Development of the prototype system for collection and integration of concept systems

Hideaki Takeda and **Masahiro Hamasaki** and **Ryutaro Ichise**
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
takeda@nii.ac.jp

Abstract

In this paper, we describe the outline and the current progress of our project called ”Collection and Integration of Concept Systems (CICS)” that aims to support development of semantics in the Internet. Since semantics is not always changing, providing fixed ontologies is not stuffiest. Our project aims to support dynamics of semantics by collecting and integrating various concept systems like Internet directories or bookmarks. We expect that integration of concept systems different in their size, shareability, purpose and so on, will yield new insight for both concept system. As the first prototype system, we build WebHical system that can integrate two concept systems by using both syntactic information (hierarchy of concepts) and semantic information (words in the contents).

Introduction

The amount of information on the Internet has been increasing with the accelerating speed. The problem we are facing is how to ensure the quality of information within the enormous amount of information. The movement of Semantic Web is the straight answer to the problem, i.e., providing the knowledge level markup to documents which are understandable intelligently both by human and machines. Since markup tags are based on ontologies, shared understanding is ensured. The mechanism is rational, but then the problem is shifted to ”how to provide ontologies” or ”how to construct semantic descriptions”? Simply providing some ”good” ontologies is not sufficient, because our semantics in the real world is dynamic in nature. Our semantics is divergent, i.e., from surely shared one to domain dependent or even private one. It is not adequate to think that all semantics should be fully shared ideally, rather divergence of shareability is important to keep semantics *alive*. Meaning sometimes comes from the domain-dependent semantics and becomes public. Meaning in public is sometimes transformed into the domain-specific one. Meaning is thus dynamic, so we should support such dynamics of meaning to realize truly useful knowledge-level descriptions (See Figure 1).

Collection and integration of concept systems

Towards support of the dynamics of semantics, we started a project called ”Collection and integration of concept systems” (CICSS). The goal is to provide an infrastructure that

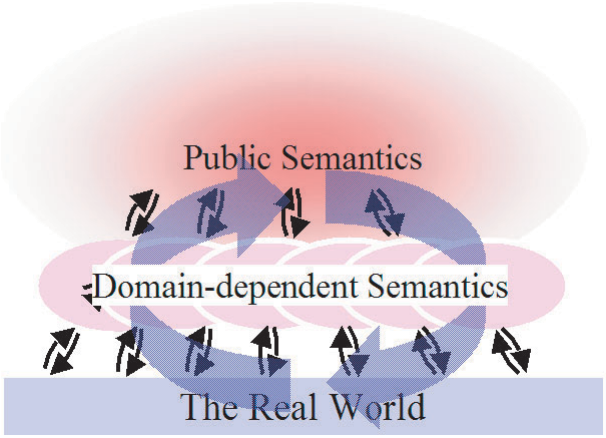

Figure 1: Dynamics of Semantics

can be used to generate conceptualization for new requirements by mixing and combining existing concept systems. To realize the goal, the mission of the project is simple, i.e., just collecting various concept systems in various fields or domains and relating them loosely. We expect that integration of concept systems different in their size, shareability, purpose and so on, will yield new insight for both concept system. One or more concept systems usually exist in each field / domain ether implicitly or explicitly. They are good sources of knowledge for that domain. All concept systems show their view of categorization at least, since the most primary nature of knowledge is categorization. Dealing with concept systems as categorization is expected to explicate the most basic level of knowledge. We propose concept systems repository as the workspace for the above research. The basic functions of concept systems repository (CSR) are as follows (Figure 2):

1. Import concept systems
2. Relate concept systems
3. Retrieve concepts or concept systems

The strategies to create CSR are (1) treating concept systems as a whole not as individual concepts, and (2) Emphasizing variety of concept systems not hiding it. The tasks to realize

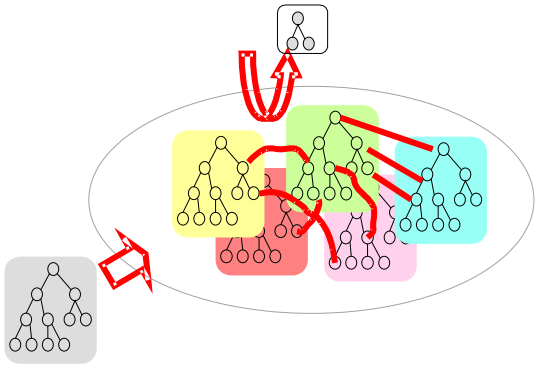

Figure 2: Concept Systems Repository

CSR are twofold. One is modeling concept systems to treat concept systems in a system, and the other is the algorithm to find relationship among them. We will consider the latter task more in the next section.

A preliminary consideration for algorithms to find relationship among concept systems

Finding relationship among different concept systems is a very difficult problem. We do not have the exact answer yet, but show some preliminary consideration and efforts for it. We adopt the instance-based or extension-based approach for discovery of relationship among concept systems. Definition-based or intension-based approach is not applicable in principle. If definitions from two concept systems are comparative, it tells us that they are essentially the same conceptualization. If they are formed by the different conceptualizations, definitions are not comparative. Since the instance-based approach does not care definition, it is suitable for our purpose. Of course the instance-based has other problems. One is how to identify instances, i.e., even definitions of instances can be different. The other problem is how to approximate instances, i.e., since instances can be enormous in nature, we will need some methods to say ”almost the same”. We have two trials based on this line.

Alignment of Internet directories

One is a method to align different Internet directories like YAHOO!, Lycos, and so on. The Internet directories are not strictly defined but highly elaborative and practical concept systems for the Internet. In this work, we proposed the method called HICAL to generate mapping rules from categories in one Internet directory to the other by the statistical method (Ichise, Takeda, & Honiden 2001c)(ICHISE, TAKEDA, & HONIDEN 2001a)(Ichise, Takeda, & Honiden 2001b). The basic idea is to find similar categories by evaluating how much they share instances (URLs). An example of the results is shown in Figure 3 (see (ICHISE, TAKEDA,

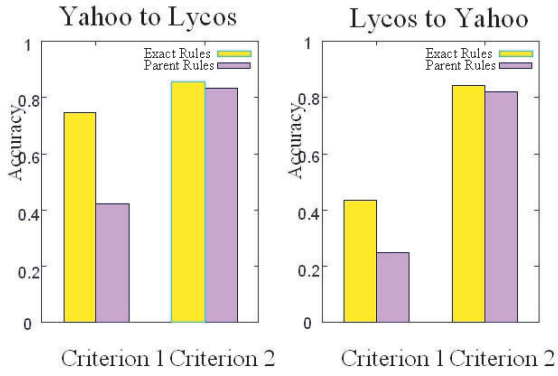

Figure 3: Results of mapping of categories between two directories (Literature case)

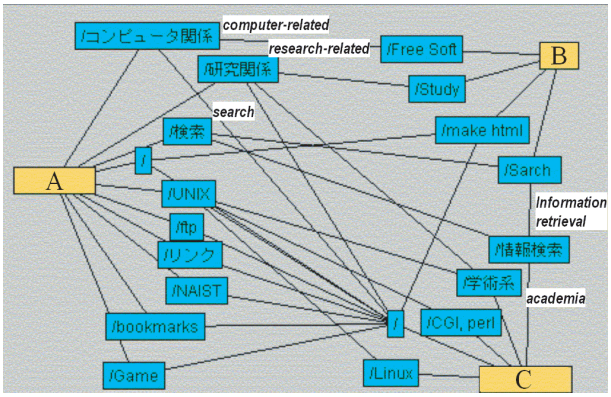

Figure 4: Mapping among WWW bookmarks

& HONIDEN 2001a) for details). The overall result is very hopeful; around 80% of instances can be mapped correctly from one directory to the other.

Finding relationship among WWW bookmarks

The other work is to find human relationship through WWW bookmarks. WWW bookmarks are results of implicit or explicit efforts to represent personal views for the Internet (Takeda, Matsuzuka, & Taniguchi 2000)(Hamasaki & Takeda 2001). Regarding WWW bookmarks as personal concept systems, we can find relationship among them with the similar method to the previous work. We call it shared topics network because concepts represent some topics which the user are interested in and they are shared among users by the relationship generated by this method. Figure 4 is an example of the generated shared topics network with three users. We can find some common relationship like (search, IR) and (academia, research-related), and community-dependent relationship like (Unix, academia). As the evaluation as recommender systems is good enough. Topics found by the system were apparently more acceptable than pages themselves.

System Architecture

The current architecture for CSR is based on the above two systems. The key function is to find similarity concepts among concept systems. We adopt HICAL as the basic algorithm to calculate similarity and expand it for more general use. The major extensions are as follows;

1. RDF and bookmark files as input:
   The system can accept bookmark files and RDF as data of concept systems. Bookmark files are used for concept systems by people or groups. RDF files are used for public directories. It can interpret files described with RDF such as RDF files provided by "The Open Directory Project."
2. Preprocessing by content-based similarity measurement:
   As we discussed in (Ichise, Takeda, & Honiden 2001c), HICAL is powerful and reliable in computing of similarity of concept systems because of dependency only on URL categorization, but is less useful when there are few shared URLs. In order to overcome this disadvantage, we find closely related URLs by examining their contents. If two URLs contain similar information enough, we add one to the category that the other belongs to, and vice versa. We can control the threshold to determine closely related URLs, e.g., the strictest case is the same as the original HICAL (no URLs are changed). We provide the preprocessing unit to find such relations.

The overall architecture is shown in Figure 5. Rel-extractor finds pairs of similar pages among pages in the given concept systems by analyzing their contents, and CS-extender modifies the given concept systems to include quasi-identical URLs. We use GETA (Generic Engine for Transposable Association)[1] to calculate matrix of page similarity

We are now testing the system with some small data and are going to apply the realistic data soon. An example with a test data is shown in Figure 6. The example is calculation of a part of Open directory ("Internet") and a personal bookmark. The upper-left part is the concept systems of the former, and the lower right is the latter. The lines between them are found links between two concept systems. In this case, there are no shared URLs between two concept systems but the system found several relationship between them.

Related Work

Integration of ontologies is getting one of the important issues in this field and researchers from different disciplines work for this theme. From knowledge engineering view, tools to realize integration are important. It is reasonable for ontology editors to provide some functions for ontology integration (for example (Noy & Musen 2000)). The other approach aims to automate ontology integration by using machine learning techniques(Wang, Zhou, & Liew 1999)(Agrawal & Srikant 2001)(Doan *et al.* 2002). But these all aim to integrate homogeneous ontologies where we

[1] http://geta.ex.nii.ac.jp/

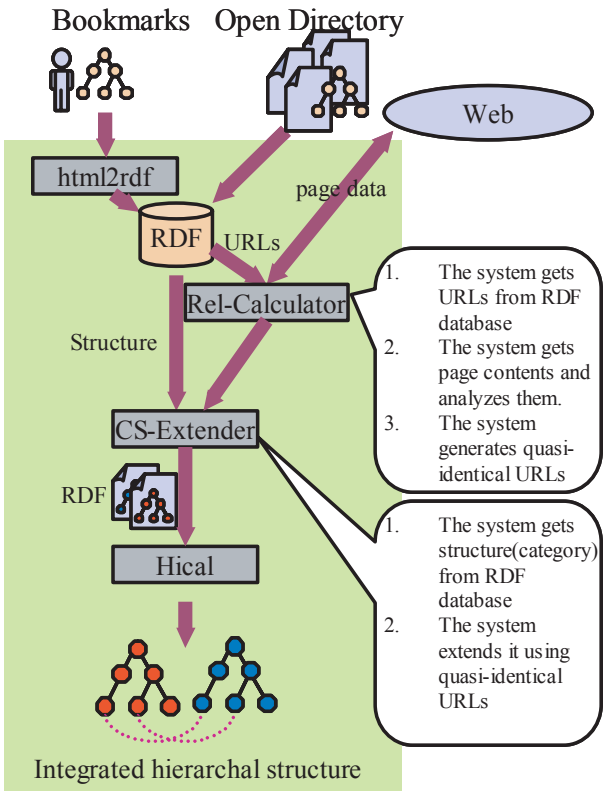

Figure 5: The architecture of WebHical

aim to integrate heterogeneous ontologies. The idea of ontology repository is firstly realized as Ontolinga server(Farquhar, Fikes, & Rice 1996) but it has no explicit functions to integrate ontologies.

Concluding remarks

We described a project called collection and integration of concept systems (CICSs). The goal is to provide a repository of concept systems that can be used to yield new conceptualization. In some sense, it is similar to Cyc, but Unlike Cyc, we do not create but collect knowledge. Our preliminary consideration leads to adopt the instance-based approach for discovery among concept systems, because concepts under different conceptualization cannot be compared directly.

References

Agrawal, R., and Srikant, R. 2001. On integrating catalogs. In *Proceedings of the 10th International WWW Conference*, 603–612.

Doan, A.; Madhavan, J.; Domingos, P.; and Halevy, A. 2002. Learning to map between ontologies on the semantic web. In *Proceedings of the 11th International WWW Conference*.

Farquhar, A.; Fikes, R.; and Rice, J. 1996. The ontolingua server: A tool for collaborative ontology construction. Technical Report KSL 96-26, Stanford University.

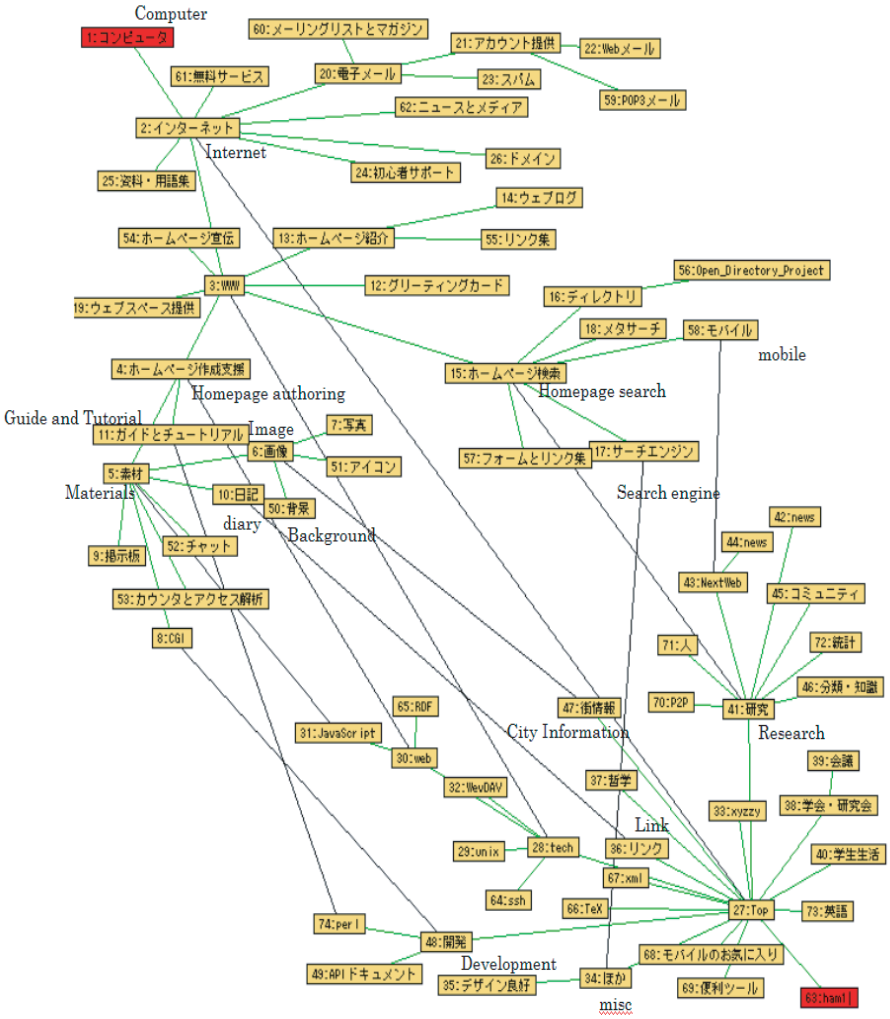

Figure 6: An example: Bookmark vs. Open Directory

Hamasaki, M., and Takeda, H. 2001. Experimental results for a method to discover of human relationship based on www bookmarks. In Baba, N.; Jain, L. C.; and Howlett, R. J., eds., *In Proceedings of Fifth International Conference on Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies (KES-2001)*, volume 2, 1291–1295. Osaka: IOS Press.

ICHISE, R.; TAKEDA, H.; and HONIDEN, S. 2001a. An alignment algorithm between concept hierarchies. Technical Report NII-2001-001E, National Institute of Informatics, Tokyo, Japan.

Ichise, R.; Takeda, H.; and Honiden, S. 2001b. Automated alignment of multiple internet directories. In *Poster Proceedings, The Tenth International World Wide Web Conference*, 194–195.

Ichise, R.; Takeda, H.; and Honiden, S. 2001c. Rule induction for concept hierarchy alignment. In *Proceedings of the IJCAI-01 Workshop on Ontology Learning (OL-2001)*, 26–29. (also available at CEUR Workshop Proceedings Vol-38 http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-38/).

Noy, N., and Musen, M. 2000. Prompt: Algorithm and tool for automated ontology merging and alignment. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.

Takeda, H.; Matsuzuka, T.; and Taniguchi, Y. 2000. Discovery of shared topics networks among people — a simple approach to find community knowledge from www bookmarks —. In *Proceedings of the Pacific Rim International Conference of Artificial Intelligence (PRICAI 00), Lecture Notes in Artificial Intelligence, No. 1886*, 668–678.

Wang, K.; Zhou, S.; and Liew, S. C. 1999. Building hierarchical classifiers using class proximity. In *Proceedings of the 25th International Conference on Very Large Data Bases*, 363–374.