Using Social Data as Context for Making Recommendations: An Ontology based Approach

Salma Noor
School of Electronics and Computer Science, University of Southampton
Southampton, UK, SO17 1BJ
sn07r@ecs.soton.ac.uk

Kirk Martinez
School of Electronics and Computer Science
University of Southampton
Southampton, UK, SO17 1BJ
km@ecs.soton.ac.uk

ABSTRACT
Web-based knowledge systems support an impressive and growing amount of information. Among the difficulties faced by these systems is the problem of overwhelming the user with a vast amount of data, often referred to as information overload. The problem has escalated with the ever increasing issues of time constraints and the extensive use of handheld devices. The use of context is one possible way out helping with this situation. To provide a more robust approach to context gathering we propose the use of Social Web technologies alongside the Semantic Web. As the social web is heavily used it could provide a better understanding of a user’s interests and intentions. The proposed system gathers information about users from their social web identities and enriches it with ontological knowledge. Thus an interest model for the user can be created which can serve as a good source of contextual knowledge. This work bridges the gap between the user and system searches by analyzing the virtual existence of a user and making interesting recommendations accordingly.

Categories and Subject Descriptors
H.3.3 [Information Storage And Retrieval]: Information Search and Retrieval – Information filtering, Selection process; H.3.1 Content Analysis and Indexing – Linguistic processing, Thesauruses; H.3.5 Online Information Services – Web-based services, Data sharing.

General Terms
Design, Human Factors, Theory, Algorithms.

Keywords
Ontologies, Context, Semantic Web, Web 2.0, user interest modeling, tagging, Recommender Systems, Cultural Heritage, linked data.

1. INTRODUCTION
Avoiding the cold-start and data overload problems are important issues common to information search and retrieval systems that tend to provide some sort of adaptation.

The cold-start problem is recommending items of interest to new users who do not have any related preferences in their profile. This especially concerns the issue that, the system may not have any trails or information useful in building a user’s interest context. The data overload problem is well known problem and is defined in many ways in literature. However we refer to it as a state when an overwhelming amount of information is available to present to the user, some of which might be highly relevant but is lost due to a lack of focus or filtering.

In practice the best known way to handle the data overload problem is by filtering information using context. While the cold start problem can be tackled by seeding then keeping track of users’ interests and intentions, to be used as context in different situations where prior knowledge about the user is required. This avoids the trickier issue of asking for preferences initially.

Context is a very broad term which can be quiet misleading. For our research context encompasses a set of interests from user profile, which are extracted from the user’s social web interactions and tagging activities, as well as the goals defined by the user. We also hope to include the current physical context of the user, like time and place/geographic location, but not at this stage.

However dealing with context is accompanied by issues like (1) making the system fully understand the context of the task in hand without tedious efforts by the user (2) finding and retrieving the desired data/information automatically, usually involving the integration of data/information from different sources to draw useful conclusions (3) and means of acquiring contextual information without breaching security and privacy issues, this might include blending relevant personal and public data on web.

We suggest that the social web can assist in context gathering tasks. We aim to use the semantic web and ontologies to help with context representation and interpretation. But the ontologies themselves need to be interpreted according to a certain context.

To address these issues in this paper we propose to:
- Define a generalized context model of user interest that serves as an interpretation of user’s intention/interest and assist during recommendation or searching processes.
- This user interest model is built initially by capturing users’ social-web data, mostly tagging activities. Which we believe will help to represent the user’s ever changing interests.
- Tags are simple text which are quiet ambiguous themselves; but this ambiguity can be clarified if tag
context is well defined and general ontologies will be applied to help ground this information.

- The tags which will now represent a concept in some higher level ontology can serve as useful context for determining user’s interests and possible intentions. This context will be used to make recommendations.

- The concept-tags can also be mapped to more fine-grained concepts in specific domain ontologies to make more domain specific recommendations. We will utilize the cultural heritage domain to test this.

The possible contributions of our work will be providing;

- A means of avoiding the cold-start problem, which is a common problem in most search and recommender systems.

- Building fairly complex contexts using strong semantics supported by known ontologies (rather than keywords). This generalized context model will help systems keep track of browsing/searching-contexts and hence aid recommendations.

2. MOTIVATION

The semantic and social web are two very different entities. However bringing them together promises to link the knowledge and expressiveness of the two domains. Their unification is an interesting arena full of possibilities on the individual as well as the community level.

In recent years, the introduction of APIs by several social websites opened a way for developers to reuse vast amounts of information on the sites to experiment and produce worthwhile applications. This was also welcomed by semantic web researchers and data from the social websites soon became a rich test-bed for the future semantic and social web technologies. Similarly Microformats and structured blogging efforts paved the way for blogging data into the semantic web.

Amongst other useful things one of the most interesting outcomes of this semantic and social web merger is the possibility of utilizing this huge amount of user-created data to understand the user better. Studies indicate that the tagging activities of an individual carries interesting information about his/her interests and therefore can play a vital role.

We believe that by linking all the different social identities of an individual over the web and by unleashing the vast amount of tagging information enclosed in them, a richer and dynamic model of user interests can be achieved. That can serve as a rich context to further assist adaptive and user oriented applications and search processes. Unified profiling and tag data portability efforts are a way forward in this direction.

Recently several Web2.0 sites started to provide links to export data from other social networks and within days the social existence of a web user became more unified e.g. Youtube for Myspace, Digg, orkut, live spaces, bebo, hi5, mix and Facebook; Orkut for Youtube and Facebook for flickr etc.

Similarly major internet players like Google, Microsoft, Yahoo, Facebook and Digg are starting to participate in data portability related activities by joining in with the Data Portability work group (DPWG). Sites like Google and Facebook are taking steps towards unified profiling through initiatives like friends connect and connect15. This is just the beginning - there is a lot to discover yet. What’s common among all these efforts is the need for a unified profiling system and cross folksonomy data sharing mechanisms.

The advantages of unified profiling and cross folksonomy data sharing mechanisms, for context oriented systems include but are not limited to: a unified user experience across different sites, easy information access for service providing agents like recommender systems and end-user applications, increased recommender productivity due to less time required to search user related information (such as user interests etc), better planning of retrieval strategies and more accurate evaluation, better equipped exchange of user information across different social networks and above all meaningful personalization.

We propose a system that gathers information about the user from social web and enriches it through the semantic web. Hence it creates an interest model for the user with the help of the best in both technologies, which can serve as a rich context base for search and retrieval systems. It finally queries over the open corpus (linked-data web) as well as a considerably closed corpus semantic data source (museum repository) to make its recommendations.

3. RESEARCH CHALLENGES

To develop a support system which can assist in providing brief and precise, high valued, context dependant, recommendations the following research challenges must be faced:

- Discovering new user information.
- Unobtrusive information gathering.
- Basic concept location.
- Removing Vocabulary gap

4. OUR APPROACH

The following approach will be followed:

- **Discovering new user information:**
  
  Our system will identify and relate user’s profiles across different social networks through Google’s Social graph
API, which has the capability of returning a set of URIs for a particular user.

- **Unobtrusive information gathering:**
  This will be ensured by gathering the publically available information about a user’s tagging activities, which doesn’t require any direct user involvement. This will help gather important information about user’s interest without requiring any help from them.

- **Basic Concept Location:**
  We will use global ontologies like YAGO [10] and DBpedia [25], as shared vocabularies and thesauri to model the user’s interest domain. It will be achieved by linking tags used by a user, to meaningful concepts in the above mentioned ontologies. This will result in an ontology-based elicitation of user interests and preferences, and could be stored as an extended overlay context model.

- **Removing Vocabulary gap:**
  By using general purpose ontologies like YAGO and DBpedia to conceptualize tags we tend to remove or minimize the vocabulary gap. The concept-tags can now more easily be mapped to more domain specific ontologies in order to support domain specific recommendations. We intend to demonstrate this by mapping these concept-tags to the Conceptual Reference Model (CRM) [20] which is an ontology developed for the cultural heritage domain.

5. **PROPOSED ARCHITECTURE**

The architecture intends to model user interests based on the user’s social-web profiles. This model is then utilized in recommending cultural heritage resources that might be of interest for the user. The use of cultural heritage as an interest domain aids our research due to the existence of semantically marked-up datasets from previous projects (e.g. eCHASE [21]). The assumption here is that the frequency of use of certain tags indicates the interest of the user. [2], [5] form the bases of our assumptions. We also assume that the user performs sufficient public-domain tagging.

The proposed architecture consists of the following main components (see Fig 1).

- **Identifying a user’s profile across social networks:**
  The first module of the system identifies the user’s social identities across the web. This will help in deciding where to extract user’s data.

- **Data Extraction:**
  Describes a set of tag extraction techniques mostly utilizing public APIs provided by the sites and some scripts.
• **Data Filtering:**
  Specifies a set of filters for cleaning the tag clouds and making it usable for the next step.

• **Concept mapping and Ontology mapping:**
  Takes the set of filtered tags and equips them with semantics by categorization and ontology mapping.

• **User interest profile modeling:**
  Takes as an input a set of ‘concept tags’ and applies a concept expansion algorithm on them to make them more suitable for recommendations.

• **Recommender system, CH repository and Open linked data:**
  The final portion is a recommender system using cultural heritage repositories as a test-bed and a query system to the linked data on web.

5.1 **Identifying A User’s Profiles across Social Networks**

The first task is to identify a user across several social networks and hence integrate the user’s profiles which are scattered across the web.

Data portability in the social networks has recently gained a lot of attention. Users shared a lot of personal data with proprietary databases in order to communicate with others in the network, this data is locked within the network, which resulted in a lot of valuable information loss; it could assist in understanding the user better. This information lock was once considered an advantage by the networks however with the advent in social network technologies and use modalities, this is now questionable.

Opening data to the world now means allowing developers to build new and interesting applications over it that in turn attracts more users to participate in the network. For example Facebook applications have played a vital role in its popularity. An interesting work here is that of Google’s Social Graph API. This API makes information about the public connections between people on the Web, expressed by XFN and FOAF markup and other publicly declared connections, easily available.

Our architecture will utilize the Google Social Graph API to identify different profiles of a person across various social networks. The “otherme” [26] method in the Google Graph API helps locate related identifiers for a person and hence can prove useful. Other techniques used to identify user profiles include matching user names and real name strings from profiles across different social web sites [13].

5.2 **Data Extraction**

The Data Extraction module is responsible for collecting user related information, mostly users tagging history from the identified Social profiles or URIs. Most of the tagging sites have public APIs that provide mechanisms to enable tag extraction, for example https://api.del.icio.us/v1/tags/get in Del.icio.us, flickr.tags.getListUser method along with several others in Flickr, the photos.getTags method in Facebook, and the user.getTopTags method in Last.fm. Some of the sites like Flickr and Last.fm have nice public APIs that help retrieve a complete history of users tagging activities. However others are not as extraction friendly so methods like screen scraping scripts need to be written. Thanks to the open source programming communities some scripts can be used off the shelf. Some other projects like [13][14] developed their own scripts for data extraction.

5.3 **Data Filtering**

The tag data from the social web comes with some inherent problems, limitations and weaknesses that need to be sorted out in order to make them usable. Some of the major issues here are those of tag ambiguity, spaces, multiple words and synonyms [4]. These issues are described in more detail below.

![Figure 2: Tag Filtering Architecture](image)
which might result in multiple words written as a single string without spaces, and is thus hard to comprehend.

5.3.3 Synonyms
In addition tags might be expressed in different word forms, or plural and singular may exist. External sources like WordNet will be used for solving the syntactic issues while Wikipedia will be consulted for the synonyms, acronyms and common names. Finally the least frequently occurring tags will be removed from the tag cloud and only a set of most frequently occurring tags will be forwarded to the next stage.

5.4 Concept and Ontology Mapping
In many recommender systems complex ontologies are utilized to model the user in order to achieve a high level of recommendation accuracy. However domain specific ontologies may not work for the data from the social web.

5.5 User Interest Profile Modeling
A tag expansion algorithm will now be applied to the set of concept tags obtained from the previous stage. The purpose of which is to enrich the set with the most related concepts in the domain. This will increase the possibility of making interesting recommendations to the user.

The resulting set of concepts is added to the system as user’s interest profile. The underlying recommender system utilizes this interest model for making relevant recommendations to the user.

5.6 Recommender system, CH Repository and Open Linked Data on Web
The user’s interest profile nurtures a recommendation system (see Fig 4.) that is an open as well as a closed corpus recommender.

5.6.1 Close corpus search
The Recommender system queries a repository of cultural heritage data from the V&A museum and the National Gallery London. The system will query this data to see how it improves the recommendation process.

5.6.2 Open corpus search
However one of the objectives of the project is improving the recommendation mechanisms for the open semantic web. Thus the system will also provide an interface to query the linked data over the web and suggest interesting things related to cultural heritage that are present on the open linked web.

6. RELATED WORK
Bobillo et al [23] addressed the problem of information overload by defining ontologies for context as well as domain knowledge. They describe a scenario for outdoor assistance of health care where context dependent information is provided for patient treatment.

Hyoung-Rae Kim et al [22], proposed a user’s interest hierarchy (UIH) for defining a users general and more specific interests. They suggest that text and phrases from users bookmarked web pages can be used to identify their general and specific interests.
Kauppinen et al. [24], addresses the problem of semantic ambiguity in geographic place names and tries to address this by designing ontologies for places (SUO and SAPO). These ontologies are published at a local server to be utilized as a mash-up service later on in their system (CULTURESAMPO portal).

Unified profiling and tag data portability among different social sites is gradually coming into the limelight, amongst research communities. The credit for this realization and initial efforts is shared among the bloggers as well as the developers of these communities. A plethora of projects are trying to answer these issues, indicating interesting results on user information in tags.

Studies on social networks indicate that users intend to tag contents they are interested in with descriptive tags that can be used to identify their interest [2]. [5] Shows how tag clouds from multiple social websites demonstrate a tendency to overlap regardless of the type of folksonomy the website uses. The work also suggests the tendency of profile enrichment through cross-linking of tag clouds. [27] Presents FLOR, a mechanism to automatically enrich folksonomy tag-sets with ontological knowledge. [3] Suggests that true collective intelligence can be achieved by linking user contributed contents and machine gathered data, and social web and the semantic web should be combined into collective knowledge systems. With this visible possibility in hand the semantic web can play a vital part in describing tags and relating them to meaningful concepts in ontologies. Significant efforts have been made to describe ontologies for the tags, taggers and tagging systems. SICO ontology [7] aims to define main concepts that are needed to describe a social community on the semantic web. The aim is to view a person’s entire contributions on the social web. The FOAF [11] Friend of a friend ontology helps describe people. SKOS [12] Simple knowledge organization system is a model for sharing and linking knowledge organizations systems like thesauri, taxonomies, vocabularies etc via the semantic web. SCOT [9] Social semantic clouds of tag ontology presents a model for main properties and concepts required to represent a tagging activity on the semantic web. Similarly MOAT [15] Meaning of a tag ontology as the name indicates is a collaborative framework to help Web 2.0 users give meanings to their tags in a machine readable format.

An interesting and promising work here is the Google Social Graph API. The API returns web addresses of public pages and publically declared connections between them which help identify and track various web identities of a user, and thus assist in the collection of tag clouds related to an individual.

We find it the right time to make an effort to utilize semantic web standards and ontologies to enrich the data from unified profiling systems in order to make it useful in semantic search and recommender systems.

Some of the projects that have tried similar approaches include [19] and [8]; Specia et al., [8] focuses on determining relations among tags in social networks to form clusters based on concepts from ontologies. This work suggests that by exploiting Wikipedia, Word Net, Google and semantic web ontologies, meaningful relations can be identified amongst tags. While [19] proposes an automated link service that uses Wikipedia as its link-base for linking data with Wikipedia concepts.

Xin Li et al., [2] suggest a mechanism to identify the social interest based on the user-generated tags. They propose an Internet Social Interest Discovery system ISID which works on the principal of clustering users and their saved URIs based on common frequently-occurring tags. These clusters identify the topics of social interest.

Jon Iturrioz et al., in [16] suggests a transition from desktop to web where more and more users are keeping their resources on the web; like pictures in flickr, bookmarks in del.icio.us and so on. Despite significant ease and advantages, this is resulting in the fragmentation of user resources and therefore a global view of resources is needed. There work is a loosely coupled federated tag management system that provides a uniform view of tagged resources across different web 2.0 sites.

Martin Szomorszor et al., [13] presents a way of determining an individual across flickr and del.icio.us assessing his/her tags, filtering them utilizing Google Search and Word Net and finally forming a FOAF based user profile. Perhaps the most related project to our work is by Ivan Cantador et al., [14]. It builds upon the tag filtering mechanism developed in [13] and moves further to design ontological profiles for tag users. This is done by matching tags with ontological concepts.

7. CONCLUSIONS AND FUTURE WORK
This work will open a way for the vast amount of structured data on cultural heritage to be exposed to the users of social networks, according to their taste and likings. However we understand that the system’s success depends on the relationships between user tags and their interests and assume some good links to semantic resources can be made. Our future work will include the
implementation of the architecture and its evaluation in comparison to the other context profiling methods used in the recommendation systems. We hope to obtain useful insight in the areas of unified social profiling, interest modeling and their applications in context intensive tasks.

We believe that by linking all the different social identities of an individual over the web and by unleashing the vast amount of tagging information enclosed in them, a richer and dynamic model of user interests can be achieved. That can further assist context sensitive and user adaptive applications and search processes.

8. ACKNOWLEDGMENTS
Our thanks to Dr monica schraefel, for her useful insight on this work.

9. REFERENCES
