

氏 名 AMAN SHAKYA

学位（専攻分野） 博士（情報学）

学位記番号 総研大 1290 号

学位授与の日付 平成 21 年 9 月 30 日

学位授与の要件 複合科学研究科 情報学専攻
学位規則第 6 条第 1 項該当

学位論文題目 Creating and Sharing Structured Semantic Web
Contents through the Social Web

論文審査委員 主 査 教授 武田 英明
准教授 Nigel Collier
准教授 相原 健郎
准教授 北本 朝展
教授 山口 高平(慶應義塾大学)

Sharing of information is important for its utilization to full potential. Information should be published with understandable semantics so that it can be used by others. It should also be accessible and properly disseminated. The Semantic Web provides structure and semantics to data making it machine understandable. The social web has made it easy for people to publish information online. It also enables collaboration and facilitates information dissemination by connecting people. These two areas complement each other to form a social Semantic Web. This is a highly promising direction but poses some major challenges.

The first challenge is to have people publish structured data on the social Semantic Web. Some specific problems for this are as follows. Systems for publishing structured data on the Semantic Web are complex and have considerable learning curve for people. It is also difficult for people to contribute due to strict constraints imposed by such systems. The second challenge is to form the models, so called ontologies, required to structure data with understandable semantics. People have a wide variety of data to share but there are limited ontologies and creating ontologies is difficult. Some specific problems for this dealt by the thesis are as follows. It is difficult to create perfect concept definitions to model things. It is not easy to cover the evolving requirements of all people. Moreover, different people may have multiple conceptualizations for the same thing due to different perspectives and contexts. It is not always possible to have consensus over conceptualizations and the collaborative process is itself difficult. Finally, proper dissemination of structured data on the web is also challenging. Information dissemination is mostly happening in a centralized and static way. There is a lack of flow of relevant structured information among people.

The thesis proposes some solutions to the specific problems. It proposes enabling people to contribute structured data by providing an easy-to-use social platform. It proposes allowing users to define their own concepts and freely contribute various types of data through a flexible and relaxed interface. Concepts contributed by people are partial definitions from their own perspective and multiple conceptualizations are allowed. These can be consolidated to form a rich unified conceptualization. This is possible by semi-automatic techniques for data integration and schema alignment supported by the community. A formalization of concept consolidation is also presented in the thesis. This serves as a loose collaborative approach that does not enforce consensus and direct interaction. Further, concepts can be semi-automatically grouped and organized by similarity. As a result of consolidation and grouping, informal lightweight ontologies gradually emerge in a bottom-up way. A system called StYLiD has been implemented to realize the proposed approach.

The thesis also proposes a decentralized approach for disseminating structured data in

communities. Relevant information can be aggregated through socially linked sources. This has been demonstrated experimentally. By combining the capabilities of publishing and aggregating, proper flow of information can be maintained in the community. A semantic blogging system called SocioBiblog has been implemented to demonstrate this for the bibliographic domain.

Experimental evaluations have been done to test the usability of StYLiD. Experimental studies have also been done to observe the multiple conceptualizations done by people and to verify that such conceptualizations can be consolidated. Methods used for concept consolidation and grouping have also been experimentally tested with some real data. The applicability and significance of the proposed approach has also been demonstrated by some real practical applications.

Through the system development and evaluation, the proposed approach has the potential for publishing and sharing information on the Internet. In particular, the following directions should be noted.

1. *Obtaining structured data from people.* The thesis proposes enabling ordinary people to author structured data for the Semantic Web by providing easy to use social web application interfaces. As extensions to existing social platforms, semantic blogging systems like SocioBiblog and OntoBlog which were implemented can facilitate easy publishing of particular types of data, for e.g., bibliographic data in case of SocioBiblog. However, it was soon realized that it is difficult to extend such systems for new types of data and even the existing types cannot evolve to accommodate requirements of people. Hence, a more flexible and generalized system, StYLiD, was implemented which enables people to share a wide variety of data of their interest by defining their own conceptual schemas. Keeping the input interface flexible and relaxed enables the users to contribute freely and easily. Freedom, ease of use and benefits are important factors for gaining social participation. Experiments showed that StYLiD is quite usable and almost requires no training to start contributing structured data. The lessons learned from the experiments can help in further refining the implementation to make it easier for people.
2. *Collaborative ontology creation.* To model the wide variety of data to be shared, new ontologies are required. Ontologies should be formed collaboratively to cover the requirements of different people. Some specific problems were identified in this area. Creating perfect concept definitions and building ontologies is a difficult process. It is difficult to achieve consensus on conceptualizations through direct collaboration. Therefore, following solutions were proposed.
 - *Defining concepts freely.* People should be allowed to define their own concepts to meet their needs and concept definitions should not be rigid and constrained. Experimental evidences were also presented supporting that people can and do express conceptual schemas and that constrained concept definitions can create problems for data contribution.

- *Allowing multiple conceptualizations.* Multiple conceptualizations should be allowed because people have different perspectives over the same thing or different contexts to be considered. Experimental evidences were also provided that different people have multiple conceptualizations
 - *Consolidation of multiple conceptualizations.* Such conceptualizations can be consolidated to form a unified model. This is possible with data integration principles and semi-automatic schema alignment methods. Consolidation serves as a new collaborative approach for creation of conceptualizations from the community. It is a loose collaboration requiring minimal interaction and consensus and facilitates collaborative knowledge formation while satisfying individual requirements. It was experimentally observed that conceptualizations of the same thing by different people overlap significantly and can be consolidated. It was also verified that satisfactory precision and coverage can be achieved even with simple methods of schema alignment.
 - *Emergence of informal lightweight ontologies.* Consolidation of concepts produces a unified common vocabulary for sharing different types of structured data. Concepts can further be grouped and organized semi-automatically. It was experimentally demonstrated that concept schemas can be grouped by similarity calculations with satisfactory precision and coverage. Concepts can evolve and emerge out of the cloud of concepts in the same manner as popular tags from a tag cloud.
3. *Structured information dissemination.* It is also important to have mechanisms for dissemination of the structured information in communities. Social web applications serve as excellent platforms for this by connecting people but are usually centralized and confine information within themselves. So a decentralized approach was proposed for dissemination of information across system boundaries. RSS feeds can easily be extended to transport structured data too. This was demonstrated by implementing SocioBiblog for sharing bibliographic information through social network links. It combines the capabilities of publishing and aggregating information into a single unit that can aggregate, filter and redistribute information. An evolving distributed network of such units can help in delivering relevant streams of information to people. The Semantic Web provides semantic structure and interoperability essential in such a decentralized environment. The proposed approach is applicable for any other system supporting RSS aggregation, including StYLiD.

博士論文の審査結果の要旨

His thesis is to show the possibility and potential of Social Semantic Web by proposing a novel way to publish and share the structured data. As the thesis is organized as follows.

In Chapter 1 and 2, the purpose of the study is identified with the explanation for Social Semantic Web. Social web where a huge amount of people participate to create and share information to each other becomes popular nowadays but it needs better ways to treat with more complicated information. Semantic Web gives the solution for it, i.e., creation and distribution of structures of metadata.

To make Semantic Web “social”, there are problems to solve, i.e., complexity and learning curve, difficulty of to create perfect concept definitions, existence of multiple conceptualizations, and difficulty of collaboration and consensus.

In Chapter 3, in order to overcome these problems, he proposes a new way of collaborative knowledge creation in which concepts can be created just by a few clicks and typing and can be shared and integrated semi-automatically. The latter feature is very unique so that multiple definitions can co-exist and even aligned to each other.

He developed system called “StYLiD” according to the above idea and demonstrated it with examples.

StYLiD system is designed to publish structured metadata and share them among people. As publishing system, it offers easy-to-use interface to create class definition of metadata and instance data based on the defined classes. Sharing is the most unique feature of the system. Users can share class definitions loosely. They can just borrow definitions by others and refine them for their purpose. Such different definitions can be integrated as a virtual class. They can map attributes of class definitions to those of the virtual class. Instance data published according to each definition are accessible through the virtual class. Thus variety and integration of class definitions are well balanced.

In Chapter 4, he focuses on how to distribute and share information. He shows to share information via social network and developed system called “sociobiblog” that enable sharing of information in the decentralized way.

In Chapter 5, he explains the results of three experiments to show how his approach is valuable and usable. The first experiment showed that the multiplicity of concepts and the alignment of them are natural. The second showed that automatic grouping and alignment is reliable. The third showed that the developed system is more usable than the existing system.

Finally he concludes his thesis in Chapter 6.

The committee evaluated the thesis highly as an important study bridging between semantic web technologies and social web activities.

After the presentation on his work, the committee members asked questions about his thesis. In particular, we discussed about the conceptualization and the conducted experiments. He explained the basic idea on how people want to represent and share concepts and how the existing approach cannot satisfy their needs. He also explained that there are no significant difference in evaluation between IT-related subjects and non IT-related subjects. It shows that the system can be acceptable not only for expert users but non-expert users who are target people for this study.

During the discussion period, he answered precisely for all the questions. We are convinced that he has the solid knowledge about computer science and semantic web technologies.