Project no. 34721

TAGora

Semiotic Dynamics in Online Social Communities

http://www.tagora-project.eu

Sixth Framework Programme (FP6)

Future and Emerging Technologies of the Information Society Technologies (IST-FET Priority)

Deliverable D5.3: A White Paper about target problems and grand challenges for Semiotic Dynamics in Online Social Communities

---

Period covered: from 01/06/2006 to 31/05/2007
Start date of project: June 1st, 2006
Due date of deliverable: May 31st, 2007
Distribution: Public

Date of preparation: 31/05/2007
Duration: 36 months
Actual submission date: May 31st, 2007
Status: Final

Project coordinator: Vittorio Loreto
Project coordinator organisation name: “Sapienza” Università di Roma
Lead contractor for this deliverable: “Sapienza” Università di Roma

Contents

Chapter 1

Web Science

Though the World Wide Web is a technology that is only a few years old, its growth, and its effect on the society within which it is embedded, have been astonishing. Its inception[CG00] was in support of the information requirements of research into high-energy physics. It has spread inexorably into other scientific disciplines, academe in general, commerce, entertainment, politics and almost anywhere where communication serves a purpose[NV04]. Innovation has widened the possibilities for communication. Weblogs and wikis allow the immediacy of conversation, while the potential of multimedia and interactivity is vast.

But neither the Web nor the world is static. The Web evolves in response to various pressures from science, commerce, the public and politics, and so pervasive and transformative is the Web that the world is evolving together with it (for instance, e-government and e-commerce are key developments). We need to understand these evolutionary and developmental forces. Without such an appreciation opportunities for adding value to the Web by facilitating more communicative and representational possibilities may be missed. But development is not the whole of the story. Though multi-faceted and extensible, the Web is based on a set of architectural principles which need to be respected. Furthermore, the Web is a social technology that thrives on growth and therefore needs to be trusted by an expanding user base - trustworthiness, personal control over information, and respect for the rights and preferences of others are all important aspects of the Web. These aspects also need to be understood and maintained as the Web changes. This is the aim of Web Science, which aims to map how decentralized information structures can serve scientific, representational and communicational requirements, and to produce designs and design principles governing such structures[BLHH+06a;BLHH+06b].

Understanding the relation between the Web and the world involves learning about a complex cycle of interactions, which (to complicate things) occur at radically different scales. Technical developments which are basically to do with computers' abilities to pass data between each other turn out to have strong social effects; computational innovations at the micro scale feed into macro-level effects on the whole of society (Figure 1.1). An idea, say for an information-sharing protocol, needs a technical engineering design that encapsulates it within a particular social context. The design in context produces a micro-level effect at the level of the individual user's control of his or her computer. But when the number of users of a design within a decentralized structure grows, macro-level effects can be detected, which raise social issues. In large part, these social issues are raised because the social effects were not only not predicted, but they were fundamentally unpredictable.

This is a scenario to which in the last decades scientists have devoted great attention, namely the study of collective phenomena and complex systems. Large systems made up of simple components (for instance atoms or molecules, animal, human or artificial agents) can in fact self-organize themselves, i.e. "acquire a functional, spatial or temporal structure without specific interference from the outside"[Hak88]. More precisely, the constituents of such systems are able to develop a

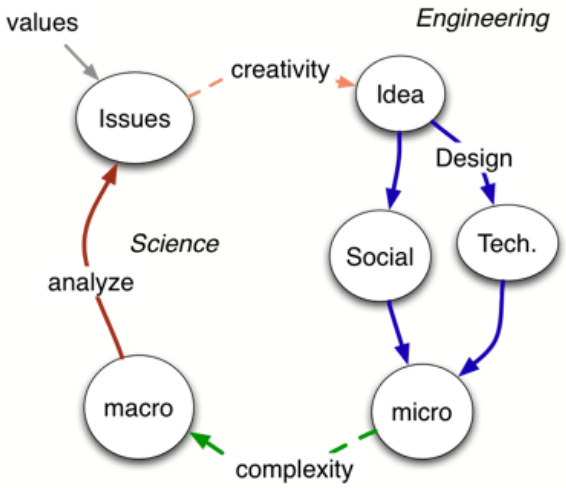

Figure 1.1: Feedback from the micro scale to the macro-level

complex collective behavior not trivially deducible from the knowledge of the rules that govern their mutual interactions [Vic01;Vic02]. We then need to analyses these effects in order to understand whether the social effects are good or bad, and to what degree, and also to help determine what new engineering designs might be capable of preserving or enhancing the good effects, or alternatively eliminating the bad. The result, therefore, is a cycle of micro-level engineering and macro-level analysis.

The Web was originally created as a means of two-way communication, writing as well as reading, but as it took off, publishing tools did not develop as strongly as browsers, and reading became more important than writing, a development deplored by many of the more idealistic who looked back to the plain text of the early Internet days with nostalgia (e.g. [Res98]). But the recent arrival of tools allowing more information-passing has led to a rebalancing of the Web, and the invention of so-called Web 2.0.

In terms of the scheme of Figure 1.1, there are micro-level effects of providing tools allowing someone, e.g. to blog, which affect the individual author. However, as the blogosphere, content-hosting sites or recommender systems grow, there are macro-level effects to be noted.

An initial state when writing, editing and publishing were difficult prompted the development of form-based editors, and the individual phenomenon of the wiki [LC01]. As more users are recruited, increasingly large groups can put together something totally unpredictable such as the extraordinary Wikipedia. Wikipedia is now the first port of call for anyone needing basic information. It is therefore an important rival to traditional undemocratically edited encyclopedias, but also now it is so widely used, its accuracy and veracity are now an issue. The social question is now about how democratic Wikipedia can afford to be.

Blogs, Wikis, and Social Bookmark Tools have rapidly emerged on the Web creating a new scenario that radically change the knowledge production process. We have virtually unbounded storage capabilities and essentially no limits in our ability to interact with other peers. This new knowledge production process is impacting on all aspects of knowledge creation on all types of knowledge and the Web is becoming the most extensive knowledge repository that ever existed. The reason for this immediate success is the fact that no specific skills are needed for participating. A new paradigm is actually gaining impact very quickly in large-scale information systems: Collaborative Tagging. In new web applications (e.g. Flickr `www.flickr.com`, Connotea `www.connotea.org`, Citeulike `www.citeulike.org`, Delicious `del.icio.us`, etc.) people no longer make passive use of online resources. They take instead an active role and enrich resources with semantically mean-

ingful information. Such information consists of terminology (or tags) freely associated by each user to resources and is shared with users of the online community. Despite its intrinsic anarchist nature, the dynamics of this terminology system spontaneously leads to patterns of terminology common to the whole community or to subgroups of it. Surprisingly, this emergent and evolving semiotic system provides a very efficient navigation system through a large, complex and heterogeneous body of information.

At the moment, there exists no foundational research for these systems, and they provide only very simple structures for organizing knowledge. Individual users create their own structures, but these can currently not be exploited for knowledge sharing. The aim of TAGora is twofold. On the one hand it is important performing analysis of Web 2.0 phenomena to determine what social effects we are now seeing from tagging structures, recommendation systems, and other content that is emergent from large-scale use of social software. On the other hand it is crucial providing theoretical foundations for upcoming Web 2.0 applications and to investigate further applications that go beyond bookmark- and file-sharing.

We strongly believe we are facing a unique opportunity to exploit and give theoretical foundations to the recent, though extremely rapid, developments of emergent semantics in Web-based applications. The common effort of researchers in many different fields could provide the right trigger to face and tame the challenges of Web Science:

> How do microscopic interactions (at the users' level) affect the macroscopic emergent behaviors of online communities, e.g., how do individual decisions to tag a Flickr photograph, say, affect the information structures that emerge at the macro-scale?
>
> How can we bridge the gap between exploiting natural intelligence (a paradigm commonly referred as Human Computing) and implementing artificial intelligence systems?
>
> In shaping large-scale IT communication systems, how can we bridge the gap between top-down and bottom up approaches? One of the big failures of user interfaces and human-machine interaction today comes from their lack of adaptivity and the assumption that ontologies and communication conventions can be fixed and imposed from outside.
>
> How will current and emerging resource sharing systems support untrained users in sharing knowledge on the Web in the next years? The knowledge acquisition bottleneck in top-down approaches, i.e., the knowledge transfer from experts to formal systems, should be rephrased here as the *wisdom of crowds* issue[Sur05]: is the knowledge aggregation and organization emerging from the uncoordinated activity of millions of users better than a centralized control of few experts?

And of course one among the most important challenging tasks of this project is the development and fostering of a new culture of interaction between IT and complex systems science. This will require that gaps are bridged from both sides.

Chapter 2

Folksonomies are our Drosophilae

Within only one year, social bookmark tools have received increased attention on the WWW. Social bookmark tools share with the Semantic Web the vision of easy knowledge sharing on the web. But a main difference lies in the fundamentally opposite approach: the Semantic Web aims at a formal knowledge representation in form of ontologies (written in XML, RDF, or OWL), whereas social bookmark tools follow a grass-root approach: there are no limitations on the kind of tags users may select. In contrast to ontologies, the resulting structures are called 'folksonomies', that is, 'taxonomies' created by 'folk'. While the Semantic Web community currently focuses, eight years after the first statement of the Semantic Web vision by Tim Berners-Lee[BL97;BL98], on theoretical issues like the definition of the adequate knowledge representation formalism, and is still looking for large-scale applications, social bookmark systems are finding increasingly large user communities on the Web in a very short time frame. The most prominent of them, for instance, del.icio.us [1] and flickr [2] have already more than one million of users. In the reference sharing systems CiteULike[3] and Connotea[4] researchers and others insert, tag, and recommend scientific references in a shared knowledge space. This indicates a currently ongoing grass-root creation of knowledge spaces on the Web which is closely in line with "the 2010 goals of the European Union of bringing IST applications and services to everyone, every home, every school and to all businesses" [5]. The reason for the apparent success of the upcoming tools for web cooperation (wikis, blogs, etc.) and resource sharing (social bookmark systems, photo sharing systems, etc.) lies mainly in the fact that no specific skills are needed for publishing and editing and an immediate benefit is yielded to each individual user (e.g. organizing ones bookmarks in a browser-independent, persistent fashion) without too much overhead. Large number of users have created huge amounts of information within a very short period of time. As these systems grow larger, however, the users feel the need for more structure for better organizing their resources. For instance, approaches for tagging tags, or for bundling them, are currently discussed on the corresponding news groups. We anticipate that resource sharing systems, together with wikis and blogs, are only first appearances of an emerging family of Web 2.0 tools.

2.1 Research Objectives

The main objective of Web 2.0 research is to answer the question: How will current and emerging resource sharing systems support untrained users in sharing knowledge on the Web within the next years? This can be split down in (at least) the following tasks:

[1] http://del-icio.us
[2] http://www.flickr.com/
[3] http://www.citeulike.org
[4] http://www.connotea.org
[5] http://www.cordis.lu/ist/workprogramme/fp6_workprogramme.htm

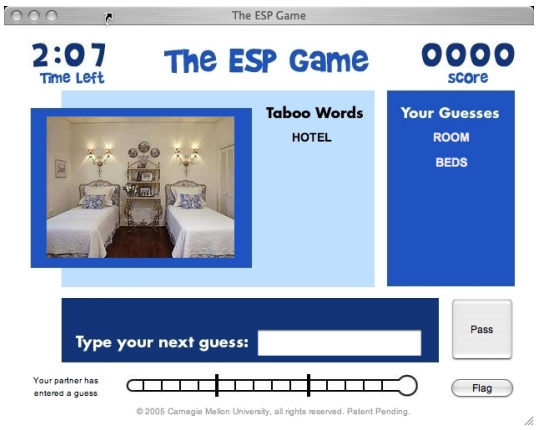

Figure 2.1: Screenshot of the ESP game

1. Which new modes for knowledge sharing can be established on the Web, that are suitable for non-experts in knowledge engineering?
2. Which other types of information (beside just tags) would increase the usefulness of knowledge sharing on the Web?
3. How can methods and techniques for knowledge engineering and management be adopted and/or developed to these modes in order to allow a minimal-invasive use?
4. Is it possible to understand the typical users' needs, motivations and behaviors?
5. How one can understand, predict and eventually control the evolution of Online Social Communities?

2.2 Tagging systems vs. Social Computing

In the Web 2.0 users acquire a completely new role: not only information seekers and consumers but information *architects*, cooperating in shaping the way in which knowledge is structured and organized, driven by the notion of meaning and semantics. In this perspective the Web is acquiring the status of a platform for *Social Computing*, able to coordinate and exploit the cognitive abilities of the users for a given task. One striking example is given by a series of Web games where pairs of players are required to coordinate to assign shared labels to pictures[?] [6]. As a side effect these games provide a categorization of the images content, an extraordinary difficult task for artificial vision systems.

From this point of view Social Bookmarking systems represent only one example of the new Web-based social interactions. Nevertheless Social Bookmarking systems display most of the complexity of Web 2.0 in a relatively controlled environment and it represents an almost clean-room version of web development driven by many users. This is the reason why TAGora project is mainly focused on Folksonomies. A cartoon of the general structure of a Social Bookmarking system is reported in Figure 2.2.

A folksonomy can be represented as a triadic structure of *(tag, user, resource)* assignments, the user interface of such a folksonomy system will typically allow the user to jump from a given tag to (a) any resource associated with that tag, or (b) any user who uses that tag, and vice versa for users and resources. Thus, the effort of getting from one node in the folksonomy to another

[6] See for instance: `http://www.espgame.org/`

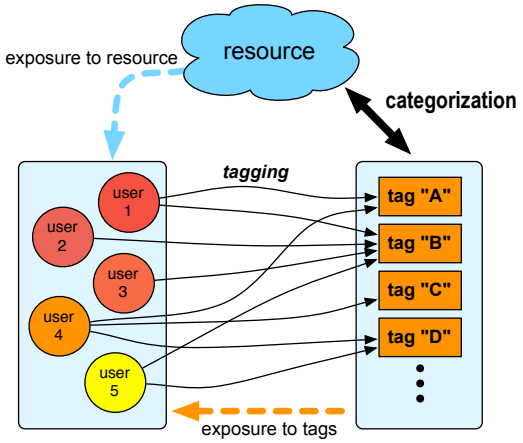

Figure 2.2: Collaborative tagging. Schematic depiction of the collaborative tagging process: web users are exposed to a resource and freely associate tags with it. Their interaction with the system also exposes them to tags previously entered by themselves and by other users. The aggregated activity of users leads to an emergent categorization of resources in terms of tags shared by a community.

can be measured by counting the *hyper*edges in shortest paths between the two. Here a path is defined as a sequence of steps linking *hyper*edges with the condition that each step has to link two *hyper*edges sharing at least a user or a resource or a tag. The triadic structure of a folksonomy can be more conveniently studied by looking at suitable projections which correspond to 'cut slices' out by selecting selecting one or two dimensions (out of user/tag/resource), and then fixing in these dimensions particular instances. Many interesting features can already be derived from one-mode projections, also called *co-occurrence networks*. A possible example is given by the set of all tags occurring in the framework of a given resource or a given user. In this one-mode projection particularly interesting is the so-called stream view. In the stream view of a folksonomy, the temporal ordering of posts and tag assignments in collaborative tagging systems is taken explicitly into account. The stream view can be used to analyze the evolution of specific aspects of a folksonomy over time (e.g. the set of tags).

Due to the tripartite character of the TAS (Tag Assignment) data, different types of streams can be defined: stream of tags, stream of users or eventually stream of resources. Furthermore, a stream might be restricted to a subset of the whole stream by picking up a selection of tags (or resources, or users) and considering only the TAS containing those selected tags (or resources, or users).

The structure of this White Paper closely reflects the main tasks/requirements in tagging systems (as on the Web):

1. Share: share knowledge between users → this points to knowledge management
2. Find: using e.g. keywords → this points to information retrieval
3. Browse & Recommend : discovering structures for organization (e.g. ontologies) → this points to knowledge discovery
4. Organize: classifying unclassified tags for easier access → this points to ontology learning
5. Understand the language-like emergent features → this points to semiotic dynamics
6. Understand the system as a whole: understand the evolution of the system in order to recognize and understand mass behavior → this points to physics of complex systems

Chapter 3

Share: Collaborative Knowledge Management

It is a truism nowadays to emphasize the dependency of organisations and individuals on knowledge in order to compete successfully and to work in a knowledge society; for example, the European Commission has set itself the goal to become “the most dynamic and competitive knowledge-based economy in the world” in its Lisbon Strategy (European Commission, 2004). Knowledge management (KM) has been recognized as an important task in order to maintain and improve an organisation’s capabilities in the knowledge-driven business environment.

In order to minimize the burden for each individual, and to allow for unforeseen knowledge exchanges and informal social networks, self-organisation has become an issue in knowledge management. Self-organization has been the research topic for efforts from many different disciplines, including physics, social sciences, cybernetics, and many others.

Social bookmarking systems are a way of collaboratively organizing collections of resources, and are thus a promising alternative to classical knowledge management approaches. Recent applications of resource sharing systems address primarily private issues like photo collections. Their high acceptance on the Web shows their high social impact. By today, the economical impact is only visible on the horizon, but it indicates a large market. IBM, for instance, announces experiments with folksonomies in their intranet, because the currently used taxonomy is too expensive to be maintained[Bud05a]. Microsoft also intends to invest in this area[Bud05b].

The easy use of resource sharing systems makes them good candidates for knowledge management applications in a commercial setting, at least in domains where stronger structured approaches like ontologies could not take hold yet, or where their maintenance is too costly. This will hold especially in domains where people with no experience in data modelling have to deal with the tools. So it seems very promising to start with a very lightweight folksonomy/ontology and make it heavier if (and only if) this is needed by the users. As their frequent use already shows, resource sharing systems avoid the knowledge acquisition bottleneck, which was one of the main reasons for the failure of many expert systems with a more sophisticated knowledge representation. The comparison of the steep rise of social bookmark systems on the Web compared to the relatively slow increase of implemented Semantic Web applications shows that apparently the former do not suffer from this bottleneck - in contrast, many people are willing to contribute. The latter, on the other hand, still suffer from the lack of data: many interesting semantic web applications can currently only be evaluated on artificial data. We strongly believe that resource sharing systems will establish themselves as shared knowledge spaces - both for private and for industrial applications. A stronger theoretical basis will provide the foundation for narrowing the gap between both approaches.

3.1 Knowledge Management with BibSonomy

Due to the system's usability, BibSonomy is a promising candidate for knowledge management applications in commerce, especially in areas in which structured intranet solutions have not been established yet, or in which the administration and update of such intranets suffer from today's fast information flow. This concerns especially applications managed by knowledge engineering beginners. In cooperation with the data process GmbH,[1] Kassel, we evaluated the implementation of BibSonomy as part of the K+S[2] group's intranet[Rez06], a worldwide supplier of potassium-and magnesium products with a coverage of 13 percent of today's potassium demand.

The implementation involved BibSonomy's integration into the company's existing SAP portal. A single sign on is offered to log in to both systems. A main aspect of being able to use a bookmarking system was its internationality. In order to meet this requirement BibSonomy was extended by offering its pages in different languages. A link validation procedure for each user's hyperlinks was implemented to increase acceptance among users and bookmarklets were created which support navigation beyond system borders. To raise the awareness among the company's intranet users, each intranet page offers a link to automatically store the page in BibSonomy. Taking advantage of the iViews-functionality of the SAP portal users are shown the top-5-tags as well as the five most recent posts. The objective is again to increase the visibility and exchange of information via the portal.

3.2 Distributed collaborative tagging infrastructure

Common collaborative tagging services like del.icio.us, flickr, or youtube are easy to use for organizing and sharing personal data but generally support just one media type (i.e. bookmark, photo, or video). Thus multiple accounts are necessary and users even have to pay for full feature access in Flickr. We argue that the next step in collaborative tagging should go towards the combination of multiple mediatypes and more independence from single centralized services. Therefore, we envision a decentralized tagging infrastructure where everyone can freely organize the personal data with tagging and browse other people's shared (tagging) data. Such a system, which basically sits on top of a peer-to-peer architecture, does not need to rely on any third party and can even provide certain anonymity for its users.

3.2.1 Requirements for a distributed tagging system

The usefulness of collaborative tagging systems is heavily influenced by the quality of the navigational support. The most widely used visual clues are tag clouds and the grouping (clustering) of related tags but also clustering of similar resources or users provides useful information for the user. In a peer-to-peer system all (meta) data is distributed among all participating users which complicates the computation of such statistics because many peers have to be contacted to gather the required data. Obviously, that does not work in a running system with frequent statistic updates. Consequently, the scalability of a distributed tagging system highly depends on the efficient computation of the included statistics. Following are some aspects concerning the applicability of a certain statistic measure in a distributed tagging system:

- can be approximated within some allowable error margin
- a few input values suffice to compute the core characteristic
- can be easily stored in a distributed data structure

[1] http://www.dataprocess.de/
[2] http://www.k-plus-s.com

- updates can be approximated with a few changed values
- updates are only necessary after a computable deviation

3.2.2 Implementation approaches for distributed statistics

Users in distributed systems are typically not permanently online but join or leave frequently. Thus the statistics computation needs to be insensible to the availability of peers. One suitable solution are distributed index structures that store for example all tags assigned to one resource or all users using one tag. Second, distributed histograms can represent frequency counts like in tag clouds or store feature vectors used for similarity computations.

The main benefit of such distributed data structures is that usually only one or a few network requests are necessary to retrieve all required information. Thus each peer can easily retrieve all statistics independent of the availability of other peers. Updates of the statistics have to be triggered by peers that encounter data changes with mayor influence on the statistics.

Chapter 4

Find: Information Retrieval

With the growth of social bookmarking systems, users address the need of enhanced search facilities. Today, full-text search is supported, but the results are usually simply listed decreasingly by their upload date.

A first step to searching folksonomy based systems – complementing the browsing interface usually provided as of today – is to employ standard techniques used in information retrieval or, more recently, in web search engines. Since users are used to web search engines, they likely will accept a similar interface for search in folksonomy-based systems.

Google's PageRank [BP98] has been a successful ranking mechanism for the Web, viewing each link between web pages as a sign of endorsement of the target page by the author of the source page. Although published seven years ago, PageRank is still the state of the art for ranking webpages. Its orientation towards directed graphs, however, does not allow to apply it directly to folksonomies.

The research question is how to provide suitable ranking mechanisms, similar to those based on the web graph structure, but now exploiting the structure of folksonomies instead. To this end, we have proposed in [HJSS06a] a new algorithm, called *FolkRank*, that takes into account the folksonomy structure for ranking search requests in folksonomy based systems. The algorithm can be used for determining an overall ranking, specific topic-related rankings [HJSS06a], recommendations, and trend detection [HJSS06b].

As discussed above, social bookmarking systems are promising for knowledge management in intranets. Applying Google-like ranking techniques in intranets and for multimedia data, however, is more difficult. Corporate intranets will consist of large collections of documents, which typically do not link to each other and are often stored in formats such as PDF or MS Office not having the idea of hypertext in mind. The hyperlink structure of intranets is often purely navigational and does not express any kind of recommendation or semantic links between contents, but will rather be engineered from scratch by a knowledge engineer or even the person who is in charge of the technical infrastructure of the intranet. With algorithms like FolkRank, one can thus exploit individual statements about resources for ranking search results, and one can additionally extract, from this additional structure, recommendations for other (intranet) users.

Chapter 5

Browse and Recommend: Knowledge Discovery

The World Wide Web has become a significant target for data mining, due to several reasons: The web is a huge resource of any kind of information, the increase of commercial applications on the Web requests the extraction of knowledge from the Web, and the immense amount of data available calls for automatic means for the extraction.

The Web differs in many regards from other mining applications. Web pages consist of (sometimes structured) natural language text, calling for text mining techniques; hyperlinks provide additional structure, that can be handled with graph mining approaches; Web servers log user activities, which also can be analyzed; and the Web is very dynamic in terms of growth, content changes, and structural changes. The combination of all of these aspects makes the Web a unique setting for data mining. During the last decade, researchers have attacked many of these challenges for Web Mining.

Recently, with the emergence of the Web 2.0, the attention of the research community has shifted to a new focus. The main emphasis of Web 2.0 systems is their easy use that relies on simple, straightforward structures. As Web 2.0 systems grow larger, however, the users feel the need for more structure for better organizing their resources. For instance, approaches in social bookmarking systems for tagging tags, or for bundling them, are currently discussed on the corresponding news groups.

The machine learning community has a long tradition in extracting structure from large scale data collections. With the Web 2.0, it faces (at least) two new challenges:

1. New data types appear, for which there exist currently no out-of-the-box data mining solutions, for instance for the triadic hypergraph structure of folksonomies or for documents in wikis that permanently change over time.

2. The majority of Web 2.0 users have no skills in knowledge engineering and data mining. Tool support targeted directly at the end user has thus to hide the complexity usually involved in the different data mining steps (e.g., data cleaning, parameter settings).

In addition to these general challenges for the machine learning community, new data mining applications arose. With the Semantic Web, a more conceptual view on (Web and other) data arose, leading to the desire to discover topics and trends (which then can be captured in an ontology); and Web 2.0 platforms facilitate simplified participation of untrained users, who started to build social and topic-oriented networks, leading to the desire to discover significant substructures and communities. We will discuss these desires and potential solutions in the sequel.

5.1 Discovery of Communities

As the data available in folksonomies grows ever larger it becomes more complicated to perform an analysis of the data, e.g. for providing recommendations to users or for ontology learning (see 6). This is because the heterogeneity of the users, their interests, behavior etc. grows with the size of the folksonomy. The discovery of communities helps to counteract this trend by identifying more homogeneous user groups to which then the analysis of the data is restricted.

In this context, one has to distinguish the discovery of communities as it is known from the social network analysis (SNA) from a broader definition of communities: In SNA one defines communities over a communication relationship between the users, e.g. if they regularly exchange e-mails or talk to each other. An equivalent in tagging systems may be found in the contact profile of users in e.g. Flickr or the comments attached to photos. But in the context of data analysis for e.g. providing recommendation strategies one is more interested in finding communities of users with homogeneous interests and behavior. Such homogeneity is independent of contacts between the users although in most cases there will be at least a partial overlap between communities defined by the user contacts and those by common interests and behavior.

Two important areas of research can be identified for the detection of communities in tagging systems. On the one hand, there is the question which observable features in tagging systems are best suited for inferring relationships between users. The selection of the feature is dependent on the application of the community detection, i.e. it may differ for the community detection in the context of recommendation systems and e.g. in the context of ontology learning. On the other hand, there is the question how one can then group several users into homogeneous communities. For this purpose, one can reuse several approaches from SNA, clustering or, in case of matrix representations of the relationships, even with methods from linear algebra.

5.2 Recommendation Systems

Recommendation systems have evolved in recent years to support users in the discovery of new items through the construction of profiles that represent their interests, and networks that connect them to other users who share similar tastes. Many of these recommendation strategies rely on the modelling of intrinsic attributes about each item (e.g. the keywords for a document or the genre of a CD) so that the items can be categorized, and the level of interest a user has can be expressed in terms of these attributes. This knowledge is usually gathered over time, by monitoring and logging various user interactions with the system (e.g. buying, browsing, bookmarking). Amazon.com, for example, provides a recommendation service that is based on collaborative filtering: if a user buys an item that has been bought by a number of other users in combination with some other items, then those other items will be recommended by Amazon.com to the user. These recommendations are entirely based on what goes on inside the system (Amazon.com in this case), ignorant of any external knowledge about the items or the users themselves.

To improve on such recommendation techniques, we believe it is useful to incorporate data from as many sources as possible. Additional information sources will help to build richer profiles that model many facets of interest that could be difficult and impractical to capture by a single system or service. In recent years, many Web 2.0 applications, such as folksonomies and blogs, have become popular places where individuals provide and share various type of information. This information may, directly or indirectly, represent the interests of those individual users. There could be much to learn about a user from analyzing their shared profile in MySpace, bookmarks in del.iciou.us, photos in Flickr, references in Connotea, and any other popular Web 2.0 applications.

Folksonomy web sites are rarely closed worlds. It is quite common for individuals to be active members of several online communities and thus one would expect certain tags to spread across such communities with time. For example one could be adding images to Flickr, bookmarking

web sites with del.icio.us, creating their music preference profiles in last.fm, and tagging articles in Connotea. In TAGora we are investigating approach and tools to enable the integration of several folksonomy sites to enable us to generate richer user profiles. These profiles will hold information about the user's interests, based on his/her activities in these folksonomies, as well as the activities of the community as a whole.

5.3 Trend Detection

Little is known so far about the behavior and dynamics of folksonomies and their relations to one another. For example, the choice of tags, or their sudden appearance or disappearance, in one folksonomy might be influenced by external sources, such as other folksonomies or newsletters. In Last.fm, user listening habits are logged over time to calculate album and song charts. A brief comparison has shown that these charts differ from the conventional ones which are based on music record purchases. However, there could be some hidden correlations between the communities driving these chart results. A big rise or fall in the ranking of a particular song or album in one community could, after a little while, influence the choice of users in the other community.

In TAGora, we have analyzed trends on the social bookmarking system del.icio.us [HJSS06b]. Furthermore, data from Last.fm and Top40Charts.com are being collected and will be used to experiment with discovering such possible trends and folksonomy dependencies.

Chapter 6

Organize: Ontology Learning

The information contained in folksonomies may be more efficiently used if one manages to bridge the gap to more formal representations like e.g. ontologies. Especially in the information retrieval and the knowledge discovery scenarios one may provide additional benefits to the users if information about relations between tags would be available. For example, current systems like Flickr and Delicious are typically browsed with the help of tag clouds or a generic “isRelatedTo” relation but it is not possible to browse along e.g. a subsumption relation between tags or certain facets like persons, locations, things or actions. Furthermore, the knowledge about semantically very similar tags and those which should be disambiguated would be beneficial.

Altogether, one can distinguish between two different approaches to ontology learning from tagging data: On the one hand, there are approaches like[Mik05], [SHJS06], [Sch06], [HRS06] and [HGM06] which only rely on the information available in the tagging system, e.g. by analyzing the tag assignment data for tags which co-occur with a higher probability than other tag combinations. On the other hand, there are recent approaches like[ASC07], [SM07] and [ASSM07] which additionally incorporate information into the analysis which are not available in the tagging system itself. For example, [ASC07] combines the tagging information with information found on the web with the help of natural language processing techniques, [ASSM07] searches the semantic web for ontologies and extracts the information about relations from there while [SM07] combine web and semantic web information in their analysis.

Both approaches can also be distinguished by the kind of learned tag relations. For the former approach, which only relies on information from the tagging system, mainly techniques for extracting the subsumption relation are proposed. This may be because only for the subsumption relation a characteristic co-occurrence pattern can be identified: A tag t_1 subsumes another tag t_2 if the set B of resources tagged with t_2 is a subset of the set of resources A tagged with t_1. With the exception of [HGM06], all of the above mentioned methods are variations of this basic principle. For other relations like instance-of, homonyms or synonyms no such co-occurrence pattern is identified yet. This problem is circumvented by the latter approach of taking tagging system external information into account, where e.g. already established ontology learning techniques like natural language processing or information from ontologies in the semantic web are used for identifying further kinds of tag relations.

There are several lines of research along which ontology learning from folksonomies can be further developed in the future:

Co-occurrence Analysis The major advantage of this approach that it only relies on the information contained in the tagging system itself. Thus, the extracted ontology only represents the knowledge and meaning implicitly available in the folksonomy. In the future, it should be researched in how far further co-occurrence patterns can be identified in order to learn other than the subsumption relation.

Hybrid Analysis The advantage of this approach is that already established ontology learning

techniques can be re-used and that it can be used for identifying several kinds of tag relations. But it has the disadvantage that it relies on further sources of information like a text corpus or ontologies which may not always be available or, like in the case of ontologies, may only contain information about a subset of the tags in the folksonomy. Nevertheless, this approach seems promising for identifying the relations which can no be identified with the help of co-occurrence analysis.

User Communities Folksonomies do not only contain information about tag relations but also about the relations between users, i.e. user communities. There may exist user communities which are specialized on a certain topic and whose tag usage significantly differs from other communities. The knowledge about the different user communities may be used for improving ontology learning results (e.g. extract the information about a topic from a specialized topic community) or to learn ontologies personalized for a specific community.

Scalability One important factor for all the approaches is that they have to be scalable to systems like Delicious or Flickr if the results from ontology learning should be used for improving the browsing and search experience of the user. Nevertheless, for most of the currently available approaches, scalability is not an issue. This has to be changed if the results should be used in real systems.

Chapter 7

Understand emergent features: Semiotic dynamics

The traditional view on ontologies (already exposed by Plato and developed extensively by logical empiricists such as Russell and Carnap) is that they are universal and statistic. It is assumed that all human beings share the same set of basic concepts, including perceptually grounded concepts (like colors or shapes). More complex concepts (like computer or telephone) are derived by composing primitives using a small set of logical primitives which include the logical connectives, the quantifiers and other operators. Differences between human languages are not due to the conceptual frameworks that they express but rather the syntactic choices or phonetic forms that have been chosen. If this universalist position is true then it should be possible to define once and for all the conceptual frameworks underlying all languages (as attempted in Lenat's Cyc project) and it should be possible to use these frameworks as a foundation for search engines and web access (as in Berner Lee's semantic web project).

But there is an alternative position, often associated with Whorf, which has argued that conceptual frameworks are relative to individuals or groups of individuals. For example, when you map out the color concepts of different populations throughout the world you find that, despite a number of general tendencies, there are also unreconciliable differences, in the sense that one group may introduce totally different regions in the color space and words to name these categories. Conceptual frameworks can nevertheless be sufficiently shared to enable communication when (1) conceptual frameworks are acquired and fine-tuned through shared sensori-motor experiences and joint actions, and (2) when they are constantly aligned in dialogue. In this view, ontologies are seen as complex adaptive systems, constantly changing under the actions of their users. This suggests a radically different way to create the semantic foundations of search engines and web information retrieval: We need to orchestrate a process of *Semiotic Dynamics* in which communities of individual users help to shape ontologies through tagging or other means. The coherence arises out of this distributed activity in a self-organized way instead of being imposed in a top-down fashion.

Over the past decade, the semiotic dynamics approach has been researched by a small but highly dedicated group of researchers, starting from the first timid experiments in the mid nineteen nineties to full-blown large-scale computer simulations and humanoid robotic experiments more recently. Also the mathematical theory of semiotic dynamics has been developing quickly. Most of this work is embedded in experiments attempting to understand the origins of human languages and meanings by evolving artificial languages. Basically three lines of works exist. Some work follows the footsteps of many artificial life experiments by using a genetic encoding of the languages and ontologies and using evolution by natural selection as the way in which communication systems and ontologies arise[? ?] . Fitness is directly coupled to communicative success, which is usually translated in a measure of similarity between lexicons and grammars. A second approach, known as iterated learning, follows theories of cultural evolution (such as those of Boyd and Rich-

erson[?]) which focus on transmission across generations[? ?] . Transmission requires learning and the learner may generalize ontologies or communication systems to make them more systematic. A third approach is based on self-organisation[? ? ? ?] . It does not rely on genetic coding nor generational transmission but purely on the peer-to-peer interaction between agents called language games. Agents need the basic scripts to play such games (which includes the ability to establish joint attention) as well as the appropriate operators for invention, acquisition, and alignment. The results achieved by the third approach are far superior to the other two approaches, in the sense that they yield much more quickly effective, shared communication systems and ontologies (within the course of a single generation), they show how ontologies and communication systems can get coordinated, and they show how natural language like grammars may emerge[?] .

Although semiotic dynamics has an obvious relation to tagging systems[CLP07], many results of this research domain have not yet been applied to web information systems, but there are obvious opportunities. For example, if tagging systems start to become richer we have to consider forms of emergent grammar which also arise in a peer-to-peer fashion just like in the case of human natural languages.

Chapter 8

Understand the system as a whole: Complex Systems Science

Statistical mechanics has proven to be a very fruitful framework to describe phenomena outside the realm of traditional physics. The last years have witnessed the attempt by physicists to study phenomena which heavily rely on human behavior, like the dynamics of financial markets and the emergence of collective organization in social systems. The latter topic, in particular, is attracting a lot of interest, as indicated by the large and rapidly increasing number of published papers. In social phenomena every individual interacts with a limited number of peers, usually negligible as compared with the total number of people in the system. In spite of that, human societies are characterized by stunning global regularities. There are transitions from disorder to order, like the spontaneous emergence of a common language/culture or the creation of consensus about a specific topic. These macroscopic phenomena have naturally called for a statistical physics approach to understand the regularities at large scale as collective effects of the interaction among single individuals, considered as relatively simple entities. This whole area goes now under the denomination of Complex Systems Science.

The paradigm of a complex system is an assembly of many interacting (and simple) units whose collective (i.e., large scale) behaviour is not trivially deducible from the knowledge of the rules that govern their mutual interactions. A classical example in Physics is the Ising model, in which many coupled spins produce the emergence of paramagnetic or ferromagnetic order. Often, however, the internal structure of interacting entities has little impact on their collective behavior, which is largely determined only by the way in which they interact. This has allowed to enlarge the field of application of statistical mechanics methods and tools to a wide range of systems, apparently much more complex than ideal ferromagnets, such as assemblies of molecules, granular particles[dG99], biological entities[AA99] (bacteria, ants, birds) or human societies (opinion formation, mass panic phenomena, hands clapping etc.).

Recently, as already observed, there has been a massive increase in the amount of autonomy and information flow and in the number of information nodes that are participating in Information Technology-based processes. Moreover there is a pressing need that information systems become ever more adaptive to user needs and rapidly expanding infrastructures. Consequently, there is a much higher interdependence of processing nodes than in the past and various properties observed in complex systems (such as self-organization) are now also observed in information systems.

On the other hand, as stated in the previous chapter, the study of the self organization and evolution of language and meaning has led to the idea that language can be seen as a complex dynamical system[Ste00]. Once relaxed the hypothesis of staticity of a language, a natural and very interesting question is how new conventions, developed in local interactions among few individuals, can become stable in the whole population[? ? ?]. The issue is of the outmost topicality since, for

the first time, the web allows for the spreading and the study of global bottom up created semiotic systems. For instance collaborative tagging systems enable users to self organize systems of tags and in that way build up and maintain social networks and share information.

We believe that the ideas and tools necessary to tackle these critical issues can come from the scientific theory of complex systems as it developed over the past decade in the natural sciences (for a first example see [CLP07]). These tools need to be adapted and made suitable to the issues confronted by IT-developers, and a serious effort needs to be undertaken to disseminate these tools in the community of IT at large.

The main lesson coming from the experience of the study of Complex Systems, in terms of methodology, procedure and tools, can be summarized as follows:

(a) identifying and defining the simplest (minimal) models (i.e. algorithmic procedures) which could capture the main features of Information Technology based systems. It is important to stress the need in this field of shared and general models to create a common framework where different groups could compare their approaches and discuss the results. On the other hand the models should exhibit the extreme level of simplicity compatible with the desired phenomenology. This has several advantages. It could allow for discovering underlying universalities, i.e., realizing that behind the details of each single model there could be a level where the mathematical structure is similar. This implies, on its turn, the possibility to perform mapping with other known models and exploit the background of the already acquired knowledge for those models. (b) identifying the most suitable theoretical concepts and tools to attempt the solutions of the models. It is important to outline how the possibility to obtain analytical and general solutions for the models proposed could open the way to a positive feedback providing further inputs for understanding and designing new experiments and devices. In general there are two main questions one should be able to answer. On the one hand, we need to find the general laws that govern the semiotic dynamics of a particular system, for example, how the maximum number of words in use is related to the number of agents in the population. On the other hand, we need to find the explanation of these laws as a mathematical property of the dynamics. (c) coupled with the theoretical activity there should always be an experimental activity with a twofold aim. On the one hand the experiments, as well as the observation of the realities one is interested in, provide inputs for the modeling and the theoretical activity. On the other hand they represent the framework where the theoretical predictions are checked. The outcome of these experiments will be compared with the theoretical and the numerical results and it will be used to better focus the modelling and the theoretical approach. There should then exist a positive feedback mechanism between the theoretical and the experimental activities in order to make the progresses robust, well-understood and concrete.

Chapter 9

Outlook

As their frequent use already shows, resource sharing systems avoid the knowledge acquisition bottleneck, which was one of the main reasons for the failure of many expert systems with a more sophisticated knowledge representation. The comparison of the steep rise of the Web 2.0 compared to the relatively slow increase of implemented Semantic Web applications shows that apparently the former does not suffer from this bottleneck - in contrast, many people are willing to contribute. The latter, on the other hand, still suffer from the lack of data: many interesting semantic web applications can currently only be evaluated on artificial data. We strongly believe that the Web 2.0 will establish itself as a shared knowledge space - both for private and for industrial applications. A stronger theoretical basis will provide the foundation for narrowing the gap between both approaches. This is one of the primary objectives of the TAGora project: bringing together researchers in various domains of complex systems and researchers in various areas of Information Technology (computer science, web technologies, ubiquitous computing), in particular those facing the challenge of Semiotic Dynamics in on line social communities. The impact that we foresee on the scientific community is potentially enormous because it goes beyond the specific scientific objectives and technologies focused on in this project. The hope is that of fostering a general movement towards the interrelation of complex systems and Information Technology by showing successful examples of cooperation and by posing and solving concrete non-trivial problems. It is only by seeing clear examples that we expect the scientific research community to follow suit.

Bibliography

[AA99] Sunny Y. Auyang and Sunny A. Auyang. *Foundations of Complex-system Theories : In Economics, Evolutionary Biology, and Statistical Physics*. Cambridge University Press, August 1999.

[ASC07] Rabeeh Abbasi, Steffen Staab, and Philipp Cimiano. Organizing resources on tagging systems using t-org. In *Bridging the Gap between Semantic Web and Web 2.0, workshop at ESWC 2007*, Innsbruck, Austria, 6 2007.

[ASSM07] Sofia Angeletou, Marta Sabou, Lucia Specia, and Enrico Motta. Bridging the gap between folksonomies and the semantic web: An experience report. In *Bridging the Gep between Semantic Web and Web 2.0 (SemNet 2007)*, pages 30–43, 2007.

[BL97] Tim Berners-Lee. Realising the full potential of the web, 1997. Presentation at a W3C Meeting, London, UK.

[BL98] Tim Berners-Lee. A roadmap to the semantic web, 1998.

[BLHH+06a] Tim Berners-Lee, Wendy Hall, James Hendler, Nigel Shadbolt, and Danny Weitzner. Creating a science of the web. *Science*, 2006.

[BLHH+06b] Tim Berners-Lee, Wendy Hall, James A. Hendler, Kieron OŠHara, Nigel Shadbolt, and Daniel J. Weitzner. A framework for web science. *Foundations and Trends in Web Science*, 1(1):1–130, 2006.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[Bud05a] Bud. Ibm's intranet and folksonomy. In *The Community Engine*. http://thecommunityengine.com/home/archives/2005/03/ibms_intranet_a.html, March 2005.

[Bud05b] Bud. xfolk: An xhtml microformat for folksonomy. In *The Community Engine*. http://thecommunityengine.com/home/archives/2005/03/xfolk_an_xhtml.html, March 2005.

[CG00] Robert Cailliau and James Gillies. *How the Web Was Born: The Story of the World Wide Web*. Oxford University Press, 2000.

[CLP07] Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences United States of America*, 104:1461, 2007.

[dG99] P. G. de Gennes. Granular matter: a tentative view. *Reviews of Modern Physics*, 71(2):S374–S382, 1999.

[Hak88] H. Haken. *Information and Self-Organization*. Springer Verlag, December 1988.

[HGM06] P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Stanford InfoLab, 2006.

[HJSS06a] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In York Sure and John Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *LNAI*, pages 411–426, Heidelberg, June 2006. Springer.

[HJSS06b] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Trend detection in folksonomies. In Yannis S. Avrithis, Yiannis Kompatsiaris, Steffen Staab, and Noel E. O'Connor, editors, *Proc. First International Conference on Semantics And Digital Media Technology (SAMT)*, volume 4306 of *LNCS*, pages 56–70, Heidelberg, dec 2006. Springer.

[HRS06] H. Halpin, V. Robu, and H. Shepard. The dynamics and semantics of collaborative tagging. In *In Proceedings of 1st Semantic Authoring and Annotation Workshop held during ISWC-2006*, 2006.

[LC01] Bo Leuf and Ward Cunningham. *The Wiki Way: Collaboration and Sharing on the Internet*. Addison-Wesley Professional, April 2001.

[Mik05] Peter Mika. Ontologies are us - a unified model of social networks and semantics. In *Proceedings of the 4th International Semantic Web Conference (ISWC)*, 2005.

[NV04] Thomas P. Novak and Alladi Venkatesh. Has the internet become indispensable. *Communications of the ACM*, 7(47):37–42, 2004.

[Res98] David Resnick. *Politics on the Internet: the normalization of cyberspace*, pages 48–68. New York: Routledge, 1998.

[Rez06] Serak Rezane. Folksonomies in intranets. Bachelor thesis, University of Kassel, Kassel, 2006.

[Sch06] Patrick Schmitz. Inducing ontology from flickr tags. In *Proceedings of the Collaborative Web Tagging Workshop at WWW2006*, 2006.

[SHJS06] Christoph Schmitz, Andreas Hotho, Robert Jäschke, and Gerd Stumme. Mining association rules in folksonomies. In *Proc. IFCS 2006 Conference*, Ljubljana, July 2006.

[SM07] Lucia Specia and Enrico Motta. Integrating folksonomies with the semantic web. In *Proceedings of the European Semantic Web Conference*, 2007.

[Ste00] Luc Steels. Language as a complex adaptive system. In M. Schoenauer, editor, *Proceedings of PPSN VI*, Lecture Notes in Computer Science, Berlin, Germany, September 2000. Springer-Verlag.

[Ste06] L. Steels. Semiotic dynamics for embodied agents. *IEEE Intelligent Systems*, 21(3):32–38, May/June 2006.

[Sur05] James Surowiecki. *The Wisdom of Crowds*. Anchor, August 2005.

[Vic01] T. Vicsek. A question of scale. *Nature*, (411):421–421, 2001.

[Vic02] T. Vicsek. Complexity: The bigger picture. *Nature*, (418):131–131, 2002.