A Scalable Framework for Collaborating Web Clearinghouses

Pei Yuen Wong
National University of Singapore
Building S-16, Level 5, Room 05/08
3 Science Drive 2
Singapore 117543
+65 9 626 7672
wongpei1@comp.nus.edu.sg

ABSTRACT
This paper describes how Web clearinghouses can be organized so that they can effectively collaborate with each other to improve both the recall and the precision of results returned for a query.

Keywords
knowledge-based, collaboration, RDF

1. INTRODUCTION
The Web has made available to users worldwide a huge number of high quality information resources scattered across the globe. With more and more resources available online, people are getting increasingly reliant on the Web for their information needs. Metadata describing these resources plays a critical role in the effectiveness of leveraging on these resources. However, as pointed out in numerous sources such as [3], current search engine technologies are not adequate for the needs of users who require precise information to be effectively gathered from specific domains.

An emerging research area that has proven promising are *Subject-Based Information Gateways*, or *SBIG*s, which are Web clearinghouses that organize information resources pertinent to a *particular domain of interest*. Many research efforts, such as *Dienst* [4], the *Resource Organization and Discovery in Subject-based Services (ROADS)* [5], *ISSAC* [6] and *OntoBroker* [1], have pursued research in this direction. This paper describes the research undertaken at the National University of Singapore in this area.

2. RESEARCH CONTRIBUTIONS
The main contribution that was accomplished by this research is the development of a scalable framework for collaborating domain specific information clearinghouses consisting of metadata describing information resources. This framework fulfills 3 objectives that we have found lacking in current work: scalability, extensibility and accessibility.

3. ABSTRACT MODEL
Figure 1 shows the top-level abstract model of a *Domain Specific Information Clearinghouse* or *DSIC*:

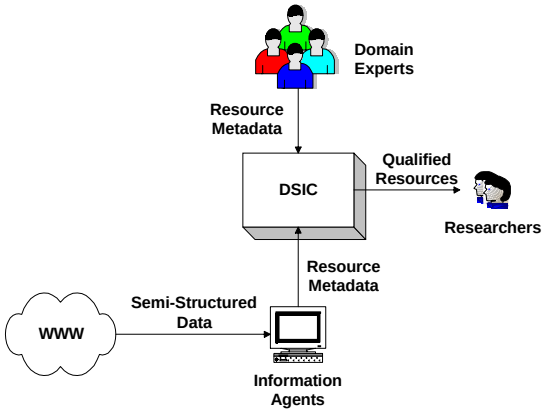

Figure 1 : The Abstract DSIC Model

A DSIC is a Web-based clearinghouse and resource repository for domain specific information resources available on the Web. It is essentially a knowledge base that stores resource metadata in the form of RDF [2] statements supplied by *human domain experts* as well as *intelligent software agents*. *Researchers*, the targeted users of the DSIC who wish to find high quality *qualified resources*, can then make use of the DSIC to locate them.

Domain experts are subject matter experts who are familiar with the DSIC's domain. These domain experts register with the DSIC as trusted information providers, after which they can submit resource metadata to the DSIC based on a set of evaluation criteria such as those presented in [7].

DSIC Information Agents are software agents that can scour the Web for relevant information resources. Unlike Web spiders that only index documents indiscriminately based on keywords, DSIC information agents are able to extract *semantic metadata* that exists in Web documents.

4. SYSTEM ARCHITECTURE
This section takes a closer look at the system architecture of the DSIC, as shown in Figure 2.

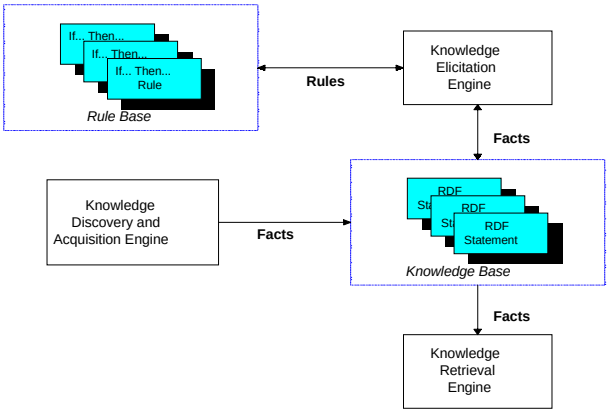

Figure 2 : System Architecture of the DSIC

Knowledge Base

The *Knowledge Base* consists of resource metadata in the form of RDF statements. These statements are asserted as facts in the knowledge-based framework.

Rule Base

The *Rule Base* comprises of *if … then* rules. Ontological specifications of the domain taxonomy are specified in the Rule Base. Additional rules given by the domain experts or discovered by the *Knowledge Elicitation Engine* are also stored in the Rule Base.

Knowledge Discovery and Acquisition Engine

The *Knowledge Discovery and Acquisition Engine* deploys DSIC information agents to discover and acquire new knowledge about relevant qualified resources. The acquired knowledge is then stored in the Knowledge Base. In addition, a Web-based interface allows human domain experts to contribute resource metadata to the Knowledge Base. This human-supplied metadata is converted from a *human-readable* form to a *machine-understandable* form before storing into the Knowledge Base as facts about resources.

Knowledge Retrieval Engine

The *Knowledge Retrieval Engine* organizes the Knowledge Base and uses it to answer users' queries. It allows structured, precise queries over the entire information space of the Knowledge Base so that relevant qualified resources can be found.

Knowledge Elicitation Engine

The *Knowledge Elicitation Engine* makes use of both the Rule Base and the Knowledge Base to generate new facts, which are asserted in the Knowledge Base. This allows the *implicit semantics* of the resources to be explicitly asserted so that the Knowledge Retrieval Engine can answer users' queries more effectively. In addition, the Knowledge Elicitation Engine can also mine for association rules in the Knowledge Base and optionally stores these rules in the Rule Base.

5. DSIC COLLABORATION

Researchers often have not just one, but multiple domains of interest. Moreover, there is often no clear boundary between domains, as shown in Figure 3. Hence, by enabling collaboration between different DSICs, we can effectively increase both the *recall* and the *precision* of the system simultaneously.

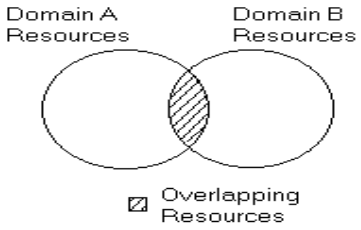

Figure 3 : Overlapping Domain Resources

Resource sharing takes place with the aid of mobile agents that communicate using the ubiquitous HTTP. Each DSIC exchanges a mobile agent with the DSIC it wishes to collaborate with. The mobile agent carries with it the Rule Base and a set of *context specifications* and resides in the destination DSIC. These context specifications, together with the Rule Base, enable the mobile agent to determine whether a *concept term* being queried for also exists as a concept term in the *same context* in the source DSIC. If so, the mobile agent requests the resource metadata that corresponds to the query from the source DSIC and returns it to the requesting DSIC. In this way, the *autonomy* of each DSIC participating in the collaboration is not affected. The number of collaborating DSICs can also scale effectively as each mobile agent functions autonomously to satisfy the users' queries.

6. CONCLUSION

In this paper, we have proposed a framework for domain specific information clearinghouses to collaborate and share resources. Resources across multiple domains of interests can be located without impacting on the recall or the precision of the results. Currently, a formal mathematical model is being formulated to give a well-founded basis for the deployment of a number of collaborating clearinghouses in different domains.

7. REFERENCES

[1] D. Fensel, et al. On2broker: Semantic-Based Access to Information Sources at the WWW. In *Proceedings of the World Conference on the WWW and Internet (WebNet 99)*, Honolulu, Hawaii, USA, October 25-30, 1999.

[2] Resource Description Framework (RDF) Model and Syntax Specification. Technical Report. World Wide Web Consortium, 1998. http://www.w3.org/TR/REC-rdf-syntax

[3] ZDNet AnchorDesk. AltaVista CTO Responds. http://www5.zdnet.com/anchordesk/talkback/talkback_13066.html

[4] The Digital Library Project – DIENST http://www.ics.forth.gr/~dienst/

[5] ROADS. Resource Organization and Discovery in Subject-based Services http://www.roads.lut.ac.uk/

[6] The ISSAC Network http://www.scout.cs.wisc.edu/research/Issac/

[7] Social Science and Information Gateway. The Internet Detective. http://www.sosig.ac.uk/desire/internet-detective.html