Politecnico di Milano
Dipartimento di Elettronica e Informazione
Dottorato di Ricerca in Ingegneria dell'Informazione

Social production of knowledge by online communities

Tesi di dottorato di:
David Laniado

Relatore:
Prof. Marco Colombetti
Tutore:
Prof. Letizia Tanca
Coordinatore del programma di dottorato:
Prof. Carlo Fiorini

XXIII ciclo - 2011

Politecnico di Milano
Dipartimento di Elettronica e Informazione
Piazza Leonardo da Vinci 32 I 20133 — Milano

Politecnico di Milano
Dipartimento di Elettronica e Informazione
Dottorato di Ricerca in Ingegneria dell'Informazione

Social production of knowledge by online communities

Doctoral Dissertation of:
David Laniado

Advisor:
Prof. Marco Colombetti
Tutor:
Prof. Letizia Tanca
Supervisor of the Doctoral Program:
Prof. Carlo Fiorini

XXIII edition - 2011

Politecnico di Milano
Dipartimento di Elettronica e Informazione
Piazza Leonardo da Vinci 32 I 20133 — Milano

Acknowledgments

First of all I wish to thank Ester, who accompanied me through all the phases of writing this thesis, patiently supporting (and "sopporting", as we would say in Italian) me, participating in the rush of long pre-deadline days and nights, and with her brightness helping my more human part to keep always alive behind the nerd. I also thank my family, who always sustained me during these years, getting to appreciate words such as "folksonomy" or "disassortativity", and my friends, who often summarize my work with implausible statements like "he is inventing the new Google, it will work much better", or "he is planning the revolution through the Internet".

I will never forget the day when I first met Davide Eynard: it was in a hacklab, and with sparkling eyes he started telling me about the wisdom of the crowds, dreaming of a peer-to-peer world... In few seconds I was hooked. This was the beginning of a beautiful friendship and, by accident, of a master thesis and a PhD. Davide is an enthusiast and a real hacker: from him I learned such important things as using linux, programming in perl, writing a paper, inspecting the logs of a server to find out which kinds of attacks it has suffered, crossing Palo Alto by bike without getting lost among thousands of identical houses, unperceivedly crawling gigabytes of data, and considering any unknown person as a friend.

The second person that will find much of himself in these pages is Riccardo Tasso, who came in turn for a master thesis, and then became a collegue and a friend. Tasso, from old greek: order. His pragmatic character is perfectly complementary to mine, so working with him was really a big deal, and I am convinced that he was sent to me as a magic helper of fairy tales. Riccardo is an innovator, and thanks to his willfulness we started studying Wikipedia, social network analysis, and python: still three pillars of my life, and three things that we continue to have in common.

Prof. Colombetti was behind all this, leaving us the freedom to follow the ideas which impassioned us and to explore realms that were mostly new also to him, but at the same time wisely leading us towards the most

promizing directions; and - the best you can expect from your advisor - always caring about us as people before anything else. In front of any apparently insuperable difficulty, easy solutions and brilliant ideas always emerge naturally after few minutes talking with him.

I'm also grateful to the reviewers of this thesis, Ciro Cattuto and Francesco Bonchi, who have taken the time to read it, and to provide precious advices on how to improve it.

I would like to acknowledge many people at Politecnico, in particular all the AirLab group and the inhabitants of our mighty PhD student office, room T11, where I spent so many days conciliating hard work with having a really great time. An essential person was Nicola Basilico, whom I first knew through his laptop in the intranet: it was named ahimsa. Indeed, beyond many PhD adventures, Nico was a companion of deep philosophical discussions, and even of hacktivism actions.

The first collaboration I had outside of Italy was for NiceTag: I learned a great deal from Freddy Limpens, Fabien Gandon and Alexandre Monnin. After years of engineering, working with a French philosopher was one of the most exciting experiences in my PhD.

When I met Peter Mika at another workshop, I already knew his work. In particular, he had written the only article about the Semantic Web which moved me: "Ontologies are us". The possibility of an internship with him at Yahoo! Research was a revolution both for my PhD and for my life. It was at the same time my Erasmus and my first experience of work out of the university; under Peter's precious supervision I had the opportunity to work on big data with huge computers, and nice people from all over the world.

In Barcelona I was also continuing the study of Wikipedia discussions with Riccardo, and by chance I met one of the most expert researchers on online discussions working just upstairs: Andreas Kaltenbrunner. So started my second and more durable adventure in Barcelona. All the improvements you may notice from all points of view towards the end of the thesis, and especially in the last chapter about Wikipedia, are due to his careful supervision, and to the influence of the other people in the Information, Technology and Society group at Barcelona Media. Having the chance to keep working with them in an interdisciplinary and intercontinental group is the best guarantee that my way has just begun.

Abstract

In recent years a new paradigm has emerged on the Web, characterized by the massive participation of users in the production of content. The results present the typical advantages of a bottom up process: information tends to cover the most various topics, keeping up to date and reflecting the point of view of the users, giving prominence to the most popular ideas but representing also the long tail of diverse views. On the downside, social Web applications suffer for a lack of organization; the absence of a single coherent point of view, in conjunction with scarcely structured content, makes it harder to retrieve and organize information. The Semantic Web offers standards and tools for the representation of knowledge in structured format, but most online communities appear as still far and sometimes reluctant to the adoption of these solutions, which can hardly deal with the simple and quick interfaces that characterize Web 2.0 applications, and with the messy heterogeneous data created by many different-minded users.

This work presents an investigation on how the new bottom up paradigm based on the participation of large masses of users on the Web can deal with production and organization of knowledge on a large scale. Starting from the observation of emerging dynamics, mechanisms and conventions adopted by online communities to manage content, this thesis offers an insight into the main challenges raised by the huge amount of heterogeneous data created by users on the social Web, focusing on three of its pillars: microblogging, social tagging and wikis. A variety of approaches, ranging from information retrieval and social network analysis to Semantic Web technologies, are leveraged to shed light on interaction patterns which characterize content production in these systems, to assess their value as sources of structured knowledge, and to propose solutions which can improve current applications.

Sommario

In questi anni stiamo assistendo all'affermarsi nel Web di un nuovo paradigma, basato sulla partecipazione massiva degli utenti alla produzione di contenuti. I risultati presentano i tipici vantaggi di un processo bottom-up: i contenuti tendono a coprire gli argomenti più vari, a restare sempre aggiornati e a riflettere il punto di vista degli utenti, mettendo in risalto le idee più diffuse ma rappresentando anche la lunga coda di punti di vista diversi. Dall'altra parte, le applicazioni del social Web soffrono di una mancanza di organizzazione; l'assenza di un unico punto di vista coerente, insieme alla mancanza di informazione strutturata, rende più difficile l'organizzazione e la ricerca dei contenuti. Il Semantic Web offre standard e strumenti per la rappresentazione della conoscenza in formato strutturato, ma le comunità online sono per lo più ancora lontane e spesso riluttanti all'adozione di queste soluzioni, che appaiono inadatte alle interfacce estremamente semplici che caratterizzano il cosiddetto Web 2.0, e alla quantità di dati eterogenei e disordinati creati dagli utenti.

La domanda che sta alla base di questo lavoro è come il nuovo paradigma bottom-up basato sulla partecipazione massiva degli utenti in rete può conciliarsi con l'organizzazione della conoscenza su vasta scala. Partendo dall'osservazione delle dinamiche emergenti, dei meccanismi e delle convenzioni adottati dalle comunità online per organizzare i contenuti, questa tesi offre una panoramica sulle principali sfide legate alla produzione collettiva di conoscenza nel social Web, concentradosi su tre dei suoi pilastri: il microblogging, i sistemi a tag e i wiki. Diversi approcci, dall'information retrieval e la social network analysis alle tecnologie del Web semantico, sono utilizzati per studiare modelli di interazione che caratterizzano questi sistemi e proporre soluzioni per migliorare le applicazioni attuali.

Contents

1 Introduction

1.1 Motivations and goal

In the last decade, we have witnessed the explosion of social applications on the Web. The low technological barriers and the development of easy interfaces based on the paradigm of the *read-write Web* have made possible the active participation of large masses of users. Web sites like Delicious[1], Flickr[2], Youtube[3], Twitter[4] and Wikipedia[5], started as obscure experimental projects, have all been characterized by an exponential growth until reaching millions of active users in few years.

The result of this wide participation is the production of huge amounts of information, with the typical advantages of a bottom up process: information tends to cover the most various topics, keeping up to date and reflecting the point of view of the users, often giving prominence to the most popular ideas, but representing also the long tail of diverse views in a democratic way.

Recently, the importance of this kind of systems and their influence on the whole society has emerged in the affirmation of so called *wiki-revolutions* of the *Arab spring* [1, 59, 49], or in the rise of the 15M movement in Spain and in the campaign for the anti-nuclear referendum in Italy, where social media such as Facebook, Twitter and Youtube supplied to the lack of information provided by TV and traditional mono-directional media, and were probably a determinant factor for reaching the quorum[6].

One of the first successful experiments which leveraged the Internet to foster the active participation of a large base of users, and can be seen as an ancestor of Web-based collaborative projects like Wikipedia, is the

[1] http://www.delicious.com
[2] http://www.flickr.com
[3] http://youtube.com
[4] http://twitter.com
[5] http://wikipedia.org
[6] See http://www.guardian.co.uk/commentisfree/2011/jun/14/silvio-berlusconi-italian-referendum

development of the Linux kernel. To describe the process which in few years allowed the project led by a 21 year old student to compete with the major commercial operating systems, Eric S. Raymond introduced the metaphor of the *bazaar*, a new bottom up model made possible by the Internet and based on the free collaboration of thousands of volunteers spread all over the world, opposed to the *cathedral*, the traditional hierarchical model [149]. Whereas to build a cathedral everything is projected in detail from the beginning by a few people, and some experts work, mainly in isolation, for the development of its single parts, in the bazaar anyone can propose tweaks and changes, which are managed by the community in a continuous spontaneous process of natural selection of best ideas. The strength of this dynamic and democratic model is given by the multiplicity of points of view: according to the so called *Linus's law*: "given enough eyeballs, all bugs are shallow"; the challenge is the ability to harmonize a huge quantity of heterogeneous contributions to create an organic result. As Reymond pointed out, the innovation was not technological, but social.

Also regarding Web 2.0, although the importance of new programming paradigms like AJAX has been often empathized, it is today a widely accepted idea that the main innovation is social. In other words, there is no specific technological advance that marks the transition from the old static Web to so called Web 2.0; instead, the ingredients were all there, and the revolution was made by the thousands of users participating in this new generation of applications, who made their success and decreed a change in the Web usage.

The model of the bazaar can be seen as the founding paradigm of many among the most successful new generation Web applications: folksonomies represent a bottom up approach to classification, opposed to taxonomies created by experts [163], while the blogosphere challenges traditional mainstream media by enabling a many-to-many communication paradigm [30], and in social news websites like Digg[7] the users propose, vote and select news; Wikipedia is an encyclopedia which can be edited by anyone, and whose entries are created, modified and discussed each day by thousands of users.

The initial general scepticism towards this kind of systems has progressively left place to a growing acceptance. The advantages of the lightweight and dynamic classification allowed by folksonomies have been largely recognized [117], while the quality of Wikipedia articles has been shown to be comparable to that of the major commercial competi-

[7] http://digg.com

tors [54]. Most importantly, the amount of content created daily by users in these systems could hardly be produced and maintained by any company or organization relying only on expert teams.

On the downside, Social Web systems suffer for a lack of organization, and it is often not easy to assess quality and provenance of content. The absence of one single coherent point of view can give place to inconsistencies and ambiguities, and makes it harder to retrieve information and to structure and organize knowledge. In particular, the absence of unique identifiers for entities is a serious issue for many of these systems, where it is often difficult to make sense of content and metadata in the form of rough text, without any explicit connection to a specific meaning. This, in conjunction with the absence of controlled vocabularies, can cause very low performance both in terms of precision and recall. The creation and use of taxonomies in this context is another challenge; hierarchies are very useful for organizing information and for helping to better browse and search, but require agreement on shared models.

Semantic Web technologies can represent a solution to many of these issues, providing models, languages and tools for knowledge representation. Unfortunately, these technologies are still far from the average internet user, and can hardly deal with the simple and quick interfaces that characterize Web 2.0 applications, and with the messy heterogeneous data created by many different-minded users. In other words, the Semantic Web appears to be still closer to the cathedral than to the bazaar; this can be seen as a reason for the scarce availability of ontologies on the Web, and for the limited diffusion of Semantic Web technologies.

Goal of this thesis is to study how the bottom up model of the *bazaar* can be applied to the production and organization of knowledge.

A first necessary step is an in depth study of existing successful applications, to understand the mechanisms and the dynamics which rule content production in this kind of systems. As we deal with a new and changing environment, which offers many challenges and for which there are no consolidated techniques, no unique approach would be suitable in this context, and focusing on one single case study could result too reductive. On the contrary, we choose several case studies from different online communities, involving on one hand different levels of participation and different kinds of activities, and on the other hand different challenges related to knowledge organization.

To understand social dynamics and detect patterns of interaction in these systems, and to assess their value as sources of structured infor-

mation, we rely on a variety of approaches ranging from information retrieval to data mining, from statistics to social network analysis. Furthermore, we propose some solutions to improve current systems by leveraging Semantic Web technologies.

1.2 Novel contributions

The first questions on which this thesis is focused are related to identifiers in the Social Web: the alphabet for the construction of structured knowledge.

As a first case study we choose Twitter, the most popular microblogging system, and we focus on hashtags, a convention adopted by users to face information fragmentation: adding a hash at the beginning of a word, this is turned into a tag. Hashtags are potentially very useful to aggregate content and conversations around a topic, a community or a thread of conversation, and their function is similar to that of URIs in the Semantic Web. However, not all hashtags are used in the same way, not all of them aggregate messages around a community or a topic, not all of them endure in time, and not all of them have an actual meaning. To identify the hashtags which show the desirable characteristics of strong identifiers, and could hence serve as identifiers for the Semantic Web, we represent them as virtual documents and we propose some metrics based on information retrieval. We introduce the notion of *nontag*, as the word corresponding to a hashtag after removing the hash, and we compare the usage of each tag with its corresponding nontag. We compute entropy of tags and we study the evolution of their usage over time. We look at the various ways in which hashtags are used, and show through evaluation that our metrics can be applied to detect hashtags that represent real world entities.

As a second case study we choose social bookmarking systems, that are one of the pillars of so called Web 2.0 and leverage the very little effort of assigning quick keywords to resources, to build collective classifications of large quantities of data. Tags here are not just inserted in short text messages as in microblogging, but they are metadata referred to a resource; on the other hand they are still just arbitrary sequences of characters freely chosen by users, which are not explicitly associated to a specific meaning. This freedom together with the simplicity of the user interfaces have done the success of this kind of systems. However, the absence of controlled vocabularies and of any kind of explicit semantics represents a strong limitation for information organization and retrieval;

the lack of an explicit meaning of tags is one of the main weaknesses of folksonomies, where on one side labels can be polysemic, and on the other one different tags corresponding to synonymous, or to different spelling or grammatical forms can refer to the very same concept. This results in a lack of both precision and recall. Also the absence of hierarchy and of explicit semantic relationships among tags affects the ease of searching and browsing in the tag-space. To address these limitations we propose an algorithm to disambiguate tags according to their context of usage, and to map them onto WordNet concepts; we use WordNet noun hierarchy to present a taxonomy of related tags, improving the users' navigation experience.

Beyond the basic level of identifiers, and the hierarchy among tags, we inspect the problems related to the lack of semantics at higher levels. In traditional tagging systems a tagging action consists just in the association of a label to a resource; there is no place for specifying anything about the kind of tag, the kind of communicative action performed, or the relation intercurring between the label and the tagged resource. This simplicity of the interfaces makes the users' vocabulary excessively poor in many cases, and limits the effectiveness of annotation retrieval. The demand for richer expressivity is witnessed for example by the diffusion of machine tags, spontaneous conventions that have become popular in tagging systems like Flickr to express specific properties of the annotated resources by means of special syntaxes. The third contribution of this thesis is the proposal of a vocabulary for tagging based on RDF named graphs. In the NiceTag ontology we represent a tag as the relation between a resource and a sign, with the possibility of typing all of these elements, including the relationship and the act of tagging itself, and we provide primitives for the description of different kinds of tagging actions as communication acts. Each tagging action is represented as a named graph, so also metadata about authorship, date and provenance can be managed. Leveraging Semantic Web technologies and standards we achieve richer expressivity, possibility of better performance in terms of information filtering, searching and browsing, and full interoperability for the aggregation of information from heterogeneous sources.

In the second part of the thesis we focus on the collaborative encyclopedia Wikipedia. Wikipedia is the largest example of collaboration on the Web, and as such covers a special interest for the study of social dynamics in online production communities. In these ten years of history, the community of Wikipedia editors has grown and evolved establishing its own norms and policies, assigning explicit and implicit

roles, until assuming the shapes of a complex "online society" based on auto-organization principles. According to the wiki paradigm, information about complete Wikipedia articles' revision history is public, giving place to one of the largest available datasets about online collaboration. The study of Wikipedia is important for us as a fundamental first step towards the comprehension of the collaboration and coordination mechanisms that drive production of knowledge by online communities on a large scale.

All content in Wikipedia is organized around articles, whose titles serve the function of identifiers; the analogy with Semantic Web URIs is straightforward, as Wikipedia articles represent already an important landmark for the Semantic Web community. Wikipedia disambiguation pages are widely used in NLP applications to associate words to the proper corresponding entities, and Wikipedia articles are among the most commonly used identifiers for entities on the Web.

Also a hierarchy of concepts is present in Wikipedia to organize content: the category structure, which is collaboratively managed by the users, analogously as the article content, merging different perspectives to produce a collective outcome. This hierarchy has been widely exploited for the construction of formal ontologies and knowledge bases; to this end, researchers have inspected the large and incoherent graph of Wikipedia categories looking for proper subsumption relations between classes and subclasses, and discarding the other ones. As an alternative approach, here we are no interested in extracting a formal ontology, but in making sense of all the associations established by the community, that result in an as imprecise as rich categorization of human knowledge. With the aim to assign Wikipedia articles to general topics, we first test the naive approach of associating with a category all the articles transitively assigned to it, which shows to be pretty good for some well delimited categories, but impracticable in most cases due to the extreme tangledness of the graph and to the conspicuous presence of inconsistencies. To deal with the entire graph we then propose and evaluate two methods, the first based on the shortest path between a page and a topic in the category graph, and the second on the probability of reaching each topic following a random path from a given article.

To study the patterns of collaboration in Wikipedia, the first challenge which we face is attributing the production of single units of content to their "real" authors. This is not always a trivial task; for example, while blog posts usually come with a single author, in the case of wikis pages are often redacted by many users, with very uneven levels of involvement

and of contribution acceptance by the community. To address this first issue, we present a method to individuate the authors of a wiki page as the editors who provided most of its accepted content, according to a metric of *edit longevity*. While previous works failed to study the network of collaborations in a wiki scaling up to the size of Wikipedia in a major language, by selecting the main contributors of each article we are able to filter out the occasional and minor editors, which constitute the large majority of users involved in the redaction of a page. As a first consequent result, we manage to represent a whole wiki as a co-authorship network, scaling up to the size of the English Wikipedia, to characterize the structure and dynamics of its community by means of social network analysis techniques, and to reveal hidden patterns of collaboration that emerge from the comparison with traditional scientific collaboration networks.

Beyond acting together on a same product, modifying and tweaking it concurrently, users in social Web applications can be provided with other mechanisms for explicit coordination through communication and discussion. In wikis, one fundamental tool employed by users for this purpose is represented by talk pages: besides each wiki page, there can be one or more special pages devoted to the discussion of its content. Previous works which have focused on the study of explicit coordination mechanisms in wikis have only taken into account the size of talk pages or their number of edits in order to generically quantify the amount of explicit communication associated to an article. In this work, a much deeper investigation is presented on the structure of Wikipedia discussions.

A first necessary contribution in this direction is the development of a tool for the extraction of discussion trees from MediaWiki talk pages. The tool has been designed, implemented and massively tested on Wikipedia, in order to make it as robust and flexible as possible, with respect to the extremely various and heterogeneous jungle of different conventions, errors and misspellings with which users have signed, dated, separated and indented comments during these ten years of Wikipedia history. The extraction of these data allows us on one hand to analyze the structure of the discussions according to different criteria; to estimate contentiousness, we count the number of chains of consecutive replies between a pair of users, while as a compact and robust indicator of depth we introduce the h-index of a discussion tree. On the other hand, we are able to study the network of explicit communications that accompany the collaborative redaction of content in Wikipedia, and to compare

patterns of discussion about articles with personal conversations.

With these two complementary approaches of studying the article content authors and the discussion structure, we are able to characterize structural units of content in a wiki along two different dimensions, corresponding to implicit and explicit coordination dynamics. Representing the networks of interactions over the whole system, we can study structure and dynamics of its community; in particular, we focus on assortativity measures to quantify the tendency of users to interact with similar or dissimilar users. For the discussion networks, where relationships are oriented, we compute directed assortativity. By comparing the interaction networks with randomly generated equivalents we assess the significance of our results, that point out the existence of distinguishing patterns of the Wikipedia community. In all networks we find evidence of the strong interaction between users in the core and in the periphery of the community, which we point out as a characterizing feature of Wikipedia community. We rely on different sociometric criteria to identify the most influential users and to analyze their role. Aggregating units of content with different levels of granularity according to semantic criteria, we characterize the sub-communities active around particular topics and areas of interest, and we find evidence of significant differences in collaboration patterns over diverse semantic areas.

1.3 Structure of the thesis

In the next chapter we describe the background of our work, focusing on a description of the two worlds that we want to bring closer to each other, the Social Web and the Semantic Web. At the light of this overview, we introduce the case studies which make up the central chapters of this thesis, and we frame them according to two dimensions: the kinds of activity, which imply different degrees of participation, and the levels of structured semantics, roughly corresponding to different layers in the Semantic Web technology stack.

In Chapter 3 we introduce the first case study: microblogging systems. We focus on hashtags, and we propose some metrics to evaluate them as strong identifiers for the Semantic Web; the results of this study have been published in [106].

In Chapter 4 we propose the approach published in [104] and [105] to map a social bookmarking site's tags onto concepts from WordNet, and to enrich the navigation interface with hierarchies derived from the ontology.

Chapter 5, based on the work described in [113] and [124], illustrates NiceTag, an ontology that leverages RDF named graphs to represent tags with a much richer expressivity than actual applications allow.

The subsequent chapters of this thesis are focused on the study of Wikipedia. The structure of Wikipedia category graph is inspected in Chapter 6, partly based on results published in [46], [107] and [108]; several approaches are proposed and examined to assign topics to articles, by leveraging the extremely rich and chaotic graph of topics and subtopics created by the users.

In Chapter 7, based on [107], the dynamics substanding the collaborative redaction of the online encyclopedia Wikipedia are studied by means of social network analysis techniques; the community of Wikipedia editors is represented as a co-authorship network and analyzed on a temporal dimension. Complementarily, the dynamics of explicit communication and coordination in Wikipedia are studied in Chapter 8, based on [108], through the analysis of discussion trees and reply networks in Wikipedia talk pages.

Finally, conclusions are drawn in Chapter 9.

2 Background

2.1 The Semantic Web

The advent of a Web of new generation, based on structured information automatically processable by machines, has been announced by Tim Berners Lee, considered as the founder of the Web, more than fifteen years ago [19]. Since then, a variety of models, languages and tools to implement this vision have been introduced and discussed, under the guide of the World Wide Web Consortium[1] (W3C).

Given the vastness of the matter, the goal of this section is not that of providing an exhaustive description of the Semantic Web; instead, it will offer a high level overview, with a special focus on Linked Data and on the challenges related to the new paradigm of the *read-write Web*.

Semantic Web basic principles have been well summarized in a W3C document, published in [97]:

1. *Everything can be identified by URIs*: resources from the physical world, such as people, objects and places, or abstract concepts, can be associated to univocal identifiers, based on namespaces; anyone controlling a part of a Web namespace can create a URI and associate it to some entity.

2. *Resources and links can have types*: while the current Web is made of simple documents and links between them, where no explicit semantics of the relationships connecting two resources is given, in the Semantic Web both resources and links are typed, making it easier for machines to make sense of them with knowledge about the classes to which resources belong, and the meaning of the relationships connecting them.

3. *Partial information is tolerated*: similarly to the current Web, the Semantic Web is unbounded: anyone can say anything about anything, so it is crucial to be able to deal with incomplete information.

[1] http://www.w3.org/

4. *There is no need for absolute truth*: like in the current Web, not all the content can be considered true, or trustworthy, so each application has to deal with this kind of environment and decide which information to rely on.

5. *Evolution is supported*: things on the Web can be defined at different times by different people, or even by the same people; there is need for tools to resolve ambiguities and clarify inconsistencies.

6. *Minimalist design*: aim of the W3C activity is to standardize no more than is necessary, to keep things as simple as possible.

These principles are implemented in the layers of Web technologies and standards commonly known as the *Semantic Web stack*, depicted in Figure 2.1 and illustrated in next Section.

2.1.1 The Semantic Web architecture stack

At the bottom of the stack there are two indispensable mechanisms: URIs and Unicode. URIs, or "Uniform Resource Identifiers" are strings of characters formatted according to a special syntax, which serve to univocally identify all kinds of resources. Unicode is a general standard for encoding, representation and handling of characters in most of the world's alphabets and writing systems.

Above these two basic standards, at the second level there is XML, which provides a surface syntax for content structure within documents. XML is completely agnostic to the semantics of documents, it is only a standard for representing structured content.

At the third level there is RDF, the "Resource Data Framework", a simple language to express relationships between resources identified by URIs, as triples <subject, predicate, object> [109]. Information is so represented in the form of graph; in fact, a collection of RDF statements inherently represents a directed, labeled graph, characterized by different kinds of relationships connecting resources. Though the standard way of representing RDF triples is through XML, other syntaxes exist; in particular, Notation3 (N3) is a serialization designed as a more compact and human readable alternative to XML. RDFa is a standard which allows to include RDF statements in XHTML documents.

A first layer over RDF is SPARQL, a query language which allows to query graph patterns along with their conjunctions and disjunctions. Through SPARQL it is possible to express queries across diverse data

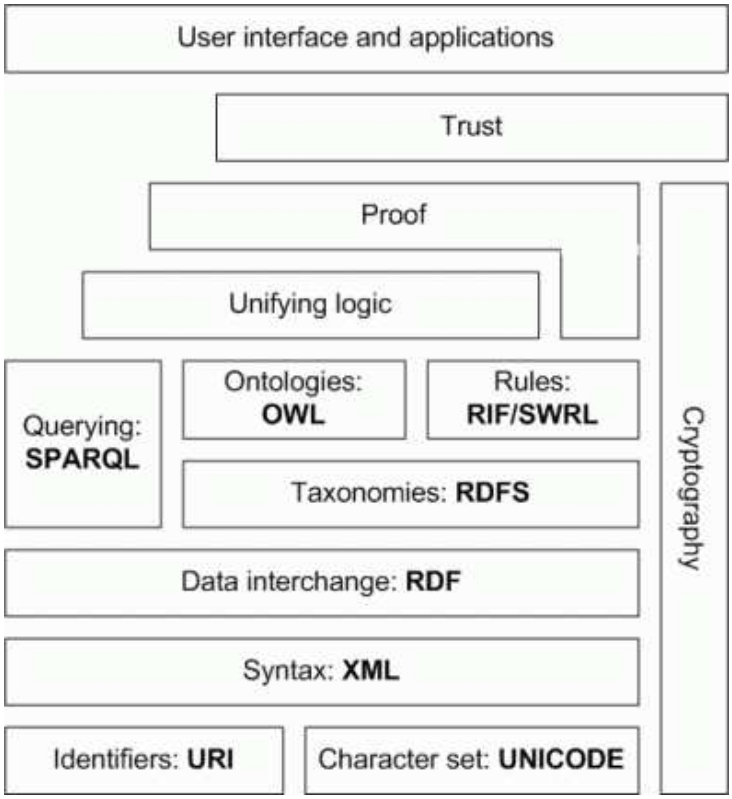


Figure 2.1: The Semantic Web architecture stack. Image from http://obitko.com/tutorials/ontologies-semantic-web/semantic-web-architecture.html. Accessed on September 7th, 2011.

sources; data can be stored natively as RDF or viewed as RDF via middleware.

RDF Schema (RDFS) provides a basic vocabulary for RDF, allowing one to define and describe properties and classes. In particular, in RDFS it is possible for example to define the domain and range of a property, or to describe hierarchies of classes and of properties.

In the above layer, the Web Ontology Language (OWL) is a standard of the W3C to extend RDFS with more advanced constructs to describe the semantics of RDF statements [10]. For example, in OWL it is possible to specify characteristics of properties such as symmetry or transitivity, to express cardinality constraints (e.g. : each movie has exactly one director) or assert that two individuals are the same, by means of the property `owl:sameAs`. An important characteristic of OWL is the "Open world assumption": if a statement cannot be proven to be true with current knowledge, we cannot draw the conclusion that the statement is false. At the same level of OWL, RIF and SWRL are rule languages, an alternative approach for reasoning, bringing a complementary expressive power. For the higher layers of the stack there are still not established standards.

2.1.2 Linked Oped Data

Two main directions can be identified in the Semantic Web community, splitting its name in two parts: one more focused on the "Web", and the second more focused on "Semantics" [45]. The first direction takes the current Web as a starting point, and aims at developing Web-based applications that use very little semantics but provide a powerful mechanism for linking data entities together relying on URIs, RDF and SPARQL to describe and query knowledge. The second direction is focused on representing more complex knowledge in machine-processable format, achieving a higher expressive power thanks to languages like OWL and RIF, and allowing for the inference of knowledge by means of automatic reasoners. The first direction, focused on the Web, and only lightweight semantics, is the one which is more relevant for this thesis, and it is the subject of this section.

The first of the six founding Semantic Web principles described at the beginning of this section, *"Everything can be identified by URI's"*, is the basis for a new paradigm, also known as the *Web of Data*: links are not simple anchors connecting HTML documents, but RDF relations between any kind of objects or concepts. Thanks to RDF, assertions can be expressed about data from different sources and from different

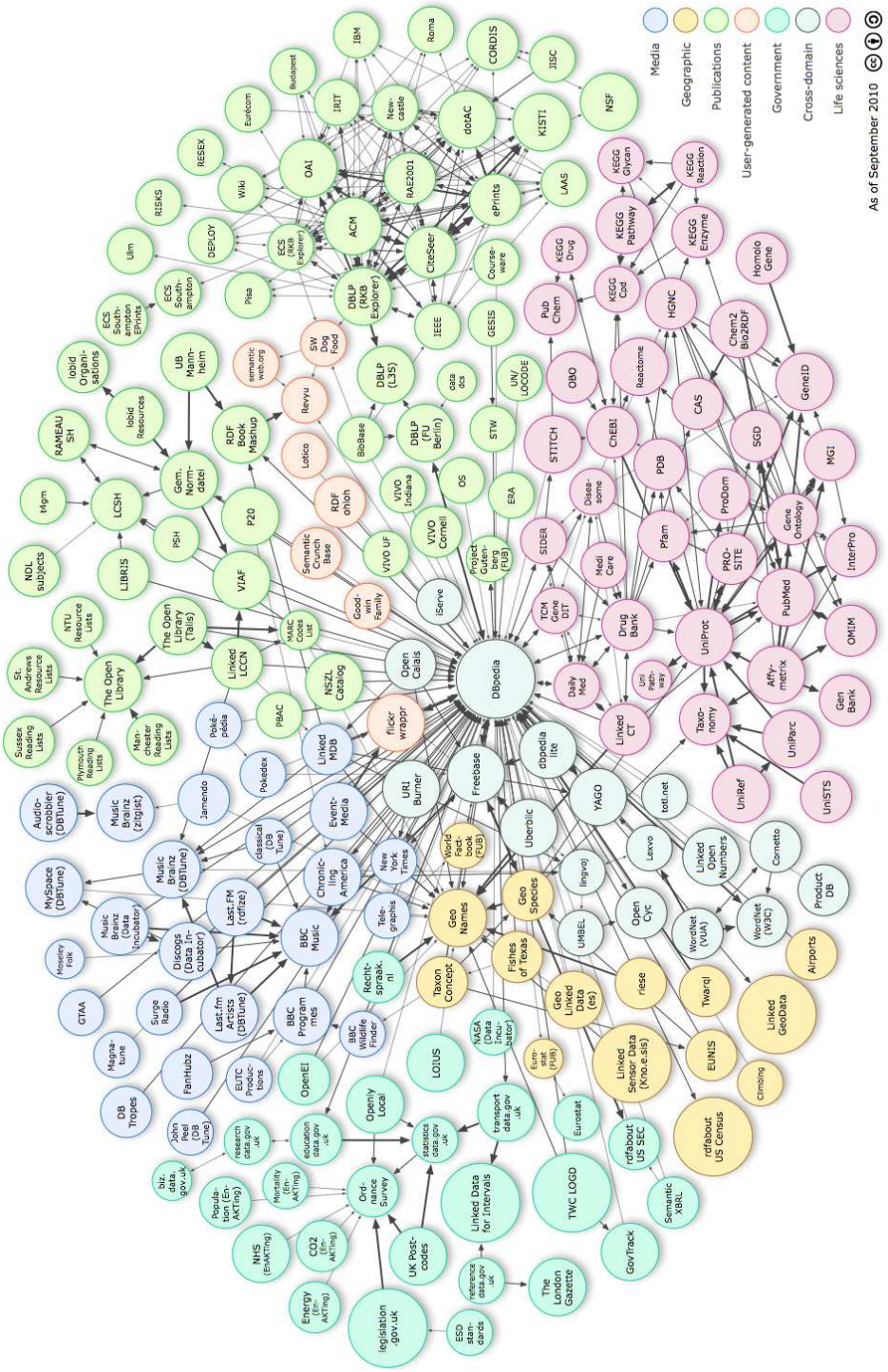


Figure 2.2: The Linking Open Data cloud diagram: circles represent data sets, and arrows the interlinkage between them; colors correspond to different domains. Image by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/. Accessed on September 9th, 2011.

namespacecs, and different vocabularies and data sets can be merged into one only Web of Linked Data [18].

The amount information published on the Web according to these principles is growing considerably: by September 2010 it had reached 25 billion RDF triples, interlinked by around 395 million RDF links. Figure 2.2 shows the Linking Open Data cloud diagram, which offers an overview of published data sets and the relationships interlinking them. As it can be seen, the central node in the cloud is DBpedia, a knowledge base derived from Wikipedia [12], which covers a variety of domains and is interconnected with many different data sets.

2.2 The Social Web

2.2.1 From the old Web to Web 2.0

As argued by Tim Berners-Lee, the Web is inherently social [144]; however, in the last decade the social dimension has gained importance in the Web, as we have assisted to a shift from static Web sites, where users are mostly just passive consumers, to more interactive Web applications, where they interact with one another and produce content, passing from consumers to *prosumers* [73].

This paradigm is also commonly known as "Web 2.0", as opposed to the old static "Web 1.0". The term "Web 2.0" got popular in 2003 with Tim O'Reilly's definition of the "Web as a platform" [134], where software applications are built upon the Web as opposed to upon the desktop, value is created by harnessing content generated by the users, and the core of applications includes data and not merely software.

A paradigmatic case of a successful application which has revolutioned the Web is Google: instead of maintaining expensive and ever aging taxonomies, redacted by some experts attempting to manually catalog the whole Web, as the major competitors were doing, Google bet on harvesting and leveraging the knowledge of Web masters, who establish links between Web pages. *Pagerank* [136] is a recursive algorithm based on mining the hyperlink structure of the Web: the more links a page receives from relevant pages, the more relevant it is considered. The core of Google success stands in focusing on links rather than on nodes, and on harnessing so called "wisdom of the crowds".

Other relevant changes include the transition from static personal Web sites to blogs, from CMSs to wikis, from centralized servers to peer to peer, from taxonomies to folksonomies, from stickyness aimed at keeping users as long as possible on a site to syndication, to make content

available through different channels.

In the following sections, we offer an insight into some relevant families of applications which represent the core of this new paradigm, and constitute the main subject of the research presented in this thesis.

2.2.2 Blogs

Weblogs, or more simply blogs, are a kind of Web sites usually maintained by an individual with regular entries. The difference with traditional Web sites is that publishing content is much easier and immediate and does not require knowledge of HTML mark up language or any technical skill; thanks to popular blogging platforms such as WordPress[2] or Blogger[3], no technical issue has to be faced to open a blog and to publish content. Blog posts are usually open to comments, enabling bidirectional communication, as well as discussion among the readers.

Started as a phenomenon in the late nineties, blogging gained momentum until reaching a large diffusion also among average internet users, turning many people into active producers of content on the Web. As of September 2011, there were about 170 million public blogs in existence [4]. Most blogs are of personal character. According to Tecnorati's 2010 "State of Blogosphere" [172], bloggers can be segmented as: hobbyists (65%), part timers (13%), corporate (1%) and self employed (21%).

The diffusion of blogs has been one of the fundamental steps in the switch from one-to-many communication, typical of mainstream media, to many-to-many communication, opening up to the neew paradigm defined by Manuel Castells as *mass self-communication*: a new medium which makes possible "the unlimited diversity and the largely autonomous origin of most of the communication flows that construct, and reconstruct every second the global and local production of meaning in the public mind" [30].

A key mechanism of blogging is content syndication, through which new content published on a blog (*feeds*) can be notified and made immediately available to other applications. Interchange of blog content is based on a common Web feed format named RSS[5]. The name means "RDF Site Summary", as it was originally designed to incorporate RDF triples, however it is also known as "Rich Site Summary", or "Really Simple Syndication", as it was soon simplified by removing RDF elements.

[2] http://wordpress.com
[3] http://www.blogger.com
[4] For live statistics about blogs, see http://www.blogpulse.com/
[5] See http://www.rssboard.org/

Due to a fork in the development of the standard, currently several different versions of RSS exist, some of which offer support for RDF and some not. The existence of a standardized format for the representation of blog content has been a key factor for the diffusion of blogging, allowing for aggregation and filtering of information from different sources and for data interchange among different applications.

2.2.3 Microblogging

Microblogging is a broadcast medium based on very short messages. The main differences with respect to traditional blogging are the brevity of messages, usually restricted to 140 characters, like SMSs, and the focus on relationships between users, as in social networking sites. Status updates in social networking sites like Facebook have generally no length limit, but in practice they tend to be very short; this scenario can in effect be considered very close to the one of microblogging, the main difference being that content is generally accessible only by a selected audience ("friends", "friends of friends", ...), while in the microblogging paradigm messages are broadcasted. While in social networking sites friendship is usually a symmetric bidirectional relation, in microblogging each user can be *follower* of any other users, to be notified of their messages.

The most popular microblogging service is Twitter[6]; created in 2006, it has known a very rapid growth and has reached over 200 million users as of 2011, generating over 200 million tweets and handling over 1.6 billion queries per day. Due to this tremendous diffusion, which made of it a social phenomenon [84], speaking of microblogging means before all speaking of Twitter, and we will refer especially to it; however, other platforms exists, such as Jaiku[7] or Tumblr[8].

While search engines present static information, updated with a relatively slow process, microblogging provides an immediate solution for real time news and real time communications. In particular, regarding the search interface, in Twitter results are presented in chronological order showing at top the most recent ones, regardless of relevance, which is the basic criterion in traditional search engines.

A paradigmatic scenario where real time communication is of crucial importance is that of emergencies; for example, the key role of Twitter during earthquakes has been reported in several occasions [119], and even

[6] http://twitter.com
[7] `http://www.jaiku.com/`
[8] `http://www.tumblr.com/`

a method to leverage Twitter users as sensors for detecting earthquakes have been proposed [153].

On the other hand the twofold nature of Twitter, news media on one side and social network on the other one [101], can raise some problems: though its strong potential for the publication and immediate retrieve of real time news, these risk of being overwhelmed by the huge amount of personal and conversational messages, which are usually not relevant for a general audience.

This issue is confirmed by studies on the content of tweets. One of the first efforts to classify Twitter messages was performed in [8]: almost half of the examined messages were classified as *pointless bubble*; many messages were also classified as *conversational* or *self promotion*; the study concluded that only 40 tweets over 2000, classified as *news*, contained interesting information. A comprehensive overview on different classifications of Twitter content in literature can be found in Stephen Dann's survey [39].

Special conventions have been adopted by users to deal with the brevity of tweets: *mentions* of other users are obtained by adding a special character "@" before the username of a tweeter, while adding a hash ("#") at the beginning of a word makes it a *hashtags*. These mechanisms, born as spontaneous conventions, have been integrated in the Twitter interface: users are notified of any message mentioning them, and hashtags inside tweets are rendered as links to the Twitter search page for that tag, where it is possible to see the last messages tagged with it. This mechanism allows for aggregation of content around events, conversations, groups of interest or topics. Another mechanism typical of microblogging is *retweet*, to forward a message written by another user.

2.2.4 Social tagging

In social tagging systems users associate freely chosen keywords to the resources that they want to bookmark or to categorize, with the double advantage of being able to retrieve them easily in the future, and to share them with the others. In fact, each user can generally explore two spaces, the one of her bookmarks and the one of everyone's bookmarks; tags can be used to filter and retrieve items. The combination of extremely simple and quick interfaces on one side, and both individual and social incentives on the other have made the success of this kind of systems, which are able to collect the very little effort performed by each user, to produce knowledge for the entire community. One of the very first applications

to adopt this paradigm was delicious[9], a social bookmarking platform started in 2003. Joshua Schachter, its founder, brilliantly synthetized the principle defining delicious as *"a way to remember in public"*.

The term *folksonomy* was coined in a discussion on the Information Architecture Institute Members Mailing List by Thomas Vander Wal [176], as the fusion of words "taxonomy" and "folk". In fact, the aggregation of tags created by many users gives place to a sort of taxonomy, or better represents an approach to classification alternative to taxonomies.

Folksonomies can be divided in two categories according to the identity of taggers: in *narrow* folksonomies only the user who published resource can annotate it, and in some case can eventually grant this right to her friends, while in *broad* folksonomies every user can annotate any item [176]. An example of narrow folksonomy is Flickr[10], a photo sharing site where only the user publishing a picture has the grant to tag it, while an example of broad folksonomy is the social bookmarking delicious, where any Web URL can be annotated and so no ownership mechanism exists.

One fundamental difference with respect to traditional taxonomic organization of knowledge, for example in a library, is that the resources annotated are not physical objects, which have to be collocated in one single place, like a book on a shelf. Library classification systems are based on the assumption that for any new book, its logical place already exists within the system, even before the book was published; instead, according to the words of Clay Shirky [163], in folksonomies "there is no shelf", and no limit to the vocabulary. This difference allows for many overlapping categorization criteria to be combined in folksonomies, where each item can be assigned to more categories, and new categories can be created at will.

As a result, folksonomies are *inclusive* and *current*, as new concepts can always be integrated, and they are *democratic*, because all points of view can find place; at the same time, they allow for *desire lines* to be drawn [121], as those associations which are established by more users tend to emerge. Semiotic dynamics typical of human languages such as crystallization of naming conventions, competition between terms and takeover by neologisms have been observed in the emergence of folksonomies [32].

On the downside, the absence of an authority and of a unique coherent point of view on the domain, combined with the simple interfaces which

[9] http://delicious.com/
[10] http://flickr.com

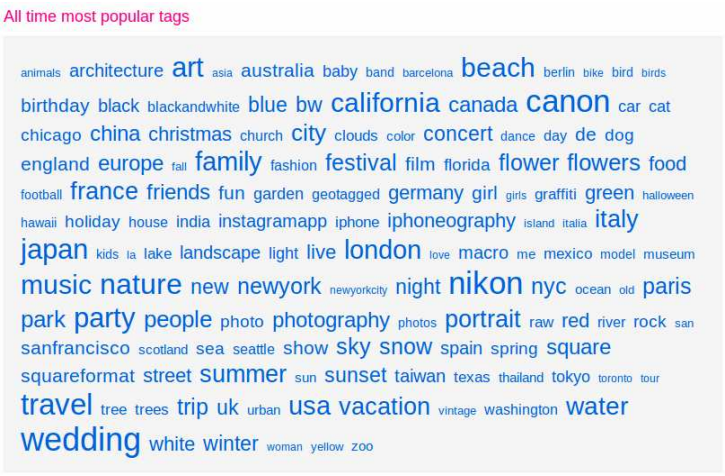


Figure 2.3: Tag cloud of the most used tags on Flickr. http://www.flickr.com/photos/tags/. Accessed: September 13th, 2011.

are devoid of any structured format to express knowledge, bears several limitations. The ambiguity of labels, which can be polysemic and are not associated to a specific meaning, causes low precision performance, while at the same time also recall can be very low due to the absence of synonym control. In fact, it is hard to retrieve all the items which have been associated to a given concept, as the same concept can be expressed through different synonyms, and even different forms of the same tag, such as singular and plurals, abbreviations, misspellings and other *low level variations* [100, 65]. Moreover, the lack of hierarchy makes difficult the exploration of the tags' flat space.

While the traditional classification schemes, based on taxonomies, favor searching and browsing, folksonomies encourage another paradigm of navigation, based on *finding* and *serendipity* [117]: as the amount of information is huge, and many meaningful paths exists among related tags, users and resources, it can be in some cases hard to search content according to traditional IR schemes, but while exploring tags it is very likely that a user will run into something interesting for her.

A common way to represent tags is by means of *tag clouds*, where keywords are usually ordered alphabetically, and variable font size is used to give prominence to the more frequent ones. As an example, Figure 2.3 shows the tag cloud of most used tags in Flickr.

Despite their strong limitations, folksonomies have rapidly emerged as

a natural low cost categorization solution, in the face of the rapid growth of content in the Web. According to Clay Shirky, *"The mass amateurization of publishing means the mass amateurization of cataloging is a forced move"*[11].

Folksonomies are nowaday common in a variety of ontext, from social bookmarking [67] to academic publications [79], from multimedia content, such as Flickr and Youtube, to geographic locations, like Foursquare[12].

2.2.5 Wikis

A wiki is basicly a Web site that can be edited by any user. The word "wiki" comes from Hawaiian and means "quick"; Ward Cunningham, pursuing the simplest mechanism to allow users modify information in a database, developed the first wiki and named it WikiWikiWeb inspired by the quick "Wiki Wiki Shuttle" bus at Honolulu airport [111].

The main characteristic of wiki systems is the ease of creating and modifying pages, directly from a Web browser, without need for any additional software; content is usually edited through some special simple syntax (*wikitext*), which is automatically turned into HTML and allows users to easily create titles of sections and paragraphs, apply different emphasis to text, such as using bold and italic, create lists of items, etc. Moreover, some wiki platforms offer WYSIWYG ("What you see is what you get") editing interfaces.

Changes introduced by users are usually made immediately effective; on the other hand, history of the edits done to a page are recorded, so it is also easy to restore a previous version, and it is possible to keep track of all interventions.

Wikis are generally open to alteration by any visitor, however different policies can be adopted for a whole wiki or for single pages; for example, the possibility to access to or to modify pages can be restricted only to authenticated users, or to the administrators.

Pages in a wiki can be easily interlinked, and users are encouraged to establish hyperlinks connecting related topics. Also links to still nonexistent pages can be created; these links are visualized in a special way, pushing users to create missing target pages.

Wikis have become popular in many contexts, ranging from soft-

[11] http://many.corante.com/archives/2005/01/22/folksonomies_are_a_forced_move_a_response_to_liz.php

[12] https://foursquare.com/

ware documentation[13] to knowledge sharing in companies and research groups[14], from map and geographical data[15] to encyclopedic knowledge.

The most popular wiki is for sure Wikipedia, an encyclopedia collaboratively redacted by users, existing in over 200 languages; thanks to the effort of thousands of users voluntarily working on it it daily, and of the improvements apported by occasional contributors and simple readers who correct imprecisions, Wikipedia is nowaday one of the most visited Web sites, and the most consulted encyclopedia. Its quality on scientific entries has been shown to be comparable to that of the major commercial encyclopedias [54], while on the other side the amount of content and the vastness of coverage offered by Wikipedia is beyond compare.

The software on which Wikipedia is based is MediaWiki [14], written in PHP and used also in other projects of the non-profit Wikimedia Foundation, such as the dictionary and thesaurus Wiktionary, Wikinews for collaboratively redacted news, Wikiversity for free learning tools and Wikiquote for quotations. The code of MediaWiki is free software, released under the GPL licence, and many other wikis are based on it. In the Wikia project, wikis about any specific topic running MediaWiki software can be created in few clicks.

Beyond MediaWiki, many other software packages offering similar features exist, such as DokuWiki[16], aimed at the needs of developer teams and small companies, and JSPWiki[17], based on Java, servlets and JSP.

Implicit coordination is the key mechanism for collaboration in a wiki, where users interact among themselves by working together on a common artifact, but they do not necessarily explicitly communicate, or know each other. To characterize this kind of interaction, some researchers have used the notion of stigmergy [173], borrowed from biology: it has been observed that, to build a nest, termites modify the environment stimulating the response of other workers, whose transformations of the nest do in turn trigger other actions. Also in wikis users communicate by modifying their local environment, editing content and thus triggering actions of other users interested on the same pages [44]. The main difference between stigmergic wiki collaboration and co-authoring is identified by Mark Elliott [44] in the lack of discourse required to initiate and par-

[13] See for example the large wiki of the Ubuntu community: `https://wiki.ubuntu.com/`

[14] For example, the wiki used at Yahoo! Research Barcelona is `http://barcelona.research.yahoo.net/dokuwiki/`.

[15] `http://www.openstreetmap.org/`

[16] http://www.dokuwiki.org

[17] http://jspwiki.org/

take in collaboration: in wikis there is no need to become acquainted and maintain relationships with fellow contributors, as it is in traditional co-authorship.

However, other mechanisms involving direct communication are provided by some wiki platforms to support collaboration. Explicit communication is supported for example in MediaWiki through *talk pages*, special pages which are designed to be used in a forum-like way; these pages can be used also for polls.

Special administrative rights can be granted to some users, such as the possibility of protecting pages to restrict editing, or blocking specific users to prevent vandalism or other undesired behaviours. In some large wikis, like Wikipedia, administrators are elected by the community [27].

Wiki communities can have policies and best practices, defined by the community itself in a continuous process of auto-organization to face the new challenges encountered; however, the first rule in Wikipedia is "Ignore all rules", meaning that everything in a wiki has to be flexible, and rules that inhibit development and improvement of content shall be ignored [183]. Polls are used in Wikipedia only as extrema ratio, as the main mechanism for taking decisions is consensus [50]. One fundamental principle in Wikipedia is "assume good faith" of the other users, respecting different points of view and being tolerant toward mistakes, both behavioural and content-based [150].

Beyond general principles, many specific policies exist, to rule for example the management of copyright issues for multimedia content, or the minimum requirements for a music band to be considered encyclopedic, and thus justifying the existence of an associated entry. As another example regarding interaction, an important rule prevents non-administrator users from performing more than three reverts to a page within a 24 hour period [182]; this policy is aimed at avoiding *edit wars*, i.e. conflicts between users over an entry, resulting in the continuous effort of each user to override the modifications made according to a different point of view.

The ability of such a large and complex online production community to auto-organize by means of implicit and explicit norms and social roles, which are costantly created and refined in a transparent and democratic process, is one of the keys of Wikipedia's success, and makes of it a milestone for the study of online collaboration dynamics on a large scale.

2.3 Approach

Given the complex and manifold nature of the environment that we are about to investigate, which as illustrated in Section 2.2 is in constant and rapid evolution with the interplay of different kinds of structuring agents, adopting one single point of observation or relying on one predefined approach would be reductive. Instead, it is important to study this chaotic and changeable world from different points of view. According to George Devereux, a method is scientific if it adapts to the object [40].

The panorama of online communities can appear as depicted in Figure 2.4 to the eyes of an explorer willing to get an overview of it: wide barely known territories inhabited by tribes having different languages, habits and customs, and delimited by uncertain and changeable boundaries.

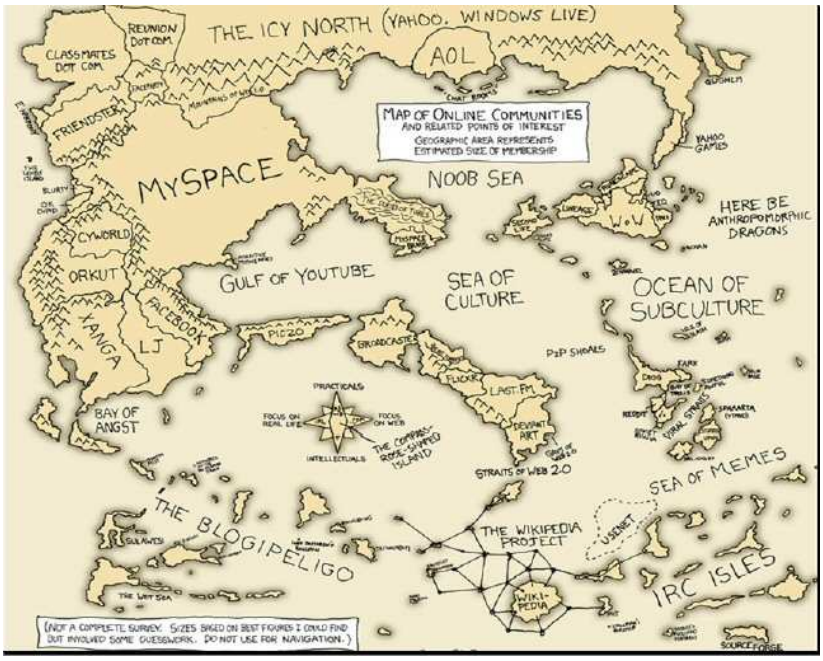


Figure 2.4: Map of online communities from the webcomic xkcd. http://xkcd.com/256/. Accessed on September 7th, 2011.

To get an overview of this composite scenario from a knowledge organization perspective, we combine two dimensions, the first based on the kind of activity, and the second on different levels of structured semantics.

In this section we briefly describe these two dimensions at the light of the concepts introduced in previous sections, then we combine them in Table 2.1 to frame the different challenges that online production communities have encountered and the most relevant solutions that they have adopted. Passing through all the cells which represent the possible combinations of user activity and structured semantics levels, we also offer an overview on the different contributions of this thesis and a key to compose them, drawing several threads that can be followed through the chapters.

2.3.1 Kinds of activity

As a first dimension we consider different activities typical of Social Web applications, ordered according to growing degree of participation.

As a first general kind of activity we consider communication finalized to *comment* something or to *notify* other users; here we are not interested in private communications involving only two or few individuals, such as emails, but in communications involving larger audience, such as blogging or public messages posted in social networking sites. As a paradigmatic case study for this activity we chose microblogging, which emerged over the last few years as one of the pillars of the Social Web. The brevity of messages makes somehow microblogging an extreme case, where the issues of fragmentation and ambiguity explode; for this reason it is a particularly relevant scenario for our investigation.

The second kind of activity which we consider is *sharing*; this is a form of participation in which usually users do not directly interact with one another, but every one bears a contribution through her individual behaviour, producing content for the whole community; individual contributions are collected, aggregated and made accessible through some interface. We are particularly interested in those systems which allow for labeling resources with text strings (*social tagging* applications), such that the aggregation of tags can provide a collective categorization of content (*folksonomy*), as explained in Section 2.2.4.

As a third kind of activity, associated to a higher degree of involvement, there is collaboration, i.e. working together on an artifact for a common outcome. According to theories of group coordination [152, 184], there are two distinct mechanisms through which communities can achieve a "unity of purpose": *implicit coordination*, based on unspoken expectations and shared mental models of the task to be accomplished, and *explicit coordination*, based on direct communication and verbal planning. We thus split collaboration into two activities, which we refer to

as *contributing*, corresponding to implicit coordination, and *discussing*, corresponding to explicit coordination.

Wikis are a straightforward example of collaboration, as users are involved in redacting Web pages together; as illustrated in Section 2.2.5, the basic mechanism which rules collaboration in wikis is implicit coordination through editing the same source text. However, explicit coordination is also fundamental, as it is witnessed by the amount of communication which accompanies collaborative writing of content in wikis [93]. In particular, the main space deputed to explicit coordination in wikis is represented by *talk pages*, special pages expressly designed for communication among the users.

2.3.2 Levels of structured semantics

On the vertical dimension we dispose different levels of structured semantics from a knowledge organization perspective.

At the lowest level we have *individuals*; the main challenge at this level is how to univocally identify and reference an individual entity. As explained in Section 2.1, in the Semantic Web this function is accomplished by URIs, which not coincidentally constitute the bottom of the Semantic Web technology stack depicted in Figure 2.1. *Identifiers* are the indispensable alphabet to construct a discourse grounding language to reality; for this reason URIs have been dubbed *Web proper names* [66]. Identifiers are not only important as a necessary step for the construction of more complex semantics, but also at a basic level for the aggregation of content.

At a second level we consider *classes*, or categories to which individuals can belong. This corresponds to RDFs classes in the Semantic Web. This level shares the challenges of the lower one, related to the univocal identification of an entity: classes have also to be identified and referenced, facing ambiguity. Grouping items in classes based on common characteristics is the basis for categorization, the first step for the organization of knowledge in semantic structures.

At the third level we consider *hierarchies*, which respond to the need of organizing information according to some coherent criterion. Hierarchies of concepts are achieved by organizing more specific classes under more general ones which can encompass them. The straightforward way of expressing hierarchical relationships according to Semantic Web standards is by means of the RDFs property `rdfs:subClassOf`.

The highest level is that of relations, or properties. It is important to observe that here we intend general properties, i.e. relations involving

individuals, so this is different from hierarchies, which instead describe a particular kind of relationship between classes. In the Semantic Web, properties are generally expressed as triples <subject, predicate, object> according to the RDF data model.

It is important to notice that, in our schema, until hierarchies higher structured semantics levels somehow imply the lower ones; in fact classes are defined to contain individuals, while hierarchies are specified between classes, and so they are also based on the lower levels. On the contrary, although we represent them at the top level, properties do not properly represent a higher level with respect to classes and hierarchies. In fact, as discussed above, properties do not necessarily depend on classes, as they can just be defined for individuals; on the other hand, hierarchies can be considered as a special kind of relationships, defined between classes. So there is a special expressive power at the level of hierarchies, and they can be considered orthogonal to properties. In fact, in the Semantic Web stack depicted in Figure 2.1, RDF is just the third layer.

2.3.3 Combining activities with structured semantics

In Table 2.1 we combine the two dimensions described above to offer a comprehensive overview of the main issues related to knowledge organization, that have to be faced by online communities. Each cell reports one or more solutions associated with managing knowledge at a given structured semantics level (on the vertical dimension), contextually with the corresponding activity (horizontal dimension). In this section we briefly comment the content of each cell of the table, passing through all the kinds of activities and the corresponding case studies, and illustrating for each of them the major challenges and solutions at each level of structured semantics.

Notify/comment

In the scenario of microblogging, which corresponds to the activity of commenting and notifying other users, the function of identifiers is mostly performed by *hashtags*, both at the level of individuals and classes. Beyond hashtags, at the level of individuals also usernames, which are usually preceded by a character "@", and URIs can be considered as identifiers. As it will be shown in Chapter 3, hashtags play an important role to face the problem of fragmentation, to which this kind of systems is naturally prone, given the brevity of messages and the potentially very wide base of users who can be involved in a conversation.

Properties	nanoformats	NiceTag (5) machine tags	semantic wikis wiki infoboxes	collab. ontology editors semantic wikis
Hierarchy		enriched folksonomies (4) emergent semantics	wiki category graph (6)	category talk pages
Classes	#hashtags (3)	tags	wiki categories	
Individuals	URIs	URIs	wiki pages (7)	wiki talk pages (8)
	microblogging	social tagging	wiki	wiki talk
	notify/comment	***share***	***contribute***	***discuss***

Table 2.1: Each cell of the table contains one or more solutions to manage structured semantics at a given level (according to the first column) in relation with different kinds of activity (indicated in the bottom row). In addition, the second last row indicates the contexts analyzed, corresponding to each kind of activity. Chapter numbers related to each topic are reported in parentheses.

Chapter 3 proposes some metrics to evaluate hashtags as strong identifiers, and to detect which ones could be mapped to real world named entities.

Given the brevity of tweets, it is apparently hard to think of structured semantics at a higher level; however, there are some experiments in this direction, based on synthetic formats variously called *nanoformats*[18], picoformats[19] or *mycrosyntaxes*[20]: among other works, TwitLogic [162] and HyperTwitter [74] propose special syntaxes which allow for the specification of structured relationships in the short space of a tweet.

Share

In social tagging systems, we can see a correspondence of the tagged resources with individuals, and of tags with classes; in fact tags are used as categories to label and organize items, in order to allow users to retrieve and browse them. The correspondence of tagged resources with instances and of tags with concepts is commonly accepted in literature, according to the classic tripartite model of <tags, resources, users> [122]. However, as shown in Table 2.1, tags can also be seen as individuals, as it is the case in the NiceTag model, described in Chapter 5.

Tagged resources are usually referenced through their URIs, while tags are just strings of characters, arbitrary keywords that are not associated by the users to any explicit meaning. Some models, such as MOAT[21] [139] and CommonTag[22], have been proposed to overcome this lack of semantics by allowing users to explicitly ground tags to ontology concepts. In Chapter 4, tags from social bookmarking site Delicious are disambiguated according to their context and mapped onto elements from the WordNet ontology, as explained in the following.

At the above level, there is usually no possibility of specifying hierarchies among tags, which constitute just a flat messy space; the most common way of visualizing a set of tags is through tag clouds, where tags are sorted alphabetically without considering any semantic criterion. An exception to this is represented by isolated cases of applications which introduced the possibility of explicitly organize tags in taxonomic structures, while some systems like Delicious slightly moved in this direction

[18] http://microformats.org/wiki/twitter-nanoformats
[19] http://microformats.org/wiki/picoformats
[20] http://microsyntax.pbworks.com/
[21] http://moat-project.org/ontology
[22] http://commontag.org/

by introducing *bundles* to group tags[23]; however, no subsumption relationships are allowed between bundles, so hierarchy is limited to one level. Moreover, bundles can only be used in the context of a single user's tag space (*personomy*), and no mechanism exists to merge categorizations created by different users.

Apart for these weak mechanisms provided by current applications, we individuate two main kinds of approaches to deal with the lack of hierarchies in folksonomies. The first approach consists in the extraction of emergent semantics; mining the tripartite graph of users, resources and tags, some researchers have proposed techniques to infer taxonomic relationships between tags [122].

The second approach instead consists in the enrichment of folksonomies by means of ontologies; a first necessary step for this task, of course, is mapping tags to ontology concepts, facing the problem of ambiguity and finding appropriate univocal identifiers at the level of *names*. In Chapter 4 we follow this approach to enrich the navigation interface of Delicious by integrating hierarchies from the WordNet ontology; to mention another relevant project based on a similar approach, FLOR is a framework for the enrichment of folksonomies by means of ontologies retrieved online [9].

An interesting experiment, which goes in the direction of integrating activities with higher user involvement in social tagging to build consensus on hierarchical relationships among tags, is presented in [112]: a framework is proposed that allows users to explicitly agree or disagree on broader-narrower relationships between tags.

At the level of general relationships, an interesting first step can be found in *machine tags*: a convention spontaneously adopted by Flickr users to express arbitrary properties of the tagged resource through a special syntax, which allows to specify both the property and the value in the space of a single tag. NiceTag, presented in Capter 5, is an ontology which offers a more general solution with full expressive richness by means of Semantic Web technologies and standards. Thanks to the use of named graphs, each tag can be represented as an RDF relation between a tagged resource and a sign, embedded in a record identified by a URI. It shall be noted that in NiceTag both the tagged resource and the tag (more precisely the sign) are treated as individuals, as arbitrary relationships can be specified to link them; tags are not considered as classes.

[23] http://blog.delicious.com/blog/2005/10/bundle_up.html

Contribute

We consider now collaboration, focusing on the context of wikis. The basic unity of content are pages, which are usually identified by their title, unique inside a wiki; page titles can hence be considered as identifiers at the level of individuals, aggregating the content which they encompass, and associated to a URL. Further, the content often offers a definition or a description of the title; this is particularly true in encyclopedic wikis like Wikipedia, where wiki pages correspond to encyclopedic entries, and so to entities in the real world. Although Wikipedia is the most relevant case, this kind of convention is followed in many other wikies: just to mention some examples, in Wikitravel[24] pages correspond to places; the Wikia project[25] offers many wikis related to the most various topics and based on the MediaWiki platform, where pages can be associated to recipes and ingredients[26], or to comic book series[27] or to camera types, models and companies[28] and so on. Also in the AIRWiki, the wiki of the Artificial Intelligence an Robotics group at Politecnico di Milano, page "Social production of knowledge"[29] univocally represents this thesis.

So, wikis offer by design a powerful mechanism to aggregate content around univocal identifiers, but on the other hand they raise issues concerning authorship: to whom can a unit of content be attributed in a wiki? There is usually no individual authorship in wikis as each page is typically the product of contributions by many users; however, it can be useful in many contexts to be able to attribute a page to its main contributors. Transparency is a key feature of the wiki paradigm and edit history of each page is usually public, so complete information about who contributed to a page, when and how, is available. However, it is hard to manually make sense of the history of edits. To address this issue, in Chapter 7 we propose a general approach to automatically identify the main authors of a wiki page, mining its revision history and selecting the users who provided most of the content which has been accepted by the community. By applying this methodology to the English Wikipedia, we are able to study it as a co-authorship network, comparing it to scientific communities; beyond investigation on collaboration patterns over the whole wiki, also analysis restricted to specific semantic areas (topical categories) is performed; this leads us towards the next level of

[24] http://wikitravel.org/
[25] http://www.wikia.com/
[26] http://recipes.wikia.com/
[27] http://comics.wikia.com/
[28] http://camerapedia.wikia.com/
[29] http://airlab.elet.polimi.it/index.php/Social_production_of_knowledge

structured semantics.

At the level of classes, MediaWiki offers a powerful instrument to group individuals (i.e. wiki pages): wiki categories, or labels which can be associated by users to pages. Like in tagging, the association between individuals (in this case wiki pages) and classes (wiki categories) does not follow a rigid semantics, and does often not correspond to an "is-a" relationship. In fact, many categories appear as more similar to individual entities (e.g. there are categories like "Berlin" or "Beatles": they are different from the articles having the same name, as they are located in a different namespace in MediaWiki, which corresponds to a different function). These categories are intended as topics which contain pages corresponding to sub-topics; for this reason they are treated as classes, having the function of grouping individual items.

In MediaWiki, categories can be in turn assigned to higher level categories, so the community can collaboratively create a hierarchy. While in most of previous work the resulting category graph has been only exploited to extract taxonomic *is-a* relationships, in Chapter 6 we present several approaches which leverage all the relationships established by the community to assign each page to one or more general topics.

At the highest level, some wikis offer mechanisms to express structured relationships by means of a special syntax; in MediaWiki this is achieved by means of *infoboxes*, special templates which allow to specify the value (or object) of properties associated with the page, which is always the implied subject of the relationship. This mechanism can be considered analogous to machine tags, with the obvious difference that, coherently with the collaboration paradigm of wikis, there is not a place for each value of a property assigned by a different users, but just one single place where the value of a property can be defined and edited by the community. Structured knowledge contained in Wikipedia infoboxes is represented according to Semantic Web standards in DBpedia [12], a knowledge base where Wikipedia articles are treated as individual instances, and the properties expressed in infoboxes are turned into RDF triples.

A more complete approach for representing properties is offered by *semantic wikis*. In Semantic MediaWiki[30] (SMW), pages are treated as instances in an ontology, and categories as classes; each link leading from a page to another one can be associated with a structured relationship involving the corresponding entities. In this way, a structured semantic layer is integrated on top of the navigational link structure. Other se-

[30] http://semantic-mediawiki.org

mantic wikis, like KiWi[31] and OntoWiki[32] allow for the specification of RDF relationships between individual entities.

In collaborative knowledge bases like Freebase[33] the wiki approach is applied for the creation of large repositories of structured data, where relationships between individual entities can be specified by users.

Discuss

We now shift to the last kind of activity, corresponding to the second founding mechanism of online collaboration, i.e. explicit coordination through discussion. In MediaWiki and other wiki platforms, each page (or *content page*) can have a *talk page* associated to it, as a space where the community can discuss about its content. A talk page is a space for explicit coordination and discussion, associated to the unit of content identified by the corresponding content page title. In Chapter 8 we present a study of discussion patterns in Wikipedia talk pages.

Though we have not focused on discussion at higher levels of structured semantics in this thesis, it is worth mentioning here some mechanisms provided by current systems. In MediaWiki there is place for discussion about categories and hierarchical relationships; each category in fact has a corresponding page, which can have an associated discussion pages. For example, in Wikipedia the inclusion of categories like "Homeopathy" or "Christal healing" into "Pseudoscience" generated intense discussion on the corresponding category talk page[34].

At the level of structured relationships, there is no specific mechanism in MediaWiki to discuss about infobox properties; as they are specified by inserting them inside the source of the page corresponding to the subject of the relationship, they can be discussed in the associated talk page. The same holds for Semantic MediaWiki, where in addition each property (but not every instance of a property) has its own page, where the meaning and the eventual restrictions of the property are defined, and there is also place for discussion. Similarly, in the collaborative ontology editor Collaborative Protégé[35], annotations can be attached to each component of an ontology, and discussion threads can be attached to an ontology [174]. However, as in Semantic MediaWiki, no annotation or discussion thread can be attached directly to an instance of

[31] http://www.kiwi-project.eu/
[32] http://ontowiki.net/
[33] http://www.freebase.com/
[34] http://en.wikipedia.org/wiki/Category_talk:Pseudoscience
[35] http://protegewiki.stanford.edu/wiki/Collaborative_Protege

a property, as only the property itself, or individual elements can be annotated. A project which allows for engaging discussion and seeking consensus about individual instance of properties is ISICIL [112]; relations are embedded in named graphs by means of the NiceTag ontology presented in Chapter 5, and so it is possible to annotate them.

3 Assessment of Twitter hashtags as strong identifiers for the Semantic Web

3.1 Introduction

Twitter, a service for publishing short messages that has been growing nearly exponentially in the past years. Twitter handled over 600 messages every second by January, 2010[1], and has become a cultural phenomenon in many parts of the world. This success can be attributed in a large part to the simplicity of system, and the resulting cleanliness of its web site and its APIs. The ease of publishing also means that Twitter inspires timely contributions and has become an important source of information for late-breaking news, and it is already being exploited by major search engines. While appealing to publishers, the simplicity of Twitter has its downsides for anyone consuming and processing Twitter data, especially when it comes to aggregating messages. Aggregation is a necessary first step for many applications of Twitter mining, including news and trend detection, brand management and customer service, and it's also a crucial first step in separating personal communications from public discussions.

Within the current system, however, the aggregation functions are limited to filtering tweets by users or restricting by keywords. Even in the latter case, tweets are organized by time, and not by relevance as is common for search engines. Without formal organization, aggregating tweets that belong to the same conversation or discuss the same topic is daunting. Table 3.1 shows ten consecutive messages retrieved for the keyword *banana*. These messages are not only posted in different languages, but are part of different ongoing conversations and refer to very different topics (the plant, a chain store, a dance, a club, and others). Keyword search is not only imprecise in aggregation, but is also missing

[1] `http://blog.twitter.com/2010/02/measuring-tweets.html`

out on a number of messages that do not contain the particular keyword. As Twitter messages are unusually short, keyword search is likely to fail in recall. As an example, during a January, 2010 earthquake in the San Francisco Bay Area, search engines have been criticized in showing only tweets that explicitly mentioned the word *earthquake*. A second, related problem is separating personal communication and news publishing, the two main cases of Twitter usage [101]. This is a crucial function for aggregators that are interested only in the conversations that concern topics of broader interests such as news or current events.

As a community solution to these problems, Twitter users have adopted the convention of adding a hash at the beginning of a word to turn it into a *hashtag*. Hashtags are meant to be identifiers for discussions that revolve around the same topic. By including hashtags in a message, users indicate to which conversations their message is related to. When used appropriately, searching on these hashtags would return messages that belong to the same conversation (even if they don't contain the same keywords), and thereby solving the aggregation problem. Coincidentally, this is the same function that strong identifiers (URIs) play in the Semantic Web. The questions we ask then is which hashtags behave as strong identifiers (if any), and could they be mapped to concept identifiers in the Semantic Web? This issue can be collocated in the bottom-left cell in 2.1: the activity in macro-blogging systems is as general as the concept of *communication*, without a strong implication of participation; also on the vertical dimension, representing the level of structure from a knowledge organization perspective, we are at the basic level, i.e. the definition of names of things.

In this chapter, strongly based on the work published in [106], we address this issue by proposing a set of metrics to measure the extent to which hashtags exhibit the desirable properties of strong identifiers. Our first contribution is thus formalizing the characteristic properties of strong identifiers in terms of usage in social media systems. We give a general description of hashtag usage according to these metrics (Section 3.2). Using a manually collected data set, we evaluate how well our metrics can identify those hashtags that represent named entities and concepts found in Freebase, a large and broad-coverage knowledge base (Section 3.3). Our contribution is in measuring the quality of hashtags as identifiers and selecting the hashtags that are candidate concept identifiers, a necessary first step in mapping hashtags to Semantic Web knowledge bases and identifying hashtags that are candidates for extending knowledge bases. We discuss related work in Section 3.4 and point

Boo368	@AvenLantz OMG I WANT A BANANA HAMMOCK XD
Endivisual	Got my dress..from banana republic..uhh im wearing dis dress once..? Thx..i dont need it to be so expensive -_-"
DevvonTerrell	World_of_Lala Fuh Sure!!RT @_RosettaStone_: Real talk DevvonTerrell grandmother needs to open up a bakery. Her Banana Pudding is on. HAHA!!
makalovesbieber	RT @bieberhechos: RT si te gusta la banana de Justin (? JAJAJA no mentira.
reidnwrite	@EDHMovement Unforgettable goes SUPER hard...he slipped like banana peels for not having you know you know on the album!
jojoserquina	Chicken Tinola with bitter melon, hot long horn and banana pepper, ginger and spices http://twitgoo.com/14sosn
Vol_Sus	RT @So_Delicious: Hot Fudge-Dipped Frozen Banana Bites wa recipe for Coconut Peanut Butter Hot Fudge Sauce! `http://bit.ly/aknbRe` YUM!
Markaw00	Eating a banana sandwich and watching Hero.
LauraRogers13	Mom asks me if I want a banana and I start doing the banana dance...I've been at cheer too much!
MissRicaRica	RT @philthyrichFOD: @MissRiCaRiCa *PHILTHY RICH* Coming Home Party And Video Shoot July 4th @ Banana Joes 950 10th St Modesto http://twitpic.com/1oh6ji PLZ RT.

Table 3.1: A consecutive sequence of Twitter message for the query 'banana'.

to future work in Section 3.5.

3.2 Metrics for hashtag evaluation

There is no special support for tagging in Twitter, and new tags are simply introduced by prefixing a word with the hash sign. Hashtags may be used for personal categorization, but in the vast majority of cases the intention of those who introduce a new hashtag is to evolve it into a symbol that is used by a community of users interested in and discussing a particular topic. The goal of such a hashtag is to help search and aggregation of messages related to the same topic, a function that is similar to the role of (shared) URIs in the Semantic Web.

There are a number of desirable criteria that a hashtag should fulfill in this role, similar to how 'cool URIs' are differentiated from poor URIs.

In the following, we formalize some of these characteristics.

1. **Frequency.** The hashtag is used by a community of users with some frequency. We measure frequency both in number of users and number of messages sent, and explore the correlations between the two ways of measuring frequency.

2. **Specificity.** The extent to which the usage of a hashtag deviates from the usage of the word without a hash.

3. **Consistency in usage.** The hashtag is used consistently by different users and in different messages to indicate a single topic or concept.

4. **Stability over time.** The hashtag should become a part of the persistent vocabulary of Twitter users, i.e. it should have sustained levels of usage and should have a stable meaning over a period time.

In the following, we formalize these notions based on a Vector Space Model (VSM) for hashtags.

3.2.1 A vector space model for hashtags

The basic model of Twitter can be represented by a set of tuples $S \subset M \times U \times P(H) \times T$ where M is a sequence of not more than 140 characters, U is the set of registered Twitter users, H is the set of hashtags and T is a set of discrete timestamps with a total order. The set of hashtags is the set of possible words that start with a hash. Hashtags form part of the message in the raw data, and we extract them using a regular expression `"#[a-zA-Z0-9_]+"`. The size limitation imposed on messages puts an upper bound on the potential length of hashtags, the number of possible hashtags as well as the number of hashtags that may appear in a single message.

In line with previous works on the analysis of folksonomy systems [31], we capture the semantics of the hashtags by their usage in the social media system. In particular, we will represent the meaning of hashtags using a Vector Space Model (VSM) [148]. VSMs are commonly used in information retrieval as a representation of documents, where each dimension corresponds to a term in the collection and each value measures the weight of that term for the document. In our case, we form virtual documents for each hashtag by considering all messages where

the hashtag appears. We don't filter messages by language, but it would be possible to build language specific representations this way.[2]

Formally, each hashtag h_j can be represented by a vector $\mathbf{h_j} = w_{1,j}, w_{2,j}..w_{N,j}$ where $w_{i,j} \in W, N = |W|$ and W is the set of unique terms in all of M. The simplest method for assigning weight is to consider term frequencies, i.e. $w_{i,j}$ is the number of messages in which term i co-occurs with hashtag j. In order to account for the different levels of specificity of terms with respect to hashtags, and to reduce the importance of the most common words, we obtain a more accurate model by applying *tf-idf* normalization: $w_{i,j} = tf_{i,j} \cdot idf_i$ where $tf_{i,j} = \frac{w_{i,j}}{\sum_{i=0}^{N} w_{i,j}}$ is the relative frequency of term i with respect to hashtag j; $idf_i = \log \frac{|H|}{|\{\mathbf{h_j} : w_{i_j} > 0\}|}$ is inversely proportional to the logarithm of the relative number of hashtags which term i appears with. For reasons of efficiency, we set elements $w_{i,j}$ lower than a threshold k to zero. In particular, this allows efficient indexing of the vectors using inverted indices.

We also introduce a bigram language model for hashtags; to do this, we define as *bigram* each pair of consecutive terms in a message, and as $\mathbf{b_j}$ the vector of all bigrams coocurring with tag j, $b_{i,j}$ being the number of messages in which bigram i and tag j co-occur. We apply tf-idf normalization in the same way as we compute it for single word co-occurrence.

Finally, we represent hashtags on a social dimension by means of their user occurrence vector $\mathbf{u_j}$, where $u_{i,j}$ is the number of messages tweeted by user u_i and containing hashtag h_j.

3.2.2 Frequency of usage

The **frequency of a hashtag** $h_i \in H$ in terms of the number of users and messages can be defined as

$$F_u(h_i) = |\{u : \exists(m, u, H_j, t) \in S \wedge h_i \in H_j\}| \tag{3.1}$$

$$F_m(h_i) = |\{m : \exists(m, u, H_j, t) \in S \wedge h_i \in H_j\}| \tag{3.2}$$

3.2.3 Specificity

While in most tagging systems tags are added as external metadata to describe the content, in Twitter tags are just words making part of the

[2]Based on previous experience, languages can be detected with good accuracy despite the short length of messages. The Twitter Search API also allows restricting tweets by language.

message, highlighted by means of a hash to assign them a special function. A hashtag can often just refer to the meaning of the corresponding word, but in some cases it can assume a very different usage. Often, the hash is added as a form of emphasis, and the user may not be aware that the word as a hashtag has a more specific or otherwise different meaning than the word itself.

It is thus interesting to observe if a hashtag has a meaning close to the one of the corresponding word without hash, that we will call a *non-tag*. As with URIs on the Semantic Web, we assume that hashtags that closely match the meaning of the corresponding non-tag will be used more frequently. On the other hand, we also expect that words that are used mostly as hashtags, or hashtags that are used with a different semantics than their non-tag, will be used more consistently, because they are re-used intentionally.

Similarly to our previous definitions, we define $\mathbf{n_j}$ as the term vector of the non-tag n_j derived from h_j by removing the hash. When building the term vector $\mathbf{n_j}$, we only consider non-tag n_j occurring in a message when the corresponding hashtag h_j is not used inside the same message. The intuition is that when a non-tag appears in a message where the corresponding hashtag has already been used, the semantics of the two are certainly not different. We apply tf-idf normalization to non-tags analogously to the one described in Section 3.2.1 for hashtags.

We compute the **specificity of a hashtag** as the similarity between the vectorial representation of the hashtag and the corresponding non-tag. For computing similarity, we use the well-known cosine similarity of the two co-occurrence vectors [154].

$$wsim(h_j, n_j) = \frac{\mathbf{h_j} \cdot \mathbf{n_j}}{\|\mathbf{h_j}\| \, \|\mathbf{n_j}\|} \tag{3.3}$$

Analogously, we define $\bar{\mathbf{u}}_\mathbf{j}$ as the model of the users of the non-tag u_j, where $\bar{u}_{i,j}$ is the number of messages in which user i used non-tag j. We measure *social specificity* by comparing the model of the users of hashtag h_j to the model of the users of non-tag n_j:

$$usim(h_j, n_j) = \frac{\mathbf{u_j} \cdot \bar{\mathbf{u}}_\mathbf{j}}{\|\mathbf{u_j}\| \, \|\bar{\mathbf{u}}_\mathbf{j}\|} \tag{3.4}$$

To be able to compare tags and non-tags also according to frequency, we define $\bar{F}_u(n_i)$ and $\bar{F}_m(n_i)$ the frequency of a non-tag in terms of users and messages, respectively.

3.2.4 Consistency of usage

An important requirement for strong identifiers on the Semantic Web is that they need to be used consistently across documents and users. As a measure of the variety of usage contexts of a hashtag, we study the *entropy* of our vectorial representations of hashtags. Entropy measures the amount of uncertainty associated with the value of a random variable, in other words how uniformly the probabilities are distributed across possible values of the variable.

We define the entropy of a hashtag j as:

$$H(j) = -\sum_{i=1}^{n} p(w_{i,j}) \log p(w_{i,j}) \tag{3.5}$$

Higher values of entropy point to more even distributions of probabilities, corresponding to tags being used in a variety of contexts, while lower values of entropy signifies more restricted usage of a tag.

Similarly, we measure entropy of bigrams co-occurring with a tag as

$$Hb(j) = -\sum_{i=1}^{n} p(b_{i,j}) \log p(b_{i,j}) \tag{3.6}$$

Non-tag entropy is measured like tag entropy: $\bar{H}(j) = -\sum_{i=1}^{n} p(\bar{w}_{i,j}) \log p(\bar{w}_{i,j})$

3.2.5 Stability over time

To study the evolution of hashtags on a temporal dimension, we chose to analyze them day by day. First of all, to be able to identify new tags emerging, we define as *new* on day d a tag not appearing in the previous k days. We will define *longevity* of a new tag $l_{d,k}(j)$ as the number of days in which tag j appears at least once, over the k days after its first occurrence on day d.

We then define $\mathbf{h_j^d}$ the vector of words appearing with tag j in some message on day d, and we measure similarity of a hashtag j on day d with respect to the previous day as

$$wsim_d(h_j) = \frac{\mathbf{h_j^d} \cdot \mathbf{h_j^{d-1}}}{\left\|\mathbf{h_j^d}\right\| \left\|\mathbf{h_j^{d-1}}\right\|} \tag{3.7}$$

Analogously, $\mathbf{u_j^d}$ is the vector of users who used tag j on day d, and $usim_d(h_j)$ is the similarity among users on day d and $d-1$.

3.3 Evaluation

3.3.1 Dataset

For this study we relied on a dataset of 539,432,680 messages, collected over the whole month of November 2009 (about 18 million per day). Slightly less than 50% of tweets are in English; to filter out messages in non-latin encoding, that we are not able to parse and study, we discarded all messages containing non-ASCII characters, reducing the size of the dataset of about 28%.

Twitter user interfaces allow for forwarding of messages written by other users; the original message is so "retweeted". As our study is based on the co-occurrence of words inside the same message, and massive retweeting that characterizes several tags might have a strong impact biasing the results, we decided to filter out all retweets. Retweets constitute 5.4% of messages, so the actual dimension of our dataset, after filtering, is of about 369 million messages.

To compute words co-occurring with a hashtag, we filtered out from the messages all Web links and Twitter usernames (words starting with "@"). To reduce the size of co-occurrence vectors, discarding items having a very low tf-idf, we used a threshold $k = 0.01$.

3.3.2 Descriptive statistics

Figure 3.1 shows the distribution of the number of hashtags per message; overall, only 31.5 million messages, corresponding to the 8.5%, have at least one hashtag. The percentage of users using at least a hashtag is higher, around 20%. Figure 3.2 shows that the number of users per tag follows a heavy tailed distribution, with some outliers tags used by hundreds of thousands of users. Both the distribution of the number of messages and of distinct tags tweeted by each user also follow a heavy tailed distribution, with a few extremely active users, tweeting up to 10 thousand messages or one thousand distinct tags in a month. The total number of distinct tags encountered is over 2 millions; however, only about 93 thousands, corresponding to 4.14%, appeared in more than 20 messages over the whole month: for our study, we considered only these tags, and discarded all the others.

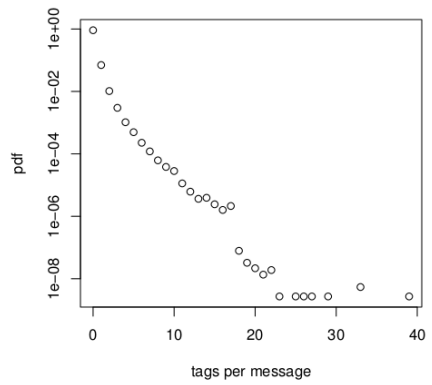


Figure 3.1: Representation of the proportion of messages having a given number of hashtags, on a logarithmic scale.

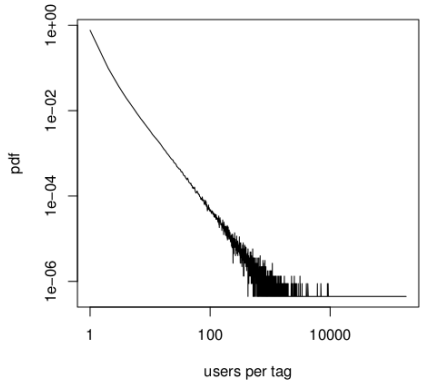


Figure 3.2: Distribution of the proportion of users using a hashtag, on a log-log scale.

3.3.3 Evaluating hashtags

In this Section we will illustrate some results obtained by applying the metrics described in Section 3.2 to evaluate hashtags contained in our dataset.

Frequency of usage

A first interesting question about hashtags is whether the corresponding non-tags also appear; about 73.5% of hashtags have the corresponding non-tag appearing at least once in our dataset. Among these, 57.8% are more frequent as hashtags than as non-tags. A "map" representing the frequency F_m of each hashtag in function of the frequency $\bar{F}_m$ of the corresponding non-tag in shown in Figure 3.3. The graphic exhibits a *glove* shape, which seems to point out the distinction between two kinds of tags: those corresponding to common words, that appear only sometimes preceded by a hash, and those on the "thumb", Twitter specific tags which are more often used with hash, and do usually not correspond to any commonly used word. Examples of this second kind of tags are `#tagtuesday`, `#iranelection`, `#sextips` and `#tcot` (acronym for "top conservatives on Twitter"). We obtained a very similar shape for user

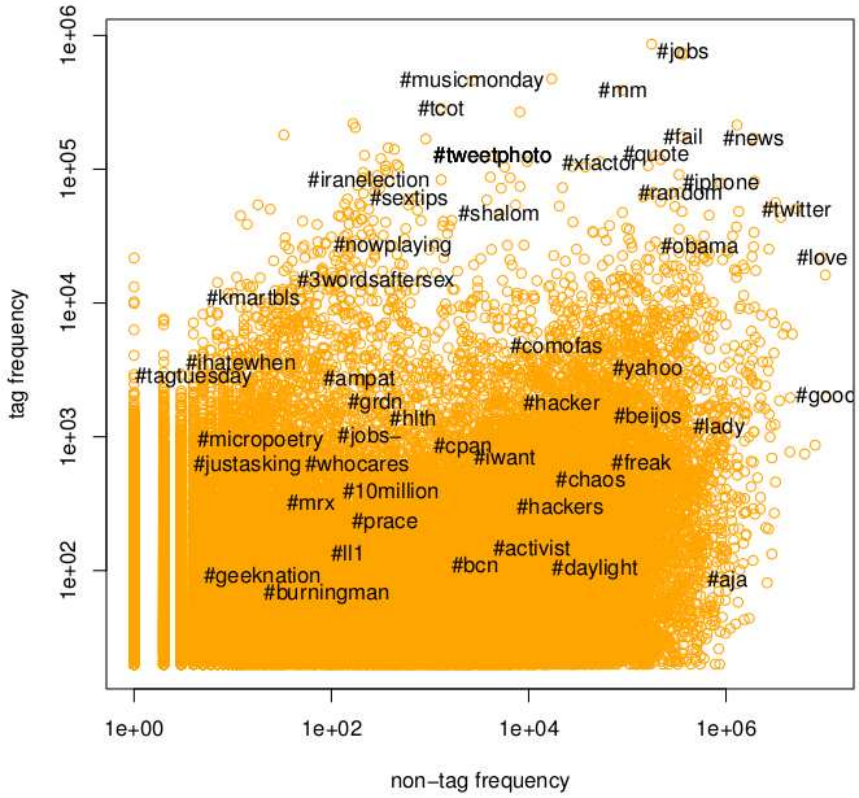


Figure 3.3: Relationship between the frequency of each hashtag and the frequency of the corresponding word with no hash.

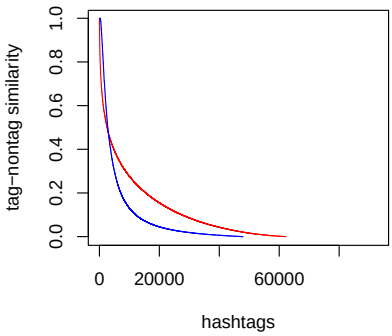


Figure 3.4: Similarities $wsim$ (red) and $usim$ (blue), in descending order.

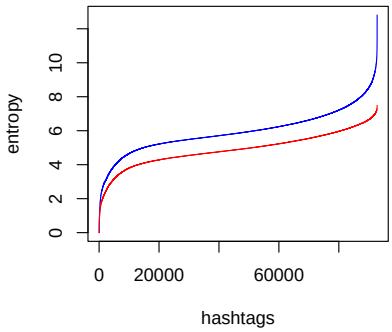


Figure 3.5: Entropies H (red) and Hb (blue) of tags, in ascending order.

frequencies F_u and $\bar{F}_u$.

Specificity

Figure 3.4 shows the similarity between tags and the corresponding non-tags, both in terms of co-occurrence vectors and of users. About a half of tags have null values of $usim$, meaning no user in common with the corresponding non-tag, while $wsim$ is null for about one third of tags; while considering this second result, it must be taken into account the fact that we have cut all values of tf-idf below a threshold of 0.01.

Among tags having the highest values of $wsim$ we find for example `#daylight`, almost always used in the context of "daylight savings", `#lady`, mostly referred to the singer Lady Gaga both as a tag and as a non-tag, and `#comofaz`, which is Portuguese slang word for "How do I do?" Among those having null or very low similarity we find tags like `#tweetphoto`, mainly found in messages generated by an application, and `#li`, that corresponds to a common word in several languages, like Portuguese, Italian and Chinese, but as a hashtag is mainly used to refer to the social network platform LinkedIn.

Figures 3.6 and 3.7 plot the relationship of similarity $wsim$ to tag and non-tag frequency, respectively. Apart from a tendency of very frequent tags to have a lower similarity, no precise relationship can be detected between $wsim$ and F_m. On the other hand, high values of similarity seem

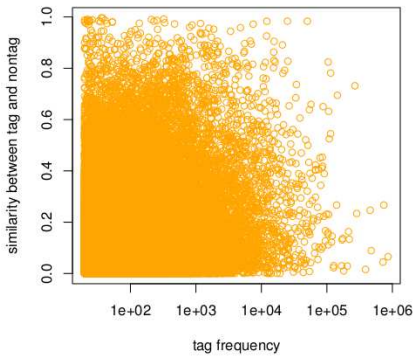


Figure 3.6: Similarity between each tag and the corresponding non-tag, in function of tag frequency.

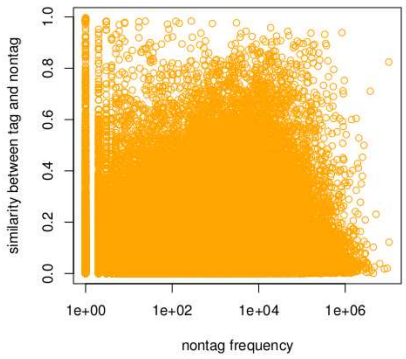


Figure 3.7: Similarity between each tag and the corresponding non-tag, in function of non-tag frequency.

to be more likely for tags corresponding to words having a frequency in the order of a few thousands, with a peak around 8000.

Consistency of usage

In Figure 3.5 we plotted the entropies of tags, in descending order. Most of the tags have values of H lying in the range between 4 and 6; entropy based on bigram co-occurrence tends to be higher, with values ranging mostly between 5 and 7.

Among tags having very high entropy we find especially tags expressing sentiments, like `#whocares`, `#argh`, `#_#`, beyond some words used in a variety of contexts, like `#freak`. Tags with a very low entropy are typically generated by applications, like `#dongdongdong` (a tweeting church), `#tweetphoto` or `#iphonebabes`.

3.3.4 Stability over time

While until here we have studied tags as static entities for the whole period of observation, in this Section we will illustrate some results based on the observation of tags over different days.

As an example, we report some statistics observed for tags appearing on November 10th, 2009; to identify new tags we based on a temporal

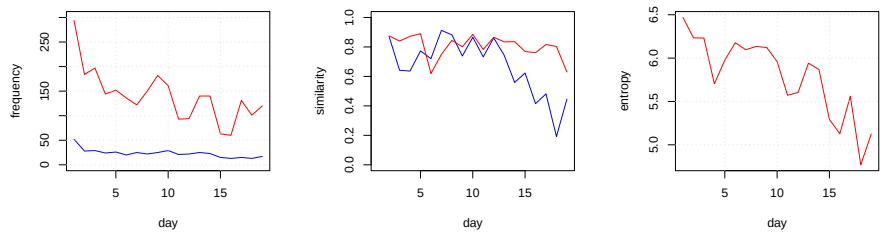


Figure 3.8: Evolution of tag `#ampat` by day (November 12th-30th). Left: Frequency F_m (red) and F_u(blue). Center: Values of $wsim_d$(red) and $usim_d$(blue). Right: Entropy H.

window of $k = 9$ days. The total number of distinct hashtags observed on November 10th is over 160 thousands, about 50% of which were not appearing in any of the 9 previous days. We looked for these *new tags* in the messages from the 9 following days to evaluate their longevity l. Most of the tags have $l = 0$ and only 36 tags (about 0.045%) appear in all days until November 19th. This is an interesting indicator of the short memory of Twitter, and of how off-handedly users do often add hashes to words.

In this way, we have selected for each day very few new tags, that are potentially new trending topics; we can now illustrate the results obtained by applying the measures defined in Section 3.2.5 to two of these tags, to characterize them.

Tag `#ampat` stands for "American patriot", and seems to have been adopted by a well defined community. Frequency of messages and users (Figure 3.8) exhibit a slow decreasing trend, after starting with about 300 messages in the first day, tweeted by 50 users; entropy tends to decrease in time (Figure 3.3.4) pointing out a convergence towards some context; both the meaning and the community behind the tag seem to be quite stable, though users tend to differentiate a bit in the last observed days (Figure 3.3.4).

`#kmartbls` stands for Kmart's blue light special offers; the extremely high similarity between consecutive days in terms of co-occurrences (Figure 3.3.4), together with the very low entropy (Figure 3.3.4), is a signal of the scarce variety of information carried by the messages; these data, contrasted with the very high frequency (Figure 3.9), can easily bring to the conclusion that the tag has been massively promoted by some au-

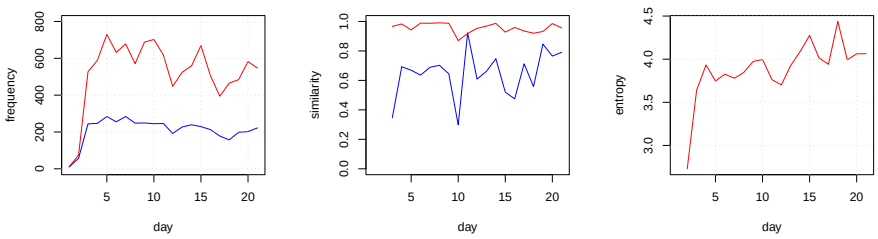


Figure 3.9: Evolution of tag `#kmartbls` by day (November 10th-30th). Left: Frequency F_m (red) and F_u(blue). Center: Values of $wsim_d$(red) and $usim_d$(blue). Right: Entropy H.

tomatic application, retweeting almost identical messages from different accounts.

3.3.5 Manual assessment

In order to assess how well our metrics are able to indicate which hashtags represent stable concepts with a unique identity, we have performed a manual evaluation on a random sample of 257 hashtags. For each tag, we collected a random sample of 100 messages with that hashtag, and asked our evaluators to answer the following questions:

1. whether they could guess the meaning of the tag just by looking at it;
2. whether the hashtag represented:
   - an event, person, organization, product, or other named entity;
   - messages generated by an application (e.g. spam);
   - messages with a common sentiment;
   - other;
   - not clear;
3. whether the tag referred to the same meaning in all messages or not.

Furthermore, the evaluators were asked to choose the closest matching concept from Freebase[3], by means of the Freebase Suggest tool[4].

In roughly 39% of cases, the messages were found to refer to a named entity; for 20% of the tags the messages were characterized by a common sentiment (e.g. `#thankfulfor`, `#grrr` or `#youknowyouareuglyif`), while 12% of times they were recognized as generated automatically by some application (e.g. `#soundcloud`, an audio distribution platform that relies on Twitter to spread notifications about users' activities, or `#shop`, massively used by spammers). In 26% of the cases, the hashtag did not represent a named entity, a sentiment or an application, but was created for some other reason, typically to discuss a general topic (e.g. `#tv`, `#politics`, `#immigration`). The meaning of the tag remained unclear in 6.7% of the cases. Among named entities, organizations were the most common (27%), followed by products, events, persons and other entities (16%, 12%, 6%, 29%).

Slightly more than half of the tags (137 out of 257) could be associated to a Freebase entry; this is higher than the number of named entities because Freebase contains also some general terms, like domains or common words, which are not named entities. As expected, most application and sentiment tags could not be mapped to Freebase. Only 33% of application and 14% of sentiment tags could be resolved, and many of these mappings are rough approximations of the intended meaning (e.g. the protest tag `#freegary` mapped to `gary_mckinnon`). We have also explicitly measured agreement on this task by reevaluating 31 judgments. 18 out of the 31 tags in this sample could be mapped to Freebase. The inter-annotator agreement on the task of determining if a hashtag can be mapped to Freebase is very high (Cohen's κ of 0.79). The judges agreed on the exact target in 12 out of 18 cases, and 4 of the 6 instances of disagreements were simply due to the same topic appearing in multiple hierarchies within Freebase. One of the other two cases was a close match (`technician` vs `technology` for the tag `#tech`), the other a broader match (`bacon` vs `food` for `#bacon`).

Using the whole set of judgements, we have also performed a logistic regression on the binary variable indicating whether there was a mapping to Freebase for a given hashtag. We have normalized the input variables by a linear transformation to the [0,1] interval, so that we obtain coefficients that are comparable in magnitude. Table 3.2 shows the coefficients of the resulting model. This model shows that tag frequency,

[3] `http://freebase.com`

[4] `http://code.google.com/p/freebase-suggest/`

Variable		Coefficient
tag frequency (#messages)	F_m	-2.00
nontag frequency (#messages)	$\bar{F}_m$	-3.45
tag frequency (#users)	F_u	-6.80
nontag frequency (#users)	$\bar{F}_u$	5.45
tag entropy (bigrams)	Hb	3.56
tag entropy (unigrams)	H	-3.68
nontag entropy (unigrams)	$\bar{H}$	0.11
word similarity	$wsim$	0.78
user similarity	$usim$	0.34
Intercept		-0.01

Table 3.2: Logistic regression coefficients of the input variables reported, for predicting output variable FBID (i.e., whether a hashtag can be mapped onto a Freebase entry).

non-tag frequency, the number of users are negatively correlated with the success of mapping to Freebase, because these frequency measures are indicators of Twitter-specific usage. Entropy is also negatively correlated, because the higher the entropy, the less consistently the tag is used. The number of non-tag users is positively correlated, because it indicates common words/sentiments. Similarities are also positively correlated, but to a smaller extent. Altogether our model achieves a 66% accuracy, a relative improvement of 25% over the baseline of choosing the majority class.

3.4 Related work

This work is strongly related to tagging in social bookmarking applications, which has a longer history; tag semantics in this context has been largely investigated over the last years. Although this context will be explored in detail in Chapters 4 and 5, in the next section we mention some studies on social tagging which are related to the work described in this chapter, before considering specific work in the field of microblogging.

3.4.1 Tag semantics in social bookmarking literature

The main difference with respect to microblogging is that tagging in social bookmarking is explicit and often serves personal categorization. Classifications of tags based on their usage are proposed in [56] and [161];

an insight into the use of non subject related tags is offered in [89]. Motivations and incentives behind tagging have been investigated in [116] and [7]. In [47] some metrics are introduced to evaluate tags, based on user behaviour. Al-Khalifa et al. [4] evaluate the potential of folksonomies to generate semantic metadata; an assessment of delicious tag vocabulary efficiency from an information theory perspective is provided in [33].

Among the studies aiming at extracting emergent semantics from folksonomy, the work described in [185] relies on a metric of tag entropy to evaluate the ambiguity of tags. In [31] some measures to compute tag relatedness are presented, and delicious tags are grounded to WordNet synsets in order to contrast semantic relations with the results of the different metrics proposed; the best semantic precision is achieved with metrics based on the cosine between each tag's context, represented as a vector of co-occurring tags. Also the study described in [17] resonates with our work for the use of information retrieval techniques to compare tags with each other: the authors build a tag-tag space based on the cosine between the co-occurrence vectors of tags, and find out significant differences in the usage of the same keywords in different tagging systems.

In [103] Körner et al. introduce the distinction between two classes of users according to their tagging behaviour and motivations: *categorizers* and *describers*. While the latter are interested in sharing, and tend to accurately choose tags in order to help other users find the resources they tag, the former are especially interested in categorizing their own stuff and so produce tags which are often hardly useful for other users. In [102] it is shown how awareness of this distinction can improve the effectiveness of algorithms for emergent semantics extraction from folksonomies.

The idea of integrating tags into the semantic Web is not new; among other works, FLOR is a framework for the enrichment of folksonomies with semantic information from existing ontologies [9], while TagOnto offers a set of techniques and heuristics to map social application tags to ontology concepts [20]. The approach of integrating tags with existing ontologies from the Semantic Web is also the basis of next chapter, where tags from a social bookmarking application are mapped onto WordNet.

Models have also been proposed to allow users to anchor tags to semantic Web URIs, such as MOAT [139] and CommonTag[5]; NiceTag ontology, described in Chapter 5, enriches this approach allowing for the representation of different kinds of tagging actions, by means of named

[5] http://commontag.org

graphs.

3.4.2 Tags in microblogging

Letierce et al. [110] investigate the use of Twitter during conferences, identifying classes of hashtags and finding out a prevalence of technical terms, and a general tendency to address especially people belonging to the same community. In [81] tagging behaviour in Twitter is compared with the one in delicious, and it is described as *conversational*; the authors in particular study the phenomenon of memes emerging around hashtags that are often abandoned after a short time, and introduce statistical metrics to detect them. A tripartite model of users, hashtags and messages is introduced in [179] to turn Twitter into a folksonomy, and to extract emergent semantics.

An alternative distributed platform for microblogging, based on semantic Web principles, is described in [138]. Special syntaxes have been proposed to allow users express structured information inside a tweet; among these we mention twitlogic [162] and HyperTwitter [74], which allows users to specify relationships among hashtags (equivalent, subtag) and express arbitrary properties between them. According to table 2.1, while we have focused on the basic level of names (bottom left cell), these works aim at introducing higher levels of semantics in microblogging systems, and correspond to upper cells.

3.5 Conclusions and future work

Since their introduction, hashtags have shown to be a successful feature of microblogging platforms, and a precious concrete solution to the problem of aggregating content in the disorganized and fragmented impetuous stream of information that characterizes these systems. However, not all hashtags are used in the same way, not all of them aggregate messages around a community or a topic, not all of them endure in time, and not all of them have an actual meaning. In this work we have addressed the issue of evaluating Twitter hashtags as strong identifiers, as a first step in order to bridge the gap between Twitter and the Semantic Web.

The first contribution of the work presented in this chapter stands in the formalization of the problem, and in the elaboration of a number of desired properties for a good hashtag to serve as a URI. We have proposed a Vector Space Model for hashtags, representing them as virtual documents; in parallel we have introduced the notion of *non-tag*, to be able to compare each tag with the corresponding word without hash.

We have defined several metrics, based both on the messages containing a hashtag and on the community adopting it, to characterize hashtag usage on a variety of dimensions: *frequency*, *specificity*, *consistency*, and *stability* over time. We have applied these metrics to a dataset of more than half a billion messages, collected over the whole month of November 2009. Beyond qualitatively illustrating the results, showing how the metrics proposed tend to correspond to actual properties of the data, we have performed manual classification of a sample of tags. Based on these data, we have tested the results obtained with the algorithms described, showing how a combination of the proposed measures can help in the task of assessing which tags are more likely to represent valuable identifiers. These results are promising, with respect to the perspective of anchoring Twitter hashtags to Semantic Web URIs, and to detect concepts and entities valuable to be treated as new identifiers. Also spam detection tasks can benefit from the metrics we have illustrated.

As a further step, this approach could be easily used to study similarity between hashtags, based both on word and user co-occurrence vectors, in order to find clusters and study emergent semantics. Also the possibility of comparing *good* hashtags with text documents is straightforward; in particular, the vectorial representations of hashtags which we have introduced in this chapter could be compared with the words appearing in each Wikipedia article. In this way it could be possible to automatically map hashtags to Wikipedia entries and so to DBPedia entities, assigning them a URI.

4 Integration of ontology hierarchies into folksonomies

4.1 Introduction

As the amount of information available on the Web grows every day faster, the task of classification is getting harder, the traditional top down approach is getting inadequate [163] and the new bottom up approach of *folksonomies* is emerging [146].

The work of categorization in folksonomies is performed by users so, as explained in Section 2.2.4, they are scalable, current, inclusive and democratic, and they have a very low cost. On the other hand, folksonomies as resulting from current tagging interfaces are characterized by many limitations, such as low performances in terms of both precision and recall, the lack of explicit semantics and the possibility of *gaming* [100, 65].

In this chapter we focus on the lack of explicit semantics in tagging applications, and in particular on the absence of hierarchy, which makes it difficult to browse a folksonomy or of a part of it, and to retrieve resources related to a topic without having to inspect several scattered labels.

As tags are just text strings, with no explicit semantics associated, it is not trivial to organize them for presentation to the user. The most common way to show a set of tags are tag clouds, visual representations where each tag is displayed with font size depending on its popularity, as shown in Figure 2.3. This kind of representation is good for getting a quick summary of the content of a folksonomy, and as a starting point for exploration, however it has been shown to be insufficient for information-seeking tasks [164]. To allow a better navigation for discovering of interesting and related items, many applications have introduced links to *related tags*, where relatedness is generally measured with some metrics based on co-occurrence data. For example in Del.icio.us, when visiting the page of bookmarks tagged with a certain tag, a list of related tags is presented in a sidebar.

These features are useful but present several limitations. First, they leave the problem of the lack of hierarchy unsolved: they build flat spaces of tags, and they provide no criterion to organize them, so only a small set of items can be displayed. There is no explicit connection with the meaning of keywords or semantic relationships among them; in tag clouds tags are usually disposed in alphabetical order, while in related tags panes they are ordered according to relevance. In both cases there is no coherent criterion related to the meaning of tags, which can help the users to browse a set of keywords.

Our purpose is to enrich the possibilities of navigation in a folksonomy by adding explicit semantics, provided by an external hierarchy of concepts, to help users orient themselves among keywords. We chose to start from Del.icio.us, a popular folksonomy for social bookmarking, and to develop an alternative tool for the suggestion of related tags, based on the WordNet hierarchy of concepts. To this end, the first necessary step is to disambiguate the meaning of tags and to map them into elements from the ontology; then a hybrid hierarchy borrowed from WordNet and containing the tags of interest can be elaborated and integrated into the navigation interface.

This chapter, based on the work published in [104] and [105], is organized as follows. In the next section we describe both the design and the implementation of the project, then in Section 4.3 we show some results of our tests and an evaluation of the application. In Section 4.4 we discuss related work, and in Section 4.5 we conclude with a summary and a discussion of future work.

4.2 Turning a tag-space into a hierarchy of concepts

The goal of this work is to investigate the possibility of integrating an ontology in the navigation interface of a folksonomy, filtering tags through a predefined semantic hierarchy to improve the possibilities of searching and browsing. In particular we chose to improve the *related tags* panel in Del.icio.us; filtering a set of related tags through WordNet noun hierarchy it is possible to display a much higher number of them, organized according to a semantic criterion. As WordNet is a semantic lexicon of English, developed to reflect the semantics of natural language and the way in which humans classify objects, the relations and categories that it contains are likely to be immediately understood by most people [48].

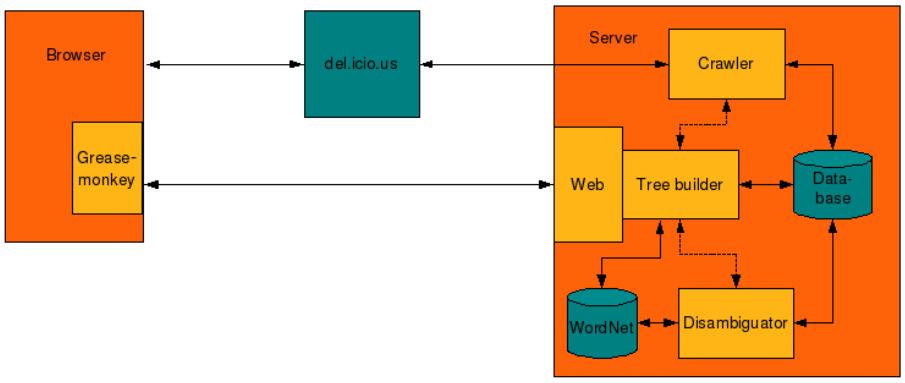


Figure 4.1: System architecture

4.2.1 System architecture

The application we have developed is based on a client-server paradigm, where all the tasks relative to the processing and storing of information are left to the server and the client has only to manage the visualization of results. The architecture of the system is shown in Figure 4.1.

The server is composed of a *scraper*, that extracts the data from Del.icio.us HTML pages and stores them on a database, a module for *tag disambiguation* and a core module that builds the *semantic tree* of tags related to a given one, based on the hierarchy of concepts of WordNet. On the client side, according to the principle of *active navigation*, a JavaScript script executed inside the browser dynamically modifies the pages visualized by the user, integrating the additional information provided by the server.

4.2.2 Mapping Delicious tags onto WordNet

The first issue when trying to map tags to WordNet is the one of tags that are not recognizable as words in the lexicon, even after a stemming process, and therefore cannot be mapped. To evaluate the relevance of the excluded data we have collected a large dataset, relative to about 30,000 Del.icio.us users and containing about 480,000 different tags. Studying these data we found that only about 8% of the different tags used are contained in the lexicon, but we also observed that the most popular tags are much more likely to belong to WordNet. This distribution in

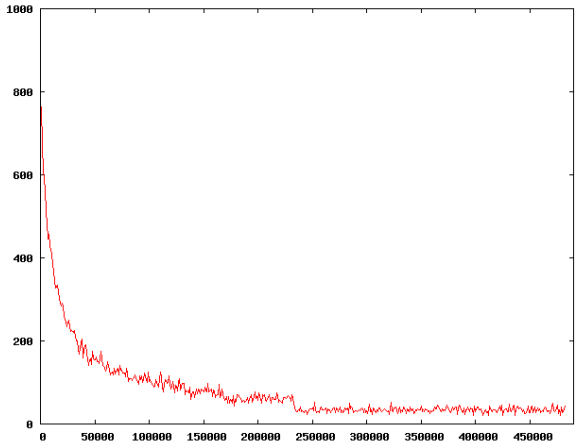


Figure 4.2: Along the X axis are represented tags from our dataset, grouped by 1000 and ordered by decreasing popularity; the Y axis shows the number of tags belonging to WordNet for each group of tags. The most popular tags are much more likely to belong to WordNet, following a power law distribution.

particular follows a power-law curve, very common in the field of collaborative systems, as showed in Figure 4.2. Of the 20 million total tagging relations present in our dataset, about 68.1% involve words contained in WordNet. We think this data might be much increased by using local wordnets in other languages and domain ontologies to cover more specific terms.

There is then the issue of words that are recognized as belonging to the lexicon, but not as nouns: these tags too cannot be mapped, as the hierarchy of WordNet is only defined on nouns. According to the distinction formulated in [56] among *factual*, *subjective* and *personal* tags, we can argue that factual tags tend to correspond to nouns, as nouns fit better to describe factual knowledge, while adjectives tend to correspond to subjective tags. Further studies about this issue can be found in [5]. From a quantitative point of view, our dataset confirms the intuition that most of the tags, and especially most of the most popular tags, are nouns. Indeed, the 85% of the different tags recognized by WordNet are nouns, while out of the over 20 million total tagging relations, about 64.9% involve WordNet nouns, and just about 3% involve words belonging to the lexicon without being nouns; in other words these data tell that, in our dataset, about 95% of the times that a tag belonging to WordNet is used it has at least one meaning as a noun: the power law

distribution is accentuated for nouns.

4.2.3 Tag disambiguation

One problem when trying to map tags on an ontology is polysemy: as no explicit semantics is associated to tags by the users, the same tag can have different meanings according to different acceptation of the word, and consequently different positions in the ontology. For example the word "turkey" may refer to the country or to the animal, and in the second case one could want to distinguish between biological and gastronomic meaning, according to the context. In WordNet semantic relationships are not defined among words, but among *synsets*, groups of synonyms that represent units of meaning; each word can belong to different synsets according to its different acceptations. The word "turkey", for example, belongs to five synsets, the first one being "turkey, Meleagris gallopavo" and the second "Turkey, Republic of Turkey" .

To properly map a tag to the corresponding position in the ontology you need first to disambiguate it, in relation with the context in which it has been used. A fair solution naturally offered by a folksonomy is to use the other tags associated by some users to the same resource as the context for disambiguation.

Our algorithm for tag disambiguation acts for each tagged resource in the following way: the C most used tags for the resource are compared among them, and for each of them the meaning that is more strictly related to the other tags is selected; semantic relatedness among tags is calculated according to a choice of metrics based on WordNet [142] (adapted lesk, Hirst and St. Onge) and disambiguation is performed using the Perl library SenseRelate [141]. In the same way the remaining tags are disambiguated using the first C as a context. This solution is effective, as it reduces the sensitivity to less used tags, and efficient, as it avoids the exponential growth of the algorithm complexity with the number of different tags associated with a resource.

4.2.4 Building the tag semantic tree

The core module, for the construction of the tree of related tags, acts in four steps: *tree building*, *compression*, *branch sorting* and result output. All the algorithms developed have linear complexity with the number of input tags.

The set of tags to be considered is selected by collecting, for each of the latest N sites associated with the given tag, the M most frequent

tags for that site; M and N are parameters that can be specified in the HTTP request. The construction of the tree is performed by an iterative algorithm; for each different tag present in the set of interest in a particular acceptation, the chain of the hypernyms is created as a path till the unique root of the noun hierarchy of WordNet and then merged with the existing tree. At the end of this process the tree is a subpart of WordNet noun hierarchy, chosen to contain all the tags of the set of interest.

As WordNet is very fine-grained, it can take more than 10 steps to descend from the root to a word; the tree has to be compressed to be useful for navigation, eliminating the useless nodes. The compression algorithm performs a breadth-first visit of the tree, in which all nodes considered unnecessary are deleted and replaced by their children. On one hand, all the nodes corresponding to high level categories in WordNet, contained in a black list, are deleted; the information content of these nodes is generally too low to be useful for navigation. On the other hand all the nodes that do not correspond to any tag and have a branching factor lower than K or have no siblings are replaced by their children. The default value for K is 2; in this way the structure of the hierarchy is preserved and at the same time the most specific terms can ascend in the tree.

The branches are ordered by weight, where the weight of a node is calculated as the number of resources in the set of interest that have been tagged with the corresponding word in that acceptation. This guarantees that the branches of the hierarchy that are most strictly related to the given tag are shown first to the user. As a last step, the tree is output by the server in HTML or XML format.

4.2.5 User interface

The system rests on Firefox Browser and Greasemonkey extension to execute some JavaScript code inside the browser. When the user is visiting the Delicious page for a certain tag, the script connects to our server to get the semantic tree of related keywords for that tag; as soon as the information is ready, a new sidebar is dynamically integrated into the page, showing an expandable tree. For each node of the hierarchy there are two links, directed one to the Del.icio.us page for that tag and one to the page of the resources tagged both with that tag and with the given one; the size of each tag's intersection with the current keyword is shown in parenthesis and represents an indicative measure of relatedness for the users. Tooltips guide users showing WordNet definitions of the

concepts corresponding to each node and indicating the destinations of links.

Figure 4.3: A screenshot from the Del.icio.us page for tag "pasta", where the inner sidebar shows an expandable hierarchy of related tags, provided by our application.

Figure 4.3 shows the result obtained for tag "pasta", where all the tags associated to the latest 300 sites tagged with "pasta" are displayed; in the picture you can see the first branches (i.e. the most related ones, in this case those about "food"), that have been expanded.

4.3 Tests and evaluation

We tested the system with different kinds of tags, according to different dimensions. The first dimension is the specificity of the tag from which the exploration starts; it's very different to display the space of a keyword situated in a specific domain or in a generic one. In the first case the resulting tree tends to be compact and to allow easier navigation, while in the second case it tends to have a high branching factor and a high

number of first level nodes; anyway, as the branches are always ordered by weight, the most interesting concepts in relation to the given one are reachable exploring the first branches, also in case of very general keywords. The second dimension is given by the popularity of a tag, while the third one is given by the semantic field; each semantic field has its specificity and some of them rest on more conventional and ordered sets of words, such as the "food" context, visible in Figure 4.3, while some others are more prone to slang and neologisms, such as the one of "software".

Figure 4.4 shows the result obtained for tag "blog"; as "blog" often refers to a kind of site more than to the content, it can be considered a particular case, and a very general tag as there are blogs almost about everything. "Blog" is also one of the most popular tags in Del.icio.us, so it is an extreme case also according to the second dimension. We obtained this result considering the latest 2000 Del.icio.us bookmarks tagged "blog", and only the 15 more used tags for each of them, to cut the *long tail* of less used tags. In the picture you can see expanded the hierarchy of scientific disciplines.

From this and other examples the main problem that emerges for scalability seams to be the high number of nodes in the first level of the tree; some improvements could be obtained by making the tree compression algorithm more dynamic.

Confronting the related tags suggested by Del.icio.us with the results we obtained, we observed that they are always somewhere in the first branches in the new sidebar. An exception must obviously be done for the words that do not belong to WordNet, that are absent in the new sidebar. Experimenting for example with tag "Greasemonkey" (the experiment is possible though the word itself is not contained in the lexicon) we found that many important related tags, like "JavaScript", are not recognized, while other important words, such as "extension", are interpreted in a wrong way as WordNet does not contain the acceptation related to software; all the tags for which there is in WordNet an acceptation related to software have instead been correctly interpreted by the system. These limitations could be addressed by resting on some domain ontologies to integrate WordNet and on Wikipedia for reconducting slang forms to more conventional ones (for example, Wikipedia recognizes "nyc" as an alternative form for "New York City", while WordNet does not).

In many cases synonyms or just different ways of spelling a word happen to be close to each other and easily recognizable in the tree provided by the new sidebar: the semantic hierarchy helps to face the problem of

Figure 4.4: A screenshot from the Del.icio.us page for tag “blog”, where the inner sidebar shows an expandable hierarchy of related tags, provided by our application.

synonym control to which a folksonomy is naturally prone.

As a last consideration we want to mention the problem of gaming. It's not unusual in Del.icio.us to see the related tags sidebar entirely mucked up by spam, as we found in some of our examples. Gamers can trick Del.icio.us to gain a good position for the tags they want to advertise and, as there are just a dozen tags suggested, the whole sidebar can easily be compromised. In the new sidebar the problem is embanked as a much higher number of tags is shown and so the presence of some spam tags does not make the whole suggestion system useless, though the order of branches can be gamed.

4.4 Related work

Other researchers have proposed to rely on ontologies, lexical resources and other sources of structured knowledge to address some limitations of tagging systems. TagOnto is a *folksonomy aggregator*, which combines information from different folksonomies mapping tags onto ontology [20]. Angeletou et al. [9] propose FLOR, a framework to semantically enrich folksonomies, based on dynamically retrieving, selecting and combining relevant knowledge from ontologies. Van Damme et al. [38] present Folksontology, another approach based on leveraging knowledge from ontologies and lexical resources to enrich folksonomies; in particular, they rely on Wikipedia entries and disambiguation pages to identify entities. They also propose mechanisms to involve the community in the process of generating hierarchies, through visualization and voting on conceptual choices.

FaceTag, proposed by Quintarelli et al. [147], is a framework aimed to integrate a top down classification paradigm with folksonomies; the system relies on both implicit and explicit semantics to organize tags in a taxonomy: implicit semantics is obtained by mining tag co-occurrence, while explicit semantics is provided by the users who can specify several kinds of relations between tags. The approach is similar to ours in the proposal of an enrichment of the browsing interface, but complementary as no knowledge from external sources is used. Another more recent work aimed at improving browsing in folksonomies by adding hierarchical relationships between tags is presented in [120], where several techniques to automatically infer the semantics of tags are proposed and evaluated.

An alternative approach to achieve better visualization and browsing interfaces is based on clustering of tags according to their usage [16]; in Flickr, this feature has been integrated also for the presentation of

sets of related tags[1]. Hassan et al. [69] proposed to improve tag clouds, presenting tags organized in clusters.

Many researchers have faced the challenge to automatically derive ontologies by automatically mining the tripartite graph of tags, resources and users [122, 75, 157, 25, 185]. Cattuto et al. [31] map tags onto WordNet synsets to contrast semantic similarity with several measures of tag relatedness. The semantic grounding of tags is leveraged in [115] to evaluation different metrics; the best semantic precision is achieved with metrics based on the cosine between each tag's context, represented as a vector of co-occurring tags. This seems to confirm the validity of our choice to leverage co-occurring tags as the context for tag disambiguation.

4.5 Conclusions

In this chapter we have faced the problems related to the lack of explicit semantics, and in particular of hierarchy, in tagging interfaces. We have proposed a new approach to enrich the navigation interface of a folksonomy adding structured knowledge provided by an ontology, and we have developed a tool that uses WordNet to build a semantic hierarchy of tags which helps users navigate and find related resources in Del.icio.us.

We have shown that in this way it is possible to combine some of the advantages of the traditional top down approach to classification with the ones of the bottom up paradigm that is emerging on the Web, providing richer possibilities of searching and browsing, and dealing with some of the limitations to which folksonomies are prone, such as lack of recall, synonym control and gaming.

As future work, it would be interesting to use the results of tag disambiguation, performed by our application, to filter resources and not only tags; in this way it might be possible for example to visualize, among the Del.icio.us bookmarks associated to the tag "turkey", only the ones that have been individuated as related to the geographical acceptation.

One strength of the solution proposed in this chapter is that it does not require any additional effort for the users, nor any change in the tagging interface: it can already be integrated on top of an existing folksonomy to improve its navigational interface; on the downside, relying on external knowledge is also the main potential limitation of this kind of approach, as the fixed and static hierarchy of concepts provided by WordNet can in many cases not reflect the most suitable categorization criteria for

[1] See `http://blog.flickr.net/en/2005/08/01/the-new-new-things/`

the mind of the users. An alternative approach, which can foster participation of the users for the specification of knowledge in structured format, is provided in the next chapter, while the possibilities and the challenges offered by collaboratively created hierarchies are investigated in Chapter 6.

5 Modeling tags as named graphs: NiceTag ontology

5.1 Introduction

Tags are one of the pillars of the Social Web. Users associate labels to resources, often for their own benefit, and this little individual effort is converted into a value for the whole community. The aggregation of tags produced by different users gives place to *folksonomies*, collective classifications of content. This paradigm has emerged in many different contexts, and people use tags for a variety of purposes and in many different ways.

The simplicity of current interfaces, where users have only to type a string of characters to assign a tag, has been a key feature for the success of social tagging applications. However, as shown in previous chapters, this simplicity and the consequent lack of explicit semantics have drawbacks for the quality of the results.

In Chapter 3 we have investigated the quality of user generated labels as identifiers and the possibility of associating them with real world entities, focusing on the scenario of hashtags in microblogging, while in Chapter 4 we have mapped tags from a social bookmarking site to the WordNet ontology and we have proposed the use of hierarchies from WordNet to organize tags and present them to the users according to a coherent semantic criterion. In this chapter we propose a general model to represent tags in a fine-grained and flexible way, thanks to the use of named graphs which allow to specify the relation intercurring between the tagged resource and the sign used to tag. The model has been drafted at the VoCamp in Nice 2009[1], and presented in [113], [125] and [124]. It is available online at the address `http://ns.inria.fr/nicetag/2010/09/09/voc.html`.

Current models of tagging allow one to link a tag to a well defined meaning; this relationship helps to face the problem of the different ac-

[1] http://vocamp.org/wiki/VoCampNiceSeptember2009

ceptations a term can have in different contexts and for different communities [140]. Still, polysemy is not the only ambiguity of tags: some meaning resides in the (so far implicit) kind of relationship between the resource and the sign used to tag. For example, the use of tag "blog", one of the most popular in Delicious, can assume at least two different meanings with respect to the same definition of the word "blog": it can mean that a resource *is about* blogs, or that it *is a* blog. As another example, if I use tag "thesis", it could mean that a resource is a thesis, or that it is relevant for my thesis. Moreover, some tags are intended for *personal* use and to only make sense for the applier, while other are used for recommending something to other specific users, and so on. In contrast, current interfaces do not offer the possibility to specify different usages of tags, and existing ontologies for tagging reflect this simplicity providing one single property to link a sign to the resource being tagged: `taggedResource` in Newman's Ontology [133] and SCOT [87], `tagged` in Common Tag[2], `hasTag` in NAO [156].

The demand for richer expressivity in tagging applications is witnessed by the experience of *machine tags*, born from spontaneous conventions among users and then integrated in the interface of some systems such as Flickr[3]. In Flickr, like in most current social annotation platforms, one is only allowed to describe a resource by associating to it one or more free character sequences, in unstructured format. In order to express well defined properties of the annotated pictures, users adopted a convention to assign a special syntax to the blank space of tags. For example, the geo-location of a picture taken in Taipei can be specified using tags like "geo:lat=25.033333" and "geo:long=121.633333", while the price of a depicted bicycle for sale can be expressed as "sell:price=100$". The syntax of Flickr machine tags resonates with RDF, as it is based on properties (e.g., "`geo:lat`", "`sell:price`") belonging to some kind of namespace ("`geo:`", "`sell:`"). This convention allows for automatic elaboration of specific properties of annotated resources and constitutes an interesting community solution, but it is subject to the limitations of not relying on Semantic Web standards: there is no formal definition of the meaning of `geo:lat`, it is just shared knowledge within a group of Flickr users; no interoperability is guaranteed and apposite tools have to be developed to process information.

The model we present in this chapter is aimed at offering a more general and complete solution, leveraging Semantic Web standards and tech-

[2] http://commontag.org
[3] http://www.flickr.com/groups/api/discuss/72157594497877875/

nologies to enrich current tagging interfaces. In NiceTag, we represent a tag as an RDF relation between a resource and a sign; we propose several properties to model different usages of tags according to literature. To model the tagged resource we rely on the IRW resource ontology [64], which allows for a fine-grained specification of the different strata of resources encountered on the Web, while signs can be modeled according to existing vocabularies for tagging.

The link between a tagged resource and a sign, which constitutes the essence of a tagging action, is embedded in a *named graph*; as the declaration of named graphs is not natively supported in RDF, we have integrated the model from Carroll et al. [29] and the RDF/XML Source declaration syntax from [53].

Thanks to the use of named graphs a tag action, identified by a URI, can also be typed itself, and this allows to distinguish different kinds of tagging and different corresponding social acts; it is also possible to express properties of the tag action, and in particular metadata such as author, date and Web container.

The chapter is organized as follows. Section 5.2 offers an overview on related work. In Section 5.3 we introduce the core of NiceTag model, i.e. the use of named graphs to represent tags and we discuss the modeling of tagged resources, signs and tag actions, while in Section 5.4 we focus on tags as social acts and we illustrate the properties that we have foreseen to express different kinds of relations between the tagged resource and the sign. In Section 5.5 we describe some use cases and in Section 5.6 we draw conclusions.

5.2 Related work

Before the emergence of social tagging and folksonomies, several tools had been proposed to allow for the annotation of Web resources in structured formats; though none of them has scaled up to a broad diffusion, it is worth giving an overview on these ancestors of semantic tagging frameworks, before exploring more recent works on modeling of tags in the context of folksonomies.

5.2.1 Semantic annotation

One of the first systems to semantically annotate Web pages was SHOE [72], a platform that allowed to mark-up HTML documents on the basis of existing ontologies; another framework supporting ontology-based annotation of Web pages is CREAM [68]. Mangrove [118] was

developed as a tool aimed to "entice ordinary people onto the semantic Web", by providing them an easy graphical interface to annotate HTML documents with semantic metadata, and on the other hand by making these metadata immediately available to a series of semantic services, such as semantic search and calendar, while Saha is an annotation editor supporting the usage of different metadata schemes and domain ontologies [175]. Kettler et al. [86] introduced a Semantic Markup Tool, based on templates to hide ontological complexity from end users and allow them to easily specify new instances in the knowledge base. The scarce diffusion of these systems can be imputed on one side to the lack or inadequacy of available ontologies, and on the other to the excessive effort required to the users.

A milestone is for sure the W3C project Annotea, aimed at providing a semantic annotation framework, to enhance collaboration via shared metadata based Web annotations, bookmarks, and their combinations [85]. It uses an RDF based annotation schema for describing annotations as metadata and XPointer[4] for locating the annotations in the annotated document. Whereas the (extensible) vocabulary allows a certain richness of expressivity for describing annotation metadata and also the type of annotation (e.g. *Comment*, *Example* and *Change*), the content of annotations is just limited to unstructured data. The possible use of Annotea for (semantic) social bookmarking is illustrated in [98].

Another tool to share annotations about any Web page (or part of a page) is CritLink [186]; of particular interest is the idea of a *mediator* in the user navigation experience, providing additional information related to the page they are visiting, and in particular showing *extrinsic* links, defined collaboratively, in addition to the *intrinsic* ones (i.e., the link embedded by the author in the source web page).

A formal model of annotations in the context of information retrieval is presented in [3], while in [28] an approach based on *social validation* of annotations for information retrieval improvement is proposed: the key idea is to study the discussion thread associated with an annotation to evaluate the consensus level it has achieved.

5.2.2 Models for social tagging

The scarce success that semantic annotation systems have encountered so far is counterbalanced in recent years by the rapid diffusion and growth of *folksonomies*, or collaborative tagging systems. A tag can be considered

[4] `http://www.w3.org/XML/Linking`

as a very simple kind of annotation, where users just assign a keyword to a resource; the semantics provided by each user is shallow, but the strength of applications like Flickr[5] or Delicious[6] resides in the high number of active users, achieved also thanks to the extremely low effort required.

There have been several proposals of vocabularies for tagging systems; a broadly accepted starting point for a formalization of tagging is a tripartite model, where a tagging action is seen as a triple *<User, Resource, Tag>* according to Gruber's conceptualization [61]. Newman proposed to implement this model through reification [133], by means of a class `tags:Tagging`, connected to the tags (property `tags:associatedTag`), the user (property `tags:taggedBy`) and the resource (property `tags:taggedResource`), as well as additional metadata such as the date (property `tags:taggedOn`). This schema was adopted also by Knerr [96] and Echart et al. [43], who extended concepts such as domain and visibility, and represented it in OWL. The MOAT project extends Newman's model allowing for the association of a *meaning*, in the form of an external URI on the Web of Data, to each occurrence of a tag [139]; this principle has also been adopted in the Common Tag ontology[7], a simplified model based on the RDFa standard, which allows to embed RDF triples in HTML Web pages. In Common Tag the class `ctag:Tag`, having as label the text of the tag, is connected to the tagged resource (property `ctag:tagged`), to the date (property `ctag:taggingDate`) and to a meaning (property `ctag:means`). Complementarily, the SCOT model [87] offers primitives to model folksonomies as aggregations of tags, by means of class `scot:Tagcloud`, and properties such as `scot:totalTags` for the number of tags in a tag cloud, and `scot:ownAFrequency` for the number of occurrences of a tag in a tag cloud. Another quite rich ontology for tags and annotations is NAO [156], defined under the NEPOMUK Social Semantic Desktop project [60]; like NiceTag, NAO makes use of named graphs, but the relationship between the resource and the sign is fixed (`nao:hasTag`), as in the other current models. A comprehensive revision and comparison of tagging ontologies is provided in [88].

Revyu[8] is a reviewing and rating Web site, built with great attention towards Linked Data principles and best practices [71]. Everything in Revyu has a URI and can be reviewed and tagged; tags are just free

[5] `http://www.flickr.com/`
[6] `http://del.icio.us/`
[7] http://commontag.org
[8] http://revyu.com/

sequences of characters, and are expressed by means of Newman's Tag Ontology. One of the most interesting features of the system is that type information is in some cases derived from tags: when a resource is tagged "book", if an ISBN number is found in the corresponding Web page, an `rdf:type` statement is added to assert that the resource is a book. Analogously, resources tagged "movie" or "film" and having the same name of a DBPedia item of type `Film` are considered movies.

5.3 The NiceTag model

In this section we introduce the core of NiceTag model, i.e. the use of named graphs to represent tags, and we detail modeling of tag actions, tagged resources and signs used to tag.

5.3.1 Tag actions as named graphs

Carroll et al. [29] noted that RDF does not provide mechanisms (apart from statement reification) for talking about graphs and relations between graphs. They introduced Named Graphs in RDF to allow publishers to communicate assertional intent and to sign their assertions. The fact that it is often useful to embody social acts with some record clearly resonates with the scenarios of social tagging. Several authors before them proposed to transform RDF triples into quads [15, 42, 83, 114] appending to them an additional URIref or blank node or ID. The definition of [29] is deliberately simpler than [62] and [165] : "A Named Graph is an RDF graph which is assigned a name in the form of a URIref. The name of a graph may occur either in the graph itself, in other graphs, or not at all. Graphs may share URIrefs but not blank nodes." [29].

Extending the class rdfg:Graph defined in Carroll et al. [29], we define a sub-class of named graphs called TagAction class and embodying the acts of tagging. The triples contained in the named graph represent the link, modeled with the property `:isRelatedTo`, between an instance of the class `irw:Resource` and a sign modeled as an instance of `rdfs:Resource`, as described in Figure 5.1.

This paradigm provides four degrees of freedom to model tags:

1. the tag action as a named graph can be typed to represent different kinds of tagging, and metadata such as creator and date can be associated to it;

2. the model of the tagged resource can be extended to represent

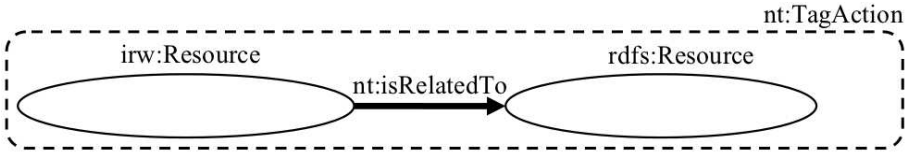


Figure 5.1: The minimal representation of a TagAction as a named graph. "nt" stands for nicetag namespace.

different kinds of resources encounterable on the Web, and to distinguish information resources from non-information resources;

3. the modeling choice of the sign used to tag is let free, and other vocabularies can be integrated on this side;

4. the relation between the tagged resource and the sign allows for a fine-grained account of the semiotics of tagging.

In the following, we address the first three aspects, showing how the tagged resource, the sign and the tag action can be modeled according to our ontology, while in the next section we focus on relationships and we illustrate how we represent tag actions as social acts on the Web taking inspiration from speech acts theory.

5.3.2 Modeling the tagged resource

Regarding the tagged resource, in current tagging applications this is usually identified by a URI, which by definition univocally identifies a resource. However, the manifold nature of URIs raises some issues and it is often not obvious to understand what exactly has been tagged. This problem is also known as the *identity crisis* of the Semantic Web [70].

Suppose one tags a photograph of the Sagrada Familia, and uses tags "800px" and "nice": the first applies probably to the photograph itself, while it could be unclear if the intended subject of the second is the picture (an *information resource*) or the monument (a *non-information resource*). As another example, let's take a tag annotating the Wikipedia entry for "Don Quixote": in this case, both the article and the book which it describes are information resources, and they can share many properties, such as the language and quality of writing, the author etc., so there is even higher ambiguity.

Hayes and Halpin pointed out the existence of two distinct relationships between names and things, which should be considered on the Web:

reference and *access* [70]. Access can be made unambiguous, as it depends on the Web architecture, but reference is inherently ambiguous as it has to do with language. While the publisher of a URI somehow owns it, and can determine what will be accessed through it, nothing can prevent someone to use that URI to refer to another arbitrary object. In the case of tagging, people who tag a URI are not constrained to refer to a specified object, but they can refer to anything.

The solution that we adopt to face this problem is the use of the IRW ontology[9], proposed by Halpin and Presutti [64], which introduces the distinction among *information resources* and *non-information resources*, and offers a comprehensive hierarchy of classes to model the different kinds of resources in the context of the Web.

We define a new class `:TaggedResource` and we declare it equivalent to the general class `irw:Resource` from the IRW ontology, while we define `:AnnotatedResource` as an equivalent of the more specific class `irw:WebRepresentation`. This is because while the tagged resource can be any resource, bound to the tag by just a *reference* relationship, the annotated resource is the resource *accessible* through the annotated URI, and therefore it has to be a Web-accessible representation of a resource. Subclasses of `irw:Resource` can be employed to describe the resource which is being tagged in each tag action.

5.3.3 Modeling the sign

Modeling of the sign in NiceTag is let free: as shown in Figure 5.1, we use for this purpose the class `rdfs:Resource`, which can subsume any other conceptualization.

As a first consequence, all currently available models of tags can be used. In particular, class `moat:Tag` from MOAT[10] and class `ctag:Tag` from CommonTag[11] can be used to anchor a sign to a specific meaning from an ontology, while the usage of `scot:Tag`[12] offers a rich set of primitives to express relationships among different labels, such as synonyms and misspellings, and to aggregate them and count their frequency and co-occurrence in some context.

As a further possibility, signs can also be treated as just literals, simple textual fields, reflecting the simplicity of current interfaces and resulting in a more synthetic representation.

[9] http://ontologydesignpatterns.org/ont/web/irw.owl
[10] http://moat-project.org/
[11] http://commontag.org/
[12] http://scot-project.net/

5.3.4 Modeling the tag action

The relation between the tagged resource and the sign used to tag is enclosed in a named graph, i.e. the corresponding tag action has a proper name, a URI which allows to make statements about it. The named graph itself can thus be typed to distinguish different kinds of tag actions; for this purpose we defined subclasses of the `TagAction` class.

Two subclasses, `:ManualTagAction` and `:AutoTagAction`, allow to distinguish tags created by humans from automatically generated tags. Other classes help in accounting for the way in which tags are expressed. `:WebConceptTagAction` is to be used when signs are computer processable by design, like URIs in MOAT and CommonTag. Subclass `:SyntacticTagAction` suits tagging involving complex signs like machine tags, tags decomposed in a plurality of elements (machine tags can thus be seen as a kind of triple tags), that sometimes have the particularity of following a specific syntax in order to be processable by APIs. Flickr machine tags, whose syntax follows the convention "namespace:predicate=value", are probably the most relevant example for this kind of tagging; we defined the corresponding `:FlickrMachineTag-Action` as a subclass of `:MachineTagAction`. For a general case of n-tuple tags which do not follow the syntax of any Machine Tag specification, we introduced the class `:N-TupleTagAction`.

Finally, the `TagAction` class is declared as a subclass of class `Item` from the SIOC ontology [24] in order to account for the shareable nature of tags, which can be seen as some sort of post in an online community platform. This, in turn, allows us to describe the place where tag actions are stored with the SIOC class `sioc:has_container`, and also the account (`sioc:UserAccount`) of the user (`foaf:Person`) of the tag with `sioc:has_creator`. We imagine a scenario in which these metadata do not have to be manually specified, but can be added automatically. Figure 5.2 shows how a tag action instance, declared as a named graph (subclass of `rdfg:Graph`), is also a `sioc:item`, and can therefore inherit the typical properties of a user generated post in the social Web.

Tag actions can also be distinguished according to the social act they embody. Beyond the subclasses of `TagAction` described in this section, we have introduced other subclasses to serve this purpose, as detailed in the next section.

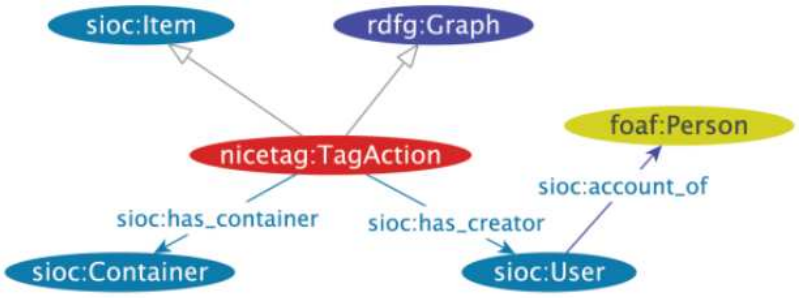


Figure 5.2: TagAction instances are declared as named graphs and as SIOC items.

5.4 Modeling nature and usages of tags

In this section we detail how the NiceTag ontology can be used to model different usages of tags, corresponding to different social acts and involving different kinds of relations between the tagged resource and the sign.

5.4.1 Tag actions as social acts

Tags can be seen as communication acts occurring on the Web. In particular, we refer to *social actions* as described by Reinach [151], who laid the foundations of speech act theory. The categories defined in speech act theory are suitable for many tag actions, such as those corresponding to *asserting* (`:Assert`) and *expressing emotions* (`:ExpressEmotion`), but in some case new categories have to be created, or existing ones have to be modified, in order to deal with the Web environment.

An example of a social act which has to be modified is *sharing*. While sharing normally involves at least two people, who know what is being shared, the `:Share` Tag Action in our model corresponds to sending a resource to someone, thus violating two condition of the traditional definition: in fact, no previous agreement and no knowledge by the receiver are implied. This choice reflects a common habit on the social Web, for which online communities have already found some conventions: in Delicious this is achieved through tags following a special syntax: "for:username", while in Twitter a symbol "@" is added at the beginning of the username.

Among categories which have to be created to describe social acts on the Web, the first one is *pointing* at a specific part of a resource (`:PointsAt`). While apposite mechanisms to support this function, such

as XPointer, would be extremely useful in the scenario of tagging, unfortunately there is currently no official standard for a concrete solution to this problem, and simple tags are often used to fulfil this need.

Another tag action which corresponds to a new social act is `:Aggregate`, for tags intended to aggregate content around some conversation, community or event. As explained in Chapter 3, this is a frequent case in microblogging, where the use of hashtags is often finalized to the inclusion of the message in some thread of conversation; the use of a hash character ("#") at the beginning of a tag, borrowed from microblogging, has been widely adopted as a convention also in other contexts to aggregate content related to a specific community. For example, tag "#icwsm" in Flickr collects pictures related to the ICWSM conference.

A last example of new social act that we introduced to deal with the Web environment is `:GrantAccessRights`, by which the publisher of a resource can determine whom it has to be accessible and not accessible to.

5.4.2 Modeling the link

One of the main innovations of NiceTag lies in the possibility of specifying the relation between the tagged resource and the sign. Inspired by previous studies, and in particular by the seminal work of Golder & Huberman [56], we modeled the different possible uses of tags with sub-properties of the most general property `:isRelatedTo`, as shown in Figure 5.3. The properties can also be grouped into three broader classes according to the categorization proposed by Sen et al. [161] who distinguish *factual, subjective* and *personal* tags.

We first consider factual tags, which are generally associated to `:Assert` TagAction. The most important property is `isAbout`, which represents the most common use of a tag, i.e. to identify the *topic* of an item. However, although many models of tagging just assume that this is the relationship by default, other relations exist, also remaining among factual tags.

The property `hasForMedium`, in rough words, is used to define *what the annotated resource is*, such as "forum", "blog", "photo" or "video". More precisely, it is used to indicate the medium which an HTTP-accessible Web representation belongs to.

Another property which allows to express an assertion and is different from `:isAbout` is `:isRelevant`; two specializations of these property allow to specify if the tagged resource is relevant to a person

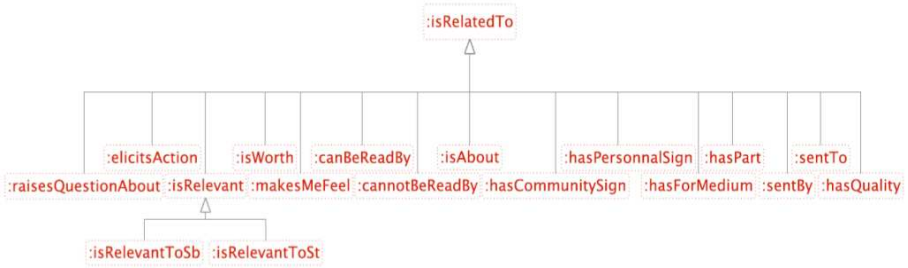


Figure 5.3: Subproperties of the property `:isRelatedTo` used to model the relation between the resource being tagged and the sign used to tag.

(`:isRelevantToSo`) or to a thing (`:isRelevantToSt`). As an example of the first case, tag "uncle" can be used to assert that something is relevant to my uncle, while for the second case I could use tag "thesis" to annotate a resource which is potentially useful for my thesis. It is clear that the tagged resources do not have to be *about* my uncle or my thesis: it is a different kind of relationship. Although this property can be associated to the `:Assert` tag action, in many cases, such as the examples mentioned above, it can be seen as personal more than factual according to the categorization by Sen et al. [161].

For subjective tags we have two subproperties. The first one is `:isWorth`, associated to the Tag Action `:Evaluate`, to associate a resource with an adjective or with any kind of sign expressing an evaluation, ranking or quality (e.g.: "nice", "bullshit", "****").

The second property corresponding to subjective tags is `:makesMeFeel`, associated to the Tag Action `:ExpresssEmotion`, for tags expressing an emotion stirred up by a resource; typical examples are exclamations and smileys (e.g.: "wow!", "<:o)").

Then we have all uses of tags intended to just make sense for the applier (personal tags). These include Golder & Huberman's classes *task organizing* (like "toread", "todo", "sendBob"), which we rendered with the property `:elicitsAction` (Tag Action `SetTask`) and *self reference* (like "mystuff"), modeled with property `hasPersonalSign`.

Similarly, we introduced the property `hasCommunitySign` to model tags which have an intended audience of a community. This property corresponds to the Tag Action `:Aggregate`, for collectively approved signs that are used to aggregate resources around a give event, community or shared interest. For example, we used the tag "#vocampnice2009"

to share resources about the VoCamp where NiceTag has been conceived across multiple social Web applications.

To model networking tasks, associated to Tag Action `:Share`, we added the two properties `:sentTo` and `sentBy`, while we introduced relations `:canBeReadBy` and `cannotBeReadBy` for tags used to grant access rights to someone. As a last property we have `:hasPart`, associated to Tag Action `PointAt`, another social act that we have introduced specifically for the Web context, as discussed in previous section.

Although our effort to account for all usages of tags reported in literature, to which we have added a few more cases, there is no limit to the users' creativity and ability to leverage existing tools to solve new problems, so we expect the model to be extended with other possible relationships intercurring between the resource and the sign. Moreover, arbitrary vocabularies can be adopted by different communities to deal with specific contexts.

5.5 Using the NiceTag ontology to represent and retrieve tags

In this section we have a look at how the described model can be used in practice to represent tags in structured format and how it can help to retrieve them. We first detail the implementation of named graphs with the RDF/XML Source declaration, showing a practical example where a tag is declared as a named graph according to NiceTag model and a SPARQL query to retrieve tags in NiceTag; then we discuss different use cases.

5.5.1 Using RDF/XML Source declaration to implement and use named graphs

In SPARQL, when querying a collection of graphs, the GRAPH keyword is used to match patterns against named graphs. However the RDF data model focuses on expressing triples with a subject, predicate and object and neither it nor its RDF/XML syntax provide a mechanism to specify the source of each triple. To serialize named graphs, Carroll et al. used TriX and TriG [29] but noted that RDF/XML is the deployed base. Therefore, Gandon et al. proposed in the W3C Member Submission "RDF/XML Source Declaration" [53] an XML syntax to associate to the triples encoded in RDF/XML an IRI specifying their origin; it uses a single attribute to specify for these triples represented in RDF/XML the

Listing 5.1: Declaration of a tag as a named graph using RDF/XML

```
1  <rdf:RDF xmlns:dc='http://purl.org/dc/elements/1.1/'
2    xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
3    xmlns:cos='http://www.inria.fr/acacia/corese#'>
4    <rdf:Resource rdf:about='http://www.yesand.com/'
5      cos:graph='http://mysocialsi.te/tag7182904'>
6      <nicetag:isAbout>improvisation</nicetag:isAbout>
7    </rdf:Resource>
8    <nicetag:ManualTagAction rdf:about='http://mysocialsi.te/
         tag7182904'>
9      <dc:creator>Fabien Gandon</dc:creator>
10     <dc:date>2009-10-07T19:20:30.45+01:00</dc:date>
11   </nicetag:ManualTagAction>
12 </rdf:RDF>
```

source they should be attached to. The IRI of the source of a triple is:

1. the source IRI specified by a cos:graph attribute on the XML element encoding this triple, if one exists, otherwise
2. the source IRI of the element's parent element (obtained following recursively the same rules), otherwise
3. the base IRI of the document.

The scope of a source declaration extends from the beginning of the start-element in which it appears to the end of the corresponding end-element, excluding the scope of any inner source declarations. Such a source declaration applies to all elements and attributes within its scope. If no source is specified, the URL of the RDF/XML document is used as a default source. Only one source can be declared as attribute of a single element.

The example in listing 5.1 shows how this applies to declare a tag as a named graph. Lines 4-7 declare the tag as a graph named `http://mysocialsi.te/tag7182904` . Lines 8-11 reuses the name of the graph to qualify the tag as a tag created manually by "Fabien Gandon" the 7th of October 2009.

Loading this RDF in a compliant triple store one can then run SPARQL queries like the one in listing 5.2. Line 2 searches for named graphs and the triples they contain. Line 3 enforces these graphs to be manually generated tags.

5.5.2 Use cases

The first and easiest scenario consists in the use of NiceTag to represent tags from existing applications; in the simplest case, where no informa-

Listing 5.2: SPARQL query to retrieve tags declared as named graphs

```
1 SELECT ?t ?a ?g WHERE {
2   GRAPH ?tag { ?t ?a ?g }
3   ?tag rdf:type nt:ManualTagAction }
```

tion is available about the relationship between the tagged resource and the sign, the general `isRelatedTo` property can be used. In this way NiceTag can provide a unified model for tagging, as data from different systems can be expressed in this general format; however, the expressive potential of the model in this scenario would not be leveraged, and other existing vocabularies could be equally useful.

A first step, which does not imply any change in tagging interfaces or any additional effort on the side of the user, but can bring an added value, is the automatic enrichment of tags, based on heuristics and knowledge from existing ontologies and eventually on machine learning. Just looking at the label of a tag, and at the context made up by other co-occurring tags, it is often possible to determine whether it is related to an information resource or not, whether it expresses an emotion or an evaluation, and so on. As in Chapter 4 we have shown techniques to disambiguate tags intended as simple labels, the same could be done to clarify other aspects of a tag action, such as the nature of the tagged resource or of its relation with the sign, or the goal of the tagger. This finer-grained representation of tags could improve performances and possibilities of search in the tag space; for example, one could choose to search for tags associated to the Sagrada Familia itself, or to pictures depicting it, and could as well restrict the search to only tags expressing an evaluation.

A potentially interesting scenario is the use of NiceTag for microblogging posts. Tweets can be seen as tag actions and represented as named graphs; the presence of URIs together with hashtags is frequent, and this combination can be modeled with property `:isRelatedTo` (or one of its subproperties) linking the two. Special conventions such as "RT: @username" for retweets, "username" for mentions or "via @username" can be also represented, while the use of nanoformats [162, 74] in conjunction with NiceTag can help to extend the expressivity of Twitter and to represent assertions and other social acts in a structured format.

As another interesting possibility which relies on current tagging interfaces, machine tags from existing applications could be converted into RDF thanks to the NiceTag ontology; conventions used in machine tags by different communities can be mapped to apposite RDF properties declared as sub-properties of NiceTag's `:isRelatedTo` property. In this

case, a set of ad hoc rules for a given context could help to achieve good results; imposing restrictions on the range and domain of specific properties it is possible to infer knowledge about tagged resources. Thanks to its robust theoretical foundations, NiceTag allows to manage also potentially complicated cases, such as the combined use of tags annotating the same URI, but referred to different resources.

The NiceTag model can also be used in combination with new interfaces, leveraging the richness of the model to provide users with full expressivity when tagging. While this can seem to contradict one of the founding principles of tagging, i.e. the extreme simplicity, we think that adding one or few clicks, for example with checkboxes, to the action of tagging, would be reasonable if it turns into remarkable advantages for the users. The immediate incentive is represented by the better possibilities of searching and browsing in the tag space; moreover, we expect that once users have got familiar with richer interfaces, they would discover new possibilities to use tags for more elaborated tasks and more ambitious purposes.

NiceTag is already being used in the ISICIL project[13], where social tagging is used inside organizations for technological watch and business intelligence, and reconciled with thesauri, information systems and business processes [112].

5.6 Conclusions

Though the success of tagging systems is due to their extreme simplicity and immediacy of use, the limitation of dealing with unstructured content appears straightforward, and users have been shown to long for more efficient and creative way of using tags to perform a wild variety of actions.

In this chapter we have proposed a general and flexible model to represent tags in all their possible flavours by means of named graphs. The essence of a tag in NiceTag is to embody in a record one or more RDF triples associating a resource with a sign and this core information can be enriched in several directions. To allow for the specification of the particular function of a tag, we have created several subproperties that can cover the different possible kinds of relationships between the sign and the resource being tagged. This relationship as a named graph is itself an instance of the class TagAction, and can thus have properties associated with it, like the user who performed the action of tagging, the

[13] http://isicil.inria.fr

date and the container. Moreover, it is possible to define the kind of a TagAction by choosing one of the subclasses we have defined. All these primitives (sign classes, function properties, tag action classes) are also designed to be extended at will.

In this way, thanks to the use of the RDF/XML Source Declaration syntax to assign a URI to a tag action, we obtain full expressive richness to represent tags from a multiplicity of facets, avoiding the burden of RDF reification. Both the Named Graphs model and the RDF/XML syntax extension provide a high-value for a small, incremental and backward-compatible change to the Semantic Web Recommendations. Combined with the tagging vocabularies this model provides us with a very flexible and extensible framework for social tagging interoperability.

On one hand, the ontology can be used as a unifying model to represent tags from current tagging applications, eventually enriched automatically thanks to rules and sources of structured knowledge about the domain; in particular, also machine tags can be turned into RDF triples and naturally represented in NiceTag. On the other hand, the model encourages the development of new interfaces, to foster a richer usage of tags.

6 Collaborative hierarchies: mining Wikipedia category structure to assign articles to macro-categories

6.1 Introduction

To organize the increasing amount of articles, in 2004 a system of categories was introduced in Wikipedia. Any user can change the categories to which a page is assigned by adding a special line in its source text; in the same way any category can be itself assigned to one or more categories, editing the corresponding page. This simple design choice allows for the creation of a collaborative hierarchy which is built and maintained by the community. While in social tagging systems every user can assign keywords at taste, and shallow semantics can implicitly emerge after harvesting metadata produced by many individual actions, here we find explicit collaboration and coordination mechanisms at the level of hierarchical relationships among items and categories. There are not many *personomies* merged into one folksonomies; instead the whole community is involved in creating one categorization structure for all the content.

The Wikipedia community has defined guidelines for the use of categories[1]; Yu et al. [187] summarized them in four main points:

- allow intersecting category structure;
- group similar articles;
- use the "right" number of sub-categories for each category;
- avoid cycles.

[1] See `http://en.wikipedia.org/wiki/Wikipedia:Category`

As a result, an article is usually not assigned directly to a general topic, but to closer low level categories, that can be in turn assigned to higher level categories. For example, "Clustering coefficient" is not assigned to "Mathematics", but to some lower level categories such as "Graph invariants" and "Graph Theory". The latter is assigned to "Mathematical relations" which is in turn assigned to "Mathematics".

Many category assignments are taxonomic and represent an "*is a*" relationship, like "Conifers" assigned to "Tree", but they may also represent other relationship types as shown in [127]; for example, "Brain" is a subcategory of "Cognitive science" as well as "Psychology". The structure can be naturally represented as a graph where nodes correspond to pages and categories, and edges to the oriented relationship "*is assigned to*". Whereas in principle the graph represents a hierarchy of topics and subtopics, with *broader* categories assigned to *narrower* ones, nothing prevents users from assigning categories following any criterion, sometimes just a "*related to*" relationship, so also loops are possible. There are also categories created for purposes related to the project, like listing stubs, articles lacking sources and so on, which do not group items according to their content.

Nowadays, more than 500 000 categories exist in the English Wikipedia, and almost all articles are assigned to at least one category. The resulting graph is a wealth of semantic relationships, one of the most complete efforts to categorize human knowledge. While most category assignments reflect coherent choices and best practices defined by the community, there are still many which slip from a coherent design, so the overall graph contains numerous inconsistencies and even cycles.

Most of the attention of previous literature has been focused on the extraction of ontologies from this pseudo-hierarchical structure, restricting the analysis on only consistent taxonomic relationships [167, 155, 143]. This approach is straightforward to derive formal ontologies or to enrich existing ones; on the other hand, it does not account for many relationships which do not follow a rigorous "is a" semantics but are often important. This is also one of the main limitations of the approach proposed in Chapter 4 when dealing with the WordNet noun hierarchy to enrich a folksonomy.

For example, the concepts of "Professor" and "University" would be hardly close to each other in a formal taxonomy, the first being subclass of "Person", and the second of "Institution" or "Building", located in different mutually exclusive branches; as a result, class "University" could be typically closer to "Shopping mall" than to "Professor", as both are

buildings. While this can be useful in scenarios like automatic reasoning, where consistency has to be guaranteed, it is probably a too restrictive approach to manage a topic hierarchy. On the contrary, it is very likely that related concepts will be close to each other in a lightweight semantics hierarchy like Wikipedia category graph, having strongly overlapping branches, multiple parent nodes and high density of relations. This kind of hierarchy has been shown to be more effective to help users in browsing tasks [187].

In this chapter we want to leverage the richness of the whole category graph as a thematic hierarchy, considering all the links established by the community between categories in order to automatically assign Wikipedia articles to general topics. More precisely, given a set of top level categories, chosen as main topics, our goal is to assign each article to one or more of these macro-categories. To this end we present several approaches and we evaluate and compare them.

The first simple method which we experiment is based on the exploration of the subgraph lying under a given category; the results obtained for some categories have been used in Chapter 7 to study collaboration networks at the level of specific categories; however, this method is not viable for the semantic areas which are more prone to the presence of multiple orthogonal or conflicting points of view. To deal with the fuzziness of these areas and to be able to study the whole graph, we propose two more sophisticated techniques, based on a set of topics (i.e. top level Wikipedia categories, or macro-categories), to which articles are assigned. One method, presented in [46], and used in Chapter 8 to aggregate discussions by topic, is based on the approach proposed by Kittur et al. [91] and consisting in searching the closest macro-categories to each article in the graph, while the other is based on the probability of reaching each macro-category starting from an article and following a random path in the graph.

The chapter is structured as follows. In the next section we illustrate an experiment to isolate a category and, recursively, its subcategories. In Section 6.3 we describe more sophisticated methodologies to associate Wikipedia articles with macro-categories and in Section 6.4 we evaluate them. Then in Section 6.5 we present related work and in Section 6.6 we draw conclusions.

6.2 Isolating the subgraph belonging to a category

The first and simplest approach to identify a set of pages belonging to a given topic is that of starting from the corresponding category, and considering all its subtree. In other words, given a category, we can select all the articles assigned to it or to a category which is directly or transitively assigned to it.

This approach would be straightforward if the category graph were a hierarchical tree: in that case in fact, starting from each of the categories at a given level in the tree, we could easily partition all the articles according to the branch in which they lie. As in our case each node can have more than one parent, i.e. our graph is not a tree, the subsets of articles found under each category would be overlapping. This fact prevents us from achieving a partition of the articles, but in principle it is still possible to assign a category to all the article of which it is an ancestor.

However, another problem makes this approach hardly applicable to the whole Wikipedia: the presence of many links established by users to connect categories which apparently have something in common, but cannot be subsumed by one another.

As mentioned in the introduction, this factor makes even possible the existence of loops in the category graph. More technically, we speak in terms of connected components, i.e. a subparts of a graph in which a path can be found between each pair of nodes, in both directions [171]. In the dump from march 2010 we found 90 directed components over the whole graph; most of them are composed of just two items being sub-categories of each other, but there are also more complex structures like the one depicted in Figure 6.1, in which the presence of several loops can be observed.

Looking at the figure, it appears straightforward that no clear relation of inclusion can be associated to many of the connections established by the users between these categories. These could hence be considered as isolated errors and corrected. However, this is not an isolated exception, and this kind of behaviour is quite usual, especially around some semantic areas. This problem seems to affect especially some wide topics of general interest, such as Society, Law, Language, History, Religion and Politics. In these fields no rigid categorization is possible, and many connections can be established between the categories according to the mind of different users. As a result, these high level categories would contain

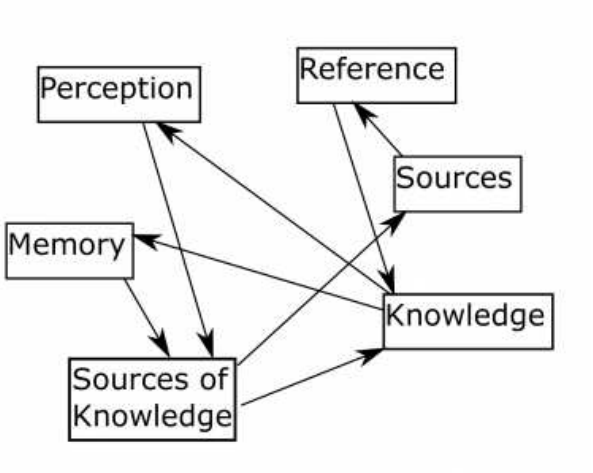


Figure 6.1: A strongly connected component in the Wikipedia category graph.

almost the whole Wikipedia, as many possible paths lead to them in the graph.

Instead, the algorithm shows to perform well for easily delimitable topics. Starting from a category like "Botany" or "Comics", in fact, and following descending links in the hierarchy, after manually excluding just few branches we obtain a set where all articles are clearly correlated with the topic. As an example, from "Botany" we had to remove some categories like "Cannabis activism" and "Cannabis culture" which would have lead to include political parties and Reggae music albums.

While suitable for some categories, in many contexts this approach is hardly viable due to the number of inconsistencies which should be faced to obtain a clean hierarchy. Alternative approaches to deal with this disordered environment are illustrated in the next section.

6.3 Assigning topics to Wikipedia articles

In this section we present two more flexible approaches to associate Wikipedia articles to general topics, leveraging the category graph. Both approaches are based on the pre-selection of a set of top level categories, or *macro-categories*, representing topics to which each article can be associated with different proportions.

6.3.1 Shortest path to a macro-category

In this section we describe the technique proposed by Kittur et al. [91], and we introduce some variations. The idea on which the technique is based is simple: if two categories are connected by an edge, they are probably semantically related. The closer two categories in the category graph, the closer their semantics. So it is possible to estimate, given a category, the macro-category in which it fits better, as the closest one in the graph. In the case of equally short paths from a category to multiple macro-categories, these are all considered suitable for the category being evaluated.

An article is assigned to macro-categories by evaluating the categories to which it is directly assigned (labels). More precisely, the degree to which an article belongs to a given macro-category is computed as the proportion of its labels which belong to that macro-category. In case of a label belonging to more than one macro-category, its contribution is split in equal parts among the macro-categories. So, suppose for example that the article "Barack Obama" is labeled with 4 categories, two of which are assigned to "Politics" and the third one to "Arts", and the remaining one is equally close to "Law" and "People": then the article will be considered related to "Politics" with a score of 0.5, to "Arts" with a score of 0.25 and to "Law" and "People" with a score of 0.125 each.

Though the category graph is based on directed relationships linking categories to super-categories, Kittur et al. [91] considered it as an undirected graph to compute the shortest paths between each category and the macro-categories, thus loosing the information carried by the assignments' direction. The simplest way to correct the algorithm would be to compute distances in the directed graph, considering only relationships followed according to the hierarchy direction, i.e. from the most specific, low level categories, up to the macro-categories. However, in this way many categories would remain disconnected from all the macro-categories, and many articles could not be assigned to any topic. Instead, we propose another way to improve the effectiveness of the algorithm by accounting for edge direction: while computing the shortest path between a category and a macro-category, we penalize by a factor w the edges followed in the wrong direction.

6.3.2 Probabilistic approach

The approach described in the previous section, based on the shortest path between an article and a topic, does not account for multiple paths

conducting to the same macro-category. In this section we present an approach based on the intuition that the more paths connect an article to a topic, the better the article will fit in the corresponding macro-category.

To develop this idea, we start from each article and we compute the probability of reaching each macro-category following a random path, no matter the number of steps required. So if an article is assigned to categories A, B and C, a probability score of 0.33 will be assigned to each of them; if both A and B have category D as their only parent node, then their contributions will be summed and a score of 0.67 will be assigned to this category. If D has 6 parent nodes, this score will be equally reparteed and each of them will result with a score of 0.11. By iterating this process until reaching the macro-categories selected as highest level nodes, it is possible to assign a probability score to each of them.

6.4 Results and evaluation

For this study we relied on a dump of the English Wikipedia dated March 12th, 2010, containing about 3.2 million articles and over 500 thousand categories. We removed all the categories which we identified as non-semantic, but project-based (e.g.: "Stubs"). While Kittur et al. [91] used 11 macro-categories, we chose to use 21 (showed in Figure 6.2), corresponding, with minor arrangements, to the current official Wikipedia top level categories[2].

We ran both the original algorithm as described in [91] and our modified version with $w = 3$, i.e. penalizing edges followed in the wrong direction in the hierarchy by a factor 3. All articles could be assigned to some macro-category, except for less than 100 pages, mostly corresponding to pages created by mistake or not yet completed when the dump was created; this result is an indicator of the attention that the community is dedicating toward article categorization, leaving almost no article uncategorized or isolated from the main component of the category graph.

The topic coverage emerging from the results of the modified algorithm are shown in Figure 6.2, where the percentages assigned to each macro-category over the whole wiki have been aggregated in order to estimate the importance of the different topics in terms of number of articles. The two largest macro-categories are "Geography and places" and "History and events". "Agriculture" is larger than it could be expected;

[2]See: http://en.wikipedia.org/wiki/Category:Main_topic_classifications

this is due to the high density of links between its subcategories, which makes it easily reachable in a few steps. Moreover, Wikipedia has a huge amount of pages about plant species. The smallest categories are "Arts" and "Computing"; this is partly due to the fact that some related low level categories are assigned to other "competitor" macro-categories, like "Culture" in the first case, and "Technology and applied sciences" in the second.

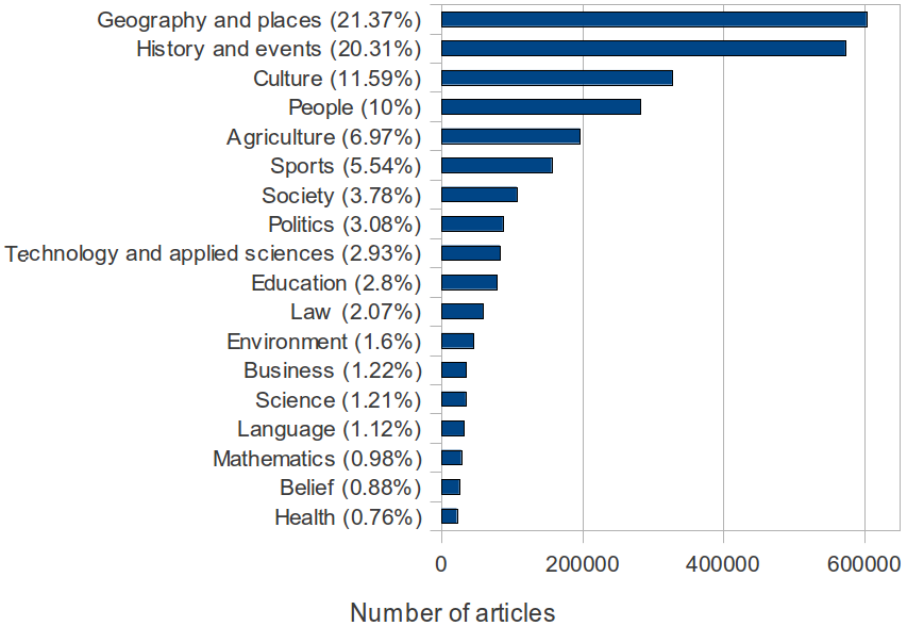


Figure 6.2: Size of the macro-categories, computed by aggregating the relatedness scores over all articles.

After executing the two algorithms, we evaluated them comparing the results with manually generated assignments. Assessment has been performed on 200 randomly selected articles, manually labeled by three human evaluators according to the 21 macro-categories. The cosine similarity between the assignments performed by human evaluators and the ones produced by the original algorithm is of 0.34; by accounting for edge direction we get a similarity of 0.37.

We also tested the probabilistic approach, which achieved a cosine similarity of 0.36 with the manually labeled sample.

6.5 Related work

6.5.1 Extraction of taxonomic knowledge from Wikipedia

The most notable project related to the extraction of structured knowledge from Wikipedia is probably DBpedia [12], a knowledge base which contains data extracted from Wikipedia and expressed in the form of RDF triples. The core of DBpedia is represented by properties of the entities corresponding to Wikipedia articles, extracted from infoboxes; Wikipedia categories are also included in the knowledge base: each category is represented as a "Concept" from the SKOS ontology (i.e., each category is treated as an instance of `skos:Concept`), and the property `dcterms:subject` is used to link each article to the categories it is labeled with. Relationships between Wikipedia categories are expressed through the SKOS property `skos:broader`, which allows to express generic brader-narrower relations, on which the Wikipedia pseudo-hierarchy is based, avoiding taxonomic "is a" properties. Taxonomic relationships are also present in DBpedia, however they are not directly extracted from Wikipedia, but from the YAGO ontology, which we describe in the following.

YAGO has been created with the goal of unifying WordNet and Wikipedia [167], Thanks to several rule-based and heuristic methods, such as identifying *conceptual categories* (i.e. taxonomic, opposed to *thematic* ones) by their label and checking the type of each entity, Wikipedia is leveraged to extract structured semantics, and especially taxonomic relationships, which are integrated into WordNet's structure. An analogous project is presented in [155], where Wikipedia category graph is used to enrich the formal ontology Cyc. Ponzetto and Strube [143] present another approach based on the extraction of "is a" relationships from Wikipedia to derive a domain-independent taxonomy; the result is evaluated through comparison with the Cyc ontology. A common feature of these works, which marks a substantial difference with respect to our approach, is that all the *thematic* part of categorization is discarded, and only taxonomic categories are considered in order to derive consistent subsunction relations.

6.5.2 Studies of Wikipedia category structure as a thematic hierarchy

Few works take into account also thematic categories. Kittur et al. [91] proposed a technique to automatically assign a semantic closeness score

between each article and a set of topics, corresponding to top level categories. The algorithm proposed is based on the shortest path between each topic and the categories which are directly assigned to an article; results are used not only to study the quantity of articles under each topic, but also to investigate the amount of conflict in articles from different topics. While in [91] the category structure is treated like an undirected graph, thus loosing relevant information about the hierarchy, we have modified this approach to account for the direction of category assignments.

Holloway et al. [77] propose an approach to compute similarity between categories according to their co-occurrence within individual articles; a map of topic coverage in Wikipedia is drawn and 8 top level categories are highlighted. Also Pang and Biuk-Aghai [137] rely on co-occurrence between categories to evaluate their similarity. An approach based on the hyperlinks between pages as one of the criteria to calculate relatedness in Wikipedia is presented in [34],

6.6 Conclusions

In this work we faced the problem of automatically assigning each Wikipedia article to one or more topics, leveraging the structure of categories and subcategories created by the community. We first tested an approach based on isolating the subgraph lying under a given category; we found out that, while this simple method is suitable and gives accurate results for topics where the categorization seems to follow a shared coherent criterion, it is unviable in many contexts in which the graph is too tangled.

So, to propose a general solution allowing to deal with the fuzziness of the whole graph, we proposed two alternative approaches. We modified the algorithm proposed by Kittur et al. [91], based on the shortest path between a category and a macro-category; by penalizing edges followed in the wrong direction with respect to the hierarchy, we are able to account for the orientation of the categories assignments, without loosing the information brought by these connections. The algorithm proposed shows to outperform the original one improving the accuracy, measured as the similarity with manually generated assignments, from 0.34 to 0.37. Also with the other algorithm proposed, based on the probability of reaching each macro-category starting from an article and following a random path in the category graph, we achieve a higher similarity score, of 0.36. These results are encouraging, though a more rigorous evaluation

process would be needed in order better assess the statistical significance of the improvements obtained.

The topic coverage computed here gives the same importance to pages of different sizes, and thus risks of overestimating categories containing many short pages, and in particular those automatically generated by bots. The count may be improved by considering, instead of the number of pages assigned to a macro-category, the number of edits or words in these pages, to obtain a more representative map of the wiki. Other article-level metrics, such as the number of polls, or of edits done by specific classes of users, can be aggregated by topic, to study how activity varies over different semantic areas.

The results described in this chapter have been used to analyze the difference among discussions in article talk pages from different macro-categories, in Chapter 8, while some of the results obtained through the simple technique of isolating a category proposed in Section 6.2 have been used in Chapter 7 to study the networks of authors active around specific domains.

7 Co-authorship 2.0: Patterns of collaboration in Wikipedia

7.1 Introduction

The idea of collecting a compendium of human knowledge in one single work can be dated back of at least 2000 years with Pliny the Elder's Naturalis Historia [159], and is recurrent in history, until the foundation of modern encyclopedias by Diderot and D'alambert [41]. However, this effort has always been faced by limited groups of experts, until the Internet and the growth of the read/write Web opened up the doors to the creation of collaborative encyclopedias, of which the most relevant example is Wikipedia. Wikipedia, now existing in more than 200 languages, is an encyclopedia that anyone armed with Internet connection and a Web browser can edit. As explained in Section 2.2, it can be seen as the model of the *bazaar* [149] applied to the redaction of encyclopedic content. This community effort has resulted in one of the largest collaborative projects in human history, and as such has attracted the attention of many researchers, who have analyzed its social dynamics from different perspectives to shed light on the process of content creation by a community.

Indeed, the analogy with a scientific collaboration community has been proposed in the literature and is straightforward, as editing of a wiki encyclopedia entry somehow resembles the collaborative writing of a scientific paper [63]. Studying Wikipedia as a co-authorship network can allow for a comparison with scientific communities widely studied in literature, and unveil patterns of collaboration that are hidden in the revision history. Nevertheless, to the best of our knowledge there is still no extensive study on the community of Wikipedia contributors as a co-authorship network. Current methods are mostly based on the assumption that just the fact that two users edited the same page is enough to establish a relationship, and fail to scale to the size of Wikipedia in a major language.

The first contribution of this work, published in [107], is the development of a general and scalable methodology to extract a co-author network from a wiki's revision history. One fundamental difference between a paradigmatic case of scientific collaboration community and a wiki is that collaboration on a wiki article has lower barriers than the process of publishing a scientific paper together, and does not imply previous agreement. Moreover, size of contributions can be strongly uneven, and not all edits are accepted by the community. Considering as co-authors all users who just edited the same article may bring to establish too many connections between people that were not really involved in writing something together; this would result in an extremely large and dense network. To select those who can be considered the "real" authors of a wiki article, and to account for the process of convergence toward a shared outcome, we rely on a metric which evaluates contribution according to the longevity of the modifications introduced [2]. According to this measure, we define a method to select the main contributors of each page as the ones who provided the most of its accepted content, and to obtain a collaboration network.

Our second contribution consists in the analysis of the co-authorship network obtained from a complete dump of the English Wikipedia, to characterize its community on a temporal dimension. The study of the network's macroscopic features and the comparison with scientific collaboration networks help understand the way the community is structured and the role of administrators and most involved users, pointing out the existence of specific patterns of collaboration.

In the next Section we offer a brief overview on the community of Wikipedia contributors, based on previous studies. Then in Section 7.3 we describe our algorithm to extract a co-authorship network from a wiki's revision history, while in Section 7.4 we show the results we obtained for the English Wikipedia, analyzing the evolution of different macroscopic properties of the network and investigating the role of the most influential users. Finally in Section 7.5 we discuss conclusions and directions for future work.

7.2 Related studies

An in-depth qualitative description of social dynamics and established rules and conventions in Wikipedia is offered in [26], where the *community of practice* of Wikipedia users is studied from an activity theory perspective; the authors investigate how new users can move from *le-*

gitimate peripheral participation to full community involvement and how their activity can change substantially over time, moving from local focus on individual articles to a concern for the quality of Wikipedia content as a whole and for the health of the community.

One of the first extensive quantitative studies on the Wikipedia community was presented in [178], where its growth is shown to follow an exponential trend, after a first linear phase; both the number of authors per articles and vice versa the number of articles per author exhibit a power law distribution. Almeida et al. [6] characterized the evolution of Wikipedia as a self-similar process growing exponentially, due especially to the continuous increase of the number of contributors. They also observed that the distribution of the number of updates per user follows two Zipf's laws with different parameters, which split the community in two groups: a small nucleus of around 5000 very active users, who contribute more than 1000 articles, and the vast majority of common contributors.

Kittur et al. [90] divide Wikipedia contributors into different categories according to their degree of participation in terms of number of edits. They observe at first the rise of an elite of very active users, who perform the most of edits, and then the decline of this "elite" in virtue of what they call the "bourgeoisie", the large majority of common users. Ortega et al. [135] found out that the 10% of contributors were responsible for more than the 90% of edits; they also noticed that this strong inequality tends to stabilize over time. The effect of contribution inequality on the quality of Wikipedia articles has been investigated in [11]: a positive effect of global inequality, measured according to the Gini coefficient of edit count distribution, is found. Kittur et al. [92] study the role of coordination, observing improvements in article quality as effect of both explicit coordination through communication, and implicit coordination through concentrating the majority of the work in the hands of a subset of users.

In [63] Wikipedia is studied as a peer review system; no evidence is found that experience helps editors avoid rejection, while the authors observe a strong tendency of users to defend their own contributions.

The first relevant attempt to study the social network of Wikipedia editors, to the best of our knowledge, was done in [99]: a directed graph is drawn to represent the network of consequent edits to a page and to evaluate the authority of authors over an article or a domain, and the degree of centralization of an article. Brandes et al. [23] represent the contributors of a page as nodes, and the different kinds of actions

linking them as edges, with attributes expressing the numbers of deleted, undeleted and restored words. By means of this kind of network, the authors study the different roles of users and the collaborative structure of pages, and they try to identify poles of opinion. Iba et al. [82] focus on the network based on consecutive edits done to a page, in order to identify editing patterns using dynamic social network analysis. The models proposed in these studies are useful to represent interactions over one or few pages, while our concern is to characterize the whole community.

Closer to our work are studies which take into account the collaboration network in a wiki as an affiliation network. Biuk-Aghai [21] proposes a visualization method which exploits co-authorship networks to compute the similarity between Wikipedia pages. In [170] a method is described to measure co-authorship relationships in MediaWiki; the model allows for the representation of weighted relationships, where the relevance of each collaboration is computed according to the temporal overlap in the activity of two authors on a same page, and to the proportion of their edits with respect to the total revisions of that page. Müller-Birn et al. [126] combine different measures in order to evaluate author activity in wikis: besides edit count, they compute for each author also a measure of *content significance* based on *tf-idf* model, and metrics of centrality in the social network. The first results, on a small collection of articles, show that the three criteria bring to quite different rankings. A model based on a tripartite network is presented in [128], the three dimensions being users, pages and categories, where categories play the same role as tags in folksonomies. Klamma et al. [95] propose a model to study wikis as social networks, taking into account articles, revisions, users and URLs, and apply dynamic network analysis to several wikis; as they consider all edits for the construction of the networks, the model cannot scale to the size of Wikipedia in a major language. All of these studies differ from ours in that they are only based on the edits done to a page, without accounting for differences in the contribution carried by different edits.

The network of replies between users in Wikipedia discussion pages is analyzed in [108], while the interplay between social ties and similarity is studied in [36], where feedback effects are found between personal communications and editing of the same articles. The network of personal communications is also studied in [55] to characterize different profiles of users.

7.3 From revision history to a co-authorship network

In the last decade, the availability of comprehensive online bibliographies has made possible the extensive study of co-authorship networks for entire fields; in particular, large-scale networks have been constructed to represent co-authorship collaborations in physics [13], mathematics, neuroscience, biology and computer science [131, 129]. The study of these networks has shown to be a useful source of information on the academic communities, both for local and global analysis.

As discussed in the previous Section, the analogy between Wikipedia and a scientific collaboration community is not new in literature as a potential useful means to study its social structure and dynamics from a sociometric perspective, and some methods have been proposed to extract a collaboration network from a wiki [99, 170, 126, 95]. However, current methods are mostly based on the assumption that just the fact that two users edited the same page is enough to establish a relationship, and fail to scale to the size of Wikipedia in a major language. In our opinion the approach of including any user who edited a page as an author is an oversimplification; in effect, we would like to extract the main contributors of a page, both in terms of quantity and quality of their interventions. In particular, while to publish a scientific paper together two researchers need to know each other in advance, and then to agree on the final version of the paper, in a wiki it is just a matter of editing the same page; by taking into account the degree of acceptance that a contribution has received by the community, we try to make up for the lack of explicit agreement between users in previous models.

We propose an algorithm that acts in three main steps: at first, for each page a *score* is computed to evaluate the contribution of its editors, then the main contributors are selected as authors of the article, and finally the co-authorship network is constructed. In the following we will illustrate these three steps.

7.3.1 Measuring contribution

The first step of our method requires the computation of *author contribution* in the scope of each wiki page. We need a function:

$$c : U \times P \to [0, +\infty) \tag{7.1}$$

which, given a user in the set of registered users U and a page in the set of pages P, has two main requirements. First it has to return a positive

numerical value. This is because with such a function we can calculate the total contribution for a page, and estimate the relative influence of each user on it. Then, in order to perform temporal studies, it is required to the function to be computable within specified intervals of time. This definition is quite general, and any measure quantifying the contribution of a user to a page can be used.

Most of the quantitative studies on the Wikipedia community just take into account the number of edits performed by a user as a measure of her activity; this naive measure is often used also inside the same community of Wikipedia (e.g., to be elected as an administrator of the Italian Wikipedia, a user needs to have performed at least 500 edits). Though it is largely used, due to its simplicity, the limitations of this approach are evident, as no importance is given either to the size or to the quality of interventions. More sophisticated approaches to compute author contribution are based on the observation of the lifespan of the changes introduced. The metric proposed in [145] takes into account the number of times a word added is viewed without being changed in the next revisions, while in [63] the lifespan of a word is measured according to the number of editors modifying the page without removing it. A set of metrics and efficient algorithms to compute author contribution to a wiki is illustrated in [2], in the framework of the WikiTrust project[1]. Among these metrics, *edit longevity* is based on the number of words edited by an author, computed with suitable heuristics, and weighted according to their longevity in the following interventions.

For this work we chose to rely on the aforementioned metric of *edit longevity* as described in [2], both for its accuracy and for the efficiency of the algorithm proposed, allowing for its computation over the whole English Wikipedia as: $el : E \to [-\infty, +\infty)$, E being the set of all edits in the wiki. While in Wikitrust edit longevity is cumulated for each author over the whole wiki, our approach is to cumulate this measure in the scope of each single page, finding as a result a score associated to each contributor, telling *how much accepted content* they have introduced in a certain article. As we are interested in cumulating a measure of the relevant contribution carried by each author to a page, we do not take into account interventions bringing a negative score (which means interventions mostly not accepted in the following revisions). We define $E_{u,p}$ as the set of edits performed by user u on page p, and we compute

[1] http://wikitrust.soe.ucsc.edu/

the contribution of user u to page p as:

$$c(u,p) = \sum_{e \in E_{u,p} | el(e) > 0} el(e). \tag{7.2}$$

7.3.2 Author selection

As pages vary substantially both in length and number of editors, it would be difficult to establish a fixed number of authors to be selected from all articles. Instead, we adopt a general and flexible strategy, which consists in selecting the first users who authored a certain percentage of the whole accepted contribution. Anonymous contributors are identified in Wikipedia revision history by their IP number, so a possible strategy would be to include them in the community as single individuals, by treating IP numbers as normal user nicknames. We are not following this approach for the fundamental reason that IP numbers are not reliable identifiers. Moreover, it makes sense to identify only users that explicitly chose to have a nickname in the community. So we discard all anonymous contribution. For each page p we define the set U_p of all registered users who edited it. We select the set of *authors* of page p as the smallest subset $A_p \subseteq U_P$ containing the first users of U_p, ordered by descending contribution, such that:

$$\frac{\sum_{a \in A_p} c(a,p)}{c_{tot}(p)} > \theta \tag{7.3}$$

where $\theta \in [0,1]$ is a relative threshold and $c_{tot}(p)$ is the total contribution to the page by registered users:

$$c_{tot}(p) = \sum_{u \in U_p} c(u,p)$$

Then we remove all the users whose contribution to that page, in absolute terms, did not reach a minimum threshold M, by imposing a further condition for each author a of page p:

$$c(a,p) > M \tag{7.4}$$

7.3.3 Network construction

As discussed in the previous Section, we select for each article a variable number of authors who have provided a significant contribution, both in

absolute and relative terms, and we obtain a bipartite network, or *affiliation network*, where each user is associated to all the articles of which she is a main contributor. To obtain a collaboration network, $G = \langle V, E \rangle$, we project this bipartite network on the users' dimension, establishing a connection between each pair of users who have collaborated on at least one article. So the set of vertices is: $V = \bigcup_{p \in P} A_p$ and the set of edges is: $E = \{(a_1, a_2) \mid \exists p \in P : a_1, a_2 \in A_p\}$.

To account for temporal dynamics, we consider slots of a fixed amount of time T, and we snapshot the wiki's revision history at different instants. For each period we build a network based only on the edits performed in that time slice ($[0, T), [T, 2T), \cdots$), and a cumulative one considering also all previous edits ($[0, T], [0, 2T], \cdots$). With the first method we can represent the network of actual interactions between users in a limited period of time; with the second approach we consider cooperation over the whole history of each article, coherently with the idea that, when editing a page, a user is working on all past contribution.

7.4 Network analysis of Wikipedia author community

We applied the algorithm described in the previous Section to the English Wikipedia, to extract its co-author network. We based our analysis on a log from the WikiTrust project, where edit longevity has been computed for all edits until February 11th, 2007.

For scientific co-author networks the usual period of time examined is one year; this is probably due to the availability of the publication year, and to the scarce relevance of a finer-grained division of time, as the process of publishing can take months. As in Wikipedia everything happens faster, and the revision history provides detailed temporal data, we chose to adopt shorter periods of $T = 3\ months$. We have constructed for each period both the cumulative and the non-cumulative network, using thresholds $\theta = 0.7$ and $M = 10$, the first telling we select as authors of an article the minimum set of top contributors responsible for at least 70% of the total contribution to it, the second establishing the minimum contribution needed to be considered an author (roughly corresponding to 10 words added and never modified in the following 10 revisions)[2].

Figure 7.1 shows the number of editors per page and the number of *authors* selected by our algorithm, for a period of three months. As it

[2]Varying the parameters we did not observe remarkable differences in the results.

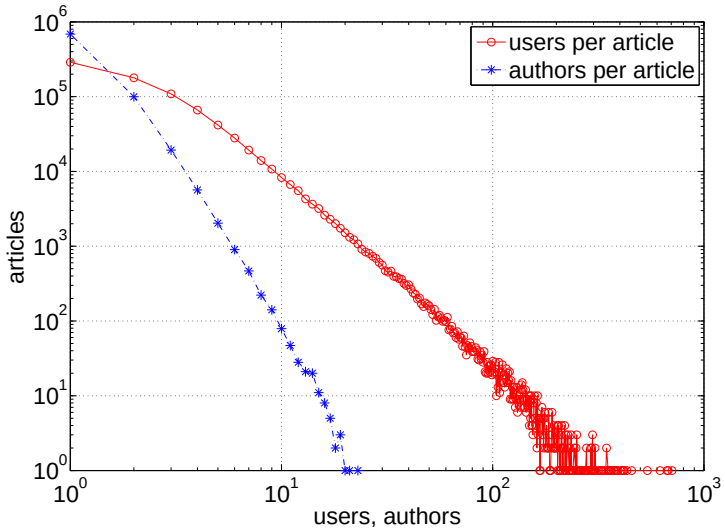


Figure 7.1: Distribution of the number of users per page observed in a three month period (November 2006 - February 2007), plotted on a log-log scale.

can be noted, though there are articles edited by up to 500 users, our algorithm does never select more than 20 editors as authors of a page. Anonymous contribution, that we discarded, adds up to 25% of edits done, but only to 10% in terms of edit longevity. These data point out the lower weight of anonymous edits in terms of size and acceptance by the community.

Figure 7.2 shows the growth of Wikipedia in terms of number of articles; together with the total number of articles, we have plotted also the number of those for which at least one and two contributors have been selected; the percentages over the whole history until February 2007 are about 97% and 39%, respectively. The graphics points out that most of Wikipedia articles have been redacted by one main editor. Analogously, besides the evolution of the total number of Wikipedia users, in Figure 7.3 we plot the number of users selected as authors of at least one article, and the number of authors who have collaborated with at least another author; the percentages are about 29% and 24% and show that the vast majority of authors have collaborated with some other authors.

In the following we analyze the networks according to several metrics to characterize the Wikipedia community and detect patterns of collaboration. For the analysis we relied on the software package Igraph for R [37].

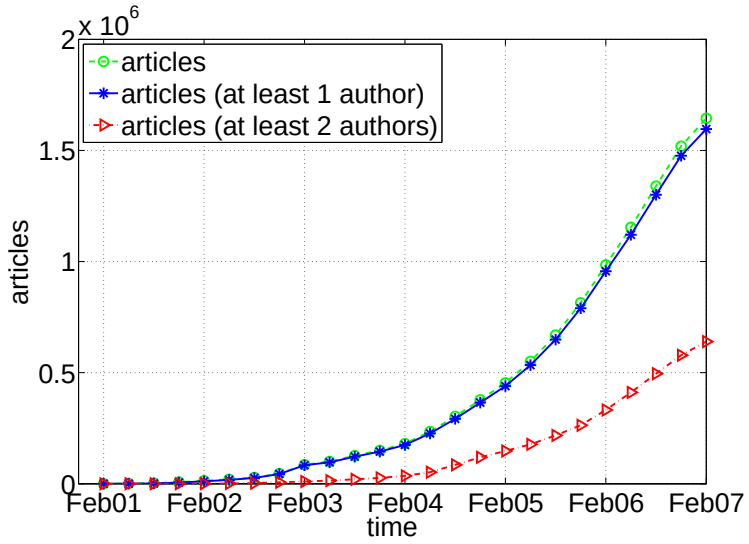


Figure 7.2: Evolution of the number of articles.

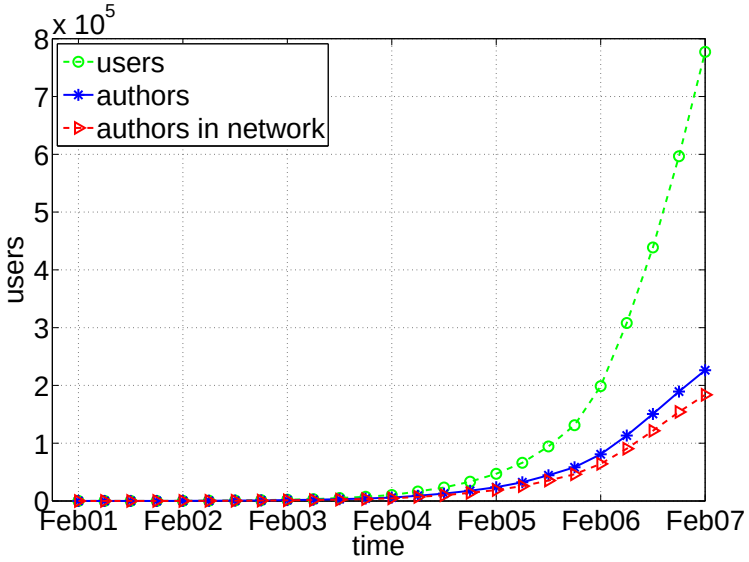


Figure 7.3: Evolution of the number of users.

7.4.1 Macroscopic network analysis

Tables 7.1 and 7.2 report the evolution of some macroscopic features of the non-cumulative and cumulative networks, respectively. The size of

the *giant component* G, the largest connected component, is always over 97% in the cumulative network, showing a very scarce fragmentation; high values are observed also in the non-cumulative network. The size of the other components does never exceed 6 or 7 nodes.

In social network analysis the number of edges k incident to a node is generally called *degree*. Looking at the evolution of *mean degree* $\langle k \rangle$ over all nodes, or *network connectivity*, in the cumulative network, we observe a rapid growth, that tends to converge around a value of 22. In the non-cumulative network, after a growth in the first periods, connectivity starts following a slowly decreasing trend; this is an interesting signal that the mean number of actual collaborations during a limited period of time remains bound, and tends to decrease as a larger base of users gets involved in the community.

The networks exhibit the *small world* property [180]: the maximum distance, or *diameter* D, tends to slowly increase over time, but no more than 10 or 12 steps are required to connect any pair of nodes. This value is considerably low, especially if compared with those of scientific collaboration networks, where the diameter can typically reach the value of 20 [129]. Analogously, also *mean distance* d exhibits a slow linear increase with time, remaining between values of 3 and 4; this result is also quite low with respect to scientific collaboration networks observed in literature, where the average values are usually over the double. These short distances can be explained in virtue of the lower barriers to the collaboration between any pair of users in a wiki; they can also be interpreted as an effect of the centralization of the network around some very active users, the so called *sociometric stars*.

Clustering coefficient

Similar conclusions can be inferred from the observation of the global *clustering coefficient* of the graph, that is computed as:

$$C = \frac{3 \cdot number\ of\ triangles}{number\ of\ connected\ triples\ of\ vertices}$$

and represents the percentage of closed triples in the network: at the extremes, a completely connected graph has $C = 1$, whereas a hierarchical tree has $C = 0$, as no loops are possible [180]. Though our networks exhibit clustering coefficients higher than the ones of an equivalent randomized graph, this value is very low with respect to scientific collaboration networks observed in literature, where it also shows to be

Table 7.1: Macroscopic features of the non-cumulative network: each row describes the network of collaborations based on the edits performed in the three month period ending on the pointed month.

Period	**N**	$\langle$**k**$\rangle$	**G%**	**C**	**d**	**D**	**r**
Feb02	124	5.8	100	0.17	2.83	6	-0.14
May02	178	6.5	98.9	0.19	2.85	6	-0.16
Aug02	214	7.0	97.7	0.22	2.88	6	-0.11
Nov02	415	9.6	99.0	0.23	2.87	6	-0.17
Feb03	585	8.3	99.9	0.17	3.07	7	-0.14
May03	723	8.9	98.1	0.18	3.07	6	-0.10
Aug03	1199	8.5	96.2	0.14	3.26	7	-0.07
Nov03	1511	8.9	92.8	0.14	3.26	7	-0.07
Feb04	2023	10.0	97.0	0.13	3.31	9	-0.06
May04	3817	10.1	95.9	0.10	3.43	8	-0.05
Aug04	5101	9.9	97.6	0.08	3.53	9	-0.05
Nov04	6781	9.5	95.9	0.06	3.46	8	-0.08
Feb05	8643	8.6	95.9	0.07	3.75	9	-0.04
May05	11678	8.3	95.3	0.07	3.83	12	-0.02
Aug05	16622	8.3	95.3	0.07	3.91	10	-0.02
Nov05	20117	8.3	94.5	0.09	3.95	11	0
Feb06	31424	9.0	94.3	0.09	3.95	11	-0.01
May06	45069	7.5	93.1	0.04	3.96	11	-0.05
Aug06	55948	7.3	92.5	0.03	4.06	12	-0.04
Nov06	62126	6.6	91.0	0.03	4.06	12	-0.04
Feb07	64318	6.9	90.2	0.03	4.08	12	-0.03

usually more stable over time [129, 35]. Among the co-authorship networks studied in [131], the only one having a similar value of C is Medline, a very large community characterized by a strongly hierarchical social structure, based on laboratories where a high number of collaborators gravitate around a "principal investigator". Comparable values of clustering coefficient have been observed in online communities [80] and message board networks [57]; a first simple consideration can be that it is easier to establish new and heterogeneous connections with other people in an online community than in the offline world.

The low and decreasing values of C in our network, shown in Figure 7.5, can be also seen as a symptom of the growing centralization of the network, that is accentuated as new users attach to the stars, central

Table 7.2: Macroscopic features of the cumulative network: each row corresponds to the network based on the whole history of pages until the pointed month. **N** stands for *network size*.

Until	**N**	$\langle$**k**$\rangle$	**G%**	**C**	**d**	**D**	**r**
Feb02	137	6.2	100	0.17	2.87	6	-0.11
May02	256	9.0	100	0.21	2.75	6	-0.16
Aug02	388	11.6	100	0.24	2.70	5	-0.19
Nov02	706	14.0	99.7	0.26	2.76	5	-0.23
Feb03	1116	14.7	99.5	0.24	2.83	7	-0.23
May03	1508	16.7	99.5	0.23	2.85	6	-0.21
Aug03	2315	17.1	98.6	0.22	2.92	7	-0.20
Nov03	3286	18.0	97.1	0.20	2.96	8	-0.19
Feb04	4542	19.5	97.4	0.19	2.99	8	-0.18
May04	7000	20.7	96.7	0.17	3.04	8	-0.17
Aug04	10033	22.0	97.8	0.16	3.08	9	-0.16
Nov04	14072	23.1	97.7	0.14	3.10	8	-0.16
Feb05	19004	23.5	97.8	0.13	3.14	8	-0.15
May05	25759	23.4	98.1	0.12	3.19	8	-0.14
Aug05	35408	23.6	98.1	0.11	3.24	8	-0.13
Nov05	46181	24.2	97.9	0.11	3.29	8	-0.12
Feb06	64268	24.3	97.8	0.10	3.33	9	-0.11
May06	90523	23.1	97.4	0.09	3.38	8	-0.10
Aug06	121461	22.6	97.2	0.08	3.40	10	-0.08
Nov06	154091	22.0	96.8	0.06	3.41	9	-0.08
Feb07	183710	22.4	96.7	0.06	3.41	10	-0.07

nodes with a very high degree. This process can be attributed to the role of some "superusers", who seem to be omnipresent: administrators, bots, and a core of very active contributors, who seem to intentionally spread themselves over the whole Wikipedia, covering all its areas. Some researchers claim that the decreasing percentage of edits performed by administrators and by the most active users suggests that the Wikipedia elite is declining and a bourgeoisie is rising [90]. Though, in our analysis we find evidence of the fundamental role that these users continue playing, by leveraging their centrality in the growing network.

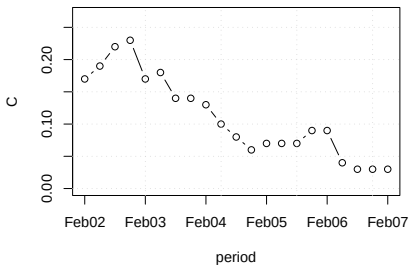


Figure 7.4: Trend of clustering coefficient in the non-cumulative network.

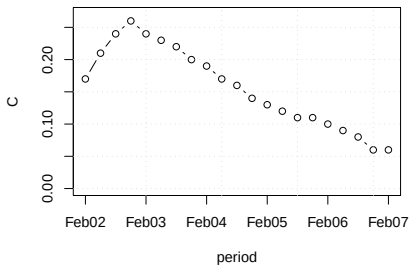


Figure 7.5: Trend of clustering coefficient in the cumulative network.

Degree distribution

The uneven level of participation is confirmed by a study of the degree distribution, that is plotted in Figure 7.7 for the cumulative network and in Figure 7.6 for the last three month period studied; both networks are *scale free* with a heavy tailed distribution. According to [26], a reason for this disparity in the number of collaborations may be found in the distinction between the *periphery* and the *core* of the community: users who feel fully involved in the project, members of the *tribe*, care about the whole content of the encyclopedia, and their activity is substantially different with respect to the majority of users who are just interested in contributing on specific topics.

Degree assortativity

An interesting question is whether these very central authors are preferentially linked with other highly connected ones or not; in other words, if the network is *assortative*. The *assortative mixing*, or *degree correlation* r of a network, measures the tendency of nodes to connect with other nodes having a similar degree [132]. Being assortative is traditionally considered a characterizing feature of social networks, in contrast with technological and biological ones, like the Internet or the WWW, which are disassortative [130]. Nevertheless, the degree assortative mixing of our networks is negative, with values increasing (decreasing in absolute value) until about -7% in the cumulative, and -3% in the non-cumulative one (Figures 7.8 and 7.9). This result marks a notable difference with respect to scientific collaboration networks, that have been shown to ex-

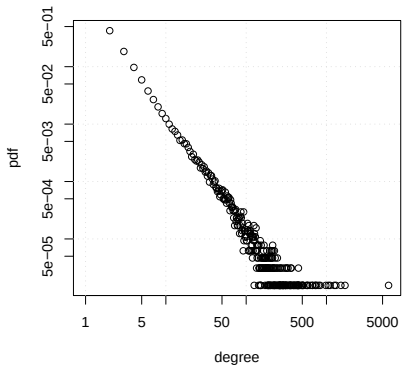


Figure 7.6: Degree distribution in the non-cumulative network (Nov2006 - Feb2007).

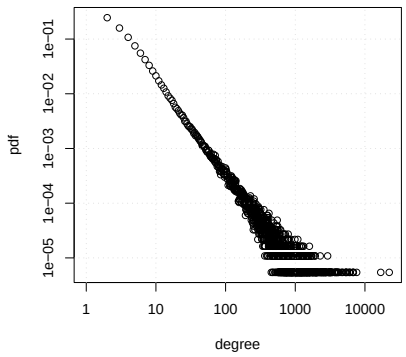


Figure 7.7: Degree distribution in the cumulative network.

hibit assortative mixing patterns [132]; instead, neutral or disassortative networks have been observed in other online communities such as Internet dating [78] and message boards [57]. This tendency has been verified for many online social networks in [80]; this recent work also highlights a transition from degree assortativity to disassortativity in the popular Chinese social network platform Wealink.

The evolution of the correlation degree coefficient we observe for Wikipedia, plotted in Figures 7.8 and 7.9, exhibits a different trend, reaching highly negative values that tend to decrease over time in absolute value. One particular reason for the disassortative mixing of Wikipedia community can be found in the tendency of more involved authors to interact with new inexpert users, correcting and improving their contributions, rather than to collaborate with each other on the same articles. Global inequality of contribution between users collaborating on a same article has been shown to be positively correlated with article quality [11]; this social dynamics can probably be considered one fundamental feature of the Wikipedia community, that has characterized it since the beginning, with a strong concern of the most involved users for the content of the whole encyclopedia. The trend of disassortative mixing in the cumulative network mirrors the one of the clustering coefficient, which also decreases in absolute value as the network grows in size and density, establishing connections also between people who authored the same pages

in different periods.

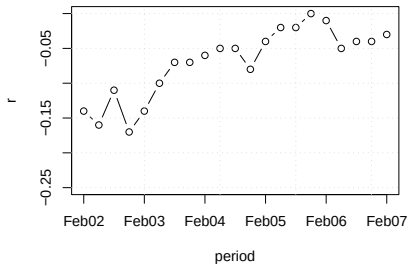


Figure 7.8: Trend of assortative mixing by degree r in the non-cumulative network.

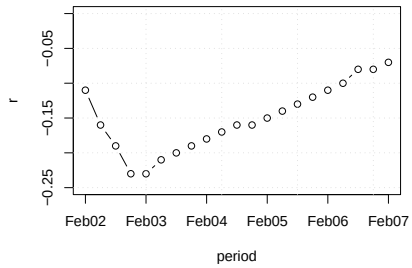


Figure 7.9: Trend of assortative mixing by degree r in the cumulative network.

7.4.2 Centrality measures

In literature, several metrics of centrality have been proposed to study the position and the influence of individuals in a network. Beside the first and simplest centrality metric of *Degree*, which just expresses the total number of collaborators of a user, i.e. the *communication activity*, others have been defined to investigate particular properties of nodes.

Betweenness is a measure based on the number of times a node occurs in the shortest path between other pairs of nodes. It is computed for node n as:

$$betweenness(n) = \sum_{i,j} \frac{|p_{inj}|}{|p_{ij}|}$$

where, for each pair of nodes (i, j) in the network, p_{ij} are all the shortest paths between them, and p_{inj} are the ones passing from node n. The idea is that the more betweenness a node scores the more influence it will have on information flow in the whole network, a sort of *control of communication* [52]. Nodes with high values of betweenness make the spreading of cutting-edge knowledge easier; in the case of Wikipedia it could be policies and best practices. Removing such nodes typically leads to the increase of the shortest path length between nodes [180].

Closeness of a node (n) is the inverse of the average length of the shortest paths to other nodes (m) in the network (p_{nm}) [22]; given N

Table 7.3: Number of *Administrators*, *Bots* and *Registered users* in the top 100 nodes according to different metrics for the cumulative network

	Admins	**Bots**	**Registered**
edit count	50	27	23
edit longevity	72	6	22
degree	70	10	20
betweenness	65	12	23
closeness	73	10	17
eigenvector	30	6	64

the number of nodes:

$$closeness(n) = \frac{N-1}{\sum_m p_{nm}}$$

This metric expresses the capability of a node to get in touch with new ideas over the network, i.e. *independence of information* [52].

Eigenvector is a metric based on degree. It corresponds to the values of the first eigenvector of the network adjacency matrix [22]. Centrality of each node is so evaluated proportionally to the sum of the centrality of nodes it is connected to.

We computed the centrality of each user, according to the different metrics mentioned. Eigenvector centrality is the only metric which was computed on the weighted network, where each edge connecting two authors is weighted according to the number of articles they co-authored.

To sketch the composition of the group of the most influential authors, we counted the number of administrators, bots and registered users appearing in the top 100 positions in the rankings according to the different centrality metrics and to the measures of edit count, or number of edits done, and to edit longevity, described in Section 7.3.1.

As it can be noted in table 7.3, all centrality metrics tend to produce results comparable to the edit longevity, on which the networks are based; an interesting exception is represented by the eigenvector, which tends to strongly penalize administrators; this result can be interpreted as a consequence of the tendency of administrators to interact preferentially with the most inexperienced users. A high number of bots emerges in the edit count ranking, but this presence is strongly reduced with edit longevity and network centrality metrics; this is probably a symptom of the high number of small edits performed, and of the scarce interactions

with other users, which characterize many bots.

7.4.3 Removing admins, bots and stars.

Given the structure of the network and its disassortative mixing, of particular interest can be the experiment of removing some classes of very central users and studying the resulting network. Table 7.4 reports the macroscopic features of the cumulative networks obtained removing various classes of users, compared with the original network.

As a first experiment we have removed administrators and bots; more precisely, we have removed all the nearly 1300 users who have been elected administrators before February 2007, and the 76 users that we have identified as bots. The peculiar role that these classes of users play inside Wikipedia is witnessed by the remarkable change in the network that is caused by their removal. As it can be noted, the size of the network decreases significantly: in fact, as we are not considering isolated individuals, almost 20 000 nodes get disconnected from the rest of the network after these special users are removed; also the giant component size percentage decreases. Mean distance and diameter increase, remarking the role of hubs that administrators and bots were playing, whereas clustering coefficient grows as the hierarchical structure of the network is partly broken with the removal of these stars. Finally, the assortative mixing coefficient increases, though the network keeps being disassortative.

The same phenomenons are observed after removing other very central authors; we removed the 1000 and 5000 nodes with the highest betweenness, obtaining the results shown in Table 7.4. Individuals having highest betweenness are the ones that are more often on the shortest path between pairs of users in the network, and correspond to Wikipedians directly connected with many heterogeneous authors; removing these hubs it is easier to understand the sub standing structure of the network. The assortative mixing coefficient gets positive after removing the 1000 most central users; the size of the network and of the giant component get smaller as many users get disconnected, but no other connected component exceeds the size of 10 nodes. After removing 5000 authors, the size of the giant component is reduced from about 180 thousand to 115 thousand nodes, meaning that more than one third of the users were connected to the rest of the network only through these stars.

Table 7.4: Macroscopic features of the cumulative network constructed removing some classes of users: Admins and Bots (**AB**), top **1000** and **5000** users having highest betweenness. Also values for the original network are reported (**none**).

	N	**⟨k⟩**	**G%**	**C**	**d**	**D**	**r**
none	183710	22.4	96.7	0.06	3.41	10	-0.07
AB	168716	13.3	94.8	0.04	3.80	11	-0.04
1000	158956	10.0	93.0	0.05	4.24	12	0.04
5000	134802	5.2	85.9	0.10	5.44	17	0.09

Table 7.5: Macroscopic features of the cumulative networks for categories Pharmacology, Botany and Comics.

	Pharm.	**Botany**	**Comics**
# of nodes N	5814	6500	11559
mean degree $\langle k \rangle$	11.22	9.79	11.35
giant comp G%	89.8	89	92.8
clustering coeff. C	0.25	0.19	0.11
mean distance d	3.59	3.5	3.57
diameter D	11	10	10
assortativity r	-0.05	-0.1	-0.06

7.4.4 Study of subcommunities

A further analysis can be performed concentrating on particular semantic areas of Wikipedia: by considering only a subset of articles it is possible to study the community of users that are active on a specific domain. Isolating a semantic area inside Wikipedia is not a trivial task. Several approaches, illustrated in Chapter 6, have been proposed to assign an article to a Wikipedia article with a certain proportion; as here we just want to identify sets of semantically related pages in order identify subcommunities active on specific topics, we rely on the approach of isolating a few well delimited lower level categories, and manually cleaning their subtrees, excluding unrelated branches. A more detailed description of this process is described in Section 6.2. We chose three categories of comparable size, from different domains: Botany, Pharmacology and Comics.

As shown in Table 7.5, the networks seem to share some macroscopic

Table 7.6: The 15 users with highest betweenness in the cumulative network for category Botany. Also the position in the global network betweenness ranking is reported for each user, together with the role.

rank	betw.	username	role	global rank
1	4392847	MPF	Admin	129
2	3050933	AntiVandalBot	Bot	1
3	2035231	Tawkerbot2	Bot	2
4	1496233	Gdrbot	Bot	41
5	603624	Wetman	Registered	23
6	395980	Ahoerstemeier	Admin	11
7	389615	JoJan	Admin	1145
8	386907	Grstain	Registered	137
9	386820	DanielCD	Admin	173
10	379741	PDH	Registered	141
11	360715	Pekinensis	Registered	1921
12	344995	VivaEmilyDavies	Registered	1915
13	311965	Badagnani	Registered	99
14	291674	Tawkerbot4	Bot	7
15	291491	Pollinator	Admin	803

features of the global one: one very large connected component, short diameter and short average distances. Clustering coefficient C reaches values remarkably higher than the ones observed over the global network. This is especially true for categories Botany and Pharmacology; the lower value observed for the category Comics seems to reflect the more occasional and sparse nature of contributions, with respect to scientific disciplines where more specific expertise on particular topics is required.

Regarding sociometric stars, we observe the prevalence of some of the same "superusers" that also emerged in the global network, but also of other users that seem to have reached a very high centrality only inside a particular area. As an example, Table 7.6 shows the first 15 users for betweenness in the Botany cumulative network. For each user also the role and the position in the betweenness ranking for the global network are reported: this information points out a certain heterogeneity in the composition of the core of the most central users in category Botany,

and offers an interesting measure of the different areas of influence of users. By discarding global stars it is possible to have an idea of the most influential contributors who focused on a given area.

7.5 Conclusions and future work

In this work we have proposed a scalable method to extract a co-author network from a wiki's revision history, based on the idea of selecting only the main contributors of a page as its authors, and we have applied it to analyze the social structure and dynamics of the English Wikipedia author community.

The results mark a considerable difference with respect to most of the scientific collaboration networks: very low values of mean distance and diameter, a quite low and decreasing clustering coefficient, and disassortative mixing by degree. We find evidence of a strong centralization of the network around some stars, a considerable nucleus of very active users, who seem to be omnipresent. The high centrality of sociometric stars points out the key role that the "elite" continue playing in the community of Wikipedia, despite the rapid growth of the number of common users. The disassortativity of the networks is a signal that the most active contributors tend to interact with the less experienced users, spreading over the whole wiki, rather than to collaborate with each other. In this continuous relationship between the core and the periphery of the community can perhaps be found one of the constituting characteristics of the Wikipedia community.

We have also shown how the community working on a particular semantic area of the wiki can be studied; the networks constructed for some categories tend to share the main features of the global ones, with some variations; in scientific disciplines we observe higher clustering, and lower values of disassortativity. An extensive study including a higher number of categories could reveal interesting patterns. By filtering out the "superusers" which have a very high centrality over the global network, it is possible to identify the most influential authors in a specific area.

The study presented in this chapter offers many directions for further investigation. Recent studies have pointed out a *plateau effect* in the growth of Wikipedia, which after 2007 seems to have significantly slowed down [168]; it would be interesting to inspect how the dynamics and the structure of the network have evolved. Different metrics could be used to compute author contribution; for example, a measure based

only on new words added could help giving prominence to the authors who provide new content. For a more complete comprehension of collaboration patterns, the coauthor networks could be compared with the explicit interactions between users in discussion pages. Finally, the bipartite network of authors and articles is a kind of *folksonomy*; it can be studied as a precious source of emergent semantics, and contrasted with the category graph. The fact that each wiki page corresponds to an encyclopedic entry, and to an entity in the Semantic Web as discussed in Section 2.1.2, makes this perspective particularly promising.

8 Network and tree structure of Wikipedia discussion pages

8.1 Introduction

Behind the most visible part of Wikipedia, i.e. the articles, there are non-encyclopedic pages which are used for coordination, discussion and personal communication among the Wikipedians. While the growth of the encyclopedia in terms of numbers of articles, edits and active users has slowed down in the last years, activity on these pages has kept increasing at a higher rate [168, 177, 166, 158]. In this chapter, which is based on the work published in [108], we focus on this less visible side of Wikipedia, in order to shed light on communication patterns that accompany collaboration on the project.

Unlike other online discussions which often only satisfy the purpose of entertainment or of defending one's point of view, the discussion on Wikipedia article talk pages has a clear objective, i.e. to reach consensus and improve the content of the corresponding article. In many cases these pages can considerably outgrow the corresponding article in size. For example, the talk page associated to the article 'Barack Obama' contains more than 22 000 comments, which is more than the 17 500 edits done to the article itself. In Wikipedia there are also talk pages associated to registered users; these pages are somehow complementary to the article discussion pages, and are used for personal communication between the Wikipedians, as a sort of public in-box.

Communications in Wikipedia are part of a complex social system, where users are involved in the project to different extents and with different roles, either explicit or implicit. Several studies have focused on the analysis of the content of talk pages [177, 166, 158], while some researchers have studied the correlation between the presence of discussion and article quality [94, 92]. Kittur et al. [94] also identify the number of edits done to a discussion page as the best indicator of conflict on the corresponding article. Though, little attention has so far been devoted to the study of interaction patterns emerging from discussions on these

pages. We claim that the study of these interactions on a large scale can reveal essential features of the Wikipedia community and its social structure.

In this chapter we offer an extensive analysis of talk pages associated to articles and to users. We analyse the structural properties of the networks derived from interactions on these pages; in particular, the study of directed degree assortativity allows us to reveal specific patterns in the communications between the Wikipedians, which differ from the results obtained for the discussion board Slashdot.

To characterize the discussions on article talk pages, we analyse their structure according to different measures, such as depth and size of the discussion threads. We show how chains of direct replies between pairs of users can be an interesting indicator of particularly contentious topics, and we report a listing of the most discussed articles, according to different criteria. Finally, we investigate the relationship between structural properties of the discussions and the corresponding semantic areas.

8.2 Experimental Setup

8.2.1 Dataset description

Wikipedians have well defined policies and conventions for using discussion pages [181]. *Article talk pages* are seen as places for coordination where editors can discuss on improving articles rather than express their personal views on a subject. On the other hand, *user talk pages* are used for communication between users, as a sort of public inbox. They can be used to give suggestions or ask questions to users, which are immediately notified by the system if new messages arrive on their own user talk page.

From a technical point of view talk pages have always remained simple wiki pages; however, their usage over the years has evolved according to the community requirements. With the growth of their content due to both their massive use in controversial pages and the need to store old discussions, many of them are now divided into many subpages containing not only archives but often also specific thematic discussions.

Another community need was to have more structured discussions, and to this end the wiki markup has been exploited to organize discussions as in a forum, with paragraph and subparagraph titles containing comments, which are normal paragraphs of text that can be indented with a special syntax, indicating the reply relationships. Comments are signed by their authors and dated at the instant of insertion; the wiki

text parser on Wikipedia provides a shortcut to sign a comment with the correct signature, including a link to the corresponding user talk page, and the date detected at the instant of insertion. In case of anonymous users (not registered), the signature reports the IP number instead of the user name.

The depicted system looks functionally similar to a typical discussion forum, but at the same time it leaves a special degree of freedom that is a guarantee to best fit the community needs which continuously evolve over time.

For this study we rely on a complete dump of the English Wikipedia from March 12th, 2010. As old comments are always archived and never removed, for the analysis of discussions we can just take into account the last version of the dump. However, to be able to contrast discussions with article edit activity, we needed to parse the whole dump including complete edit history of each page.

8.2.2 Data preparation and cleaning

The freedom which is left to editors in the usage of discussion pages is a drawback for our analysis. Each entire discussion page is stored in the Wikipedia dump as one only block of wiki text. There is no structure surrounding a single comment nor an always valid schema to detect its start and end. Moreover, signing and dating comments is left to users so, though there are bots in charge of automatically adding missing signatures and dates, many comments are unsigned. To extract the thread structure with comment indentation, signatures and dates, we had to deal with many different explicit and implicit conventions, changing over years and not always attended by users.

We grouped all subpages associated to discussion on the same article or user into a single unit of content, to analyze discussions aggregated by their topic. Comment identification is not straightforward; we chose to recognize a comment as a fragment of text followed by the pattern of a signature. This pattern is expressed by a flexible regular expression designed to detect a user name and eventually a date at the end of a line, following one of many different possible conventions. The end of an unsigned comment can be recognized with formerly used separators, or with the start of a new thread or of a new comment nested on a different level. We discarded all the talk pages with no signed comments.

In Table 8.1 we report some basic quantities of the data extracted from the Wikipedia dump. The percentages of unsigned comments encountered in article and user talk pages are very similar, 16.1% and 14.7%,

#articles	3 210 039	
#edits of article pages	402 851 686	
#articles with talk page (ATP)	871 485	(27.1%)
#total comments in ATP	11 041 246	
#signed comments in ATP	9 421 976	(85.3%)
#anonymous (ip signed) comments in ATP	1 000 824	(9.1%)
#users who comment articles	350 958	(2.8%)
#registered users	12 651 636	
#user talk pages (UTP)	1 662 818	(13.1%)
#comments in UTP	13 670 980	
#signed comments in UTP	13 493 254	(98.7%)
#anonymous (ip signed) comments in UTP	2 009 658	(14.7%)

Table 8.1: Basic quantities of the data analysed.

while the percentages of anonymous comments are 10.0% and 1.3%, respectively.

8.3 Wikipedia discussion networks

There are no explicit networks between users in Wikipedia. In order to study the patterns of communication and discussion, we extracted three implicit directed networks according to different types of interactions between users:

Article reply network (reply-NW) direct replies between users in article discussion pages.

User talk network (talk-NW) direct replies in user talk pages.

Wall network (wall-NW) personal messages posted on the talk page of another user.

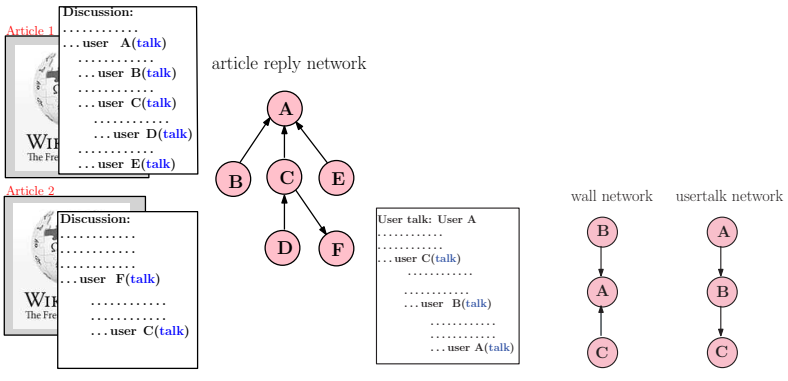
In all networks we discard anonymous users, as IP numbers are not reliable identifiers. In Figure 8.1 we schematically explain the idea of how these networks are constructed. In the article reply network (Figure 8.1(a)) we establish a directed edge from a user B to a user A if B has written at least one comment indented under an entry by user A in any article discussion page. The user talk network (Figure 8.1(b)) is analogously defined, but based on the comments in user talk pages, while the wall network establishes a link from user B to user A if user B has written something on the talk page of user A.

8.3.1 Basic network parameters

In Table 8.2 we report some macroscopic features of the networks. Besides the dimension in terms of number of nodes, we report for each network the number of nodes having at least one outgoing or one ingoing link, respectively. Interestingly, these quantities vary significantly from network to network. In the article discussions around 90% of users have replied to at least one user, while nearly 60% have received replies. On the contrary, in the other two networks almost all (wall: 98.4%, talk: 90.3%) users have at least one incoming link, while many do not have any outgoing link. In particular, only less than one over ten users in the wall network have written on another user's talk page. This result is due to the presence of *welcomers*, users and bots who write a welcome message on the wall of newly registered users; this also explains the larger size of the wall network, which contains many users who are not active. For this reason also reciprocity is lower in the wall network.

8.3.2 Network comparison

For the network comparison we modify the metric proposed in [169], based on Jaccard coefficient of the link overlap, such that it takes values within $[0, 1]$ and it is independent of the network densities. It measures the co-occurrence of links between users in different social networks. More formally, let $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ be two networks with the same set of nodes V, and with the sets of edges E_1 and E_2,



(a) Article reply NW (b) User talk and wall NWs

Figure 8.1: Schema of the networks construction.

variable	reply-NW	talk-NW	wall-NW
#nodes with edges	204 017	114 258	1 861 702
w. in-degree ≥1	121 682	103 147	1 832 168
w. out-degree≥1	182 881	63 334	177 331
#edges M	1 489 734	852 065	4 412 212
size of giant comp.	88.5%	89.2%	96.3%
mean distance	4.10 (0.75)	3.86 (0.69)	4.06 (0.68)
maximal distance	15	11	12
Clustering coeff.	0.083 (0.19)	0.053 (0.16)	0.035 (0.14)
mean in-degree	7.30 (29.6)	7.46 (32.8)	2.37 (15.75)
mean out-degree	7.30 (35.2)	7.46 (41.5)	2.37 (103.79)
network density	$3.58 \cdot 10^{-5}$	$6.53 \cdot 10^{-5}$	$1.27 \cdot 10^{-6}$
reciprocity	0.44	0.45	0.15

Table 8.2: Global measures of the Wikipedia discussion and talk network. Values within parenthesis indicate stdv.

respectively. Then,

$$C_{jaccard} = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2|} \cdot \frac{\max(|E_1|, |E_2|)}{\min(|E_1|, |E_2|)},$$

where we denote as $|\cdot|$ the number of elements in the set.

	reply-NW	talk-NW	wall-NW
reply-NW	1	0.11	0.09
talk-NW	0.11	1	0.35
wall-NW	0.09	0.35	1

Table 8.3: Jaccard coefficient between the networks.

We compare the three networks and present the results in Table 8.3. In the wall network, we discarded all users who are not present in any of the two reply networks, to keep only active users. As it could be expected, the highest overlap is between the two networks extracted from user talk pages. Though, it is important to point out that these two networks capture different kinds of interaction, and none of the two is subsumed by the other. The overlap between edges in these networks of personal communications and the one extracted from articles is of about 10%, indicating substantially different networks.

8.3.3 Directed assortativity by degree

As explained in Section 7.4.1, assortativity by degree is a basic measure of diversity in networks, quantifying the tendency of nodes to link with

other having similar number of edges [132].

To compute degree assortativity accounting for the direction of edges, we rely on the Assortativity Significance Profile (ASP), a novel approach proposed in [51].

To account for the direction of edges we first need to introduce a notation in which $\alpha, \beta \in \{in, out\}$ are used to index the degree type. For each edge e, i_e^α and j_e^β are the α- and β-degree of the source node and the target node. A set of four assortativity measures can now defined using the Pearson correlation:

$$r(\alpha, \beta) = \frac{E^{-1} \sum_e [(i_e^\alpha - \bar{i^\alpha}) * (j_e^\beta - \bar{j^\beta})]}{\sigma^\alpha \sigma^\beta} \tag{8.1}$$

where E is the number of edges in the network, $\bar{i^\alpha} = E^{-1} \sum_e i_e^\alpha$, and $\sigma^\alpha = \sqrt{E^{-1} \sum (i_e^\alpha - \bar{i^\alpha})^2}$; $\bar{j^\beta}$ and σ^β are analogously defined.

The absolute values of the correlation coefficients are dependent on the network degree distribution; to compute statistical significance, degree-degree correlations are compared with an ensemble of 100 randomised networks with the same in- and out-degree sequence as the original network. The statistical significance of each correlation $r(\alpha, \beta)$ is computed as the difference between the value observed in the original network and its average in the randomised ensemble $r_{rand}(\alpha, \beta)$ in units of the standard deviation $\sigma_{rand}(\alpha, \beta)$:

$$Z(\alpha, \beta) = \frac{r(\alpha, \beta) - r_{rand}(\alpha, \beta)}{\sigma_{rand}(\alpha, \beta)} \tag{8.2}$$

Values of $|Z| > 2$ can be considered statistically significant. As a last step, as Z scores are dependent on the network size, they are normalised by defining an Assortative Significance Profile (ASP) for each network: $ASP(\alpha, \beta) = Z(\alpha, \beta) / [\sum_{\alpha,\beta} Z(\alpha, \beta)^2]^{1/2}$.

Results are reported in Table 8.4, together with the results for the reply network extracted from the Slashdot discussion board. We added these results to be able to compare discussions in Wikipedia with discussions from another large online community. The Slashdot reply network contains about 80 000 users and 1 million connections; for a detailed description of this dataset, see [57]. The resulting ASPs of networks are plotted in Figure 8.2.

None of the assortativity values computed for the talk network is statistically significant ($|Z > 2|$). The wall network exhibits dissortativity according to all four measures, which points a general tendency

	type	r	$\langle r_{rand} \rangle$	σ_{rand}	Z	ASP
Slashdot	(out, in)	-0.035	-0.046	0.00059	**17.677**	0.329
	(in, out)	-0.016	-0.033	0.00063	**26.613**	0.495
	(out, out)	-0.015	-0.038	0.00063	**35.843**	0.667
	(in, in)	-0.027	-0.040	0.00057	**24.143**	0.449
Reply	(out, in)	-0.025	-0.019	0.00063	**-8.629**	-0.485
	(in, out)	-0.018	-0.018	0.00061	0.062	0.003
	(out, out)	-0.027	-0.018	0.00062	**-14.179**	-0.797
	(in, in)	-0.015	-0.019	0.00063	**6.385**	0.359
Talk	(out, in)	-0.045	-0.030	0.00998	-1.526	-0.655
	(in, out)	-0.025	-0.026	0.00753	0.109	0.047
	(out, out)	-0.042	-0.028	0.00848	-1.753	-0.753
	(in, in)	-0.028	-0.029	0.00894	0.076	0.033
Wall	(out, in)	-0.126	-0.087	5.1e-5	**-769.81**	-0.936
	(in, out)	-0.039	-0.020	0.00020	**-93.51**	-0.114
	(out, out)	-0.063	-0.043	7.5e-5	**-26.04**	-0.317
	(in, in)	-0.061	-0.039	0.00026	**-84.21**	-0.102

Table 8.4: Directed assortativity results for the three networks of Wikipedians and for the Slashdot reply network. Values in bold are significant ($|Z > 2|$).

of socially active users to interact preferentially with users having few connections. In particular, the remarkably high value observed for the (out, in)-assortativity shall be imputed to the activity of users and bots who massively welcome new registered users writing on their personal talk page.

The reply network extracted from Wikipedia articles shows to be (out, out)- and (out, in)- dissortative, with significant Z-scores, pointing out a marked tendency of users having many outgoing links to interact preferentially with users having few connections, and vice versa. On the contrary, (in, in) assortativity is positive, revealing a tendency of users to reply more often to others having a similar *in*-degree. We do not observe this pattern in the Slashdot reply network, which is assortative according to all four measures[1]. The difference could be due to the peculiar nature of Wikipedia article talk pages, where discussions are usually aimed at taking decisions about content production according to the community policies. While a high *out*-degree is the result of an active behaviour, replying to many users, a high *in*-degree is achieved getting many replies from different users. These two measures seem to capture two distinct

[1] Note that this is different from what has been reported previously in [57], where no comparison with randomised networks was taken into account.

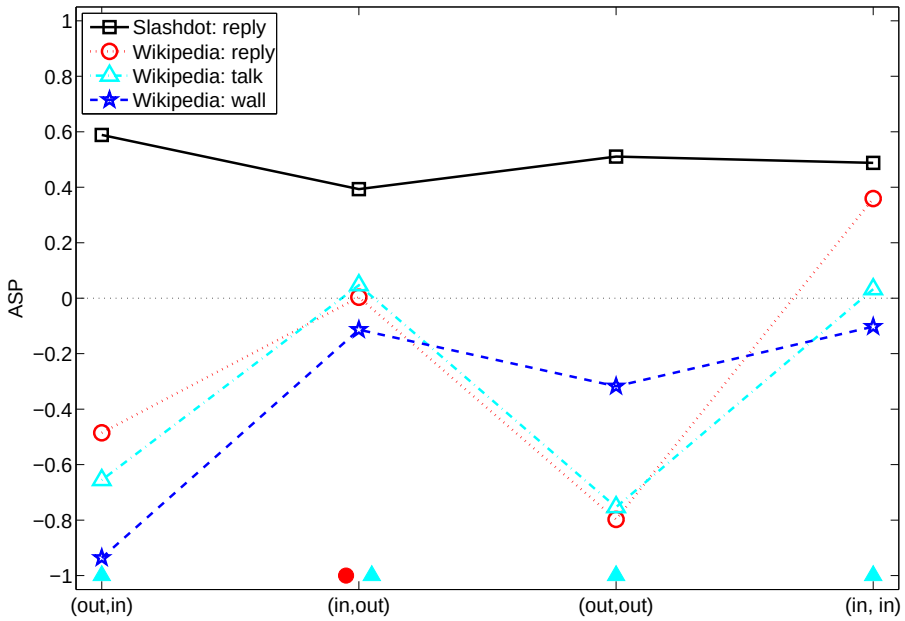


Figure 8.2: Comparison of the ASP score for the comment networks of Slashdot and Wikipedia. In cases where $|Z| < 2$ and the ASP score is not significant the corresponding ASP is marked with the corresponding symbol at the figure bottom.

characteristics of Wikipedia influential users, resonating with a distinction between *hubs* and *authorities*. The Wikipedians who reply to many other users in article talk pages tend to interact mostly with users having few connections, i.e. newbies and inexperienced users, while the Wikipedians who receive replies from many users tend to interact preferentially with each other.

8.3.4 Mixing by k-coreness

As a further step to investigate the social structure of Wikipedia communication networks, we rely on the notion of k-core decomposition [160] to detect the core of individuals all highly interconnected with each other. In simple words, the *k-core* of a graph is the maximal subgraph in which each vertex is adjacent to at least k other nodes of the subgraph. Formally, for a given graph $G = (V, E)$ we define a subgraph $H = (C, E|C)$ induced by the set $C \subseteq V$. We call H as *k-core* of G if and only if for every vertex $v \in C$: $deg_H(v) \geq k$, and H is the maximum subgraph with such property. Here $deg_H(v)$ is the degree of the vertex v in the

Table 8.5: Results of the k-shell decomposition and of k-coreness assortativity for Slashdot and for the three networks of Wikipedia users, compared with 100 randomized equivalents.

net	measure	value	avg_{rand}	σ_{rand}	Z
Slashdot	k_{max}	99	93.9	0.51	**9.85**
(dense)	k_{max}-shell	771	771.5	72.4	-0.01
	r_k	0.027	-0.0179	0.0010	46.13
Article	k_{max}	66	65.0	0.17	5.84
(dense)	k_{max}-shell	930	1369.7	126.4	-3.48
	r_k	0.016	0.0142	0.0291	0.07
Usertalk	k_{max}	97	94.0	0.6	5.00
(dense)	k_{max}-shell	649	568.9	59.7	1.34
	r_k	0.010	-0.0221	0.0013	24.59
Wall	k_{max}	129	255.9	0.72	-175.16
(dense)	k_{max}-shell	898	706.7	67.6	2.83
	r_k	-0.394	-0.1628	0.0005	-488.51
Wall	k_{max}	44	41.5	0.52	4.76
(sparse)	k_{max}-shell	361	568.6	79.3	-2.62
	r_k	0.006	0.0051	0.0016	0.81

subgraph H in the case of undirected graph A vertex v has *shell index* k if it belongs to k-core but not to $(k+1)$-core. The set of all vertices with shell indexes equal k is called *k-shell*.

To study k-cores we first turn the directed networks into undirected ones. From each directed network we derive an *undirected dense* network, containing a link between each pair of users connected in at least one direction. Given the particular structure of the wall network, with many asymmetric links directed from few "welcomers" to many inactive registered users (as discussed in Section 8.3.1), it is useful to also consider the *undirected sparse* network, derived from the original one keeping only reciprocal edges. To measure the tendency of nodes in the core of the network to mix with peripheral ones, we introduce a measure of *assortativity by k-coreness* r_k, defined as the correlation between the k-index of each pair of nodes connected by a direct edge. This measure is defined analogously to degree correlation in (8.1), replacing degree with the k-index of each node. It must be noted that in this case we are dealing with undirected networks, so each edge has to be counted twice, for both directions. To assess the statistical significance of the results we contrast each undirected network with an ensemble of 100 random networks

having the same degree sequence.

Table 8.5 shows for each network the maximum k-index, the size of the corresponding k-shell, and the assortative mixing by k-coreness. Besides the values observed in each real network, Table 8.5 reports also the average value and the standard deviation over the randomized equivalents, and the Z-score of the original network's value.

The significantly high values of k_{max} in the real networks tell of the higher level of organization observed. Values of k_r vary substantially over networks, revealing different interaction patterns. In the user talk network, which captures personal conversations, users in the core tend to interact with each other. This result is similar to the one observed for replies in Slashdot. In contrast, the article reply network exhibits neutral k-coreness assortativity, as the small value of the Z-score tells there is no significant difference with respect to the randomly generated networks. This resonates with the results observed for degree assortativity in Section 8.3.3 in that diversity in interactions seems to be higher in conversations about articles than in personal ones. A further study of directed k-coreness assortativity could allow for a more complete comparison with the ASPs obtained for degree-degree correlations.

In the wall undirected dense network we find a marked tendency to dissortativity; however, the sparse network exhibits neutral assortativity. While in the other cases the values observed for sparse and dense networks are in line with each other (though the latter are here omitted, for reasons of space), in the wall network the difference is substantial, due to the huge presence of asymmetric relationships. Given the nature of many of these asymmetric connections, which do not represent real interactions, we consider the sparse network a more reliable model.

8.4 The discussion trees

In this section we focus on the shape and size of interactions in the discussion pages on Wikipedia. These interactions can be modelled in the form of discussion trees, where the root node corresponds to the article page on Wikipedia, and child nodes to comments or structural elements of the discussion pages. Unlike other online discussions, for example observed in blogs [123] or at Slashdot [57], the Wikipedia discussion pages do not only consist of comments, which represent the actual interactions between the users, but may also contain many structural elements such as a separation of the total number of the comments into several sub-pages, or titles and subtitles to organize the content of the discussion. We model

each of these different elements as a separate node in the discussion tree. The structure of the tree reflects the hierarchy of the pages. A reply to a comment is a child node of this comment and comments which are placed below a title or a new page are child nodes of the corresponding structural node unless they reply to another comment. Note that there can be several nested levels of structural nodes as there can be several levels of titles and subtitles.

To help in the comprehension of the following analysis we show in Figure 8.3 one of these trees. It corresponds to the Wikipedia article "Presidency of Barack Obama" (represented by the red node in the centre of the radial tree) and contains 989 nodes of which 254 are structural nodes (in blue in Figure 8.3) and the rest comments. Note that this article is different from one just on "Barack Obama" cited earlier.

8.4.1 Size of the discussions

Out of the approx. 3.2 million articles in our dataset nearly 870 000 have an associated discussion page (about 27%), which contain more than 9.4 million signed comments, created by more than 350 000 users (See Table 8.1 for details). As one would expect the distribution of the number of comments and users among the different articles follows heavy tailed distributions as shown in Figure 8.4 (left).

Although more than 85% of all articles have discussions with only 10 or less comments, there is still a considerable number of articles (approx. 15 000) with more than 100 comments and 826 discussions even contain more than 1000 comments. The largest discussions reach more than 30 000 comments involving several thousand users.

What are the shapes of these discussions? The example of Figure 8.3 suggest that we can basically identify two patterns: comments that are placed directly after a structural node (a headline etc.) and do not receive any replies; and large chain-like subthreads of comments, containing a sequence of replies between several users. Only occasionally a comment receives more than just one reply in these subthreads, which contain about 65.4% of all comments. The remaining 36.6% of the comments correspond to isolated unanswered comments who themselves are also not replying to another comment.

To investigate those subthreads further we focus on the number of such chain-like subthreads that can be found per discussions and on their lengths. To formalise the concept of chains we consider only subthreads where exactly two users interact subsequently. We define as n-chains (or simply chains, if not stated otherwise) all sequences which include

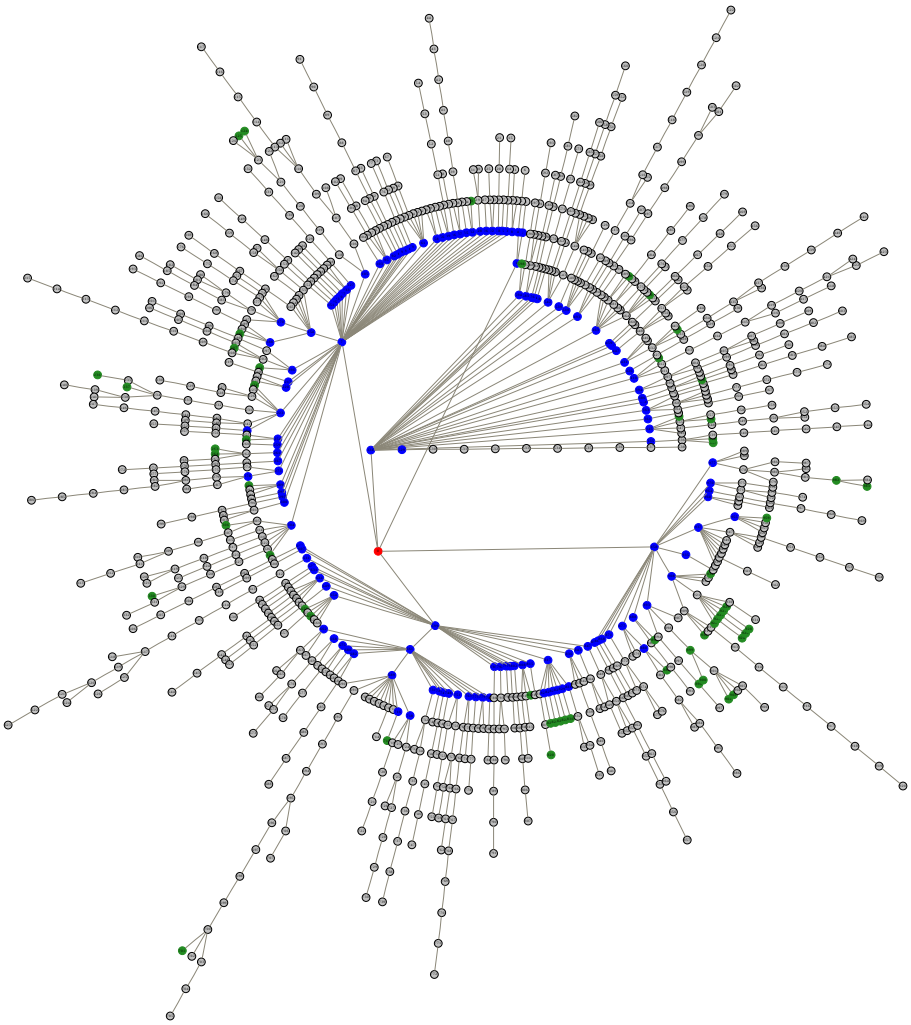


Figure 8.3: The structure of the discussion page of "Presidency of Barack Obama" (generated with Graphviz). Blue nodes are structural, green nodes are unsigned comments.

at least three comments. So the shortest n-chains are of the form $A \leftarrow B \leftarrow A$ where A and B are two different users and the arrows indicate a reply of B to A and a back-reply from A to B. These chains can grow considerably. The longest chain in our dataset is of length 31 in the discussion page of "Central Bosnia Canton"[2]. Figure 8.4 (right) shows the number of chains of different lengths. Given a chain of length k,

[2]See `http://en.wikipedia.org/wiki/Talk:Central_Bosnia_Canton`

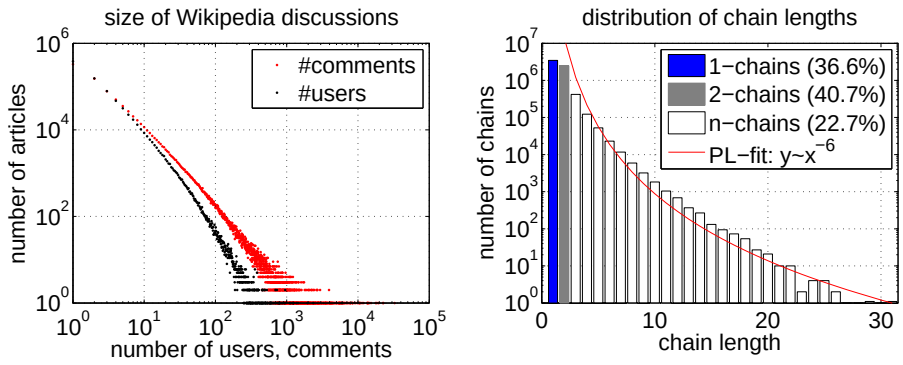


Figure 8.4: Distributions of the number of **(left)** comments and users per discussion page, and **(right)** discussion chains of different lengths in the dataset. Percentages indicate the proportion of comments in the different types of chains.

we do not count any sub-part of it as a chain. We find that 22.7% of all comments form part of chains of length of at least 3 (n-chains), while 40.7% belong only to 2-chains (are either parent or reply but not part of an n-chain). The remaining comments are isolated (1-chains). The distribution of the number of n-chains with different lengths follows roughly a power-law with exponent 6 (red line in Figure 8.4).

In Figure 8.5 we show the distribution of the number of n-chains in the discussion pages. Again we find a heavy tailed distribution, with some discussions containing several thousand chains. The distribution can be fitted with both, a power law distribution (with cut-off) with exponent 2.23 and a truncated log-normal distribution. Both fits are not rejected by a Kolmogorov Smirnov test (see figure legend for the corresponding p-values).

The number of chains gives us an idea about how many times a controversy arises in the article discussion. In Table 8.6 we list the top 20 articles according to this measure and compare it with the total number of comments, registered users who comment and edits of the corresponding article page. The numbers in parenthesis indicate the rank number of the article when ordering by the corresponding variable. Note that nearly in all cases the number of comments is much larger than the number of actual edits on the corresponding article page. Most of the topics in the list represent also highly disputed subjects in realarticles life, either due to political, ideological, religious or scientific disputes. They seem to be a good barometer of contentious discussion topics in the last

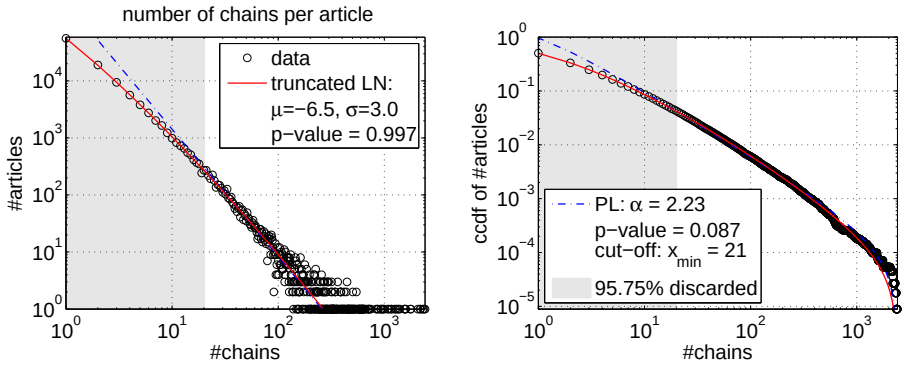


Figure 8.5: Number of discussion chains per discussion page.

few years.

We furthermore list as well the depth of these discussion trees which we will treat separately in the next subsection.

8.4.2 Depth of the discussions

We study the depth of the discussions using two measures: its maximal depth (i.e., the level of the deepest comment in the discussion tree) and its h-index, a balanced depth measure which was introduced in [57] on the base of the h-index of [76]. To calculate the h-index of the tree structure we count the number of nodes per level, starting at level one (the root node) and descending the tree. The h-index of the tree is then the maximum level, for which the corresponding number of nodes is greater or equal to the level number (and all previous levels fulfil the same condition). Note that we also consider structural nodes in these calculations.

In Figure 8.6 we show the distributions of the maximal depths and the h-index for all discussion and only for the ones with more than 100 nodes (in the insets). We observe that the two distributions have a similar shape but slightly different modes.

The deepest discussion can be found about the article “Liberal democracy” . It reaches a depth of 42, while its h-index is only 12. The maximal h-index is observed for “Anarchism” (h-index = 20).

In the Columns 6 and 7 of Table 8.6 we present the depth and h-index of the 20 most discussed articles. From the rank-values within parenthesis we can conclude that the rank by h-index is closer to the rank by number of chains than the rank of the maximal depth of these discussions. The maximal depth is very sensitive to the presence of isolated

#	Title	chains	comments		users		h-index		max. depth		edits	
1	Intelligent design	2413	22454	(3)	954	(13)	16	(20)	20	(358)	9179	(53)
2	Gaza War	2358	17961	(6)	607	(47)	19	(2)	27	(28)	11499	(29)
3	Barack Obama	2301	22756	(2)	2360	(2)	18	(6)	21	(245)	17453	(6)
4	Sarah Palin	2182	19634	(4)	1221	(9)	17	(10)	25	(56)	12093	(24)
5	Global warming	2178	19138	(5)	1382	(5)	17	(10)	20	(358)	14074	(15)
6	Main Page	2065	32664	(1)	5969	(1)	15	(34)	22	(169)	4003	(674)
7	Chiropractic	1772	13684	(13)	243	(389)	18	(6)	29	(17)	6190	(204)
8	Race and intelligence	1764	13790	(12)	410	(126)	17	(10)	24	(74)	7615	(100)
9	Anarchism	1589	14385	(9)	496	(76)	20	(1)	28	(22)	12589	(19)
10	British Isles	1556	12044	(16)	576	(56)	17	(10)	23	(113)	4047	(658)
11	CRU[a] hacking incident	1551	11536	(17)	474	(88)	17	(10)	20	(358)	2346	(2364)
12	Jesus	1397	17916	(7)	1239	(7)	13	(119)	16	(1383)	17081	(7)
13	Circumcision	1356	10469	(21)	436	(113)	17	(10)	26	(42)	7354	(117)
14	Homeopathy	1323	13509	(14)	516	(68)	17	(10)	25	(56)	6902	(151)
15	George W. Bush	1281	15257	(8)	1969	(3)	14	(65)	18	(676)	32314	(1)
16	September 11 attacks	1250	13830	(11)	1244	(6)	16	(20)	26	(42)	11086	(30)
17	Evolution	1165	13404	(15)	942	(16)	13	(119)	23	(113)	9780	(44)
18	Catholic Church	1162	14104	(10)	620	(43)	15	(34)	18	(676)	14082	(14)
19	Cold fusion	1098	8354	(29)	359	(174)	15	(34)	20	(358)	4320	(557)
20	2008 South Ossetia war	1075	10596	(20)	853	(20)	17	(10)	23	(113)	9930	(43)

Table 8.6: Several structural measures of the top 20 Wikipedia discussions ordered by the number of n-chains (length ≥ 3).

[a] Climatic Research Unit

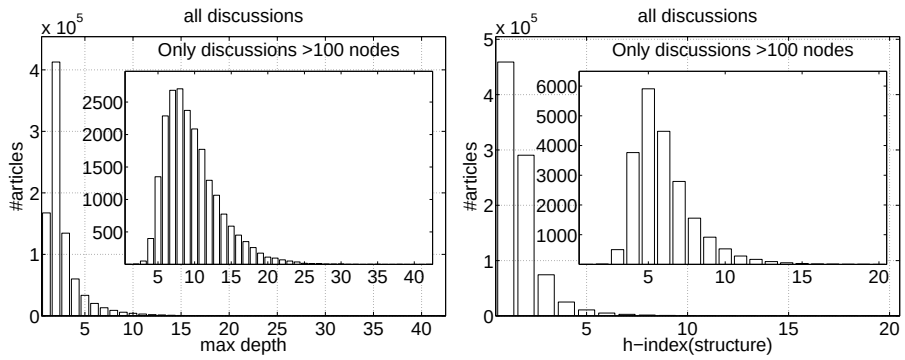


Figure 8.6: Distribution of the max. depth (left) and the h-index (right) of the article discussion pages.

individual discussions between a small number of users, which can reach considerable depths while not being representative for the entire discussion. The h-index overcomes this limitation and we will use it therefore in the next section to account for the depth of the discussions.

8.4.3 Comparison with categories

In this section we investigate whether the structure of the article discussions differs for the different topic categories of Wikipedia articles. For the assignment of articles to macro-categories we rely on the results obtained with the approach based on the shortest path, described in Section 6.3.1, where each article is assigned with different proportions to different macro-categories.

Structural differences between the categories

We investigate the proportion of pages with discussions among the different categories. We use the category weights for this calculation. So, if an article has a 60% weight in category A and a 40% weight in category B it contributes with the corresponding proportions to these two categories. The black bars in Figure 8.7 show these general proportions of articles with discussions (the corresponding %-value is written on the right y-axis). We observe a large heterogeneity among the different categories. "Geography and Places" and "History and events" are nearly of the same size and account together for more than 46% of all discussion pages. The next two categories "Culture" and "People" account for another 20%. Interestingly, if we restrict this analysis to only the top 1% or 0.1% of the

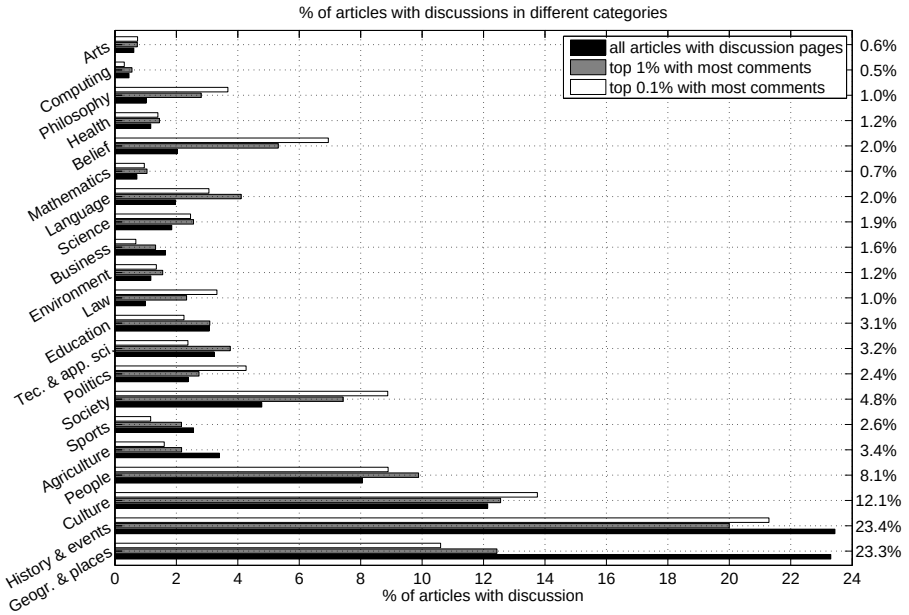


Figure 8.7: Proportion of articles within different categories for all discussion pages

article discussion pages (according to their number of comments) we observe rather different distributions (indicated by the grey and white bars in Figure 8.7). The "Geography and Places" proportion decays to a only half of its original value, while some other categories like "Belief", "Society", "Philosophy" or "Law" and "Politics" approximately double their share.

This change seems to indicate that these categories, although less frequent among the entire set of articles with discussions, attract more than an average number of comments. To investigate this further we calculated for every category (using the category weights of every article) the weighted average of several structural metrics presented in the previous subsections. The outcome of this analysis is presented in the form of two cross-plots in Figure 8.8. To verify the results we performed a bootstrap test ($N = 1000$) and depict the 95% interval of the observed average value with grey areas. In the cases where the area is absent the symbol size is larger than the corresponding confidence interval.

We can extract several interesting conclusions from the two cross-plots. From the right cross-plot we see a clear correlation between the average values of the depth of the discussion measured with the h-index and the

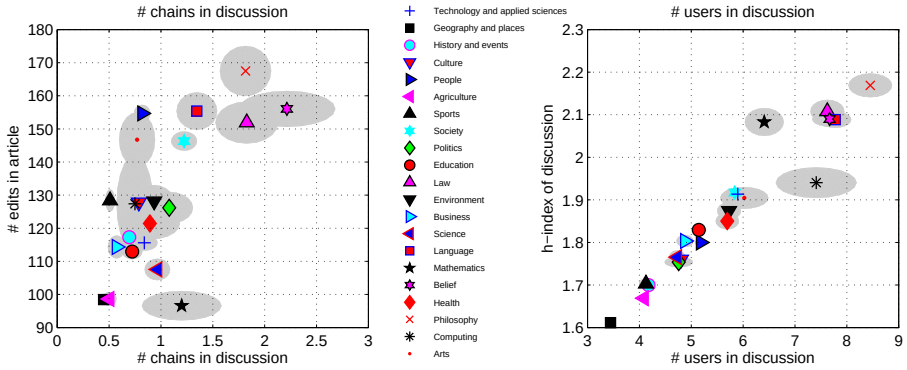


Figure 8.8: Differences between Categories: (left) the average number of chains in the article discussion pages vs. the average number of article edits, (right) the average number of users versus the h-index of the tree structure. Gray areas indicate (when larger than the corresponding symbol size) the 95% confidence interval.

number of users who left at least one comment in the discussion. As we discussed above the category "Geography and Places" is on average the one with the flattest discussions. On the other hand, the top categories according to these measures are "Philosophy", "Law", "Language" and "Belief". These categories trigger, on average, the deepest discussions involving the largest amount of users. They are also the top 3 categories if we use the number of discussion chains as a measure as can be seen from the left sub-figure of Figure 8.8 where we compare the number of edits with the number of discussion chains. This seems to agree with [91], where the amount of conflict in different macrocategories is estimated according to a page-level heuristics, and "Belief" and "Philosophy" are identified as the most contentious categories. While the authors find the lowest level of conflict in category "Mathematics", from Figure 8.8 we can observe that, although this category has the lowest average number of edits per article, it still reaches a considerable amount of debate in the form of discussion chains. Some article categories like "People", "Arts" or even "Sports" are on average less discussed than their number of edits would suggest.

In the right subplot we also observe two outlayers to the otherwise quite correlated averages. The categories "Computing" and "Mathematics" obey a different behaviour. Discussions on "Computing" articles involve more users, than their average h-index would suggest, while ar-

ticles of the "Mathematics" category have the opposite behaviour. They are deeper but involve less users than expected.

To summarise, we observe quite different relations between the discussion structures and the number of edits among the different categories. We have also found that the size and shape of the discussion varies significantly among the different categories. A more detailed analysis involving the study of individual user activities in the different categories might shed further light on whether these differences are community or content based.

8.5 Related Studies

Kittur et al. [94] describe the growth of the *hidden side* of Wikipedia, comprehending talk pages and all Wikipedia-specific pages deputed to conflict and coordination. A longitudinal study to investigate the role of coordination in the improvement of Wikipedia articles' quality is described in [92]; positive improvements are observed as effect of discussion only on small and "young" pages. Presence of discussion is just measured in terms of size of the article talk pages, and patterns of communications are not considered.

Few researchers focused on Wikipedia talk pages to study social relationships between users. Crandall et al. [36] investigate the interplay between social ties, modelled as interactions in discussion pages, and similarity, modelled as editing activity on the same articles. They find evidence of a feedback effect between the two phenomenons. Only a 15% overlap is found between the graph of social interactions and the one of similarity; interestingly, properties of the social network reveal to be better predictors of future behaviour than properties of the similarity network. A qualitative description of different patterns corresponding to different social roles in Wikipedia is offered in [55], based on the local network of personal communications around single users.

As reported in Section 7.2, several studies have focused on extracting networks of interactions from the edit history of a wiki, and Chapter 7 is dedicated to the study of Wikipedia co-author network. Both the approaches of extracting co-authorship networks and edit networks are complementary to the one described in this chapter, and could be integrated to contrast direct replies in discussion pages with relationships emerging from editing activity.

To the best of our knowledge, in this work we propose the first extensive study of Wikipedia as a discussion space. Similar analysis have been

performed for blogs [123] and online discussion boards [57]. A generative model for the structure of the discussion threads analysed here has been presented in [58]. The model parameters show important structural differences between the discussions in Wikipedia and those of other social media platforms.

8.6 Conclusions

In this chapter we have focused on Wikipedia talk pages to detect structural patterns of interaction which accompany collaboration on the project. The study of directed assortativity reveals the existence of a characterizing pattern in the reply network extracted from article discussion pages. Users who reply to many other users tend to reply preferentially to inexperienced users, while the Wikipedians who receive comments by many users are more likely to interact with each other. This pattern is not observed in the Slashdot reply network neither in personal conversations in Wikipedia. We suggest that it derives from the nature of discussion on article talk pages, focused on solving issues and controversies according to codified community policies, and reflects the existence of different social roles among the more influential users.

The study of shape and size of the discussions at the article level reveals interesting patterns and suggests some metrics to characterize different talk pages. The number of chains of direct replies between pairs of users seems to be a good indicator of contentious discussion topics, while h-index of the tree is a compact measure to capture the actual depth of a discussion. We found evidence of significant differences in discussions from different semantic areas. For example, discussions about Mathematics tend to reach a much higher depth than the number of users involved and of edits in the corresponding articles would suggest.

This work proposes a first insight into Wikipedia as a space of discussion and offers many directions for improvement and for future investigation. The comparison of users' behaviour in the different networks (and maybe also in networks derived from interactions in article editing) could help in the identification of social roles. A more fine grained analysis involving the time-stamps of the comments may allow for a better understanding of social dynamics on a temporal dimension, and to detect contentious topics during a certain interval of time.

9 Conclusions

Along the chapters of this thesis we have studied different online communities, focusing on the mechanisms, conventions and interaction patterns which are at the basis of the production of knowledge in the Social Web.

While the Semantic Web community has proposed standards and tools for the representation of knowledge in structured format, most online communities appear as still far and sometimes reluctant to adopt these solutions.

In order to elaborate solutions to bridge these two worlds, the starting point of this work has been a careful observation of spontaneous dynamics and emergent trends, with a special attention toward those novel conventions and solutions adopted by communities to face issues related to the production and the representation of knowledge on the Web, like hashtags in microblogging, machine tags in social bookmarking, policies for categorizing articles or managing conflict in Wikipedia.

Starting from this point, we have drawn a map of issues concerning different levels of structured semantics, in combination with different kinds of activities, investigating emergent trends and dynamics, and proposing solutions to improve current systems, pursuing a synthesis between habits and needs of the users, and effective tools and standards for knowledge representation.

The first bunch of contributions of this work concern the use of identifiers in the Social Web. In Chapter 3 we have investigated the nature and usage of hashtags in microblogging, as a community solution to the problem of fragmentation. We have proposed an approach based on information retrieval techniques to represent tags as virtual documents, and we have introduced several metrics to evaluate them as strong identifiers, according to some desired properties. We have studied *specificity* as the difference between the usages of a hashtag and of the corresponding word with no hash (*nontag*), *consistency* by means of the entropy of the vectorial representation of tags, and *stability over time* observing the evolution of the usage of a tag, both in terms of the co-occurring words and of the community adopting it. We have illustrated how these measures can help to automatically distinguish different kinds of tags, and

we have shown through manual evaluation that combining these metrics it is possible to achieve improvements in the identification of labels that can be mapped to real world entities.

While the main application of this approach is harvesting information produced by users to identify new valuable categories and labels in order to extend knowledge bases and improve search engines, in Chapter 4 we have somehow followed the same road in the inverse direction, borrowing structured knowledge from existing ontologies to enrich navigation in a folksonomy.

To this end we have mapped Delicious tags onto the WordNet lexicon; while only 8% of the tags in our dataset could be mapped to concepts from WordNet, this percentage is largely over 50% for the most common tags. To correctly map each occurrence of a tag, we have performed disambiguation considering as context the other keywords employed by the community to tag the same resource. This allowed us to perform a combination of bottom up and top down approaches to categorization, through the creation of hybrid hierarchies, made of related tags selected on the basis of usage but organized in the structure provided by WordNet noun hierarchy. We have illustrated through an expressly designed browser extension how the folksonomy navigation interface can be extended by integrating this kind of hierarchies to improve the possibilities of browsing in the tag space.

A strong advantage of this approach is that it does not imply any overhead for the users, as the enrichment is performed in an entirely automated way; the interface for tagging items remains unchanged in its simplicity and immediacy of use, while the navigation interface is improved thanks to knowledge from an external source. In this aspect resides also the main limitation which inherently characterizes this approach: though the richness and generality of the ontology adopted (WordNet), this offers by definition one single rigid categorization scheme, which may in some cases be inadequate, when other categorization criteria would better suite the mind of the users of a given community.

To address this limitation, a largely pursued solution which we have not experimented in this thesis consists in deriving ontologies from folksonomies [122]. As a relevant case of alternative approach, where collaboration reaches the level of hierarchies, and the community is involved in defining a shared conceptual structure, in Chapter 6 we have studied the Wikipedia category structure.

Beyond semantics at the levels of individual entities, categories and hierarchies, we have also focused on how to add expressive power to ex-

isting tools at the level of relations. Starting from the observation of the many different and creative usages of tags in current systems, we developed a general model based on named graphs to allow users specify the relationship intercurring between the tagged resource and the sign used to tag (Chapter 5). Thanks to its flexibility, we have shown how the NiceTag ontology can be leveraged to represent tag actions from different existing systems in one unified model, bridging folksonomies to the Semantic Web; without adding overhead for the users, annotations in NiceTag can be enriched through automatic processing based on background knowledge, allowing for searching and browsing of tags at a more fine-grained granularity. On the other hand, NiceTag provides a robust and flexible framework for the development of richer interfaces, to capture and aggregate more specific knowledge from users.

In tagging systems the contributions of many users are aggregated to create folksonomies, as collective classifications of content; still, the contribution of each user is well distinguishable, and for this reason we bet on named graphs as a key promizing approach to easily determine the provenance of an annotation in the Social Semantic Web, without limiting expressive power, and avoiding the burden of RDF reification.

Taking a step forward, in the second part of thesis we have shifted from collective to collaborative systems, where users work on a common artifact, and the focus is not on many single users, but on one community. To shed light on the social dynamics which rule collaboration in the Social Web, we have focused on the emerging paradigm of wikis, and on the English Wikipedia as a case study.

The wiki approach to the creation of concept hierarchies gives place in Wikipedia to a very rich and tangled graph, where many different criteria and points of view are represented in one only category structure. The main limitation of this kind of categorization is that it lacks of coherence and consistency, so for example transitivity of hierarchical relations is not guaranteed. This makes difficult some tasks that would be straightforward in a taxonomic hierarchy, such as identifying the subgraph corresponding to a given semantic area.

In fact, the simple and natural approach of isolating a category and including iteratively its subcategories, which we tested in Section 6.2, has shown to be viable only in the context of some well-delimitable categories, like Botany or Comics, which seem to be characterized by the presence of a coherent shared categorization criterion. In these cases this method can give highly accurate outcome with just a little effort to manually remove few branches. On the contrary, in most cases the graph is too dense and

tangled to be able to isolate a specific semantic area. Categories in areas like "Society", "Culture", "Politics" and "Religion" are too interconnected with one another and have too many links which respond to different and sometimes contradicting criteria, to be able to make sense of the category structure as a coherent hierarchy.

To deal with the tangledness and fuzziness of the whole graph we have proposed two alternative approaches, based on different heuristics to assign each article to one or more predetermined macro-categories. The first one consists in an improvement of the algorithm proposed by Kittur et al. [91], based on the identification of the closest macro-categories to the categories assigned to each article; we have modified the algorithm to account for the orientation of category assignments, obtaining a slight improvement. Also the second technique proposed, based on the probability of reaching each macro-category starting from an article and following a random path in the graph, achieved promising results. The results obtained mining the category graph have been instrumental to study the community of Wikipedia along a semantic dimension, i.e. analyzing patterns of collaboration and interaction over different semantic areas.

In Chapter 7 we have focused on collaboration patterns, proposing a general method to extract a co-authorship network from a wiki's revision history. Leveraging a metric of edit longevity, we select as authors of a page the users who contributed to most of its accepted content; in this way we are able to scale up to the size of the English Wikipedia and represent it as a collaboration network.

Co-authorship in wikis is based on implicit interaction; complementarily, in Chapter 8 we have studied explicit coordination and communication mechanisms, extracting and analyzing the network of interactions in article talk pages and user talk pages.

Macroscopic analysis of both co-authorship and discussion networks depict Wikipedia community as a *small world*, where the vast majority of active users are connected to each other through some path a few steps long. The most interesting feature emerging in both analyses is disassortativity, i.e. a tendency of users having many interactions to connect preferentially with poorly connected users, and vice versa. This tendency emerges in all networks and suggests the tight relationship between the core and the periphery of the community as a characterizing pattern of Wikipedia, due to the effort of the more involved users to help, correct and address newbies and occasional editors. For discussions, disassortativity is more marked in article talk pages than in user

talk pages, while it is not observed in other online discussion spaces, confirming that this dynamics is probably associated to the production of content in Wikipedia .

As reply relationships in discussions are directed, they allow for a more fine-grained analysis of assortativity, which reveals that Wikipedians replying to many users interact more with inexperienced users, while the ones who receive replies from many users tend to interact preferentially with one another. This significant difference in the behaviour points out the existence of two distinct profiles of very active Wikipedians, and gives an important hint for a more in-depth study of social roles in Wikipedia.

On the other side, the analysis of the co-authorship network's evolution points out the strong and increasing centralization around some stars, an elite of very active users who are only in part administrators, and seem to spread all over the wiki to diffuse experience and policies, and never leave newbies working alone on some article.

In both contexts we have contrasted patterns of interaction with semantic areas, to investigate the existence of correlations between the social and the semantic dimensions. By restricting the analysis of co-authorship to only a set of pages, we have shown how it is possible to identify the most influential authors in a given semantic area according to distinct centrality metrics; shedding light on this information hidden in revision history can be useful for many tasks, like expert finding. Comparison of the subcommunities active around specific topics has pointed out remarkable differences in structural properties such as the clustering coefficient, showing the existence of diversified work organization patterns for different topics.

For the analysis of discussion patterns at the article level we have introduced some metrics such as the h-index of the discussion tree as a robust and compact measure of depth, and the number of discussion chains, or consecutive replies between a pair of users, as an indicator of conflictive conversations. The aggregation of data by semantic areas reveals significant differences in the discussion on different topics, such as a marked presence of contentious conversations on "Philosophy", "Law", "Religion and belief systems", and of deeper threads in "Mathematics", or the little amount of discussion about "People", "Arts" and "Sports" with respect to the number of edits done to the corresponding articles.

Results from both scenarios show evidence that the topic which is object of collaboration has an influence on social dynamics; knowledge of these differences can be precious for the design of collaborative platforms.

Interdependence between the *social* and the *semantic* dimension is a

fascinating characteristics of social Web systems, which we have observed all along this thesis: wiki pages, as well as micro-blogging hashtags and folksonomy tags, play the function of aggregators in both dimensions, bridging the semantics that they convey and the communities active around them; similarly, hierarchies and relations which best suite the needs of the users are the ones emerging from the community itself, either collaboratively defined or emerging from the users' behaviour. Social patterns and semantics are inherently bound and can hardly be studied and understood separately; on the contrary, in this thesis we have shown how combining them can be a key factor for the comprehension and for the improvement of existing systems, as well as for the design of new paradigms.

Online communities have been shown to be extremely creative in finding effective solutions to the problems that they encounter; however, the need for more powerful tools to represent knowledge appears as straightforward from the analysis of the usage of current Web applications. The Semantic Web will hardly represent a viable solution as long as it will merely reflect a top down approach; in this work we have instead privileged adherence to the state of art in current social Web applications, and to the observed habits and needs of the users. The solutions proposed, based on the improvement of current systems by means of little steps of lightweight structured information, can represent an effective approach to ferry online communities towards the adoption of richer explicit semantics. How to bring a bottom up approach at higher levels of structured knowledge, enabling automatic reasoning and definition of formal ontologies, is a crucial question, which constitutes an open and challenging issue for future research.

Bibliography

[1] D. Activism. Social media and the revolutions. *PersPectives*, page 80, 2011.

[2] B. Adler, L. de Alfaro, I. Pye, and V. Raman. Measuring author contribution to the wikipedia. In *WikiSym 2008: International Symposium on Wikis*, Porto, Portugal, September 2008. ACM Press.

[3] M. Agosti and N. Ferro. A formal model of annotations of digital content. *ACM Trans. Inf. Syst.*, 26(1):3, 2007.

[4] H. Al-Khalifa and H. Davis. Exploring the value of folksonomies for creating semantic metadata. *International Journal on Semantic Web and Information Systems*, 3(1):12–38, 2007.

[5] H. S. Al-Khalifa and H. C. Davis. Towards better understanding of folksonomic patterns. In *HT '07: Proceedings of the 18th conference on Hypertext and hypermedia*, pages 163–166, New York, NY, USA, 2007. ACM Press.

[6] R. Almeida, B. Mozafari, and J. Cho. On the evolution of wikipedia. In *Int. Conf. on Weblogs and Social Media*, 2007.

[7] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 971–980, New York, NY, USA, 2007. ACM.

[8] P. Analytics. Twitter study-august 2009. *Online: http://www.pearanalytics. com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009. pdf (Geraadpleegd 28-02-2011)*, 2009.

[9] S. Angeletou. Semantic enrichment of folksonomy tagspaces. In A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, and K. Thirunarayan, editors, *International Semantic Web*

Conference, volume 5318 of *Lecture Notes in Computer Science*, pages 889–894. Springer, 2008.

[10] G. Antoniou and F. Harmelen. Web ontology language: Owl. *Handbook on ontologies*, pages 91–110, 2009.

[11] O. Arazy and O. Nov. Determinants of wikipedia quality: the roles of global and local contribution inequality. In *Proceedings of CSCW*, New York, NY, USA, 2010.

[12] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. *The Semantic Web*, pages 722–735, 2007.

[13] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590 – 614, 2002.

[14] D. Barrett. *MediaWiki*. O'Reilly Media, 2008.

[15] D. Beckett. Redland notes - contexts. http://www.redland.opensource.ac.uk/notes/contexts.html, 2003.

[16] G. Begelman, P. Keller, and F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, pages 15–33. Citeseer, 2006.

[17] D. Benz, M. Grobelnik, A. Hotho, R. Jaschke, D. Mladenic, V. D. P. Servedio, S. Sizov, and M. Szomszor. Analyzing tag semantics across collaborative tagging systems. *Dagstuhl Seminar 08391 ? Working Group Summary*, 2008.

[18] T. Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

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

[20] S. Bindelli, C. Criscione, C. Curino, M. Drago, D. Eynard, and G. Orsi. Improving search and navigation by combining ontologies and social tags. In *On the Move to Meaningful Internet Systems: OTM 2008 Workshops*, pages 76–85. Springer, 2008.

[21] R. Biuk-Aghai. Visualizing co-authorship networks in online wikipedia. *Communications and Information Technologies*, pages 737–742, 2006.

[22] P. Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182, 1987.

[23] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. Network analysis of collaboration structure in wikipedia. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 731–740, New York, NY, USA, 2009. ACM.

[24] J. Breslin, A. Harth, U. Bojars, and S. Decker. Towards semantically-interlinked online communities. *The Semantic Web: Research and Applications*, pages 500–514, 2005.

[25] C. H. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 625–632, New York, NY, USA, 2006. ACM.

[26] S. L. Bryant, A. Forte, and A. Bruckman. Becoming wikipedian: Transformation of participation in a collaborative online encyclopedia. In *GROUP '05: Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 1–10, New York, NY, USA, 2005. ACM Press.

[27] B. Butler, E. Joyce, and J. Pike. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1101–1110. ACM, 2008.

[28] G. Cabanac, M. Chevalier, C. Chrisment, and C. Julien. Collective annotation: Perspectives for information retrieval improvement. In D. Evans, S. Furui, and C. Soulé-Dupuy, editors, *RIAO*. CID, 2007.

[29] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler. Named graphs, provenance and trust. In *14th Int. Conference on World Wide Web WWW*, pages 613–622, New York, NY, USA, 2005. ACM.

[30] M. Castells. Communication, power and counter-power in the network society. *International Journal of Communication*, 1(1):238–266, 2007.

[31] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic grounding of tag relatedness in social bookmarking systems. *The Semantic Web-ISWC 2008*, pages 615–631, 2008.

[32] C. Cattuto, V. Loreto, and L. Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences*, 104(5):1461, 2007.

[33] E. H. Chi and T. Mytkowicz. Understanding the efficiency of social tagging systems using information theory. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 81–88, New York, NY, USA, 2008. ACM.

[34] D. Cosley, D. Frankowski, L. Terveen, and J. Riedl. Suggestbot: using intelligent task routing to help people find work in wikipedia. In *Proc. of IUI*, 2007.

[35] C. Cotta and J. J. M. Guervós. Where is evolutionary computation going? a temporal analysis of the ec community. *Genetic Programming and Evolvable Machines*, 8(3):239–253, 2007.

[36] D. J. Crandall, D. Cosley, D. P. Huttenlocher, J. M. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *Proceedings of KDD*, 2008.

[37] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal Complex Systems*, 1695(1695), 2006.

[38] C. V. Damme, M. Hepp, and K. Siorpaes. Folksontology: An integrated approach for turning folksonomies into ontologies. In *Bridging the Gap between Semantic Web and Web 2.0 (SemNet 2007)*, pages 57–70, 2007.

[39] S. Dann. Twitter content classification. *First Monday*, 15(12-6), 2010.

[40] G. Devereux and W. La Barre. *From anxiety to method in the behavioral sciences*, volume 3. Mouton The Hague, 1967.

[41] D. Diderot, J. Le Rond d'Alembert, and O. Diodati. *Encyclopédie*, volume 1. Hermann, 1976.

[42] E. Dumbill. Tracking provenance of rdf data. Technical report, ISO/IEC, 2003.

[43] F. Echarte, J. Astrain, A. Córdoba, and J. Villadangos. Ontology of folksonomy: A new modeling method. *Proceedings of Semantic Authoring, Annotation and Knowledge Markup (SAAKM)*, 2007.

[44] M. Elliott. Stigmergic collaboration: The evolution of group work. *C Journal*, 9(2), 2006.

[45] D. Eynard. *A Virtuous Cycle of Semantics and Participation*. PhD thesis, Politecnico di Milano, 2009.

[46] J. Farina, R. Tasso, and D. Laniado. Automatically assigning Wikipedia articles to macrocategories. In *Proc. of Hypertext'11*, 2011.

[47] U. Farooq, T. G. Kannampallil, Y. Song, C. H. Ganoe, J. M. Carroll, and L. Giles. Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. In *GROUP '07: Proceedings of the 2007 international ACM conference on Conference on supporting group work*, pages 351–360, New York, NY, USA, 2007. ACM.

[48] C. Fellbaum. *WordNet – An Electronic Lexical Database*. MIT Press, 1998.

[49] M. Ferron and P. Massa. Collective memory building in Wikipedia: the case of North African revolutions. In *Proc. of WikiSym 2011*, 2011.

[50] A. Forte and A. Bruckman. Scaling consensus: Increasing decentralization in wikipedia governance. In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, pages 157–157. IEEE.

[51] J. G. Foster, D. V. Foster, P. Grassberger, and M. Paczuski. Edge direction and the structure of networks. *PNAS*, 107(24):10815–10820, 2010.

[52] L. C. Freeman. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35–41, 1977.

[53] F. Gandon, V. Bottolier, O. Corby, and P. Durville. Rdf/xml source declaration, w3c member submission. http://www.w3.org/Submission/rdfsource/, 09 2007.

[54] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438:900–901, 2005.

[55] E. Gleave, H. T. Welser, T. M. Lento, and M. A. Smith. A conceptual and operational definition of 'social role' in online community. In *Proc. of HICSS 2009*, 2009.

[56] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32(2):198–208, 2006.

[57] V. Gómez, A. Kaltenbrunner, and V. López. Statistical analysis of the social network and discussion threads in slashdot. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 645–654, New York, NY, USA, 2008. ACM.

[58] V. Gómez, H. J. Kappen, and A. Kaltenbrunner. Modeling the structure and evolution of discussion cascades. In *Proc. of Hypertext 2011*, 2011.

[59] S. Goodman. *Social Media: The Use of Facebook and Twitter to Impact Political Unrest in the Middle East through the Power of Collaboration*. PhD thesis, California Polytechnic State University, 2011.

[60] T. Groza, S. Handschuh, K. Moeller, G. Grimnes, L. Sauermann, E. Minack, C. Mesnage, M. Jazayeri, G. Reif, and R. Gudjonsdottir. The nepomuk project-on the way to the social semantic desktop. *Proceedings of I-Semantics*, 7:201–211, 2007.

[61] T. Gruber. Ontology of folksonomy: A mash-up of apples and oranges. *International Journal on Semantic Web Information Systems*, 3(1):1–11, 2007.

[62] R. M. R. Guha and R. Fikes. Contexts for the semantic web. In *Int. Sem Web Conf, ISWC*, 2004.

[63] A. Halfaker, A. Kittur, R. Kraut, and J. Riedl. A jury of your peers: quality, experience and ownership in wikipedia. In *Int. Sym. Wikis*, 2009.

[64] H. Halpin and V. Presutti. An ontology of resources: Solving the identity crisis. 5554:521–534, 2009.

[65] H. Halpin, V. Robu, and H. Shepard. The dynamics and semantics of collaborative tagging. In *Proceedings of the 1st Semantic Authoring and Annotation Workshop (SAAW'06)*, 2006.

[66] H. Halpin and H. Thompson. Web proper names: Naming referents on the web. In *Proceedings of The Semantic Computing Initiative Workshop at the World Wide Web Conference*. Citeseer, 2005.

[67] T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools (i): A general review. *D-Lib Magazine*, 11, Apr 2005.

[68] S. Handschuh and S. Staab. Authoring and annotation of web pages in cream. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 462–473, New York, NY, USA, 2002. ACM.

[69] Y. Hassan-Montero and V. Herrero-Solana. Improving tag-clouds as visual information retrieval interfaces. In *International Conference on Multidisciplinary Information Sciences and Technologies*, pages 25–28. Citeseer, 2006.

[70] P. J. Hayes and H. Halpin. In defense of ambiguity. *Int. J. Semantic Web Inf. Syst.*, 4(2):1–18, 2008.

[71] T. Heath and E. Motta. Revyu.com: A reviewing and rating site for the web of data. pages 895–902. 2008.

[72] J. Heflin, J. Hendler, and S. Luke. Shoe: A knowledge representation language for internet applications. Technical report, University of Maryland, 1999.

[73] A. Hemetsberger. When consumers produce on the internet: the relationship between cognitive-affective, socially-based, and behavioral involvement of prosumers. *The Journal of Social Psychology*, 2003.

[74] M. Hepp. Hypertwitter: collaborative knowledge engineering via twitter messages. *Knowledge Engineering and Management by the Masses*, pages 451–461, 2010.

[75] P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. *Stanford InfoLab Technical Report*, (2006-10):1–5, 2006.

[76] J. E. Hirsch. An index to quantify an individual's scientific research output. *PNAS*, 102(46):16569–16572, 2005.

[77] T. Holloway, M. Bozicevic, and K. Börner. Analyzing and visualizing the semantic coverage of wikipedia and its authors. *Complexity*, 12(3):30–40, 2007.

[78] P. Holme, C. R. Edling, and F. Liljeros. Structure and time-evolution of an internet dating community. *Social Networks*, 26(2):155–174, 2004.

[79] A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme. Bibsonomy: A social bookmark and publication sharing system. In *Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, pages 87–102. Citeseer, 2006.

[80] H. Hu and X. Wang. Disassortative mixing in online social networks. *Europhys. Lett.*, 86(arXiv:0909.0450):18003. 6 p, 2009.

[81] J. Huang, K. Thornton, and E. Efthimiadis. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 173–178. ACM, 2010.

[82] T. Iba, K. Nemoto, B. Peters, and P. Gloor. Analyzing the creative editing behavior of wikipedia editors:: Through dynamic social network analysis. *Procedia-Social and Behavioral Sciences*, 2(4):6441–6456, 2010.

[83] Intellidimension. Rdf gateway - database fundamentals. http://www.intellidimension.com/pages/dfgateway/dev-guide/db/db.rsp, 2003.

[84] S. Johnson. How twitter will change the way we live. *Time Magazine*, 173:23–32, 2009.

[85] J. Kahan and M.-R. Koivunen. Annotea: an open rdf infrastructure for shared web annotations. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 623–632, New York, NY, USA, 2001. ACM Press.

[86] B. P. Kettler, J. Starz, W. Miller, and P. Haglich. A template-based markup tool for semantic web content. In Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, editors, *International Semantic Web Conference*, volume 3729 of *Lecture Notes in Computer Science*, pages 446–460. Springer, 2005.

[87] H. Kim, S. Decker, and J. Breslin. Representing and sharing folksonomies with semantics. *Journal of Information Science*, 36(1):57, 2010.

[88] H.-L. Kim, S. Scerri, J. Breslin, S. Decker, and H.-G. Kim. The State of the Art in Tag Ontologies: A Semantic Model for Tagging and Folksonomies. In *International Conference on Dublin Core and Metadata Applications*, Berlin, Germany, 2008.

[89] M. E. Kipp. @toread and cool : Subjective, affective and associative factors in tagging. In *Proceedings Canadian Association for Information Science/L'Association canadienne des sciences de l'information (CAIS/ACSI)*, 2008.

[90] A. Kittur, E. Chi, B. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World Wide Web*, 1(2):19, 2007.

[91] A. Kittur, E. H. Chi, and B. Suh. What's in Wikipedia?: mapping topics and conflict using socially annotated category structure. In *Proc. of CHI 2009*, 2009.

[92] A. Kittur and R. E. Kraut. Harnessing the wisdom of crowds in Wikipedia: quality through coordination. In *CSCW '08: Proceedings of the ACM 2008 conference on Computer supported cooperative work*, pages 37–46, New York, NY, USA, 2008. ACM.

[93] A. Kittur and R. E. Kraut. Beyond wikipedia: coordination and conflict in online production groups. In *Proceedings of CSCW*, 2010.

[94] A. Kittur, B. Suh, B. Pendleton, and E. Chi. He says, she says: Conflict and coordination in Wikipedia. page 462, 2007.

[95] R. Klamma and C. Haasler. Dynamic network analysis of wikis. In *Proceedings of I-KNOW '08 and I-MEDIA '08*, 2008.

[96] T. Knerr. Tagging ontology - towards a common ontology for folksonomies, 2006. http://tagont.googlecode.com/files/TagOntPaper.pdf.

[97] M. Koivunen. W3c semantic web activity. *Semantic Web KickOff in Finland*, pages 27–41, 2001.

[98] M.-R. Koivunen. Semantic authoring by tagging with annotea social bookmarks and topics. In *Proc. of the 1st Semantic Authoring and Annotation Workshop (SAAW2006)*, 2006.

[99] N. Korfiatis, M. Poulos, and G. Bokos. Evaluating authoritative sources using social networks: An insight from wikipedia. *Online Information Review*, 30(3):252–262, 2006.

[100] E. Kroski. The hive mind: Folksonomies and user-based tagging, December 2005. http://infotangle.blogsome.com/2005/12/07/the-hive-mind-folksonomies-and-user-based-tagging/.

[101] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 591–600, New York, NY, USA, 2010. ACM.

[102] C. Körner, D. Benz, M. Strohmaier, A. Hotho, and G. Stumme. Stop thinking, start tagging - tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th International World Wide Web Conference (WWW 2010)*, Raleigh, NC, USA, Apr. 2010. ACM.

[103] C. Körner, R. Kern, H. P. Grahsl, and M. Strohmaier. Of categorizers and describers: An evaluation of quantitative measures for tagging motivation. In *21st ACM SIGWEB Conference on Hypertext and Hypermedia (HT 2010)*, Toronto, Canada, June 2010. ACM.

[104] D. Laniado, D. Eynard, and M. Colombetti. A semantic tool to support navigation in a folksonomy. In *Proc. of Hypertext'07*, 2007.

[105] D. Laniado, D. Eynard, and M. Colombetti. Using WordNet to turn a folksonomy into a hierarchy of concepts. In *Semantic Web Application and Perspectives - Fourth Italian Semantic Web Workshop*, 2007.

[106] D. Laniado and P. Mika. Making sense of Twitter. In *The Semantic Web – ISWC 2010: 9th International Semantic Web Conference*, 2010.

[107] D. Laniado and R. Tasso. Co-authorship 2.0: Patterns of collaboration in Wikipedia. In *Proc. of Hypertext'11*, 2011.

[108] D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner. When the Wikipedians talk: network and tree structure of Wikipedia discussion pages. In *Proc. of ICWSM*, 2011.

[109] O. Lassila and R. Swick. Resource description framework (rdf) model and syntax. *World Wide Web Consortium, http://www. w3. org/TR/WD-rdf-syntax*.

[110] J. Letierce, A. Passant, J. Breslin, and S. Decker. Understanding how twitter is used to widely spread scientific messages. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, March 2010.

[111] B. Leuf and W. Cunningham. The wiki way: quick collaboration on the web. 2001.

[112] F. Limpens, F. Gandon, and M. Buffa. Helping online communities to semantically enrich folksonomies. In *Web Science 2010: Extending the Frontiers of Society On-Line*, Raleigh, NC, USA, april 2010. http://webscience.org, http://webscience.org.

[113] F. Limpens, A. Monnin, D. Laniado, and F. Gandon. Nicetag ontology: tags as named graphs. In *SNI'9: 1st International Social Networks Interoperability Workshop*, 2009.

[114] R. MacGregor and I.-Y. Ko. Representing contextualized data using semantic web tools. In *Practical and Scalable Semantic Systems (ISWC workshop)*, 2003.

[115] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th international conference on World wide web, April*, pages 20–24. Citeseer, 2009.

[116] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM.

[117] A. Mathes. Folksonomies – cooperative classification and communication through shared metadata, December 2004. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

[118] L. Mcdowell, O. Etzioni, S. D. Gribble, A. Halevy, H. Levy, W. Pentney, D. Verma, and S. Vlasseva. Mangrove: Enticing ordinary people onto the semantic web via instant gratification. In *2nd International Semantic Web Conference*, volume 2870, pages 754–770, 2003.

[119] M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the First Workshop on Social Media Analytics*, pages 71–79. ACM, 2010.

[120] P. D. Meo, G. Quattrone, and D. Ursino. Exploitation of semantic relationships and hierarchical data structures to support a user in his annotation and browsing activities in folksonomies. *Information Systems*, 34(6):511 – 535, 2009.

[121] P. Merholz. Metadata for the masses, 2004. `http://www.adaptivepath.com/ideas/e000361`.

[122] P. Mika. Ontologies are us: A unified model of social networks and semantics. *The Semantic Web–ISWC 2005*, pages 522–536, 2005.

[123] G. Mishne and N. Glance. Leave a reply: An analysis of weblog comments. In *Proc. of WWE 2006*, 2006.

[124] A. Monnin, F. Limpens, F. Gandon, and D. Laniado. Speech acts meet tagging: Nicetag ontology. In *Proc. of I-SEMANTICS '10*, New York, NY, USA, 2010.

[125] A. Monnin, F. Limpens, D. Laniado, and F. Gandon. L'ontologie nicetag: les tags en tant que graphes nommés. 2010.

[126] C. Müller-Birn, J. Lehmann, and S. Jeschke. A composite calculation for author activity in wikis: Accuracy needed. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 1:84–91, 2009.

[127] V. Nastase and M. Strube. Decoding wikipedia categories for knowledge acquisition. In *Proc. of the 23rd national conference on Artificial intelligence - Volume 2*, pages 1219–1224. AAAI Press, 2008.

[128] F. Nazir and H. Takeda. Extraction and analysis of tripartite relationships from Wikipedia. pages 1–13, 2008.

[129] M. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(suppl_ 1):5200–5205, 2004.

[130] M. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3):36122, 2003.

[131] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):404–409, January 2001.

[132] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701+, 2002.

[133] R. Newman, D. Ayers, and S. Russell. Tag ontology, December 2005. http://www.holygoat.co.uk/owl/redwood/0.1/tags/.

[134] T. O'Reilly. What is web 2.0. design patterns and business models for the next generation of software. September 2005.

[135] F. Ortega, J. M. Gonzalez-Barahona, and G. Robles. On the inequality of contributions to wikipedia. In *HICSS '08: Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, page 304, Washington, DC, USA, 2008. IEEE Computer Society.

[136] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[137] C. Pang and R. Biuk-Aghai. A method for category similarity calculation in wikis. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, page 19. ACM, 2010.

[138] A. Passant, T. Hastrup, U. Bojars, and J. Breslin. Microblogging: A semantic web and distributed approach. In C. Bizer, S. Auer, G. A. Grimmes, and T. Heath, editors, *4th Workshop on Scripting for the Semantic Web co-located with ESWC2008*, Tenerife, Spain, June 2008.

[139] A. Passant and P. Laublet. Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China, Apr*, 2008.

[140] A. Passant, P. Laublet, J. Breslin, and S. Decker. A uri is worth a thousand tags: from tagging to linked data with moat. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):71–94, 2009.

[141] S. Patwardhan, T. Pedersen, and S. Banerjee. SenseRelate::TargetWord - A Generalized Framework for Word Sense Disambiguation. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 73–76, Ann Arbor, MI, June 2005.

[142] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet: : Similarity - measuring the relatedness of concepts. In *AAAI*, pages 1024–1025, 2004.

[143] S. Ponzetto and M. Strube. Deriving a large scale taxonomy from wikipedia. In *Proceedings of the national conference on artificial intelligence*, volume 22, page 1440. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.

[144] J. Porter. *Designing for the social web*. New Rider Pr, 2008.

[145] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in wikipedia. In *GROUP '07: Proceedings of the 2007 international ACM conference on Supporting group work*, New York, NY, USA, 2007.

[146] E. Quintarelli. Folksonomies: power to the people. June 2005. http://www-dimat.unipv.it/biblio/isko/doc/folksonomies.htm.

[147] E. Quintarelli, L. Rosati, and A. Resmini. Facetag: Integrating bottom-up and top-down classification in a social tagging system. In *IA Summit 2007*, 2007.

[148] V. V. Raghavan and S. K. M. Wong. A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science*, 37(5):279–287, 1986.

[149] E. Raymond. The cathedral and the bazaar. *Knowledge, Technology & Policy*, 12(3):23–49, 1999.

[150] J. Reagle Jr and L. Lessig. *Good faith collaboration: The culture of Wikipedia*. Mit Pr, 2010.

[151] A. Reinach. The apriori foundations of the civil law. *Aletheia*, 3:1–142, 1983.

[152] W. Rouse, J. Cannon-Bowers, and E. Salas. The role of mental models in team performance in complex systems. *Systems, Man and Cybernetics, IEEE Transactions on*, 22(6):1296–1308, 1992.

[153] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[154] G. Salton. *Automatic Text Processing – The Transformation, Analysis, and Retrieval of Information by Computer*. Addison–Wesley, 1989.

[155] S. Sarjant, C. Legg, M. Robinson, and O. Medelyan. "all you can eat" ontology-building: Feeding wikipedia to cyc. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '09, pages 341–348, Washington, DC, USA, 2009. IEEE Computer Society.

[156] S. Scerri, M. Sintek, L. van Elst, and S. Handschuh. Nepomuk annotation ontology specification 2007. http://www. semanticdesktop.org/ontologies/nao/, 2007.

[157] C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme. Mining association rules in folksonomies. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Žiberna, editors, *Data Science and Classification. Proceedings of the 10th IFCS Conf.*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 261–270, Heidelberg, July 2006. Springer.

[158] J. Schneider, A. Passant, and J. Breslin. A qualitative and quantitative analysis of how Wikipedia talk pages are used. *Proc. of the WebSci 2010*, 2010.

[159] G. Secundus and J. Sillig. *Naturalis historia*. Number v. 1 in Naturalis historia. Teubner, 1831.

[160] S. B. Seidman. Network structure and minimum degree. *Social Networks*, 5(3):269–287, 1983.

[161] S. Sen, S. K. Lam, A. M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. M. Harper, and J. Riedl. Tagging, communities, vocabulary, evolution. In *20th conf. Computer Supported Cooperative*

Work, CSCW, pages 181–190, New York, NY, USA, November 2006. ACM.

[162] J. Shinavier. Real-time# semanticweb in<= 140 chars. In *Proceedings of the Third Workshop on Linked Data on the Web (LDOW2010) at WWW2010*, 2010.

[163] C. Shirky. Ontology is overrated: Categories, links, and tags, 2005. `http://www.shirky.com/writings/ontology_overrated.html`.

[164] J. Sinclair and M. Cardew-Hall. The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 34(1):15, 2008.

[165] M. Sintek and S. Decker. Triple - a query, inference, and transformation language for the semantic web. In *Intl. Sem. Web Conf., ISWC*, 2002.

[166] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Information quality work organization in Wikipedia. *JASIST*, 59(6):983–1001, 2008.

[167] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203 – 217, 2008. World Wide Web Conference 2007Semantic Web Track.

[168] B. Suh, G. Convertino, E. H. Chi, and P. Pirolli. The singularity is not near: slowing growth of Wikipedia. In *Proc. of WikiSym 2009*, 2009.

[169] M. Szell, R. Lambiotte, and S. Thurner. Multirelational organization of large-scale social networks in an online world. *PNAS*, 107(31):13636–13641, 2010.

[170] L. V.-S. Tang, R. P. Biuk-Aghai, and S. Fong. A method for measuring co-authorship relationships in mediawiki. *Procceedings of the 2008 International Symposium on Wikis (WikiSym) Porto, Portugal*, 2008. WikiSym '08, September 8-10, Porto, Portugal. Copyright 2008 ACM 978-1-60558-128-3/08/09.

[171] R. E. Tarjan. Depth-first search and linear graph algorithms. *SIAM J. Comput.*, 1(2):146–160, 1972.

[172] Tecnorati. State of the blogosphere 2010. Online: `http://technorati.com/blogging/article/`

`state-of-the-blogosphere-2010-introduction/` accessed 23-August-2011.

[173] G. Theraulaz and E. Bonabeau. A brief history of stigmergy. *Artificial life*, 5(2):97–116, 1999.

[174] T. Tudorache, N. Noy, S. Tu, and M. Musen. Supporting collaborative ontology development in protégé. *The Semantic Web-ISWC 2008*, pages 17–32, 2008.

[175] O. Valkeapaa, O. Alm, and E. Hyvonen. Efficient content creation on the semantic web using metadata schemas with domain ontology services (system description). *The Semantic Web: Research and Applications*, pages 819–828, 2007.

[176] T. Vander Wal. Folksonomy. *Information Architecture Institute Members Mailing List*, 2004.

[177] F. Viégas, M. Wattenberg, J. Kriss, and F. van Ham. Talk Before You Type: Coordination in Wikipedia. In *Proc. of HICSS 2007*, 2007.

[178] J. Voss. Measuring wikipedia. In *Proceedings International Conference of the International Society for Scientometrics and Informetrics*, Stockholm, Sweden, 2005.

[179] C. Wagner and M. Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proc. of the Semantic Search 2010 Workshop (SemSearch2010)*, april 2010.

[180] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[181] Wikipedia. Help:Wikipedia: The missing manual/collaborating with other editors/communicating with your fellow editors, 2010. Online; accessed 28-October-2010.

[182] Wikipedia. Wikipedia:edit warring, 2011. Online; accessed 21-August-2011.

[183] Wikipedia. Wikipedia:list of policies and guidelines, 2011. Online; accessed 21-August-2011.

[184] G. Wittenbaum, S. Vaughan, and G. Strasser. Coordination in task-performing groups. *Theory and research on small groups*, pages 177–204, 2002.

[185] X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 417–426, New York, NY, USA, 2006. ACM.

[186] K. P. Yee. Critlink: Advanced hyperlinks enable public annotation on the web, 2002. http://zesty.ca/pubs/cscw-2002-crit.pdf.

[187] J. Yu, J. Thom, and A. Tam. Ontology evaluation using wikipedia categories for browsing. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 223–232. ACM, 2007.