Intelligent Resource Discovery Based on Web Service in Information Security Resource Grid

Hu Ruo¹, Chen Chao Tian², Xie Zan Fu²
¹Department of Computer Science, GuangDong Polytechnic Normal University, GuangZhou city, China
Hu68@163.com
²Department of Computer Science, GuangDong Polytechnic Normal University, GuangZhou city, China

Abstract—Intelligent discovery model with the concept expression for information security under the information security resource grid is proposed in this thesis, in which Concept Expression is adopted in the formal method of description on information security knowledge, this method is used to support reasoning and semantically intelligent discovery. Characterization and resource optimization are two main technologies in resource intelligent discovery. Through normative concepts and semantic correlation between concepts, resource characteristics are dealt with in terms of characterization and in regard to the disunity of concepts while algorithm of resource characteristics is proposed. During the optimization process of the resource discovery, the optimized algorithm based on concept expression of reasoning is sought for realizing the automatic expansion of concepts and enhancing coverage and accuracy.

Index Terms—Information Security, Resource Grid, Intelligent Discovery, Concept Expression, Unification Concept Algorithm

I. INTRODUCTION

Information security resource refers to the software, hardware and relevant data dealing with threats on information security. Such threats include virus, hacking, Trojan Horse, system vulnerability, network junk and other malicious software, etc. The information security resources include anti-virus software, intrusion discovery software, firewall, service pack and anti-spam products, etc.

When new requests are received, the user (Web service developer) will firstly search whether there are any Web service resources for information security satisfying such requests that can be reused. If not, further exploration is made to find out whether there’s a set of service resources that can be made to fulfill the requests through synergy. When there’s no suitable service or service combination satisfying the requests, such requests will be further decomposed to continue searching for new usable services. Figure 1 shows the Discovery and combination processes regarding user requests on Web service resources for information security.

How to describe Web service for information security and realize automatic discovery and dynamic combination of information security service is a hot topic and difficulty in the field of Web service for information security research. Currently, WSDL (Web service description language) is an accepted norm for interface description, which describes Web service as a set of ports that mainly describes the dynamic performing characteristics of Web service, realizable for mutual operations between and among services. However, the Web service described in WSDL is in lack of semantics, which makes it hard for computers to understand, resulting in inability to support the automatic discovery of Web services.

This article proposes a kind of intelligent discovery model based on concept expression of information security by adopting OWLDL language to structure concept expression of information security, to realize smaller granularity characterization, to propose unified algorithm for main contents as well as to provide concept expression and concept image mostly approximate to the semantic contents of resource characteristics; automatic extension and optimization of user discovery concepts are realized based on OWL language reasoning to arrive at an in-depth semantic realization. Experiments have shown that in comparison to the conventional main-word discovery, intelligent discovery based on concept expression can gain higher coverage and accuracy.

II. INTELLIGENT DISCOVERY MODEL BASED ON INFORMATION SECURITY CONCEPT EXPRESSION

Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-
Type 1 or Open Type fonts are preferred. Please embed symbol fonts, as well, for math, etc.

In the traditional resource discovery models, no normalized concept set with semantic characteristics is deployed to make meaningful characterization on resources, but often index items are exclusively abstracted from the perspective of grammars and certain item weighting tactics (like tf-idf weighting tactics[11]) is used to grant weight values to such index entries. Such indexing dependant on concept grammar expressions can not effectively reflect the semantics of resources, thus it fails to effectively represent the resources; meanwhile, due to the lack of guidance of normative concept set with semantic characteristics, multiple and inaccurate user information requests, it fails to effectively reflect the semantics of user information requests. Pure concept matching ignores the generally existing problems of “disunity of concepts”, “multiple words for the same meaning” and “multiple meanings for the same word”. Separation of grammatical expression and semantic contents always leads to unsatisfactory discovery performance.

The intelligent discovery model we proposed based on concept expression of information security is illustrated in Figure 2.

Firstly, semantic modules are set up for information security knowledge to establish the concept expression of information security; pre-treatment is done to resources to segment them into several resource characteristics information based on the integrity of meanings expressed. Based on resource characteristics, information-oriented is done to resource characteristics. Based on the results of such information-oriented of characteristics, one or several resource characteristics will form resource objects, with the only labeling and characteristic concept set used as metadata to be stored to the database; the user discovery requests are optimized and extended under the guidance of concept expression for information security. The detecting task does not directly detect from original resources and data, but rather detecting from the metadata and the matching result will be reflected to relevant resources through labeling and return back users.

This article focuses on studying the architecture of concept expression for information security, characteristics and discovery optimization technologies with the aim of abstracting resources and user requests into a group of concept set that is semantically representative with machines and detecting based on semantics.

III. ARCHITECTURE OF CONCEPT EXPRESSION FOR INFORMATION SECURITY

A. Basic Concepts

Definition 1 Concept expression for information security can be expressed in a five-tuple array:

\[
O = \{C, R, Hc, rel, A_o\}
\]

In which, \(C\) means set of concepts; \(R\) means set of relations; \(Hc\) means concept level; \(rel\) means relations between/among concepts; and \(A_o\) means axiom of concept expression.

Definition 2 Concept \(C\) in concept expression can be expressed in a triple array:

\[
C_i = \{Name_i, ID_i, SynList_i\}
\]

In which, \(Name_i\) refers to the name of concept \(i\); \(ID_i\) refers to the only identification of concept \(i\) in the concept expression for information security; and \(SynList_i\) refers to the synonym list of concept \(i\).

This article mainly considers the basic semantic relations among three types of concepts, i.e., Kind-of, Instance-of and Part-of.

Definition 3 Set of relations \(R\) includes three kinds of relations:

\[
R = \{\text{Kind-of, Instance-of, Part-of}\}
\]

In which, Kind-of relation refers to inclusive relations among concepts, e.g., “anti-virus software” and “information security software”; “anti-virus software” is a kind of “information security software”. Apart from that, “information security software” also includes “intrusion discovery software”, “firewall software”, etc.

Instance-of relation refers to relations between concepts and instances, e.g., “Anti-virus Software” and “Kingsoft Anti-virus”; “Kingsoft Anti-virus” is an instance of “Anti-virus Software”.

Part-of relation refers to relations between part and whole of the concepts, e.g., “Kingsoft Vulnerability Scanning” and “Kingsoft Antivirus”; “Kingsoft Antivirus” is a whole which is made up of “Kingsoft Vulnerability Scanning” and “Kingsoft Killer” jointly. “Kingsoft Vulnerability Scanning” is part of “Kingsoft Antivirus”.

B. Concept Expression Language

This article utilizes descriptive logic to build up concept expression. In comparison to semantic networks, frame and other non-logic formal methods, descriptive logic can accurately depict semantics and in comparison to the one-stage predicate logic reasoning from which
only partial judgment is deduced, descriptive logic does not only possess stronger knowledge expression ability, but also guarantees determinant reasoning.

Nowadays, the often utilized concept expression languages include XOL, SHOE, OML, RDF (S), OIL, DAML + OIL and OWL, which are all based on descriptive logic and in which OWL makes DAML + OIL as the starting point and stands as the latest research results on concept expression language.

OWL has three subsets, namely OWL Lite, OWL DL and OWL Full, of which OWL Lite has the most limited expression ability and high reasoning efficiency; OWL DL has the possibly strong expression ability provided that completeness and judgment of reasoning are safeguarded; and OWL Full has the strongest expression ability without giving any guarantee to reasoning. During the structuring of concept expression, this article explores OWL DL as the descriptive language of concept expression to take both highly efficient reasoning and expression ability into full consideration and the following definitions are given based on OWL DL language.

In OWL DL language, concept match along with class while instance match along with individual. Class is a group of individual sets with public features and individual is the instance for class. During characterization and discovery optimization, operations to concept and instance are the same, therefore, no distinction is made temporarily and they are together called concept. However, when describing based on OWL DL, the two are rigidly distinguished.

Definition 4 (Instance-Of) For concept Ci and instance d i in the concept expression, if d i is defined as the “Individual” of Ci, d i is regarded as the instance of Ci.

Definition 5 (Kind-Of) For two concepts Ci and Cj in the concept expression, if Cj is defined as the “subClass Of” of Ci, the semantics of concept Ci is regarded as inclusive of Cj.

Definition 6 (Part-Of) For concept Ci and instance d i in the concept expression, if d i is defined as the “subProperty Of” of Ci, d i is regarded as a part of Ci.

The following takes “Kingsoft Vulnerability Scanning” and “Anti-virus Software” as examples to specifically describe concept and instance in concept expression.

C. Concept Expression for Information Security

Concept expression for information security is the accurate description on relations between/among concepts and concepts itself within the field of information security. In accordance with the above-mentioned basic definitions, information security concept is taken as nodes and relations between/among concepts are taken as the edge linking such nodes, therefore, concept expression for information security can be illustrated in the Directed Acyclic Graph (DAG). Figure 4 shows an example of the concept expression for information security, which is set up based on common information security terms and knowledge of information security experts.

IV. CHARACTERIZATION

A. Concept Characterization

Concept characterization is the process for establishment linkages between information with resource characterization and information security concept in order to combine the new resource object generated from information after characterization based on the concept sets of characterization and the existing resource object. The algorithm for characterization is shown in Figure 4.

B. Concept Unification

Due to the universally existing concept disunity in natural language, i.e., one main word may correspond to multiple concepts or synonym list of different concepts may include the same main word, the concept of resource characteristics through characterization may result in different concepts, namely, the contents of resource characteristics not truly reflected in concepts. For instance, a resource characteristics with the contents of “At 8:30, Feb. 10, 2008, Kingsoft Anti-virus Software is used at University of Shanghai for Science and Technology (USST) and besides killing viruses, Kingsoft Anti-virus Software will also deploy the Kingsoft Vulnerability Scanning to scan the vulnerabilities in the system with the

Figure 3. An example of Concept Expression for Information Security
Some relevant concepts in concept expression for information security constitute a certain class, which is referred to as “Main Contents” in this article, e.g., “Anti-virus Software”, “Vulnerability Scanning” and the relevant concepts and instances of “Anti-virus Software” constitute the main contents of a description “Anti-virus Software”; also, the relevant concepts and instances of “Vulnerability Scanning” constitute the main contents of a description “Vulnerability Scanning”.

Generally, characteristics of a resource should correspond to a certain main content. Therefore, concept unification should firstly be the unification of “main contents”, i.e. defining that concepts suitable for characterization of the resource characteristics belong to what kind of main contents of concept expression for information security; then excluding disunity of concepts due to such disunity in the main contents. Based on the above-mentioned concept, this article proposes a unified algorithm “Main contents-Concept” (in abbreviation, “M-C” algorithm) and through calculating weighting value, formal description of algorithm is realized.

Assuming that concept Ci has a number of (n) synonyms and its synonym sets is expressed as SynLi = {li1, li2, li3, …, lij, …, lin}. Definition 7 Synonym weighting (SWij) Weighting of synonyms lij of concept Ci is the ratio between the number of appearance of lij in resource characteristics of concept characterization (Num of lij of Ci Matched) and the total number of appearance of this concept in the concept synonym sets expressed in concept expression (Total Num of lij) as expressed in the following formula:

$$SW = \frac{\text{Num of lij of } C_i \text{ Matched}}{\text{Total Num of lij}}$$

Definition 8 Concept weighting (CWi) Weighting of concept Ci is the maximum value of all its synonym weighting.

$$CW_i = \max_{1 \leq j \leq n} SW_{ij}$$

Definition 9 Semantic distance SD (Ci, Cj) Semantic distance between concept Ci and Cj is the shortest path between two concepts in the concept expression.

Definition 10 Correlated concept (Cor-Concept) semantic distance with concept Ci. in the concept expression is not equal to an unlimited concept.

Correlated concepts appeared during the characterization process of resource objects provide a meaningful context for correct understanding of resource meaning and effective unification. For example, “Kingsoft Anti-virus Software” is correlated with “USST” and “Anti-virus Software”, so during unification, defined enhanced weighting affects correlated concepts in terms of quantification.

Definition 11 Enhanced weighting (EW) Assuming that the correlated concept set of concept Ci is \( \{ C_k \mid k = 1, 2, \ldots, r \} \), the enhanced weighting of Ci is concept weighting of Ci plus the sum of all correlated concept weighting divided by semantic distances between C1: SD (C1, Ck), namely,

$$EW_i = \text{Weight}_i + \sum_{C_k} \frac{\text{Weight}_k}{\text{SD}(C_i, C_k)}$$
It can be known from the definition of enhanced weighting that the more correlated concepts of C_i are gathered in characterization concept, the smaller the semantic distance would be; the bigger its enhanced weighting is, the more correlated it would be with meaning of resource objects.

Based on the above-mentioned definitions, unified algorithm M-C can be described as follows:

\begin{algorithm}
\begin{algorithmic}
\State \text{Begin}
\State \text{// algorithm M-C is adopted to unify discovery concepts}
\State ConceptSet CS = M-C Disam (DO) ;
\State for each ( \( C_i \in CS \))
\State \text{OntoClass class = getOntologyClass ( C ) ;}
\State \text{if ( Individual ( class ) is not null)}
\State \quad CS = CS \cup \text{Individual ( class )} ;
\State \text{if ( subClassof ( class ) is not null)}
\State \quad CS = CS \cup \text{subClassof ( class )} ;
\State \text{if ( Individual ( subClassof ( class ) ) is not null)}
\State \quad CS = CS \cup \text{Individual ( subClassof ( class ) )} ;
\State \text{return CS ;}
\State \text{end ;}
\end{algorithmic}
\end{algorithm}

In which, description about Kingsoft Vulnerability Scanning comes from the OWL description of the concept expression of “Information Security Software” on “Kingsoft Vulnerability Scanning”:

\begin{verbatim}
\verb+( owl:Class rdf:ID ="Vulnerability Scanning Software"+)\end{verbatim}

Reasoning device learns “Information Security Software” through reasoning, inclusive of “Anti-virus Software”, “Firewall Software”, “Vulnerability Scanning Software” and other subsets. During discovery, such concepts will automatically extend to discovery concepts. This article uses RACER as the descriptive logic reasoning device and nRQL as discovery language.

V. EXPERIMENTAL ASSESSMENT

Validation of the practicality of information discovery models generally uses the generally accredited test reference resource sets (e.g., TREC, CACM, CF, etc.) to compare accuracy and coverage of different models, and analyze advantages and disadvantages of models. However, generally accredited test reference resource sets of vocabulary marking in concept expression are not used by now, so this article collects 4100 information security resources from Internet, Chinese Periodicals Database, network news and Vip Periodicals Database as the data sources for experiment, mainly focusing on information security field. Specific data is illustrated in Table 1.

Concept expressions for information security is established to include over 600 concepts in combination of correlated knowledge in the information security field. Accuracy and coverage are used as the norms for assessing discovery performance, in which accuracy is defined as the ratio between the number of correlated resources in discovery results and total number of resources included in discovery results while coverage is defined as the ratio between the number of correlated

<table>
<thead>
<tr>
<th>No.</th>
<th>Resources</th>
<th>Sources</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Network resources</td>
<td>Internet</td>
<td>1200</td>
</tr>
<tr>
<td>2</td>
<td>Academic paper</td>
<td>Chinese Periodicals Database</td>
<td>600 articles</td>
</tr>
<tr>
<td>3</td>
<td>Reports</td>
<td>Network message</td>
<td>850 articles</td>
</tr>
<tr>
<td>4</td>
<td>Periodicals</td>
<td>Vip Periodicals Database</td>
<td>900 articles</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>User Discovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>“Finding news on information security software”</td>
</tr>
<tr>
<td>2</td>
<td>“Finding cases of Kingsoft Anti-virus software used at USST”</td>
</tr>
<tr>
<td>3</td>
<td>“Finding cases where information security software are used”</td>
</tr>
<tr>
<td>4</td>
<td>“Finding uses of Kingsoft Vulnerability Scanning software”</td>
</tr>
<tr>
<td>5</td>
<td>“Finding uses of old information security software”</td>
</tr>
<tr>
<td>6</td>
<td>“Finding normal working principles of anti-virus software”</td>
</tr>
<tr>
<td>7</td>
<td>“Finding the development trend of anti-virus software”</td>
</tr>
<tr>
<td>8</td>
<td>“Finding effects of Vulnerability Scanning software”</td>
</tr>
</tbody>
</table>
resources in discovery results and total number of correlated resources in the test sets. In order to safeguard the universality of validation, professionals and non-professionals are invited to propose 8 discovery requests for experimental validation. User Discovery is illustrated in Table 2.

Based on selection of test sets, discovery experimental results based on concept expression are compared with discovery results using discovery tools based on main words. The comparison is illustrated in Figure 5 and 6.

REFERENCES
