Data Mediation and Interoperation in Social Web: Modeling, Crawling and Integrating Social Tagging Data

Ying Ding
University of Innsbruck
Technikerstraße 21a
6020 Innsbruck, Austria
ying.ding@sti2.at

Ioan Toma
University of Innsbruck
Technikerstraße 21a
6020 Innsbruck, Austria
ioan.toma@sti2.at

Sin-Jae Kang
Daegu University
South Korea
sjkang@daegu.ac.kr

Michael Fried
University of Innsbruck
Technikerstraße 21a
6020 Innsbruck, Austria
michael.fried@student.uibk.ac.at

Zhixian Yan
Swiss Federal Institute of Technology
Station 14, CH-1015
Lausanne, Switzerland
zhixian.yan@gmail.com

ABSTRACT
Data mediation and interoperation have already become one of the central topics of IT for decades. Since the Web appears, this problem has been exploded due to the increasing amount of data and Web users. On the one hand, the current Web makes this problem complicated; on the other hand, it also provides space for potential solutions. This paper instances it in social tagging system. It proposes Upper Tag Ontology (UTO) to model and integrate different tagging data.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Search and Retrieval

General Terms
Algorithms, Experimentation, Human Factors.

Keywords
Social tagging, upper tag ontology (UTO), data integration

1. INTRODUCTION
The previous Web is a syntactically structured Web weaved by un-typed hyperlink and mainly a read-only Web. It has built an exciting multimedia world for the users but provides little support for them to share and collaborate directly. Instead of one-way communication by presenting information on the Web, the Web has changed from the place to read to the place to write and share. It creates a platform for knowing people and sharing information.

The term "Social Web" was introduced in 1998 by Peter Hoschka [1] who tried to stress the social medium function of the Web. From Wikipedia, the Social Web is defined as an open global distributed data sharing network which links people, organizations and concepts. Here, Social Web is extended to include any Web related technologies, phenomena and development which aim to enhance the social feature of the Web. Current Web 2.0 is the main stream of the Social Web which provides platform and technologies (such as wiki, blog, tag, RSS feed, etc.) for online collaboration and communication. Photos, bookmarks, news, diaries, music, videos and many other data are shared directly on the Web now and the new data is generated daily.

Data mediation and interoperation have already become one of the central topics of IT for decades ([2], [3]). Since the Web appears, this problem has been exploded due to the increasing amount of data and Web users. There are already some efforts aiming to provide machine supported meditation on the Web ([4], [5]). Among them is adding metadata. There are different ways of adding metadata. Well-defined formal way driven by the Semantic Web is to first build up ontologies and then annotate the Web data based on existing ontologies. Data mediation problem has been shift from data level up to ontology level. Since ontology reflects the shared understanding and conceptualization in a domain and is represented in formal and machine processable ontological languages, certain level of mediation can be automated within certain domains. But ontology generation, annotation, and maintenance are very time consuming and hardly scalable [6]. Social-driven approach mainly from Web 2.0 is to allow users to tag anything in anyway they like and this leads to various tag clouds, folksonomies, and wikipedias. This gives the full freedom to the end users and provides sufficient tool supports. But user added metadata are not well-defined to reflect community consensus and are not formally represented in machine understandable manner, data mediation can be hardly achieved automatically [7].

Besides those, standards-based approach tries from different perspective to realize the compatibility between systems, databases and services. The standard organizations such as the World Wide Web Consortium (W3C) have directed major efforts at specifying, developing, and deploying standards for sharing meaning. These efforts certainly march a crucial step to promote the wide spread deployment to enhance the Web functionality and interoperability [4]. Vertical and horizontal domain metadata have been gradually established, for example, FOAF (metadata for friends), SKOS (metadata for taxonomies), DOAP (metadata for project), RSS (metadata for news), SIOC (metadata for social networks), Dublin Core (metadata for documents), GEO (metadata for geographic coordinates),

GeneOnt (metadata for human genes), microformat (metadata for Social Web) and so on. All these efforts are trying to establish their roles to alleviate the interoperation problem.

This paper takes major social tagging systems as examples, namely delicious, flickr and youtube, to analyze the social phenomena in the Social Web in order to identify the way of mediation and integration of social data. The main contributions of our work include:

- Modeling social tagging data based on proposed Upper Tag Ontology (UTO). Mediating UTO with other related social metadata (such as FOAF, DC, SIOC, SKOS, etc.)
- Crawling tag data from major social tagging systems and integrating them according to UTO.
- Searching tagging data across different tagging systems.

According to above, this paper is organized as follows. Section 2 gives the detailed description of how to model social tagging data, how to crawl social tagging data and how to integrate and search them across different tagging systems. Section 3 discusses the related work. Section 4 provides discussion and Section 5 concludes the paper and presents some future work.

2. SOCIAL TAGGING

Tag is a keyword used to categorize online objects. The goal of tagging is to make a body of information increasingly easier to search, discover, share and navigate over time. Social tagging is not simply just tagging, tags are social metadata generated from collective intelligence. The consensus of tags forms social semantics which are called folksonomies. It is bottom-up approach and reflects collective agreement. It speaks the same language as the users and makes the things easier to find.

2.1 Modeling Social Tagging Data

We can tag bookmarks (del.icio.us), photos (flickr), videos (YouTube), books (LibraryThing), Music (Last.fm), citations (CiteULike), blogs (Technorati), etc. Tag is nothing special than a typed hyperlink. We can use "rel" attribute to create typed hyperlink. There are many social networks providing tagging services, here we take three major social tagging systems, namely delicious, flickr, and youtube, to analyze their social tagging behavior. Based on this analysis, we propose Upper Tag Ontology (UTO) which is originated from Tag Ontology proposed by Tom Gruber [8]. In his tag ontology, he proposed five key concepts which are object, tag, tagger, source and vote. Here in UTO, we add another two concepts: comment and date. Furthermore, we add has_relatedTag relationship to tag concept itself. More details about modeling social tagging data were discussed in [9].

Let O be UTO ontology, $O = (C, \Re)$ (1)

Where $C = \{c_i, i \in N\}$ is a finite set of concepts

$\Re = \{(c_i, c_k), i, k \in N\}$ is a finite set of relations established among concepts in C.

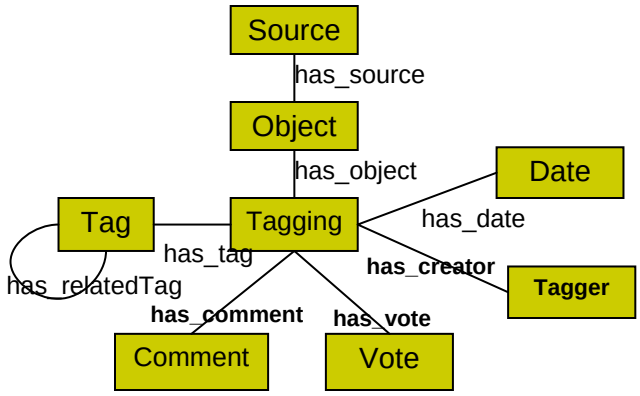

Figure 1. Upper Tag Ontology (UTO)

UTO is different comparing to folksonomy which is focusing on the meaning of the tags. With the basic ontology design idea of "making it easy and simple to use", UTO is designed to capture the structure of the social tagging behavior rather than the topic or meaning of the tags. It aims to model the structure of the tagging data in order to integrate different tagging data and mediate them with existing social metadata. Furthermore, the alignment between UTO and other social metadata, such as FOAF, DC, SIOC and SKOS can be easily established.

2.2 Crawling Social Tagging Data

ST crawler (Social Tagging crawler) is a developed multi-crawler designed for crawling major social tagging systems including del.icio.us, flickr and youtube [10]. This crawler is based on the "Smart and Simple Webcrawler"[1]. The ST crawler is written in Java with Eclipse IDE 3.2 on Windows XP and Ubuntu 6.04. Data has been cleaned up using linux batch commands. ST crawler can start from one or a list of links. Here shows one example of using UTO to model tag data in del.icio.us. The instance data is represented in RDF triple according to UTO. For example:

One user has tagged http://www.deri.org on 20.07.2007. The page has been saved by 2467 other people, on del.icio.us with the tags web2.0, semanticweb, deri and innsbruck. He gave it the comment "Deri home". The entry would be crawled via the http://del.icio.us/tag/web2.0 page. The tags tools, blog, design and community are stored as related to web2.0. The output tag data according to UTO in RDF would look like following.

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:j.0="http://uto.deri.at/" >
  <rdf:Description rdf:about="http://uto.deri.at/b47145e3-5417-
4d6e-885e-43469bc18ca4">
   <j.0:has_tag rdf:resource="http://del.icio.us/tag/web2.0"/>
   <j.0:has_tag rdf:resource="http://del.icio.us/tag/semanticweb"/>
   <j.0:has_tag rdf:resource="http://del.icio.us/tag/deri"/>
   <j.0:has_tag rdf:resource="http://del.icio.us/tag/innsbruck"/>
   <j.0:has_comment>Deri home</j.0:has_comment>
   <j.0:has_vote>2468</j.0:has_vote>
   <j.0:has_tagger>MichaelFried</j.0:has_tagger>
   <j.0:has_date>Jul 07</j.0:has_date>
```

[1] https://crawler.dev.java.net/

```
    <j.0:has_object rdf:resource="http://www.deri.org"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://del.icio.us/tag/web2.0">
            <j.0:has_related_tag
rdf:resource="http://del.icio.us/tag/tools"/>
      <j.0:has_related_tag
rdf:resource="http://del.icio.us/tag/blog"/>
      <j.0:has_related_tag
rdf:resource="http://del.icio.us/tag/design"/>
      <j.0:has_related_tag
rdf:resource="http://del.icio.us/tag/community"/>
    </rdf:Description>
    <rdf:Description rdf:about="http://www.deri.org">
      <j.0:has_source>http://del.icio.us</j.0:has_source>
    </rdf:Description>
  <rdf:RDF>
```

For del.icio.us, the whole UTO is filled out. For youtube, there are no related tags objects to store. For flickr, the source is the URL of the page containing the picture instead of flickr.com. You need the source to identify the type (bookmark, picture, video) of the object. Has-vote means a link has been tagged multiple times (del.icio.us), a photo has been favoured (flickr.com) or a video has been voted for (youtube.com). Finally, ST crawler has crawled social tagging data from delicious, flickr and youtube and modelled them according to UTO. These data are represented in RDF triples and stored in Jena. In total, the crawled output contains several RDF files with a complete file size of 2,10G B. In detail:

- 16 del.icio.us data files at a size of 1,64GB
- 3 flickr data files at a size of 233MB
- 3 youtube data files at a size of 234MB

2.3 Integrating Social Tagging Data

The integrated tagging data from these three social tagging systems have been stored in Jena. Based on those data, some interesting queries can be performed. For example, we take one tag as input and returns a list of objects and their votes ordered descendent by vote's value (see Figure 2).

SPARQL query:

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
select distinct ?object ?vote where {
{ ?x <http://uto.deri.at/has_object> ?object .
?x <http://uto.deri.at/has_vote> ?vote .
?x <http://uto.deri.at/has_tag> <http://del.icio.us/tag/" +
tag_text.getText() + ">}
UNION
{ ?x <http://uto.deri.at/has_object> ?object .
?x <http://uto.deri.at/has_vote> ?vote .
?x <http://uto.deri.at/has_tag>
<http://flickr.com/photos/tags/" + tag_text.getText() + ">}
UNION
{ ?x <http://uto.deri.at/has_object> ?object .
?x <http://uto.deri.at/has_vote> ?vote .
?x
<http://uto.deri.at/has_tag><http://youtube.com/results?
search_query=" + tag_text.getText() +
"&amp;search=tag> }
}order by desc(xsd:integer(?vote))
```

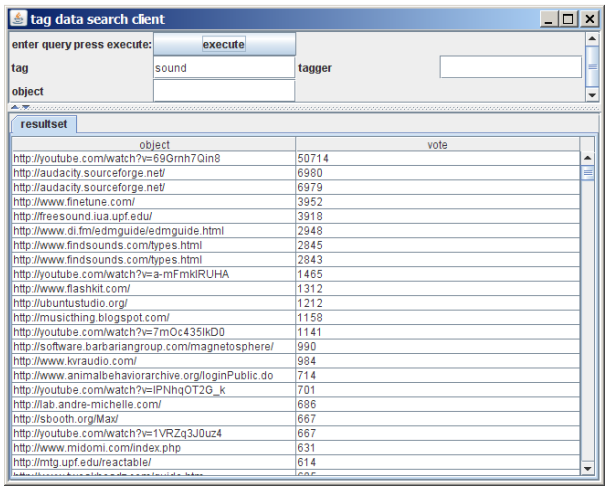

Figure 2. Scenario 1 search frame

3. RELATED WORK

In 2005, Tom Gruber proposed the idea of using ontology to model tagging data. His idea has been further formalized and published in 2007 [8]. His tag ontology contains tagging (object, tag, tagger, source, + or -). He introduced vote to tag ontology and uses it for collaborative filtering. UTO contains more concepts and relations comparing to his tag ontology, such as date, source, comment, etc. Furthermore, UTO also focuses on integration with other existing social metadata in order to achieve data integration. UTO is based on Gruber's idea and goes a bit further on ontology alignment and data integration.

SCOT[2] (Social Semantic Cloud of Tags) Ontology semantically represents the structure and semantics of a collection of tags and to represent social networks among users based on the tags. While UTO does not care much of tagcloud and it is defined in such a way which can be further aligned with many other social metadata, such as DC, microformat, etc.

Holygoat Tag Ontology[3] models the relationship between an agent, an arbitrary resource and one or more tags. Taggers are linked to foaf:agents. Taggings reify the n-ary relationship between tagger, tag, resource and data. They can perform some simple subsumption inference. This approach goes a bit deep to semantic web by utilizing ontology reasoning and inference. UTO aims to keep things simple and easy to use therefore ontology reasoning and inference is not considered at this stage.

MOAT Ontology[4] is a lightweight ontology to represent how different meanings can be related to a tag. MOAT assumes that there exists a unique relationship between a tag and a label that a tag can have a unique MOAT identifier. UTO cares more about the structure of the tagging behavior rather than the meaning of the tags. But provide unique identifier to tag is always a helpful and important issue to social tagging and furthermore to web in general.

[2] http://scot-project.org/

[3] http://www.holygoat.co.uk/projects/tags/

[4] http://moat-project.org/ontology

4. DISCUSSIONS

The current Web has experienced tremendous changes to connect information, knowledge, people and intelligence. Meanwhile, Web 2.0 represented Social Web has successfully motivated users to share information and collaborate each other directly via the Web [11]. Web 2.0 is not completely different from the Semantic Web. As Sir Tim Berners-Lee mentioned "the Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation[5]". Web 2.0 not only extends the communication dimensions (publishing, commenting and arguing) but also tries to add extra contextual information (we can call it "social metadata") to the current Web data in a social and informal way (e.g. tagging, bookmarking and annotating). Web 2.0 provides scalable community-powered information sharing platform, while the Semantic Web adds valuable machine understandable metadata to enable efficient and automatic way of heterogeneous information sharing and cross-portal communication and collaboration.

This paper takes social tagging systems as examples and aims to identify some pragmatic ways of utilizing Semantic Web and Social Web phenomena to realize data mediation and integration. A simple Upper Tag Ontology (UTO) is proposed to integrate different social tagging data and mediate with other related social metadata. It has the following important features:

- *Community driven mediation based on collective intelligence:* Social Web changes the current Web into a community platform where ordinary users participate daily for communication and collaboration. This social synergy can be used for data mediation as mediation itself is a kind of activity supporting communication and collaboration. Community driven mediation based on social collective intelligence can be an appropriate approach for data mediation.
- *Instance-based metadata mediation:* There are already some existing researches on instance-based metadata mediation from the Semantic Web and database area. But they are more focusing on the formal transformation problem between schema and instances. Ideas on how to advance the data mining techniques to mediate metadata based on instances and contextual information around the data and metadata can be further explored.
- *Efficient mashing-up of Social Web services and metadata semantics:* Web is often described as being in the Lego phase, with all of its different parts capable of connecting to one another. Properly mashing-up social services can assist the mediation process and further enable the browsing and querying of the mediated data.

5. CONCLUSIONS AND FUTURE WORKS

Social aspect of the Web indeed influences fundamentally the usage and sharing of the web information. The Web relies on people serving useful content, linking them and providing trust and feedback. The massive participation of the web users has significantly increased the heterogeneity of the Web. On the other hand, it has created the additional way for data integration, namely integration by collective intelligence. By tagging and sharing data, intuitively they also enrich the contextual information of the concepts and relations. Here we take social tagging systems as examples to identify some pragmatic ways of utilizing Semantic Web and Social Web phenomena to realize data mediation and integration. In the future, we would like to put some efforts to mine some associations among these tagging data in order to portray tagging behavior in current social networks. We can also build up recommender systems based on these associations. Furthermore, some efficient statistical methods can be identified to extract mediation rules based on instances and contextual information.

[5] http://www.w3.org/2001/sw/EO/points

6. REFERENCES

[1] P. Hoschka. CSCW research at GMD-FIT: From basic groupware to the Social Web. *ACM SIGGROUP Bulletin*, 19(2): 5-9, 1998.

[2] C. Batini, M. Lenzerini and B. Navathe. A Comparative Analysis of Methodologies for Database Schema Integration. *ACM Computing Surveys*, 18(4): 323-364, 1986.

[3] E. Rahm and P. A. Bernstein. A Survey of Approaches to Automatic Schema Matching. *VLDB Journal: Very Large Data Bases*, 10(4): 334-350, 2001.

[4] G. Antoniou and F. van Harmelen. *A Semantic Web Primer*. MIT Press (ISBN 0-262-01210-3), 2004.

[5] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, 2001.

[6] A. Gomez-Perez, M. Fernandez-Lopez, and O. Corcho. *Ontological Engineering: Advanced Information and Knowledge Processing*. Springer. 2003.

[7] P. Mika. Social Networks and the Semantic Web. Springer (ISBN: 978-0-387-71000-6), 2007.

[8] T. Gruber. Ontology of Folksonomy: A Mash-up of Apples and Oranges. *International Journal on Semantic Web & Information Systems*, 3(2), 2007. Available at: http://tomgruber.org/writing/ontology-of-folksonomy.htm

[9] Y. Ding, S. J. Kang, I. Toma, M. Fried, O. Shafiq and Z. Yan (submitted). Adding Semantics to Social Tagging: Upper Tag Ontology (UTO). The 70*th Annual Meeting of the American Society for Information Science & Technology (ASIS&T)*, Oct 24-29, 2008, Columbus, Ohio, USA.

[10] M. Fried. *Social Tagging Wrapper*. Bachelor Thesis, Institute of Computer Sciences, University of Innsbruck, Austria, 2007.

[11] D. Hinchcliffe. The State of Web 2.0. *Web Services Journal*. 2006. Available: http://web2.wsj2.com/the_state_of_web_20.htm