Information dynamics in web-based social systems

Vittorio Loreto[1], Ciro Cattuto[2], and Andrea Baldassarri[1]

[1] Dipartimento di Fisica, "Sapienza" Università di Roma and SMC, INFM-CNR, Piazzale A. Moro 2, 00185 Roma I-ITALY
vittorio.loreto@roma1.infn.it
[2] Centro Studi e Ricerche Enrico Fermi
Compendio Viminale, 00184 Rome, Italy

1 ICT[3]-mediated social interactions

Understanding the relation between the Web and the world involves learning about a complex cycle of interactions, which 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. 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.

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. In particular, a new paradigm is actually gaining impact very quickly in most of such new large-scale information systems: Collaborative Tagging. In web applications like *Flickr* (`http://flickr.com`), *Connotea* (`http://www.connotea.org`), *BibSonomy* (`http://www.bibsonomy.org`) or *del.icio.us* (`http://del.icio.us`), people

[3] Information and Communication Technology

no longer make passive use of online resources. Instead, they take on an active role and enrich resources with semantically meaningful 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.

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" [1]. More precisely, the constituents of such systems are able to develop a complex collective behavior not trivially deducible from the knowledge of the rules that govern their mutual interactions [2, 3].

We believe we are facing a unique opportunity to exploit and give theoretical foundations to the recent, though extremely rapid, developments of self-organization and 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 upcoming challenges of Web Science:

- How do microscopic interactions (at the users' level) affect macroscopic emergent behaviors of online communities?
- 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 ICT systems, how can we bridge the gap between the top-down and bottom up approaches? One of the big failures of user interfaces and human-machine interaction today comes from their lack of adaptability 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 few years? The knowledge acquisition bottleneck in top-down approaches, i.e., the knowledge transfer from experts to formal systems, should be rephrased here in terms of *wisdom of crowds* [4]: is the knowledge aggregation and organization emerging from the uncoordinated activity of millions of users better than the centralized control of a few experts?

And of course one among the most important challenging tasks is the fostering and development of a new culture of interaction between ICT and Complex Systems Science. This will require efforts on both sides, at the level of methodology and communication, but even more at the research planning level, where incentives are needed to promote and support an entirely new research community.

2 Folksonomies as laboratories

During the last 18 months, social bookmarking tools have received surging attention in academic and industrial communities. Social bookmarking tools share with the Semantic Web vision the idea of facilitating the collaborative organization and sharing of knowledge 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 the "folk". The most prominent collaborative tagging systems, for instance, *del.icio.us* (`http://del.icio.us/`) and *Flickr* (`http://flickr.com`) have already more than one million of users. In the reference sharing systems *BibSonomy* and *Connotea* researchers and others insert, annotate, 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*. 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 one's 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. In the framework of these tools 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 [5] (see for instance `http://www.espgame.org/`). In such a case, the (ludic) activity of Internet users have been used to enrich of metadata images in a database. More generally, the idea that the individual, selfish activity of users on the web can possess very useful side effects, is far more general than the example cited. Just to mention another very relevant case: logs of queries submitted by users to search engines represent in principle an incredible wealth of information,

precious to refine future searches and, more generally, to organize knowledge and content of the web. The techniques to profit from such an unprecedented opportunity are, however, far from trivial. Specific technical and theoretical tools need to be developed in order to take advantage of such a huge quantity of data and to extract from this noisy source solid and usable information. Such tools should explicitly consider how users interact on the web, how they manage the continuous flow of data they receive, and, ultimately, what are the basic mechanisms involved in their brain activity. In this sense, it is not excluded that folksonomies and, more generally, the new social platforms appearing on the web, could rapidly become a very interesting laboratory for psycholinguistics and social sciences.

3 Theoretical foundations

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 [6, 7, 8, 9, 10, 11]. These macroscopic phenomena have naturally called for a Complex Systems approach to understand the regularities at large scale as collective effects of the interaction among single individuals, considered as relatively simple entities.

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 [12, 13]). These tools need to be adapted and made suitable to the issues confronted by ICT researchers and practitioners, and a serious effort must be undertaken to disseminate these tools in the ICT community 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:

- identifying and defining the simplest (minimal) models (i.e., algorithmic procedures) which are able to capture the main features of the investigated systems. It is important to stress the need of shared, general and validated models, in order to create a common framework where different groups can 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 allows to uncover underlying universalities, i.e., realizing that behind the details of each single model there can be a level where the mathematical structure is similar. This implies, in turn, the possibility to perform mapping with other well-known models, and to exploit existing background knowledge about the latter.

- identifying 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 insights to understand and design new experiments and devices.
- planning experiments and analyzing their outcome. Coupled with the theoretical activity there should always be an experimental activity with a twofold aim: on the one hand the experiments provide data for the modeling and the theoretical activity. On the other hand, they represent the framework where the theoretical predictions can be tested. The outcome of these experiments can be compared with theoretical models, and this comparison can be used to feed back into the modeling activity. As with every true scientific investigation, there should always exist a feedback cycle between the theoretical and the experimental activities in order to make the progresses robust and well-understood.

4 Interfaces and Impacted Fields

The vision outlined above gains momentum from – and feeds back into – several different areas of expertise. One of the challenges that has to be tackled is defining and opening communication channels to existing disciplines and research fields, in order to leverage existing know-how and avoid duplication of effort where a simple and informed "translation" of established results is possible. In the following we discuss some of the interfaces that could play the role of enablers, and convey impact to other fields.

4.1 Collaborative Knowledge Management

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 [14]. Microsoft also intends to invest in this area [15].

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 modeling 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.

4.2 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 [16] 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 web pages. 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, a new algorithm has been proposed in [17], 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 [17], recommendations, and trend detection [18].

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.

4.3 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:

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

4.4 Social network analysis

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 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 Outlook

In this paper we explored one of the possible reserch avenues related to the call IST-2007.8.4: Science of complex systems for socially intelligent ICT. We focused in particular on the information dynamics in web-based social systems, a unique opportunity to foster social partecipation to the management of the huge amount of resources available through the web. We addressed in particular the focal points related to “theoretical tools/models” as well as “data-driven simulation” while the issue of predictability stands on a longer

timescale. We also highlithed different areas of expertise which could be impacted by these studies.

References

1. H. Haken. *Information and Self-Organization*. Springer Verlag, December 1988.
2. T. Vicsek. A question of scale. *Nature*, (411):421–421, 2001.
3. T. Vicsek. Complexity: The bigger picture. *Nature*, (418):131–131, 2002.
4. James Surowiecki. *The Wisdom of Crowds*. Anchor, August 2005.
5. L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, New York, NY, USA, 2004. ACM Press.
6. P. Clifford and A. Sudbury. A model for spatial conflict. *Biometrika*, 60(3):581–588, 1973.
7. K. Sznajd-Weron and J. Sznajd. Opinion Evolution in Closed Community. *Int. J. Mod. Phys. C*, 11:1157–1165, 2000.
8. G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. Mixing beliefs among interacting agents. *Adv. Compl. Sys.*, 3(1-4):87–98, 2000.
9. R. Axelrod. The Dissemination of Culture: A Model with Local Convergence and Global Polarization. *J. Conflict Resolut.*, 41(2):203–226, 1997.
10. L. Steels. A self-organizing spatial vocabulary. *Artificial Life Journal*, 2(3):319–332, 1995.
11. A. Baronchelli, M. Felici, E. Caglioti, V. Loreto, and L. Steels. Sharp transition towards shared vocabularies in multi-agent systems. *J. Stat. Mech.*, (P06014), 2006.
12. C. Cattuto, V. Loreto, and L. Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences United States of America*, 104:1461, 2007.
13. C. Cattuto, C. Schmitz, A. Baldassarri, V.D.P. Servedio, V. Loreto, A. Hotho, M. Grahl, and G. Stumme. Network properties of folksonomies. *AI Communications Special Issue on "Network Analysis in Natural Sciences and Engineering" (to appear)*, 2007.
14. Bud. Ibms intranet and folksonomy. In *The Community Engine*. http://thecommunityengine.com/home/archives/2005/03/ibms_intranet_a.html, March 2005.
15. Bud. xfolk: An xhtml microformat for folksonomy. In *The Community Engine*. http://thecommunityengine.com/home/archives/2005/03/xfolk_an_xhtml.html, March 2005.
16. S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.
17. A. Hotho, R. Jäschke, C. Schmitz, and G. 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.
18. A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Trend detection in folksonomies. In Y.S. Avrithis, Y. Kompatsiaris, S. Staab, and N.E. O'Connor, editors, *Proc. First Int. Conf. on Semantics And Digital Media Technology (SAMT)*, volume 4306 of *LNCS*, pages 56–70, Heidelberg, dec 2006. Springer.