An Automatic Semantic Relationships Discovery Approach[1]

Hai Zhuge, Liping Zheng, Nan Zhang and Xiang Li
China Knowledge Grid Research Group (http://kg.ict.ac.cn),
Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences, China
01062562703, 086
zhuge@ict.ac.cn

ABSTRACT
An important obstacle to the success of the Semantic Web is that the establishment of the semantic relationship is labor-intensive. This paper proposes an automatic semantic relationship discovering approach for constructing the semantic link network. The basic premise of this work is that the semantics of a web page can be reflected by a set of keywords, and the semantic relationship between two web pages can be determined by the semantic relationship between their keyword sets. The approach adopts the data mining algorithms to discover the semantic relationships between keyword sets, and then uses deductive and analogical reasoning to enrich the semantic relationships. The proposed algorithms have been implemented. Experiment shows that the approach is feasible.

Categories & Subject Descriptors: H.3.1 **[Information Storage and Retrieval]**: Content Analysis and Indexing – *indexing methods, linguistic processing*.

General Terms: Algorithms, Experimentation.

Keywords: Algorithm, analogical reasoning, Semantic Link Network, Semantic Web, data mining.

1. INTRODUCTION
Semantic Web is to overcome the shortcomings of the current Web by making the semantics of web pages machine-understandable.

The Semantic Link Network (*SLN*) is a directed graph where nodes are semantic components and arcs are semantic links. A semantic component can be another *SLN*, a concept, or an atomic semantic component like a piece of text or an image. A semantic link is an arc between two semantic components. Each arc is marked with a type and a certainty degree between 0 and 1 that reflects the inexactness of semantic relationship. There are eight basic types of the semantic relationships: cause-effect (*ce*), implication (*imp*), sub-type (*st*), instance (*ins*), reference (*ref*), similar-to (*sim*), sequential-to (*seq*) and undefined (*und*).

Previous studies on *SLN* include *SLN* definition, ranking of semantic components and *SLN*-based reasoning [2,3,4,5,6]. However, they are confined to an existing *SLN* manually constructed. How to effectively build an *SLN* for large-scale resources is still an open issue. Thus the requirement for automatically or semi-automatically generating semantic relationships between web resources arises. This paper presents an approach that combines data mining and *SLN*-based reasoning to solve this issue.

2. RELATED WORK
The proposed approach in this paper is similar to web mining in principle and adopts frequent pattern mining, an important technique of data mining to investigate those simultaneous and frequent occurrences of data [1], as one of its major parts.

Semantic link deduction is used to find out the implied semantic relationships between semantic components from existing semantic links in *SLN*, which is based on a set of semantic link connection rules. It can be realized through the adjacent matrix representation of *SLN* and the special matrix self-multiplication.

Two semantic components are determined to have a particular semantic relationship if there are another two semantic components similar to them respectively that have such relationship. However, such discovered relationship has to be verified manually on whether it is valid.

3. TECHNICAL PATH
We only discuss the semantic relationships between web pages and leave alone the certainty degree of semantic relationship.

For each web page we extract a number of keywords (*keyset*) that can properly describe its subject. Then we construct a small set (*training set*) that includes a number of web page pairs among which each pair of web pages has a certain type of semantic relationship. Apriori-like algorithm can be used to mine relationship between keyword sets.

Reasoning is used to complement the mined frequent rules so as to further find out more relationships between keywords sets:

(1) $st \bullet st = st,\ \sim sim = sim$
(2) $l \bullet ins = ins,\ l \in \{ins, st, imp\}$
(3) $imp \bullet l = imp,\ l \in \{imp, st\};\ l \bullet imp = imp,\ l \in \{st, ins\}$
(4) $l \bullet ref = ref,\ l \in \{ref, ins, st, imp\}$
(5) $ce \bullet l = ce,\ l \in \{ce, imp, st, sim, ins\};\ l \bullet ce = ce,\ l \in \{imp, st, ins\}$
(6) $seq \bullet seq = seq;\ \sim l \bullet seq \bullet l = seq,\ l \in \{ce, imp, ref, ins, st\}$

WWW 2004, May 17-22, 2004, New York, NY USA.
ACM 1-58113-912-8/04/0005.

[1] This work was supported by the National Science Foundation of China (NSFC).

(7) $(l_1+l_2)\bullet l_3 = l_1\bullet l_3+l_2\bullet l_3$, and $l_3\bullet(l_1+l_2) = l_3\bullet l_1+l_3\bullet l_2$

(8) $sim\bullet l\bullet sim = l$, $l\in$ {*ce*, *imp*, *ref*, *ins*, *st*, *imp*, *seq*}

In an *SLN* environment comprising only web pages, reasoning can be realized by our proposed approach through the multiplication and addition operations on matrix entries defined by the above eight operation laws.

The general system architecture of the proposed approach is shown in Figure 1. Herein we construct a small training set by *SLN* definition tools. After using the Apriori algorithm on the training set to generate frequent antecedents and consequents, we connect them to generate candidate rules. For each candidate rules we compute its support degree and confidence degree, and compare them with *min_s* and *min_c* respectively. Those candidate rules with support degree and confidence degree more than *min_s* and *min_c* respectively are output as frequent rules. Reasoning then acts on frequent rules to generate complementary rules. At last, frequent rules together with complementary rules are used on test sets (a deposit of web pages not within the training set) to divide the web pages in test sets into corresponding web pair groups with particular semantic relationship. Web users will verify the validity of the classification results.

4. MINING FREQUENT RULES

Mining frequent rules include two consecutive algorithms: *Apriori-like algorithm* to generate frequent antecedents and consequents; and *connection algorithm* to connect them into candidate rules, compute the support degrees and confident degrees of them, and output those candidate rules with support degrees more than min_s and confident degree more than min_c as frequent rules.

In Apriori-like algorithm, we store the training subsets in a binary tree structure so as to save the I/O overhead and make mining efficient and highly scalable. The algorithm consists of two consecutive phases: *tree construction* organizing the training subsets in a compact structure similar to the data structure of the *FP-growth* method, and *rule generation*.

In the rule generation phase, Apriori-like checks each keyword in the binary tree. If one keyword β passes the support degree threshold, Apriori-like investigates the subsequent keywords following β in those paths starting with β. If the combination of β and some of its subsequent keywords passes the support degree threshold, Apriori-like outputs the combination as frequent item. Note that if a keyword cannot pass the support degree threshold, then the keywords combinations including the keyword will not be checked.

5. CONCLUSIONS

To automatically discover the semantic relationship between web resources is the key to the success of Semantic Web. This paper proposes an approach to automatically build a Semantic Link Network through automatically finding out semantic relationships between web pages. The proposed approach integrates the data-mining algorithm and the semantic link reasoning, and reduces the issue of automatically discovering the semantic relationship to establishing the semantic relationships between the keyword sets of web pages. The approach could avail effective organization of web resources and favor current web data mining research.

Ongoing work includes extending the approach by taking into account the relationship between keywords, the certainty degree of semantic link and dealing with other types of web resources other than web pages.

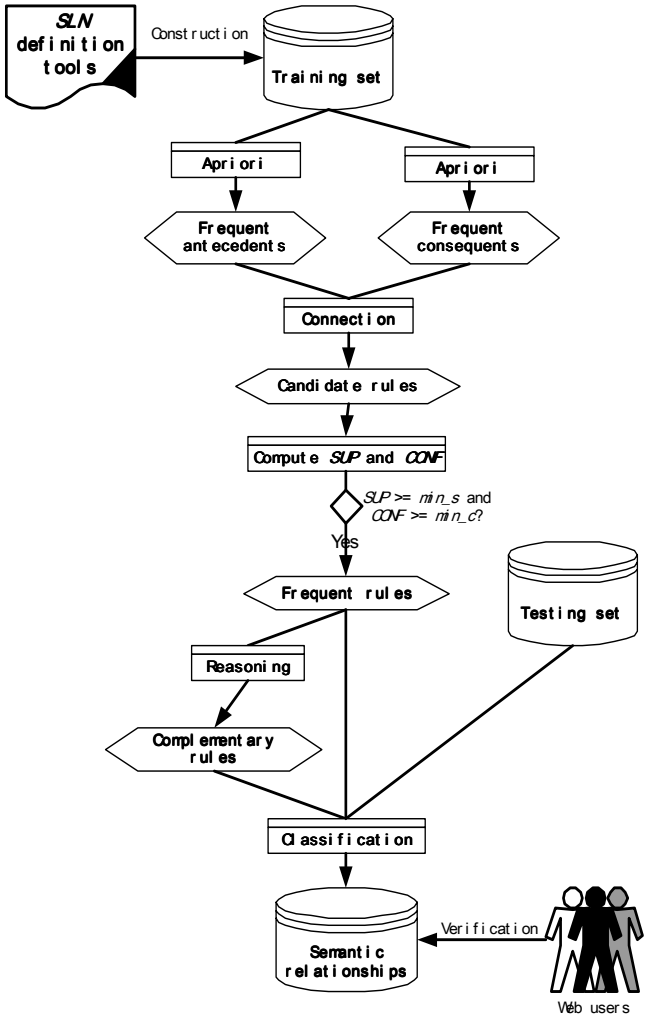

Figure 1. The system architecture of the proposed approach.

6. REFERENCES

[1] Agrawal, R., and Srikant, R. Fast algorithms for mining association rules. Proceeding of the 20th International Conference on Very Large Data Bases, 1994, 478-499.

[2] Zhuge, H. Active e-Document Framework ADF: Model and Tool. Information & Management, 2003, 41: 87-97.

[3] Zhuge, H and Zheng, L. Deductive and Analogical Reasoning on Semantic Link Network. Technical Report of China Knowledge Grid Research Group.

[4] Zhuge, H. and Zheng, L. Ranking Semantic-linked Network. The 12th International World Wide Web Conference, 2003, Budapest.

[5] Zhuge, H., China's e-Science Knowledge Grid Environment, IEEE Intelligent Systems, 2004, 19(1): 13-17.

[6] Knowledge Grid Forum, http://www.knowledgegird.net