The Massive Knowledge Web

Xiaoping Sun
China Knowledge Grid Research Group
Institute of Computing Technology
Chinese Academy of Sciences
(+86)1062562703, Beijing, China
sunxp@kg.ict.ac.cn

Hai Zhuge
China Knowledge Grid Research Group
Institute of Computing Technology
Chinese Academy of Sciences
(+86)1062562703, Beijing, China
zhuge@ict.ac.cn

Qing Li
Department of Computer Science
City University of Hong Kong,
(+852)27889695, Hong Kong, China
itqli@cityu.edu.hk

ABSTRACT
As a vision for the future Web, this paper proposes *Massive Knowledge Web (MKW)*, a referential architecture that will support effective sharing of versatile knowledge in a large-scale decentralized network, and establish effective and efficient means to manage massive knowledge which may be created, derived and shared by various sources including individual end-users on the Web 2.0. The architecture incorporates a virtual P2P overlay into the server side platform to support efficient knowledge management, semantics-rich and personalized query processing for huge amounts of dynamic and personalized Web contents in a large-scale distributed network environment. Three major components are discussed herein: (1) semantic data model for massive Web2.0 content, (2) semantics-rich query model, and (3) scalable distributed semantic indexing model. Among other advantages, *MKW* provides a feasible architectural solution to build more accessible, scalable and intelligent services on the future Web.

Categories and Subject Descriptors
H.3.4 [**Information Systems**]: Systems and Software –Distributed systems, Information networks.

General Terms
Design.

Keywords
Web, Web 2.0, massive knowledge sharing, architecture.

1. INTRODUCTION
The emergence and popularization of Web 2.0 have significantly changed the logic architectures of the Web application systems. It not only hatches out new applications on the next generation Web platform, but also imposes new challenges on effective and efficient means towards making the massive knowledge on the Web more accessible. Search engine techniques have evolved from text information retrieval to global link-analysis-based ranking, achieving significant advances in terms of keyword-based query accuracy and recall rate when handling the vast amount of Web pages [5]. However, facing the rapid growth of Web 2.0, we argue that search engines need to reconsider the whole technical architecture to meet the new requirements of managing massive information/knowledge originated from personalized Web sources. Three major challenges are envisioned:

1. The current prevailing keyword-based Web information search is not expressive nor accurate enough to support semantics-rich personalized queries.
2. The core techniques of current Web information services like search engines are not fully prepared to support individual/personalized information and knowledge sharing on Web 2.0. In particular, personalized resources management and sharing will not only require a variety of query semantics but also call for different mechanisms of measuring, filtering, ranking and explaining the returned results.
3. Existing data models and system architectures of current Web information services are inadequate to provide scalable, intelligent and personalized information and knowledge sharing services in large-scale environments. The existing Web information organization models lack extensibility to accommodate more semantics and support semantics-rich search and queries; many important semantics of the original Web contents cannot be even retained, making it harder to extract richer semantics. On the other hand, logic based semantic model faces a critical problem of scalability when dealing with the dynamic and large-scale data on the Web.

These challenges drive us to reconsider the technical architecture of current Web information searching systems, and we propose a Web 2.0-oriented large-scale distributed system architecture called *Massive Knowledge Web (MKW)*. Among other advantages, *MKW* will provide a semantic/knowledge overlay over the Web, allowing users to publish, manage, share, and query their personalized information on the Web. The decentralized nature of Web 2.0 content sources inspires us to promote the integration of P2P computing techniques [2] with Web to enable scalable Web information sharing services on the future Web.

2. BACKGROUND
"*The World Wide Web (WWW, or simply Web) is an information space in which the items of interest, referred to as resources, are identified by global identifiers called Uniform Resource Identifiers*" − Architecture of the World Wide Web, www.w3.org. This simple architecture enables the Web to be easily, freely and quickly developed and boomed into the vastest information pool on the earth in less than ten years time. People must resort to machines to locate resources in the information space of the Web containing hundreds of billions of Web pages. Google and other search engines have indexed about ten billions of Web pages [12].

From the viewpoint of users, search engines are almost the only tool to search information on the Web when no direct information sources are available. However, when people want to find some specific information or knowledge, searching results are hardly satisfactory. In many cases, a searching process becomes even frustrating when users have to carefully manipulate the search process to narrow down the possible resource ranges. And finally, we have to make the decision on whether the ranked list should be trusted. A study shows that using Google, about 65% information

WWW 2008, April 21--25, 2008, Beijing, China.

queries are truly satisfactory, far less than the satisfaction rate of 85% of simple navigation queries [4].

From the viewpoint of system developers, we believe that several key factors lead to above situation. Firstly, URI is not semantically rich enough to support searching in the information space of the Web. Secondly, Web page techniques such as HTML/XHTML are mostly presentation-oriented not content-oriented. Thirdly, the content/data of Web pages is mixed with the presentation of content/data. The DOM tree structure is used more often to describe the presentation layout of Web pages than to describe the content of Web pages. Thus, Web pages are more like free text than semi-structured data. This feature enables developers to quickly write human-readable Web pages but makes it hard for a machine to search and query.

Most current search engines support only keyword-based search/query operations on Web pages. Although it is easy to use, the semantics of keyword-based queries can be too vague to accurately express the requirements of users, and search engines also cannot accurately represent Web page content by keyword combinations. Vague query semantics and incomplete Web content semantics motivate researchers to seek for more scalable and semantics-rich Web search solutions.

XML (htttp://www.w3.org/XML/) and Semantic Web techniques (http://www.w3.org/2001/sw/) are two complementary components to the Web semantics. Semi-structured data models and formal semantics are defined to enable more strict content-oriented description of resources on the Web. Structured queries are supported on XML data [8]. Logic reasoning can be performed on RDF documents (www.w3.org/TR/2004/REC-rdf-mt-20040210/). Involving more semantic structures will inevitably increase the computational complexity of searching [7]. This to some degree has hindered the extensive application of Semantic Web techniques on the Web. Search engines like Swoogle on Semantic Web provide a scalable indexing on Semantic Web documents [10]. Its index does not reflect the semantic contents of the indexed documents. RDF stores such as Sesame [6] and Jena [24] provides native or database based storage solutions to support semantic query languages including RDL [25] and SPARQL [26]. Those RDF storage and query solutions face a critical performance challenge upon query operations [18].

Recent emergence and popularization of Web 2.0 applications have dramatically changed the Web content publishing model. Web2.0 applications enable large amount of individual users to actively participate in publishing, sharing and managing their own data, information and knowledge on the Web. Blog systems (http://www.blog.com), Wiki (http://www.wikipedia.org) and Web tag systems (http://del.icio.us/) allow users to publish their data/information/knowledge on the Web in a managed way, so that the information can be shared and utilized more explicitly and purposefully. Those personalized information contents on the Web impose new challenges on search engines. User queries on Web 2.0 tend to be more specific and personalized than previous simple keyword-based queries.

To provide personalized search, existing solutions based on search engines mainly fall into two kinds, viz., search history based and user preferential based [9][10]. Such solutions do not focus on the inherent vague semantics of keyword-based queries. A few Web query languages and models were also studied to support more complex queries on [15][16][20]. These works mainly focus on querying the topology of the Web, rather than on the content of Web pages. Various information retrieval techniques are studied to extract more complex semantic information from free text on the Web [14][17]. Many question answering systems on the Web also provide more intelligent information searching services by using natural language process (NLP) method to acquire complex semantics from Web pages [19]. However, NLP methods and information retrieval methods are still not intelligent enough to discover the deep semantics of free text.

The core problem is the semantics on the Web. To solve this problem requires a compromising between rich semantics and scalable processing. It can be also seen that the problems existed on Web 1.0 are becoming even more challenging on Web 2.0 by using the same technical strategies of information searching. Traditional architectures of Web information services fall short in providing one-method-fits-all solutions. However, Web 2.0 provides a new chance to help tackle previous problems in an alternative way. Web 2.0 provides a platform to utilize the power of mass to make Web 1.0 and 2.0 more accessible. For example, Web 2.0 information can be utilized to improve search engines [3]. This paper aims at envisioning and proposing a distributed architecture that can enable users to publish, search and query the Web in a more intelligent and scalable way.

3. SYSTEM ARCHITECTURE

3.1 General Methodology

The MKW architecture aims at providing a distributed resource model, extensible indexing mechanism, and scalable query model over massive knowledge on the Web. Based on these, we are going to build a distributed large-scale knowledge Web to support massive personalized information and knowledge sharing and management. The decentralized nature of the massive personalized resource sharing and management drives us to advocate combining Peer-to-Peer (P2P) computing [2] with Web techniques to support scalable personalized knowledge management in large-scale distributed environments on Web 2.0. The core methodology is to utilize massive user power to divide and conquer the hard problems. The scalability and extensibility of P2P infrastructure make it a promising enabling technique to support various dynamic and personalized services in a large-scale network. It can help release the load of the centralized servers and make the system self-organized.

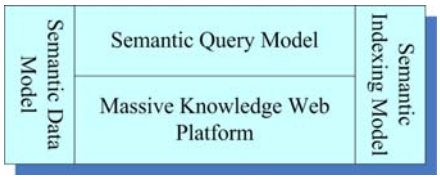

Figure 1. Massive Knowledge Web (MKW) Model.

The core of the MKW architecture should consist of at least four major components for supporting scalable Web content publishing, indexing and querying/searching (see Figure 1).

1. A distributed semantic data model that supports semantics-rich information and knowledge publishing, storage and querying on Web 2.0.
2. An extensible semantic query model that supports semantics-rich and computable query operations and expressions.
3. Scalable semantic indexing mechanisms that support effective and efficient query processing.

4. The MKW platform that incorporates above models to support resource management and sharing.

3.2 Platform Structure

To utilize the power of mass over the Web 2.0, a P2P overlay network based infrastructure will be adopted in the system architecture. On P2P overlay networks, peers act as both clients and servers, and are connected to form a certain type of network where queries can be forwarded by peers on the overlay to reach the target peer. Self-organization and scalable query routing in large-scale networks are two prominent features of P2P networks [2]. However, P2P networks face critical problem of resilience in highly-dynamic environments.

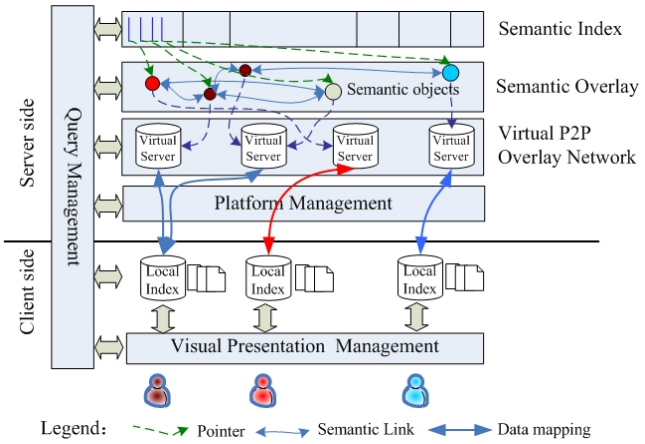

Figure 2. Massive Knowledge Web Platform Architecture

Instead of building a real P2P overlay network for users on the Web, embedding a virtual P2P overlay network into the server side architecture will be an alternative way to utilize P2P computing power while avoid its disadvantage. Figure 2 shows the proposed architecture. A real Web information source can publish one or several virtual servers on the virtual P2P overlay that are physically managed by the centralized servers of the platform. In this way, the problem of resilience can be effectively solved. The logic feature of P2P overlay networks can be retained and utilized to support scalable information querying.

On the MKW platform, semantic information is published as semantic objects which are to be discussed in next section. A virtual server is responsible for holding the semantic objects published by the owner of that virtual server, i.e., the real information source on the Web. Each virtual server is assigned with a unique semantic address on the virtual structured P2P overlay so that it can be efficiently located.

To enable semantic queries, semantic objects are also connected by their semantic relationships to form a semantic overlay. To support efficient and scalable query processing, a semantic index is built over the semantic overlay based on the hierarchical clusters of the published semantic objects.

The Query Management module will allow users to perform queries on different semantic levels. A local index at each peer on the Web supports direct information retrieval on the local contents. Users can also directly search other peers' contents by the virtual server addresses without consuming the server side computation. A set of semantic query operations will be also supported on the semantic overlay formed by the semantic objects. Semantic indexing can extend and facilitate scalable searching on the semantic overlay. User queries at different levels of semantic abstraction can be combined and integrated to provide users with more systematic results.

The Visual Presentation Management module will allow users to separate content from the user-interface by providing a set of presentation language and tools for defining specific presentation of data. Separating data from presentation enables users to flexibly manipulate the visual presentation without affecting the content structure of data. Changing content structure of data becomes more easily without concerning about the presentation formats.

By incorporating the P2P overlay techniques at the server side architecture, local semantic features of Web content can be well kept, which serves as the most important feature enabling scalable and extensible personalized information services on the future Web.

3.3 Scalable Semantic Data Model

The semantic data model is the main trunk that eliminates semantic gap between different levels of the system. To make the system scalable, selecting a proper semantic data model is critical. Formal logic models such as Description Logic [1] own strict and rich expressiveness but often have high computational complexity. Instead, a simple model that allows users to contribute their own semantic information in a scalable way is desirable. Here, we suggest the Semantic Link Network (SLN) model to support semantics representation, management and sharing [23].

The core of the semantic data model SLN is the semantic object defined as a tuple $SO(A, L)$, where A is the attribute set containing description of the semantic object and L is the link set containing semantic links to other semantic objects. Each attribute is defined as a triple $a(n, p, v)$ where n is the name of the attribute, p is the predicate indicating the semantic interpretation of the attribute value v of a. For example a("name", "is", "Johnson") represents an name attribute of an object. The semantic link is defined as a tuple $l(t, p, r)$ where t is the name of the link, p is the semantic interpretation of the link and r is the pointer to a remote semantic object. Semantic objects are taken as a piece of information with certain attributes and links. Specific interpretations are left to user-defined semantic models. Semantic objects can encapsulate both HTML/XHTML Web contents and the RDF-based semi-structured Semantic Web data. That is, semantic objects can support different user-defined semantic models.

The semantic data model should have the following features:

1. Opening structure that allows different or even conflicting user-defined semantic models coexists.
2. Index-able semantic data structure that allows semantic information in different user-defined semantic models can be indexed under one unified indexing structure.
3. Unified semantics searching model that allows queries/searches in different user-defined semantic models.

3.4 Extensible Semantic Query Model

Based on the semantic link network model, a distributed semantic query model is required to provide a simple and extensible query interface for users. The query model will provide basic query operation framework based on the SLN semantic data model. The query model will also allow users to define the query operations

on their own (user-defined) semantic models with the basic query operation framework; this is because query semantics depends on the semantics of contents published by users themselves. Two basic types of query operations are to be provided, namely, attribute-based and link-based. Both are to return the semantic objects that satisfy certain computable query conditions. Although more complex queries such as join can be defined, we leave such user-defined operations to users at the client-side for the sake of system scalability.

3.5 Scalable Semantic Indexing

To support scalable semantic query processing, a three-dimensional resource space scheme [22] can be used to hold the attribute triples and link triples based on the hierarchical classification semantics. Previous high-dimensional indexing structures such as R* tree can be used to support efficient tuple searching in the resource space. Moreover, the semantic paths of the semantic overlay can be also indexed based on user-defined query history, so as to facilitate fast extraction of structures of the semantic objects in the semantic overlay.

4. CONCLUSION

In this paper, we have advocated the MKW architecture with the enabling characters that will make information publishing and sharing on Web 2.0 more flexible, efficient and effective and intelligent in large-scale distributed environments. It will be an important step towards our Knowledge Grid endeavor [21].

5. ACKNOWLEDGMENTS

This research work was supported by the National Basic Research Program of China (Project No. 2003CB317000).

6. REFERENCES

[1] Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P.F., editors. The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press, 2003.

[2] Balakrishnan, H., Frans Kaashoek, M., Karger, D., Morris, R., and Stoica, I. “Looking Up Data in P2P Systems”, Communications of the ACM, vol.46, no.2, pp. 43 – 48, 2003.

[3] Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., and Su, Z. Optimizing web search using social annotations. In Proceedings of WWW '07. ACM, New York, NY, 501-510. 2007.

[4] Liu, B. Personal Evaluations of Search Engines: Google, Yahoo! and MSN, http://www.cs.uic.edu/~liub/searchEval /SearchEngineEvaluation.htm. 2007.

[5] Brin, S. and Page, L. The anatomy of a large-scale hypertextual Web search engine. Comput. Netw. ISDN Syst. 30, 1-7 (Apr. 1998), 107-117, 1998.

[6] Broekstra, J., Kampman, A., Harmelen, F., Sesame: Ageneric Architecture for Storeing and Querying RDF and RDF Schema, In Proceeding of the 1st Semantic Web Conference, pp. 54-68, 2002.

[7] Bojańczyk, M., David, C., Muscholl, A., Schwentick, T., and Segoufin, L. Two-variable logic on data trees and XML reasoning. In Proceedings of PODS '06. ACM, New York, NY, 10-19, 2006.

[8] Chien, S., Vagena, Z., Zhang, D., Tsotras, V. J., and Zaniolo, C. Efficient structural joins on indexed XML documents. In Proceedings of Very Large Data Bases 2002, 263-274, 2002.

[9] Coyle, M. and Smyth, B. Supporting intelligent Web search. ACM Trans. Inter. Tech. 7, 4, 20. 2007.

[10] Ding, L., Finin, T., Joshi, A., Pan,R., Scott Cost, R., Peng, Y., Reddivari, P., Doshi, V., Sachs, J., Swoogle: A Search and Metadata Engine for the Semantic Web, CIKM 2004, 652-659, 2004.

[11] Dou, Z., Song, R., and Wen, J. A large-scale evaluation and analysis of personalized search strategies. In Proceedings of WWW '07. ACM, New York, NY, 581-590, 2007.

[12] Gulli, A. and Signorini, A. The indexable web is more than 11.5 billion pages. In Special interest Tracks and Posters of WWW '05. ACM, New York, NY, 902-903, 2005.

[13] Kanza, Y. and Sagiv, Y. Flexible queries over semistructured data. In Proceedings of PODS '01. ACM, New York, NY, 40-51, 2001.

[14] Luo, G., Tang, C., and Tian, Y. Answering relationship queries on the web. In Proceedings of WWW '07. ACM, New York, NY, 561-570, 2007.

[15] Mendelzon, A. O. and Milo, T. Formal models of Web queries. In Proceedings of PODS '97. ACM, New York, NY, 134-143, 1997.

[16] Mihaila, G. A. WebSQL-An SQL-like query language for the world-wide web. Master's Thesis, Department of Computer Science, University of Toronto. 1996.

[17] Nie, Z., Ma,Y., Shi, S. Wen, J., Ma, W., Web Object Retrieval, In Proceedings of WWW '07. ACM, New York, NY, 81-89, 2007.

[18] Pérez, J., Arenas, M., Gutierrez, C, Semantics and complexity of SPARQL. ISWC 2006, 2006. Springer.

[19] Radev, D., Fan, W., Qi, H., Wu, H., and Grewal, A. Probabilistic question answering on the web. In Proceedings of WWW '02. ACM, New York, NY, 408-419, 2002.

[20] Raghavan, S. and Garcia-Molina, H. Complex queries over web repositories. In Proceedings of VLDB 2003, 33-44, 2003.

[21] Zhuge, H. The Knowledge Grid, World Scientific, 2004.

[22] Zhuge, H. The Web Resource Space Model, Springer, 2007.

[23] Zhuge, H. Autonomous Semantic Link Networking Model for the Knowledge Grid. Concurrency and Computation: Practice and Experience 19(7) (2007) 1065-1085.

[24] Jena, A Semantic Web Framework for Java, http://jena.sourceforge.net/.

[25] SeRQL, Sesame RDF Query Language, http://www.openrdf.org/doc/users/ch05.html, 2003.

[26] SPARQL Query Language for RDF, W3C Recommendation, http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/, 2008.01.