_{12}—*partOf*$\rightarrow c_{21}, \ldots$
subCluster	c_{11}—*subCluster*$\rightarrow c_{21}$, c_{11}—*subCluster*$\rightarrow c_{31}, \ldots$

Table AV. Semantic links between papers.

Semantic links	Paper pairs
similar	(d_3, d_{38}), (d_3, d_{39}), $(d_{38}, d_{39}), \ldots$
sameTopic	$(d_3, d_{38}) \rightarrow c_{12}$, $(d_{38}, d_{39}) \rightarrow c_{12}, \ldots$
sameAuthor	$(d_3, d_{39}), \ldots$
refer	(d_{38}, d_3), $(d_{38}, d_{39}), \ldots$
cocited	(d_{38}, d_{39})—$cocited \rightarrow d_3, \ldots$
sequential	(d_{38}, d_3), $(d_{38}, d_{39}), \ldots$

Figure A2 shows the semantic abstract degree and the semantic cover degree with data in Table AII.

The number of document clusters is controlled by the number of the semantic links. Several groups of document clustering experiments have been done in [23]. The document clustering results are different by adopting different document similarity calculation formulas.

Table AIII lists several clusters and the related papers by using the similarity calculation formula 0.5cosine+0.5Jaccard with 4-times iteration. The numbers of edges for iterative clustering are 15, 30, 50 and 100, respectively. Each paper links to the corresponding cluster by the *instanceOf* link. At the initiation stage, the semantic link degrees are set as 1.

Table AIV shows some discovered semantic links among paper clusters. *Similar* or *equal* links may exist between clusters at the same clustering level. Clusters at different levels may have semantic links *partOf* or *subCluster*. During the evolution of SLN-D, semantic links such as *similar*, *partOf*, *subCluster* and *equal* may be transformed into each other.

DOI: 10.1002/cpe

Table AV shows some discovered semantic links according to the citations among papers. The *partOf* link implies the similar link. By comparing keyword sets, the *partOf* link can be discovered.

REFERENCES

1. Zhuge H, Zheng L. Ranking semantic-linked network. *Proceedings of the 12th International Conference on World Wide Web* (*WWW 2003*), Budapest, May 2003.
2. Zhuge H. Socio-natural thought semantic link network—A method of semantic networking in the cyber physical society. *Keynote at IEEE AINA 2010*, Perth, Australia, 20–23 April 2010; 19–26.
3. Harabagiu S, Lacatusu F, Hickl A. Answering complex questions with random walk models. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR 2006*). ACM: New York, NY, U.S.A., 2006; 220–227.
4. Bayardo RJ, Ma Y, Srikant R. Scaling up all pairs similarity search. *Proceedings of the 16th International Conference on World Wide Web* (*WWW 2007*). ACM: New York, NY, U.S.A., 2007; 131–140.
5. Menczer F. Mapping the semantics of Web text and links. *IEEE Internet Computing* 2005; **9**(3):27–36.
6. Yin X, Han J, Yu P. LinkClus: Efficient clustering via heterogeneous semantic links. *Proceedings of the 32nd International Conference on Very Large Data Bases*. VLDB Endowment: Seoul, Korea, 2006; 427–438.
7. Attardi G, Gull A, Sebastiani F, Hutchison C, Lanzarone G. Automatic web page categorization by link and context analysis. *Proceedings of the European Symposium on Telematics*, *Hypermedia and Artificial Intelligence* (*THAI-99*), Varese, Italy, 1999; 105–119.
8. Getoor L, Diehl C. Link mining: A survey. *SIGKDD Explorations* 2005; **7**(2):3–12.
9. Huang Z. Link prediction based on graph topology: The predictive value of generalized clustering coefficient. *Proceedings of KDD 06 Workshop on Link Analysis*: *Dynamics and Static of Large Networks* (*LinkKDD2006*), Philadelphia, U.S.A., 2006.
10. Cohn D, Hofmann T. The missing link—A probabilistic model of document content and hypertext connectivity. *Advances in Neural Information Processing Systems*, Vancouver, British Columbia, Canada, 2001; 430–436.
11. Mahler D. Holistic query expansion using graphical models. *New Directions in Question Answering*, Standford, CA, U.S.A., 2004; 203–214.
12. Luo G, Tang C, Tian Y. Answering relationship queries on the web. *Proceedings of the 16th International Conference on World Wide Web*. ACM Press: New York, NY, U.S.A., 2007; 561–570.
13. Zhuge H. Communities and emerging semantics in semantic link network: Discovery and learning. *IEEE Transactions on Knowledge and Data Engineering* 2009; **21**(6):785–799.
14. Zhuge H. Interactive semantics. *Artificial Intelligence* 2010; **174**(2):190–204.
15. Zhuge H. *The Knowledge Grid*. World Scientific: Singapore, 2004.
16. Aleman-Meza B, Halaschek-Wiener C, Arpinar IB, Sheth AP. Context-aware semantic association ranking. *Semantic Web and Databases Workshop Proceedings*, Humboldt-Universität, Berlin, Germany, 2003; 33–50.
17. Anyanwu K, Maduko A, Sheth A. Semrank: Ranking complex relationship search results on the semantic web. *Proceedings of the 14th International Conference on World Wide Web* (*WWW 2005*). ACM: New York, NY, U.S.A., 2005; 117–127.
18. Sowa JF (ed). *Principles of Semantic Networks*: *Exploration in the Representation of Knowledge*. Morgan Kaufmann Publishers: San Mateo, CA, 1991.
19. Strehl A, Ghosh J, Mooney R. Impact of similarity measures on web-page clustering. *Proceedings of the AAAI Workshop on AI for Web Search* (*AAAI 2000*), Austin, 2000; 58–64.
20. Robertson S, Walker S, Beaulieu M. Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive. *Proceedings of the Seventh Text Retrieval Conference* (*TREC-7*), Gaithersburg, Maryland, 1998; 242–500.
21. Singhal A. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin* 2001; **24**(4):35–43.
22. Zhao H. Semantic matching across heterogeneous data sources. *Communications of the ACM* 2007; **50**(1):45–50.
23. Zhuge H, Zhang J. Automatically discovering semantic links among documents. *SKG08*: *Proceedings of the Fourth International Conference on Semantics*, *Knowledge*, *and Grid*, Beijing, China, 3–5 December 2008. IEEE Computer Society: Silver Spring, MD, 2008.

DOI: 10.1002/cpe