CRAWLING BIG DATA IN A NEW FRONTIER FOR SOCIOECONOMIC RESEARCH: TESTING WITH SOCIAL TAGGING

Juan D. Borrero
Estrella Gualda

ABSTRACT

Tags and keywords, freely chosen by users for annotating resources, offer a new way for organizing and retrieving web resources that closely reflect the users' interests and preferences, as well as automatically generate folksonomies. Social tagging systems have gained increasing popularity as a method for annotating and categorizing a wide range of different web resources. They also attract researchers in social sciences because they offer a huge quantity of user-generated annotations that reveal the interests of millions of people. To date, the study using digital trace data methods continues to lack a theoretical framework, particularly in social science research. This paper presents a methodology to use big data from Web 2.0 in social research. At the same time, it applies a method to extract data from a particular social bookmarking site (*Delicious*) and shows the sort of results that this type of analysis can offer to social scientists. The illustration is made around the topic of globalization of agriculture. Using data crawled from a large social tagging system can have an important impact in the discovering of latent patterns, which is needed to provide effective recommendations to different actors. In this paper, a sample of 851 users, 526 URLs and 1,700 tags from the *Delicious* classification system on the subject of globalization were retrieved and analyzed. Through the analysis, main users and websites around globalization issues in Delicious emerged, along with discovering the most important tags that were applied by users to describe the globalization of agriculture. The implications of these methodology and findings for further research are discussed.

Keywords: Information Retrieval; Social Network Analysis; Collaborative Tagging; Web 2.0

JEL Classification: C81, D85, F60

1. INTRODUCTION

The advent of the social web has significantly contributed to the explosion of web content and, as a side effect, to the consequent, explosive growth of the information overload. Thus, in recent years, there has been a substantial growth in social computing systems (Parameswaran and Whinston, 2007a; Kwai Fun and Wagner, 2008) that serves as intermediaries for social relations (Schuler, 1994) and are characterized by online community formation and user content creation (Parameswaran and Whinston, 2007b; Duan *et al*., 2008). Some of the best known social computing systems are content sites such as Wikipedia, Flickr, YouTube, social networking sites such as Facebook, microblogging such as Twitter, and social bookmarking services such as Delicious (Marlow *et al.,* 2006; Parameswaran and Whinston, 2007a). Overall, these social computing systems are characterized by a high

heterogeneity of information sources and make large amounts of information available to their users (Schueler *et al*., 2007; Vickery and Wunsch-Vincent, 2007). For some authors, certain data collection methods commonly applied in social studies like interviews or surveys have yielded inconclusive results, especially when it is in a web environment (Herring *et al.,* 2004; Nardi *et al.,* 2004; Yao, 2009). However, some studies are already deploying automatic data extraction techniques (Garrido and Halavais, 2003; Jones *et al.,* 2008; Shumate and Dewitt, 2008; Xu *et al*., 2008; Carmel *et al.,* 2009; Wang and Jin, 2010; Ackland and O'Neil, 2011) from big data. These techniques are of interest to social researchers; but, to date, the study using digital trace data methods continues to lack a theoretical framework (Janetsko, 2009: 170). Some authors clearly point out the need for flexibility and adaptability of the methodology (quantitative and cualitative techniques) to the object (Domínguez *et al.,* 2010: 10), and also focus on discovering collaborative process methods and tools (Zacarias and Ventura, 2011: 45). Our first objective will be to present a methodology to use big data from Web 2.0 in social research, and then to show the sort of results that this type of analysis can offer to social scientists.

Currently, the extensive use of the social web is emphasizing the central role of users and their (cor)relations. The focus is on the profile, preferences, needs, feedbacks, reputation, relationships, and, last but not least, the personal way each user classifies the huge amount of information at her/his disposal in the form of tags (Golder and Huberman, 2006; Marlow *et al.,* 2006; Cattuto *et al.,* 2007a; Rattenbury *et al.,* 2007; Dattolo *et al.,* 2012). Tags are keywords freely chosen by users (e.g. "globalization", "agriculture", "trade"), employed to annotate various types of digital content including images, bookmarks, blogs, and videos (Golder and Huberman, 2006; Shneiderman *et al*., 2006; Rattenbury *et al.,* 2007; Trant, 2009). The idea behind the concept of tagging is simple: a user enjoys a resource – e.g. an image or a website – and, according to her/his mental model, identifies those terms that better describe the information conveyed by that resource. The same resource can be annotated by several users: some of them will reuse the tags already assigned to that resource, while others will adopt new tags. Social tags produced by users are usually regarded as high quality descriptors of the web pages' topics and a good indicator of web users' interests and preferences. This process also allows building of a socially-constructed classification schema, called folksonomy (Vander Wal, 2007).

Social tagging systems have recently begun receiving increasing attention from the scientific community. The growing number of scientific publications concerning this issue and the development of real, social tagging systems, such as social networks (Twitter), social bookmarking applications (Delicious), sharing systems (Flickr), and in the e-commerce field (Amazon), confirm this tendency. The popularity of tagging is attributed, at least in part, to the benefits users gain from effectively organizing and sharing very large amounts of information (Cattuto *et al.,* 2007b) and users' interests (Golder and Huberman, 2006).

Some prominent examples of a tagging-intensive social computing system are social bookmarking sites such as Digg, StumbleUpon, Reddit, Pinterest, and Delicious[1]. These services are an inestimable source of information for scholars, as they produce a huge amount of user annotations (tags) and reflect the interests of millions of users. The social aspects of these systems derive from the fact that the resources (mainly websites) are tagged by the community, a feature known as collaborative tagging, which provides important metadata for investigators and others practitioners.

Delicious (www.delicious.com) – formerly *del.icio.us* in 2003 – is a free, social bookmarking web service for storing, sharing, and discovering web bookmarks. Delicious uses a non-hierarchical classification system in which users can tag each of their bookmarks with freely chosen index terms. Its collective nature makes it possible to view bookmarks added by

[1] http://www.ebizmba.com/articles/social-bookmarking-websites (retrieved 10.09.2012).

other users. All bookmarks posted to Delicious are publicly viewable by default. Tagging in systems like Delicious is an important change in the way web bookmarks are organized and shared (Ames and Naaman, 2007).

Introducing folksonomies as the basis for social network analysis means that the usual binary relation between users and resources, which is largely employed by traditional Recommender Systems, changes into a ternary relation between users, resources, and tags which is more complex to manage. Nevertheless, very few works highlight how to employ folksonomies in the field of social research. This leads us to think that further researches, evaluation studies, and insights are needed. Hence, our second objective in this paper will be to use data crawled from a large social tagging system to discover latent patterns, which will form a basis in order to provide effective recommendations to different actors.

Due to the current lack of a theoretical framework in retrieving automatic data and analyzing digital data in social science research, this paper presents a methodology to use big data from Web 2.0 in this field. At the same time, it applies this method to automatically extract data from a particular social bookmarking site (Delicious) and to show the type of results that this kind of analysis can offer to social scientists.

We focus our study on the Delicious site, specifically, in its user community around the issue of globalization of agriculture. According to Stiglitz (2006), issues such as bilateral trade are impeding development in the world's poorest countries. The globalization of the agriculture system is at the centre of this debate, because so many poor people depend on agriculture as an income source and because they spend a large proportion of their resources on food. Given that the majority of the poor inhabit rural areas and earn a living as small farmers, the effects of globalization on employment and small-farm competitiveness are central to determining its impact on poverty.

This paper begins by reviewing the literature around collaborative tagging, paying specific attention to meta-knowledge and networks perspectives. Then we expound the methodology, laying out the empirical data and describing in detail the data extraction process applied. The next sections analyze and provide the results of a study that involved 851 users on 1,077 URLs and 1,700 tags, before concluding with a discussion of several implications for research and practice.

2. LITERATURE REVIEW

2.1. Web 2.0 and Collaborative tagging

The Web 2.0 concept was developed by Tim O'Reilly (O'Reilly, 2007). According to O'Reilly, ''Web 2.0 is the business revolution in the computer industry caused by the move to the Internet as platform, and an attempt to understand the rules for success on that new platform[2].'' Web 2.0 is combined with the programmable Web and was designed to be dynamic, peer to peer, and an online storage of knowledge.

Collaborative – or social – tagging is the activity in Web 2.0 of annotating digital resources with keywords, or so-called tags (Golder and Huberman, 2006; Trant, 2009). This process is easy to indulge in, because it does not require any professional background; all that is needed is to freely choose keywords from an individual's vocabulary to annotate a Web resource. This process of annotation has converted ordinary people into metadata generators. Hence, collaborative tagging is a form of a user-centric, social, and democratic method of indexing. The use of tags creates a collective classification scheme and provides a snapshot of the current mindset of the user.

[2] http://radar.oreilly.com/2006/12/web-20-compact-definition-tryi.html (retrieved 01.10.2012).

Collaborative tagging has two purposes: first, it can quickly generate personal categorizations for later information retrieval; second, the collective use of tags makes inferences about related resources and tags possible. A resource can be tagged with an unlimited number of tags. The collaborative tagging of websites allows for the organization and sharing of digital resources. These websites allow users to publicly tag available resources and share content; therefore, users can categorize information by themselves and browse or search for the information by using these tags (Golder and Huberman, 2006). In this sense, the collaborative tagging of websites works as a shared resource for a given community of actors that could be used for different motives and in different moments.

A collaborative tagging system is mainly composed of three interconnected components – users, tags, and resources (Smith, 2008) – which can be described as follows:

- Users: They employ a tagging system to create tags, and sometimes they add resources. Users – who have a variety of different interests, needs, goals, and motivations – try to share or label a resource so they can find it later.
- Resources: They are the items that users tag such as the Web pages in Delicious and the photos in Flickr.
- Tags: They are the keywords added by users. Tags are essentially metadata about the resource. Users can tag just about any kind of term to resources, and different users have different tagging patterns.

2.2. Tagging and Folksonomy

Social tagging systems aggregate the tags of all users and describe the resources in a so-called folksonomy (Vander Wal, 2004; Trant, 2009). The word ‘‘folksonomy” (Vander Wal, 2004) is a combination of ‘‘folks” and ‘‘taxonomy”. Recently, the use of folksonomies gained more attention because of their simplicity: using tags, users can freely model the information without the constraints of a predefined lexicon or hierarchy (Mathes, 2004). Folks are the common people of a society; taxonomy means a hierarchical structure of classification. However, the simplicity of the approach also has an important drawback: the information managed by folksonomies is modeled in a simple, syntactical way. Therefore, collaborative tagging systems suffer from the vague-meaning problem when users retrieve or present resources with keyword-based tags. The vague-meaning problem is created by the following causes (Kroski, 2005; Golder *et al.,* 2006; Hope *et al.,* 2007; Marchetti *et al.,* 2007):

- Synonyms: It is when multiple tags share the same meaning. For example, resources tagged as *Web site* and *Website*, or *global warming* and *climate change* could have the same meaning: the first ones are semantically similar, and in the second example the words are differents. However, collaborative tagging systems do not understand it.
- Term variations: There is no standard for the structure of tags; for instance, a noun can be singular or plural, uppercase or lowercase. In collaborative tagging systems, we can also have simple, morphological variations. Moreover, mis-tagging due to spelling errors occurs often. Also, spacing is not allowed in a tag in most collaborative tagging systems; therefore, both the underscore and the hyphen are typically used to separate words by a single tag. Additionally, different possible spellings of the same word and tags using different languages generate term variations such as *globalization* and *globalisation*.
- Lack of relationships: Relationships between tags cannot be structured in existing, collaborative tagging systems. For instance, resources might be labeled with the tags *fast food* or *hamburger*, and there is no mechanism that might indicate that hamburger is a sub-class of fast food.

These drawbacks hinder the use of folksonomies for tasks more complex than the simple browsing of resources. In order to avoid these problems, in recent years many tools have

been developed to facilitate the user in the task of tagging, by also speeding up the tag convergence (Cattuto *et al.,* 2007b).

2.3. The collective knowledge inherent in social tags

Social tagging is the process by which many users add metadata in the form of keyword-based tags to shared resources. In social tagging systems, users can annotate a variety of digital resources with tags, for instance, bookmarks (e.g. www.delicious.com), pictures (e.g. www.flickr.com), or products (e.g. www.amazon.com). In most applications, users are free to choose any tags for describing their resources in order to structure, organize, and re-find their own stored Web material. The tags that are used will reflect individual associations with regard to resources, and they will describe a specific meaning or relevance for the respective users. The social aspect of social tagging systems lies in the opportunity to use other people's tags as navigation links for one's own search processes. The folksonomy process is developed in a bottom-up process of individual tagging in which the tags of many different users are aggregated and the resulting collective tag structure – such as tag cloud – depicts the collective knowledge of Web users (Cress *et al.,* 2012) – although, in some cases, such as by laziness, users could have aggregated tags suggested by the social tagging system. The individual users' tags establish a network of connections between resources and tags, and among those tags themselves. The more frequently tags are used for one resource, the stronger the connection becomes among them. Analogously, the more often two tags co-occur for one resource, the stronger they are related to each other. Social tagging systems can be considered shared, external knowledge structures of communities (Fu, 2008) and augment the collective structure of a community with the individual knowledge representations of individual users. When aggregating all tags from a community, a collective representation of the connections between related tags and their strengths of association will emerge. These associations are typically visualized by tag clouds, in which different font sizes represent the strength of association of tags to a related tag or a resource. Tag clouds externalize the community's associations between tags and the strengths of associations. In this way, social tags are able to provide visual representations of the conceptual structure of a domain that is built upon the knowledge of individuals who belong to a large Web community. One study (Kammerer *et al.,* 2009) showed that tags as *navigational signposts* are able to provide a kind of scaffold to learn new topics, leading to better understanding of a knowledge domain. Although a few studies have investigated the influence of tag clouds on visual attention, recognition, and tag selection (Rivadeneira *et al.,* 2007; Bateman et *al.,* 2008), this research addresses how users could use this collective knowledge representation.

Much research on social tagging has focused on the description of regularities in user activities (e.g. Golder and Huberman, 2006; Millen *et al.,* 2007). Research has also investigated the motivation of people to tag and the research question of how to design Web sites and platforms in order to motivate users to annotate content with social tags (Van Velsen and Melenhorst, 2009). However, surprisingly little is known about how these new technologies directly interact with individuals at the knowledge and cognitive level (Fu, 2008).

The first study that investigated the interplay between collective and individual knowledge was presented by Fu (2008). He introduced a rational model of social tagging in Delicious and provided evidence for the interaction between social and cognitive systems. Further studies addressed the emergence of stable tagging patterns (Kannampallil and Fu, 2009) or investigated how the use of social tags affects search performance (Fu *et al.,* 2010), tag choice, and the individual interpretation of documents through processes of imitation (Fu *et al.,* 2009; Kang *et al.,* 2009), users' behaviours (Kang and Fu, 2010), or innovation impacts in organizations (Parise and Iyer, 2011), with Delicious as well. In these studies,

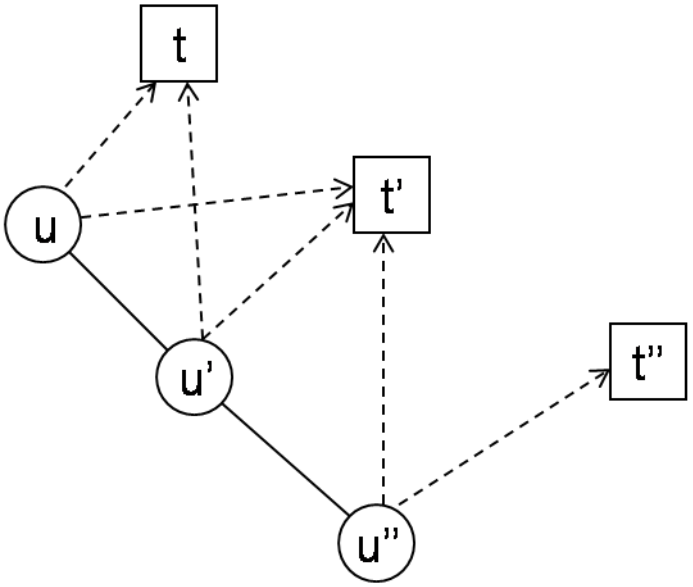
model simulations were used to demonstrate the exchange process between the collective knowledge that is inherent in social tags and the individual knowledge of users. These studies show the potential of social tagging systems.

2.4. Tagging and Social networks

The structure of Social tagging websites can be viewed as a network consisting of three parts, or a network of three different node types: the **U** users, the **R** resources (web sites – URLs), and the **T** tags that the **U** users deploy to tag the **R** web sites. A user can attach one or more tags to a URL. The network that emerges can be graphically illustrated by means of the *1* link between the *u* user and the *r* website that passes through the *t* tag.

A particular class of networks is the bipartite networks, whose nodes are divided into two sets (e.g. users and tags), and only the connection between two nodes in different sets is allowed, as illustrated in Figure 1.

Figure 1. A Bipartite Network made of three users U=(u,u',u''), three tags T=(t,t',t'') and two kinds of links: between users R^U (straight lines), and between users and tags R^T (dashed lines)



Source: Authors

Two kinds of bipartite networks are important because of their particular significance in social, economic, and information systems. One is the so-called *collaboration network*, which is generally defined as a network of actors connected by a common collaboration (Wasserman and Faust, 1994; Scott, 2000). Examples in the social systems are numerous, such as scientists connected by coauthoring a scientific paper or movie actors connected by co-starring in the same movie, or in other fields like on the technological collaboration of software and urban traffic systems. The other one is called the *opinion network* (Maslov and Zhang, 2001; Blattner *et al.,* 2007), in which users connect to the objects that they gather. For example, listeners are connected with the music groups they collected from a music-sharing library such as iTunes, web-users are connected with the webs they collected from a bookmark site such as Delicious, or customers are connected with the books they bought from a site such as Amazon.

A central problem closely related to the opinion network is how to extract the hidden information. The exponential growth of the Internet confronts people with an information

overload. Two landmarks for social research are the use of digital trace data methods and the data analysis in the context of social networks analysis. This paper is an approach to these references.

2.5. Social Web and its impact on Information Retrieval (IR) and Recommender Systems (RS)

During the last few years, the advent of the Social Web has greatly changed the role of Web users, providing them with the opportunity to become key actors and to share knowledge, opinions and tastes thanks to the interaction through online media.

2.5.1. Recommender Systems and Social Web

Introducing folksonomies as the basis for recommendations means that the usual binary relation between users and resources, which is largely employed by traditional RS, changes into a ternary relation between users, resources, and tags, thus becoming more complex to manage.

Different surveys (Dattolo et al., 2012; Kumar and Thambidurai, 2010) analyze the use of social tagging activities for recommendations, focusing their attention particularly on the following aspects:

- RS improvement due to tags: an interesting overview on social tagging systems and their impact on RS is presented in Milicevic *et al.* (2010), while a methodology to improve RS due to Web 2.0 systems, and particularly to social bookmarking platforms, is offered by Siersdorfer and Sizov (2009); moreover, Xia (2010) provides a recommender system model based on tags.
- Role of tag recommendation: the system presented in Rendle and Lars (2010) exploits a factorization model to propose personalized tag recommendations, while Niwa *et al*. (2006) illustrate a strategy used in a Web page recommender system exploiting affinities between users and tags. In addition to these affinities, Durao and Dolog (2009) propose a recommender system exploiting tag popularity and representativeness to recommend web pages.
- Tags and User modeling: since RS rely on a user model to generally personalize recommendations, Wetzker *et al.* (2010) propose an original way to enhance modeling to improve tag recommendation. In a general context, Carmagnola *et al.* (2007) and Simpson and Butler (2009) also illustrate how tag activity can improve user modeling.

2.5.2. Information Retrieval and the Social Web

From a Social IR point of view, i.e. IR that uses folksonomies, tags and particularly the relations between tags have been studied as a novel knowledge base related to information exploited in the IR process:

- As a pull approach, users retrieving information need to understand what information is available to identify which one is relevant to their need. A tag cloud has been used in this context to offer an original and improved visual IR interface (Hassan-Montero and Herrero-Solana, 2006; Bar-Ilan *et al.,* 2010) which allows user browsing information. A more powerful visualization based on tag clusters (Knautz *et al.*, 2010) is considered as better than a tag cloud.
- FolkRank (Hotho *et al*., 2006) is a new search algorithm for folksonomies. It can also be used to identify communities within the folksonomy that are used to adapt information ranking. This algorithm is inspired from the famous PageRank model from Google. Information ranking (scoring) has also been studied according to query (Liu *et al.,* 2009). Another document ranking based on relationships extracted from the different node types - user, tag, and resource - is illustrated in Bender *et al.* (2008).

- IR have also been improved thanks to folksonomies and two original measures (Bao *et al.,* 2007): SocialPageRank, which computes the popularity of web pages, and SocialSimRank, which calculates the similarity between tags and queries.
- Query expansion based on tag co-occurrence has been studied in Wang and Davison (2008), Biancalana and Micarelli (2009), and Jin *et al.* (2009). Results show that such an approach consistently improves retrieval performance.

In summary, this paper aims to exhibit a methodology to retrieve big data from Web 2.0 and use social network analysis in order to represent the main users and websites around the globalization of agriculture issue in a particular social bookmarking site – Delicious –, along with the most important tags that were employed by users around this topic. An additional aim is examining if it is possible to discover latent pattern links to the activity of collaborative tagging, which could be key in order to provide effective recommendations to different actors.

3. METHODOLOGY

The setting chosen for this study is Delicious (www.delicious.com). Delicious is a prominent example of a social bookmarking system whose content is created, annotated and viewed by its users. Delicious uses a non-hierarchical classification system in which users can tag each of their bookmarks on the Delicious website, and it provides knowledge about the URL marked (Golder and Huberman, 2006; Marlow *et al.,* 2006). Its collective nature makes it possible to view bookmarks added by other users. Delicious also allows users to organize existing tags into groups, called tag bundles. In addition, a Delicious user can follow the latest discoveries from people who share their interests. Hence, we believe that Delicious would be a good setting to investigate how to discover latent structures by using data crawled from a large, social tagging system.

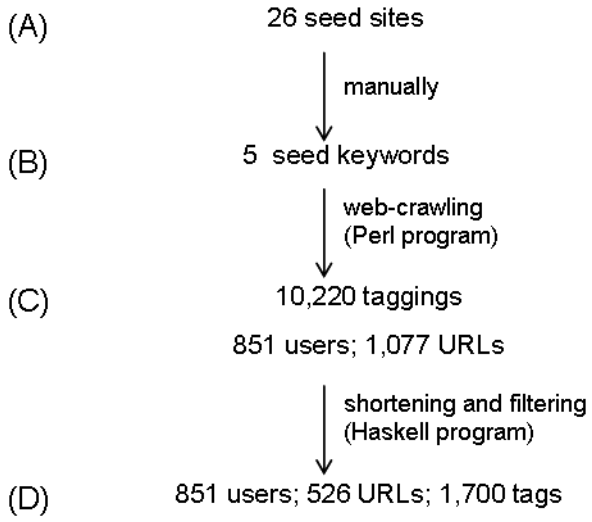
3.1. Data Collection procedure

In Social Bookmarking Services, an annotation typically consists of at least four parts. The link to the resource (e.g. to the website), one or more tags, the user who makes the annotation, and the moment when the annotation is made: user, resource, tag, and time. A user labels a resource with a specific tag at a given moment. This paper is less interested in when the annotation took place than in the co-occurrence of users, resources and tags (*user, resource, tag).* The dataset collected is written as: $\mathbf{U} = \{u1; u2; : : : ; uK\}$, $\mathbf{R} = \{r1; r2; : : : ; rM\}$, and $\mathbf{T} = \{t1; t2; : : : ; tN\}$ as the set of K users, *M* URLs, and *N* tags, respectively.

We built the network of globalization of agriculture using a combination of search techniques proposed for researching "issue networks" (Rogers and Zelman, 2002): associative reasoning, whereby educated guesses are made about relevant issues and related websites; public trust logics, finding groups commonly linked to by players trusted to be important in the debate; media stories, following links from an authoritative news source; and search engine crawls of key words.

The process to retrieve the data and of representing the *Delicious* community as a network follows a procedure that we present in Figure 2:

Figure 2. Data Collection Procedure



Source: Authors

Firstly (A), following links from an authoritative news source, we identify the search attributes on the basis of an original sample of a set of 26 web pages (Appendix 1), according to the Wikipedia definition of "critics of globalization[3]." We could have randomly selected them from other sites or sources, but we focus on this page because it is one of the most popular Web 2.0 pages, and because it has a high reputation[4]. On the other hand, we could have chosen another starting point, and it may have changed the keywords, but that was not relevant at the time of this study. The important thing in this phase was to have an authoritative news source as baseline to find, as a first step, keywords connected to globalization, and, as a second step, the idea of 'globalization of agriculture' as the main issue for the present illustration. We propose future research considering other starting points.

Based on a detailed study of site content, we selected main concepts from external links to these webpages (B). The search attributes were extracted manually from the website homepages and from the tag clouds or the topics that appear on the homepage. Following Rogers and Zelman (2002), we decided to identify these keywords through associative reasoning, whereby we made educated guesses about relevant issues and finding key concepts commonly linked to all seed websites. Finally, we found a set of attributes related with agriculture – agriculture, food, organic, and GMO – that had been grouped along with the word globalization under the denomination of "globalization of agriculture". Other different concepts were rejected at this step as they were not directly associated to agriculture, though could be linked to globalization. The decision to proceed at this stage with the manual extraction of these keywords, as opposed to using another automatic selection, was taken due to the importance that we give to the researcher in this stage, due to his/her expertise.

In a third stage (C), we gather the raw data sample of all the users' records, URLs and tags available for the four tag pairs around the globalization main tag – globalization+agriculture; globalization+food; globalization+organic; globalization+GMO – identified by crawling through the social bookmarking website Delicious using a Perl-developed[5] web crawler[6]. The data-gathering process from the four attributes covered the period between 22 April 2011

[3] http://es.wikipedia.org/wiki/Categor%C3%ADa:Cr%C3%ADticos_de_la_globalizaci%C3%B3n (retrieved 02.04.2011).
[4] http://www.alexa.com/topsites (consulted 13.09.2012).
[5] José Carpio, Intelligent Systems and Data Mining research group from University of Huelva, Spain (TIC-198).
[6] A web crawler is a program that automatically traverses a web site (e.g. Delicious) by retrieving all users, URLs and tags that match the search criteria.

and 21 May 2011 (one completed month), and produced 10,220 taggings that involved 851 users on 1,077 URLs and 1,720 tags.

Finally, we developed a program in Haskell[7] to reduce the amount of data (D) by cutting the URLs and using key words, including the identification of synonyms, and eliminating words with capital letters and derivatives such as words in plural. Both software programs, Perl and Haskell, are free software and they are in line with Web 2.0 philosophy. The definitive data constituted 851 users, 526 URLs and 1,700 tags.

Table 1 shows the key words and the frequency with which they occurred around the topic of globalization of agriculture.

Table 1. Keywords Used in the topic "Globalization of agriculture"

Search attributes used	Number of resulting tags (I+II)	More frequent Tags / Main Tags
Globalization (I) + agriculture (II)	1,116	Food (268), economics (176), environment (145), politics (85), trade (81), sustainability (70)
Globalization (I) + food (II)	1,682	Economy (180), economics (171), environment (122), sustainability (78), politics (60)
Globalization (I) + organic (II)	22	Business (3), fair-trade (3)
Globalization (I) + GMO (II)	54	Food (13), agriculture (12)

Source: Authors, from Delicious dataset (from 22-4-2011 to 21-5-2011)
Note: Each user can label each URL with a different number of tags

3.2. Analysis procedure

We are interested in computing the proportion of links preferentially created towards some kind of agents, relative to the proportion of these agents in the whole network (Barabási *et al.,* 2002).

Node centrality, or the identification of the nodes that are more "central" than others, is a fundamental part of network analysis (Freeman, 1979; Bonacich, 1987; Borgatti, 2005; Borgatti *et al.,* 2006). It is a network level property which gives a rough idea of the node's social power based on how well it "connects" to the network.

The literature on social networks conceptualizes centrality in many different ways (Freeman, 1979). The degree of a node is the number of ties it has; specifically, the number of direct connections individuals have with others in the group, which reflects the level of activity. The node with the highest degree exerts influence (or authority). The directed networks differentiate between *In-degree* and *Out-degree*. *In-degree* is the number of incoming ties that reflect the popularity of a website. As a result, the prominent, well-connected members (those with a high degree of centrality) are usually the opinion leaders. *Out-degree* is the number of outgoing ties which determine if a particular user is an active or passive participant within the network.

Our aim is also to describe, in a simple manner, the resulting tag structure as a tag cloud that depicts the interests of Web users. In this way, social tags are able to provide visual representations of the conceptual structure of an issue, which is built upon the knowledge of individuals who belong to a large Web community.

[7] Antonio Regidor, Agricultural Economics research group from University of Huelva, Spain (SEJ-110).

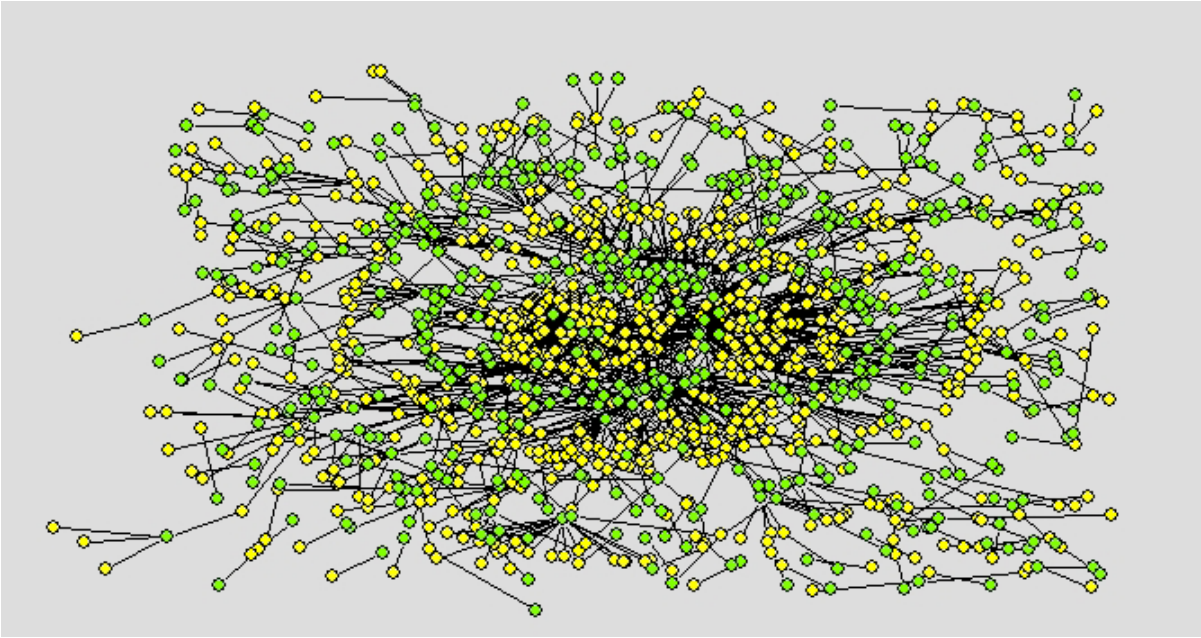
4. RESULTS

In this section, we present some empirical analysis of the network for the globalization of agriculture and a first approach to the tags associated to URLs found after the retrieval of information process. The next section reviews that analysis with reference to the three elements that compose the network: users, URLs, and tags. We use Social Network Analysis techniques (with the help of the software Pajek, which is better for big series of data than others such as Ucinet) to build the network[8] that we have called "globalization of agriculture"[9]. Through the connections among three key elements (Users write Tags to characterise URLs), different calculations were made. The following pages focus on two different approaches that allow us to find visible and invisible patterns when a Delicious bookmarking system's user is simply using Delicious. In that sense, we discovered latent structures. Firstly, we pay attention to power that emerges from the network – main users and websites. In the second section, we focus on concrete tags that were elaborated by users describing URLs and their importance.

4.1. Centralization: Authority

Centralization is a network-level property that broadly measures the distribution of power or prominence amongst actors in a given network (Hanneman, 2005). We calculate centralization by first computing a particular node-level degree centrality. Each time a user labels a particular URL, the intersection between user and URL was coded by 1. In the "useràURL" directed network (Figure 3), we calculate the indegree from each URL as the sum of total inbound links, and, in the same way, the outdegree from each user as the sum of the outbound links.

Figure 3. Hyperlink Network Energy Kamada-Kawai Map. Bipartite Network useràurl



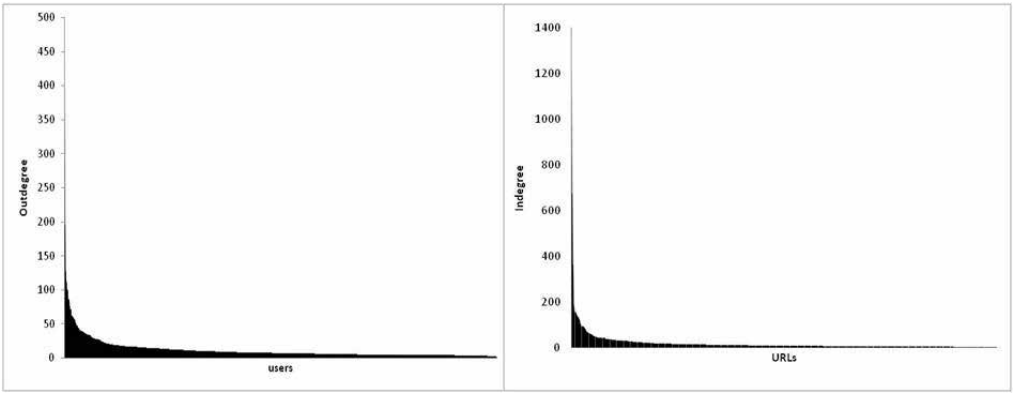
Source: Authors by Pajek
Note: Users in Yellow color; URLs in Green color

Most important Users and URLs are placed in the middle of the Figure 3, where density of connections is higher. Figure 4 shows the degree of variability in the website and user

[8] In this work, we use Social Network Analysis for calculating some measures, but we do not show the visualization of the network.
[9] This network is based on the original extraction of data from Delicious that took into account search attributes connected to globalization and agriculture (see Table 1).

centrality scores according to indegree and outdegree. As expected, the network is highly centralized within a few nodes because only 10 URLs from 526 (1.90%) account for 32.29% of the links to URLs[10] and only 10 users from 851 (1.17%) account for 14.05%[11]. This imbalance is not unusual given what we know about the long-tail distributions and the scale-free properties of the web. The power law is a defining characteristic of large-scale networks such as the web (e.g. Barabási and Albert, 1999), which implies a high degree of network centralization; it also proffers the empirical starting point for the question: Why?/ How come a few users and websites are better connected than the majority?

Figure 4. Hyperlink Network. 851 users arranged in rank order by number of outbound links and 1,077 URLs arranged in rank order by number of inbound links



Source: Authors, from data retrieved from Delicious dataset

Table 2 shows the 10 most centralized websites, and we can see that nine are media-based (online newspapers such as The New York Times, BBC, The Guardian, Washington Post, Financial Times, Reason, The Nation, Spiegel and The Economist).

Table 2. Top Authoritative Sites in the hyperlink network

	Indegree		Outdegree	
1	1203	http://www.nytimes.com/	433	/mritiunjoy
2	674	http://news.bbc.co.uk/	195	/laura208
3	365	http://www.guardian.co.uk/	127	/rd108
4	186	http://www.washingtonpost.com/	112	/amaah
5	158	http://www.ft.com/	111	/thepouncer
6	154	http://www.reason.com/	100	/anilius
7	147	http://www.thenation.com/	100	/emmarlyb
8	137	http://www.spiegel.de/	87	/adorngeography
9	136	http://www.foodfirst.org/	86	/pagolnari
10	130	http://www.economist.com/	85	/freemanlc

Source: Authors, from data retrieved from Delicious dataset

[10] These ten URLs got 3,290 inbound links from a total of 10,190.
[11] These ten Users got 1,436 outbound links from a total of 10,219.

This table also shows the users with a greater degree of centrality. We observe that the user, *mritiunjoy*, plays a very important role in the network. We could take it a step further to know more about this central user and his possible connection to the links. For example, we discover on a Delicious web page that *mritiunjoy* joined Delicious on 12 March 2007, and, to date, he has 10,020 links and is following 38 users. However, on the internet, we also discover that *mritiunjoy* – Mritiunjoy Mohanty – is a professor at the Indian Institute of Management Calcutta, India, and his Research Interests are the Political Economy of growth and development.

Remembering Diani (2003), the analysis has identified valuable nodes (websites). Its value is not only due to the links that they receive (its instrumental nature) but also to the profile of these organizations (newspapers that channel big quantities of resources – information), due to the quality of the links. This last particularity (quality of links), added to the first (instrumentality), determines that these URLs are central and have some authority. As a consequence, they could be relevant to produce currents of opinion. Most URLs bookmarked are singulars, because they could create or modify opinions.

In addition, the results suggest that the most centralized users (those with the highest number of links) do that because they have other interests than simply bookmarking, sharing, or labelling a resource.

4.2. Node Tags: Users producing Tags

In this section we explore the collective tag structure (excluding the key search words, such as globalization, agriculture, food and organic, and GMO) in an attempt to identify topics around our main theme. A natural approach to identifying the topical groupings in a tag network is to use tag clouds. Thus, Figure 5 shows a selection of highly descriptive keywords for the globalization of agriculture system in Delicious. Cluster keywords were automatically identified.

The clouds were produced with Wordle (Viégas *et al.,* 2009), where the sizes of the terms in the tag clouds are proportional to the weights, with the top 25 highest weighted tags included. The resulting key topics were *economics* and the *environment*, which were the main keywords used by users to describe or characterise in Delicious the topic 'globalization of agriculture'.

Figure 5. Tag Cloud for Agriculture Globalization Network Identified on the Delicious Data Set

Source: Authors

After this brief description, we want to clarify that these are not the unique results that could have been exposed here, after the complex process described for the retrieval of information. We have chosen them as good examples that give sense to the operation of crawling big data, as first points of departure for knowing a bit more about the topic of globalization of agriculture, and for demonstrating the way that people describe and share

websites about this issue through a modern and collaborative process of tagging. In the next section, a discussion about the results will be presented in order to know more about some alternative analyses, reflections, etc.

5. DISCUSSION

5.1. Centrality and Power

Hanneman (2005) reminds us: "a very simple, but often very effective measure of an actor's centrality and power potential is their degree". In our case, as indegree concern URLs and as these represent some kind of collective actor, the determination of centrality measures make sense. Higher indegrees mean that the URLs are chosen by more users (they received more links). It is evident that the New York Times, in this network of globalization of agriculture in Delicious, greatly surpasses other URLs (with 1,203 inbound links, followed by the BBC website with 674 links). Most cited, recommended, or considered websites with regards to a topic occupy a central place and have an important role in the process of dissemination of news, events, trending topics, ideology, culture, etc. Knowing this previously hidden hierarchy is also very useful for different socioeconomic reasons. At the same time, this identification of key collective actors (represented here through URLs) allows a better comprehension of leadership, influence process, and power-related structures. For social practitioners, it is a good way to identify key informants in a community through whom disseminating useful and important information occurs.

Indegree in Table 2 also shows a very unequal distribution of power of the URLs cited by users in the topic of globalization of agriculture, represented by an important accumulation of inlinks. Only 10 URLs represent an important, accumulated indegree).

Regarding other actors in the networks, the users, for the identification of key actors that disseminate and share URLs, as the previously cited *Mritiunjoy*, it is important to determine from where key elements that structure the network emerge. Is it possible to explain why 'that' greatly important actor is in the network of globalization of agriculture? Key actors in this type of network could configure and reconfigure the evolution of the network, structure, and even manipulate the type of interchange of resources in Delicious or in similar bookmarking sites.

Their prominence has something to say to social researchers, practitioners, etc. Is it by chance? Are most prominent actors in the type of website like Delicious corresponding to a profile of very active and participative people? Do they usually work (or have as a hobby) in this area, which could explain the accumulation and tagging of so many URLs in Delicious? These and other questions could be answered in further steps of the research, depending on the concrete goals at each moment.

5.2. Central Tags: Users producing Tags

In the process of linking URLs in Delicious, the majority of users selected tags suggested by the website or added new tags in a creative way for describing or qualifying the URLs that they were recommended. A 'tag cloud' was built in order to have a visual approach to the language that was employed by users in their descriptions. As we have focused on the retrieval of information regarding the topic of globalization of agriculture, the question now is to wonder what we could know about this topic through the extracted tags. From a total of 1,700 tags, two words were the main ones, as most cited when users labelled URLs. It is important to note that each user could label a URL with an unlimited number of tags (average 12 tags per user, max 433 and min 2). The most frequently used tags were the words: 'economics' (350 citations out of 1,700 tags, 20.6%) and 'environment' (273, 16%).

Other, very frequent tags were: sustainability (153), politics (152), economy (144), trade (131), business (99), poverty (97), culture (84), farming (84), africa (83), health (78), and development (76); in relative terms, these 13 tags represent one out of four labelled tags surrounding the topic (25.9%).

Discovering the importance of these words make us wonder not only of the reasons for the prominence of the first two tags regarding the globalization of agriculture but also for the rest. In addition, as 1,700 tags were also found qualifying and describing webpages regarding the topic of globalization of agriculture, some analyses are possible to know if some tags are used on an interchangeable basis, considered as synonyms, as was reported as one of the problems of collaborative tagging or the suggested vague-meaning problem (Kroski, 2005; Golder *et al.,* 2006; Hope *et al.,* 2007; Marchetti *et al.,* 2007).

The same thing could be done regarding 'term variations'. For instance, economy and economics are two important words in the topic of globalization of agriculture. Are these tags used in a similar or equivalent way at tagging? Why is the word economics sometimes used, and why, at other times, is economy used? Are they used in the same way at classifying the URLs?

By limiting the analysis to a particular period of time, tags could be associated to the use of language at a particular moment. They could even be a good representation of the ideological and terminological approach to the topic in the international arena, at that moment, and be useful for the study of the evolution and usage of language in a topic over time. On the other hand, the use of some tags at classifying URLs connected to the globalization of agriculture, and the distinction among users in the way they use some words as labels, could yield other types of results. Are scientific users utilizing the same tags as other professionals or general users? Perhaps different scientists or other users produce different labels around the same topic. Perhaps the first people tagging a topic are influencing the following tags that are incorporated in Delicious, etc. Nevertheless, some of this analysis can be limited by the information available from users. Other possible studies, going into more depth, may retrieve the pages that were labelled and undertake a content analysis to determine what kind of content is labelled through concrete tags. This is a cognitive way to see how users summarize and represent in short and definite words what could be broad and detailed content. It could be a way to see keywords that remind them about something. Through this, different applications could be suggested (for instance, in advertising, mobilizing, etc.).

Although this article has been more focused on the retrieval and illustration aspects, we have not shown networks of tags linked by users or by URLs tagged by users; a complementary, detailed analysis could help to identify users that have the same patterns at tagging or URLs that were similarly labelled. This opens a door to study structural equivalences and considering, for instance, applications for particular types of users.

Other questions emerge, as to why some labels are present but not others? Is it a question of language usage? Is it a question of traditions at tagging in Web 2.0? Is it a fashion? Supposedly, if people use Delicious for collaborative reasons, tags must at least be understandable for other users, unless the user prioritizes their own usage of Delicious.

6. CONCLUSIONS AND FUTURE RESEARCH

The main objective of this paper was presenting a methodology to use big data from Web 2.0 in social research. We had an interest in illustrating the extraction of data from a social bookmarking site (*Delicious*) and showing the type of results that this type of analysis could offer to social scientists. As it concerned big series of data crawled from a large social tagging

system, the analysis could have an important impact in the discovering of latent patterns, which form the basis in order to provide effective recommendations to different actors.

Our approach represents an important first step towards the development of empirical techniques capable of automatically differentiating groups of individuals with common interests, and individuals who occupy a more central position. This research is also of interest to make recommendations on the knowledge base of individual interests.

In addition, our analysis offers a previously unavailable understanding in the definition of recommendation services. To be able to identify a short list of the most centralized users and ties is extremely useful for researchers attempting to understand a community of more than a thousand links. This is particularly important for researchers interested in formulating strategies for intervention and mobilization, but practitioners and companies could also make use of this. The discovery of the central elements in a network (users and URLs) and the tags employed by users could be a key to the design of future strategies for the dissemination of messages, while also helping to achieve more success in communications, such as making use of important keywords to attract greater attention, etc. At the same time, if we know other interests of the users belonging to a network - through, for instance, other webpages that they link, and others tags that they label - we would be able to make recommendations, as done by other systems such as Amazon.

With regards to the process of retrieval of information, the method presented here was somewhat complex but easy to apply if there is some computer knowledge. Nevertheless, working in interdisciplinary teams could greatly help to develop this kind of knowledge, as it was in our case. Though the technical process described was successful, improvements are necessary in the future, at least regarding the retrieval methods and the implementation of IR and RS techniques in social commerce and social media contexts.

On the other hand, the relation between users and resources, which is largely employed by traditional Recommender Systems, changes into a ternary relation between users, resources, and tags, which is more complex to manage. This article has laid the first stone in the difficult process of understanding and discovering patterns in the process that characterizes users tagging URLs for collaborative reasons. The application was made under the topic 'globalization of agriculture'.

Some of the first contributions in the area of globalization of agriculture were that tags used to describe URLs in the Delicious's social bookmarking site were mostly concentrated around a few terms. The approximation to this topic in the future through other bookmarking sites (for instance, dominated by Spanish-speaking users) will allow the researchers to know if the recommended URLs are again media-based or are even the same webpages; or, for instance, if there is a semantic change concerning tags used for describing and classifying URLs.

Lastly, we do not want to close the article without clarifying that researching the topic of globalization of agriculture in a systematic and broad way may require the consideration of other starting points for the retrieval of information, at least to compare and contrast results. This is a limitation of this work. Nevertheless, the search yielded 1,700 different tags that have been used in the period of only a month to qualify and describe the phenomenon. It is a large number of tags, but we found a great concentration of them, as the centrality measure showed. The same argument could be made regarding URLs, in that they were extremely concentrated in mass media. Of course, other analysis in the future could be made with a longer period of time, along with other explorations. This is only a beginning.

ACKNOWLEDGEMENTS

Special thanks to José Carpio (University of Huelva) for his help in collecting the data used for this study by the Perl program. We are also deeply grateful to Antonio Regidor, who helped us with filtering data with his expertise in Haskell computing. We also have received some interesting comments to a draft of this work from Ainhoa de Federico (University of Toulouse 2) and Teresa González (University of Huelva). The methodology was presented as working paper at the CIEO, Centre for Spatial and Organizational Dynamics, at the University of the Algarve. At the Universidad of Huelva, a preliminary version of this text was discussed with other colleagues. We want to thank all of them for their suggestions, especially to Marielba Zacarias and Paula Ventura Martins (University of the Algarve), and Andrea Capilla and Mónica Carmona (University of Huelva).

REFERENCES

Ackland, R., and O'Neil, M. (2011) Online Collective Identity: The Case of the Environmental Movement. *Social Networks.* **33**: 177-190.

Ames. M. and Naaman, M. (2007). Why we tag: motivations for annotation in mobile and online media. *In: Proceedings of the SIGCHI conference on Human factors in computing systems.* San Jose, California, USA.

Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., and Su, Z. (2007). Optimizing web search using social annotations. *In: Proceedings of the 16th International Conference on World Wide Web, WWW 2007.* New York: ACM, pp. 501–510

Bar-Ilan, J., Zhitomirsky-Geffet, M., Miller, Y., and Shoham, S. (2010). Tag, cloud and ontology based retrieval of images. *In: Proceeding of the Third Symposium on Information Interaction in Context, IIiX 2010*. New York: ACM, pp. 85–94.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*. **286**: 509–512.

Barabási, A.L., Jeong, H., Ravasz, E., Neda, Z., Vicsek, T., and Schubert, A. (2002). Evolution of the Social Network of Scientific Collaborations. *Physica A.* **311**: 590-614.

Bateman, S., Gutwin, C., and Nacenta, M. (2008). Seeing things in the clouds: The effect of visual features on tag cloud selections. *In: Proceedings of the ACM conference on hypertext and hypermedia*. New York: ACM Press, pp. 193-202.

Bender, M., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Parreira, J.X., Schenkel, R., and Weikum, G. (2008). Exploiting social relations for query expansion and result ranking. *In: Data Engineering for Blogs, Social Media, and Web 2.0*, ICDE 2008 Workshops. Cancun, Mexico, pp. 501–506.

Biancalana, C., and Micarelli, A. (2009). Social tagging in query expansion: A new way for personalized web search. *In: Proceedings IEEE CSE 2009, 12th IEEE International Conference on Computational Science and Engineering*. Vancouver: IEEE Computer Society, pp. 1060–1065.

Blattner M., Zhang Y.-C., and Maslov S. (2007). Exploring an opinion network for taste prediction: An empirical study. *Physica A*. **373**: 753-758.

Bonacich, P. (1987). Power and Centrality: A Family of Measures. *American Journal of Sociology.* **92**: 1170-1182.

Borgatti, S.P. (2005). Centrality and Network Flow. *Social Networks.* **27**(1): 55-71.

Borgatti, S.P., Carley, K., and Krackhardt, D. (2006). On the Robustness of Centrality Measures Under Conditions of Imperfect Data. *Social Networks.* **28**(2): 124-136.

Carmagnola, F., Cena, F., Cortassa, O., Gena, C., and Torre, I. (2007). Towards a tag-based user model: How can user model benefit from tags? In: C. Conati, K. McCoy and G. Paliouras (eds.), *UM 2007. LNCS (LNAI).* **4511**: 445–449. Springer, Heidelberg.

Carmel, D., Zwerdling, N., Guy, I., Ofek-Koifman, S., Har'el, N., and Ronen, I. (2009). Personalized social search based on the user's social network. *In: Proceeding of the 18th ACM conference on information and knowledge management.* Hong Kong, China, pp. 1227–1236. 1646109: ACM. doi: 10.1145/1645953.1646109.

Cattuto C., Loreto, V., and Pietronero, L. (2007a). Semiotic dynamics and collaborative tagging. *In: Proceedings of the National Academy of Sciences.* **104**: 1461–1464.

Cattuto, C., Schmitz, C., Baldassarri, A., Servedio, V.D.P., Loreto, V., Hotho, A., Grahl, M., and Stumme, G. (2007b). Network properties of folksonomies. *AI Communications.* **20**(4): 245–262.

Cress, U., Held, C., and Kimmerle, J. (2012). The Collective Knowledge of Social Tags: Direct and Indirect Influences on Navigation, Learning, and Information Processing. *Computers and Education.* http://dx.doi.org/10.1016/j.compedu.2012.06.015

Dattolo, A., Ferrara, F., and Tasso, C. (2012). On social semantic relations for recommending tags and resources using folksonomies. In: Z.S. Hippe, J.L. Kulikowski and T. Mroczek (eds.), *Human-Computer Systems Interaction. Backgrounds and Applications*. **98**: 311-326.

Diani, M. (2003). 'Leaders' or Brokers? Positions and Influence in Social Movement Networks. In: M. Diani and D. McAdam (eds.), *Social Movements and Networks. Relational Approaches to Collective Action.* Oxford University Press. New York.

Domínguez J. A, Merino, B., and Aledo, A. (2010). The Identity of Sociology or what to do when the Universe is Unknown: Qualitative solutions against the quantitative obsession, *Spacial and Organizational Dynamics,* **5**: 7-22.

Duan W., Gu, B, and Whinston, A.B. (2008). Do online reviews matter?—an empirical investigation of panel data. *Decision Support Systems.* **45**: 1007–1016.

Durao, F., and Dolog, P. (2009). A personalized tag-based recommendation in social web systems. *In: Proceedings of International Workshop on Adaptation and Personalization for Web 2.0* (AP-WEB 2.0 2009). Trento, Italy.

Freeman, L.C. (1979). Centrality in Social Networks: Conceptual Clarification. *Social Networks.* **2**: 215-239.

Fu, W. (2008). The microstructures of social tagging: A rational model. *In: Proceedings of the ACM conference on computer supported cooperative work*. San Diego: ACM, pp. 229–238.

Fu, W., Kannampallil, T. G., and Kang, R. (2009). A semantic imitation model of social tag choices. *In: Proceedings of the IEEE international conference on computational science and engineering*. Los Alamitos: IEEE Computer Society, pp. 66–72.

Fu, W., Kannampallil, T. G., and Kang, R. (2010). Facilitating exploratory search by model-based navigational cues. *In: Proceedings of the international conference on intelligent user interfaces*. Hong Kong: ACM, pp. 199–208.

Garrido, M., and Halavais, A. (2003). Mapping Networks of Support for the Zapatista Movement: Applying Social Network Analysis to Study Contemporary Social Movements. In: M. McCaughey and M. Ayers (eds.), *Cyberactivism: Online Activism in Theory and Practice*, London: Routledge, pp. 165-184.

Golder, S., and Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*. **32**(2): 198–208.

Hanneman, R.A, and Riddle, M. (2005). *Introduction to social network methods*. University of California, Riverside, CA (published in digital form at http://faculty.ucr.edu/~hanneman/)

Hassan-Montero, Y., and Herrero-Solana, V. (2006). Improving tag-clouds as visual information retrieval interfaces. *In: InScit 2006: International Conference on Multidisciplinary Information Sciences and Technologies*. Mérida, Spain.

Herring, S., Kouper, I., Scheidt, L., and Wright, E. (2004). Women and Children Last: The Discursive Construction of Weblogs. In: L. Gurak, S. Antonijevic, L. Johnson, C. Ratliff and J. Reyman (eds.), *Into the blogosphere: Rhetoric, community, and culture of weblogs* (available at http://blog.lib.umn.edu/blogosphere/).

Hotho, A., Jaschke, R., Schmitz, C., and Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In: Y. Sure and J. Domingue (eds.), *ESWC 2006. LNCS*. **4011**: 411–426. Springer, Heidelberg.

Janetsko, D. (2009). Nonreactive Data Collection on the Internet. In: N. Fielding, R. M. Lee and G. Blank (eds.), *SAGE Handbook of Online Research Methods.* London: Sage.

Jin, S., Lin, H., and Su, S. (2009). Query expansion based on folksonomy tag co-occurrence analysis. *In: 2009 IEEE International Conference on Granular Computing*. IEEE, Los Alamitos, pp. 300–305.

Jones, S. Millermaier, S., Goya-Martínez, M., and Schuler, J. (2008). Whose Space is MySpace? A Content Analysis of MySpace Profiles. *First Monday.* **13**(9).

Kammerer, Y., Nairn, R., Pirolli, P., and Chi, E.H. (2009). Signpost from the masses: Learning effects in an exploratory social tag search browser. *In: Proceedings of the SIGCHI conference on human factors in computing systems.* New York: ACM, pp. 625–634.

Kang, R., and Fu, W. (2010). Exploratory information search by domain experts and novices. *In: Proceedings of the international conference on intelligent user interfaces*. Hong Kong: ACM, pp. 329–332.

Kang, R., Kannampallil, T., He, J., and Fu, W. (2009). Conformity out of diversity: Dynamics of information needs and social influence of tags in exploratory information search. In: D. Schmorrow, I. Estabrooke, and M. Grootjen (eds.), *Foundations of augmented cognition. Neuroergonomics and operational neuroscience*. Berlin/Heidelberg: Springer, pp. 155–164.

Kannampallil, T., and Fu, W. (2009). Trail patterns in social tagging systems: Role of tags as digital pheromones. In D. Schmorrow, I. Estabrooke, and M. Grootjen (eds.), *Foundations of augmented cognition. Neuroergonomics and operational neuroscience*. Berlin/Heidelberg: Springer, pp. 165–174.

Knautz, K., Soubusta, S., and Stock, W.G. (2010). Tag clusters as information retrieval interfaces. *In: Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, HICSS 2010. IEEE Computer Society, Washington, DC, pp. 1–10.

Kumar, A., and Thambidurai, P. (2010). Collaborative web recommendation systems a survey approach. *Global Journal of Computer Science and Technology.* **9**(5): 30–36.

Kwai Fun, I.P.R., and Wagner, C. (2008). Weblogging: a study of social computing and its impact on organizations. *Decision Support Systems.* **45**: 242–250.

Liu, D., Hua, X.S., Wang, M., and Zhang, H. (2009). Boost search relevance for tag-based social image retrieval. *In: Proceedings of the 2009 IEEE International Conference on Multimedia and Expo, ICME 2009*. IEEE Press, Piscataway, pp. 1636–1639.

Marlow, C., Naaman, M., Boyd, D., and Davis, M. (2006). HT06, tagging paper, taxonomy, Flickr, academic article, to read. *In: Proceedings of the seventeenth conference on Hypertext and hypermedia*. Odense, Denmark.

Maslov, S., and Zhang, Y.-C. (2001). Extracting Hidden Information from Knowledge Networks. Physical Review Letter. **87**: 248701-248705.

Mathes, A. (2004). Folksonomies - cooperative classification and communication through shared metadata (Acceded in 13th of August 20111, on the Web site of: http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Milicevic, A., Nanopoulos, A., and Ivanovic, M. (2010). Social tagging in recommender systems: a survey of the state-of-the-art and possible extensions. *Artificial Intelligence Review.* **33**: 187–209. 10.1007/s10462-009-9153-2

Millen, D., Yang, M., Whittaker, S., and Feinberg, J. (2007). Social bookmarking and exploratory search. In: L. Bannon, I. Wagner, C. Gutwin, R. Harper and K. Schmidt (eds.), *Proceedings of the European conference on computer-supported cooperative work.* London: Springer, pp. 21–40.

Nardi, B., Schiano, D., and Gumbrecht, M. (2004). Blogging as Social Activity, or, Would You Let 900 Million People Read Your Diary?. In: *Proceeding CSCW '04 Proceedings of the 2004 ACM conference on Computer supported cooperative work*. ACM: New York, pp. 222-231.

Niwa, S., Doi, T., and Honiden, S. (2006). Web page recommender system based on folksonomy mining for ITNG 2006 submissions. *In: Proceedings of the Third International Conference on Information Technology: New Generations*. IEEE Computer Society, Washington, DC, pp. 388–393.

O'Reilly, T. (2007). *What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. International Journal of Digital Economics*. **65**: 17-37.

Parameswaran, M., and Whinston, A. (2007a). Research issues in social computing. *Journal of the Association for Information Systems*. **8**: 336–350.

Parameswaran, M., and Whinston A. (2007b). Social computing: an overview. *Communications of the Association for Information Systems*. 19: 762–780.

Rattenbury T., Good. N, and Naaman, M. (2007). Towards extracting Flickr tag semantics. *In: WWW '07: Proceedings of the 16th international conference on World Wide Web.* ACM Press. New York, USA, pp. 1287-1288.

Rendle, S., and Lars, S.T. (2010). Pairwise interaction tensor factorization for personalized tag recommendation. *In: Proceedings of the Third ACM International Conference on Web Search and Data Mining,WSDM 2010*. ACM. New York, USA, pp. 81–90.

Rivadeneira, A.W., Gruen, D.M., Muller, M.J., and Millen, D.R. (2007). Getting our head in the clouds: Toward evaluation studies of tagclouds. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM Press. New York, USA, pp. 995–998

Rogers, R., and Zelman, A. (2002): Surfing for knowledge in the information society. In: G. Elmer and L. Rowman (eds.), *Critical Perspectives on the Internet*. Lanham, MD.

Schueler, B., Sizov, S., and Staab, S. (2007). Management of Meta Knowledge for RDF Repositories. *In: International Conference on Semantic Computing*, pp.543-550. http://doi.ieeecomputersociety.org/10.1109/ICSC.2007.79

Schuler D. (1994). Social computing. *Communications of the ACM.* **37**: 28–29.

Scott J. (2000). *Social Network Analysis.* Sage Publication. London.

Shneiderman B., Bederson,BB,and Drucker, S.M. (2006). Find that photo!: interface strategies to annotate, browse, and share. *Communications of the ACM*. **49**(4): 69–71.

Shumate, M., and Dewitt, L. (2008). The North/South Divide in NGO Hyperlink Networks. *Journal of Computer Mediated Communication.* **13**(2): 405–428 (available at http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2008.00402.x/pdf).

Siersdorfer, S., and Sizov, S. (2009). Social recommender systems for web 2.0 folksonomies. In: Proceedings of the 20th ACM Conference on Hypertext and Hypermedia, HT 2009. ACM. New York, USA, pp. 261–270.

Simpson, E., and Butler, MH. (2009). Analyzing Communal Tag Relationships for Enhanced Navigation and User Modeling. *Collaborative and Social Information Retrieval and Access: Techniques for Improved User Modeling.* IGI Global, pp. 43-64. doi:10.4018/978-1-60566-306-7.ch003

Smith, G. (2008). *Tagging: People-powered metadata for the social web*. New Riders Press. Berkeley, CA.

Stiglitz, J.E. (2006). *Making Globalization Work.* WW Norton and Company. New York, NY.

Trant, J. (2009). Studying social tagging and folksonomy: A review and framework. *Journal of Digital Information* **10**(1). <http://journals.tdl.org/jodi/article/view/ 269/278> Retrieved on 01.01.12.

Van Velsen, L., and Melenhorst, M. (2009). Incorporating user motivations to design for video tagging. *Interacting with Computers.* **21**: 221–232.

Vander Wal, T. (2004). *You down with folksonomy?* Acceded in 17th of July 2011, on the web site of: http://www.vanderwal.net/random/entrysel.php?blog=1529

Vander Wal, T. (2007). Folksonomy coinage and definition. Acceded in 17th of July 2011, on the web site of: http://vanderwal.net/folksonomy.html

Vickery G., Wunsch-Vincent, S. (2007). *Participative web and user-created content: Web 2.0, wikis and social networking*. OECD Publishing.

Viégas, F.B., Wattenberg, M., and Feinberg, J. 2009. Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics.* **15**(6): 1137-1144.

Wang, J., Davison, B.D. (2008). Explorations in tag suggestion and query expansion. *In: Proceeding of the 2008 ACM Workshop on Search in Social Media, SSM 2008*. ACM. New York, USA, pp. 43–50.

Wang, Q., and Jin, H. (2010). Exploring online social activities for adaptive search personalization. In Proceedings of the 19th ACM international conference on information and knowledge management ACM. Toronto, ON, Canada, pp. 999–1008. doi:10.1145/1871437.1871564.

Wasserman, S., and Faust, K. (1994). *Social Network Analysis.* Cambridge University Press. Cambridge.

Wetzker, R., Zimmermann, C., Bauckhage, C., and Albayrak, S. (2010). I tag, you tag: translating tags for advanced user models. In: Proceeding WSDM '10 Proceedings of the third ACM international conference on Web search and data mining. ACM. New York, NY, USA, pp. 71–80.

Xia, X., Zhang, S., and Li, X. (2010). A personalized recommendation model based on social tags. In: Database Technology and Applications (DBTA), 2010 2nd International Workshop on. IEEE Conference Publications, pp. 1–5

Xu, S., Bao, S., Fei, B., Su, Z.,and Yu, Y. (2008). Exploring folksonomy for personalized search. In Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval. ACM. Singapore, Singapore, pp. 155–162. ACM. doi:10.1145/1390334.1390363.

Yao, A. (2009). Enriching the Migrant Experience: Blogging Motivations, Privacy and Offline Lives of Filipino Women in Britain. *First Monday.* **14**(3).

Zacarias, M., and Ventura, P. (2011). Collaborative Methods for Busines Process Discovery, *Spacial and Organizational Dynamics,* **7**: 45-55.

Appendix 1
Seed sites

	Wikipedia	**External links**
1	http://es.wikipedia.org/wiki/Walden_Bello	http://waldenbello.org/
2	http://es.wikipedia.org/wiki/Jos%C3%A9_Bov%C3%A9	http://www.jose-bove.eu/
3	http://es.wikipedia.org/wiki/Noam_Chomsky	http://www.chomsky.info/
4	http://es.wikipedia.org/wiki/Michel_Chossudovsky	http://www.globalresearch.ca/
5	http://es.wikipedia.org/wiki/Enrique_Javier_D%C3%ADez_Guti%C3%A9rrez	http://diezgutierrez.wordpress.com/
6	http://es.wikipedia.org/wiki/Christian_Felber	http://www.christian-felber.at/
7	http://es.wikipedia.org/wiki/Ram%C3%B3n_Fern%C3%A1ndez_Dur%C3%A1n	http://laexplosiondeldesorden.wordpress.com/
8	http://es.wikipedia.org/wiki/Eduardo_Galeano	http://eduardogaleano.org/
9	http://es.wikipedia.org/wiki/Susan_George	
10	http://es.wikipedia.org/wiki/Ha-Joon_Chang	http://hajoonchang.net/
11	http://es.wikipedia.org/wiki/Naomi_Klein	http://www.naomiklein.org/
12	http://es.wikipedia.org/wiki/Michael_Moore	http://www.michaelmoore.com/
13	http://es.wikipedia.org/wiki/Raj_Patel	http://rajpatel.org/
14	http://es.wikipedia.org/wiki/Nicanor_Perlas	http://www.nicanor-perlas.com/
15	http://es.wikipedia.org/wiki/James_Petras	http://petras.lahaine.org/
16	http://es.wikipedia.org/wiki/Ignacio_Ramonet	
17	http://es.wikipedia.org/wiki/Dani_Rodrik	http://rodrik.typepad.com/
18	http://es.wikipedia.org/wiki/Arundhati_Roy	http://www.weroy.org/arundhati.shtml
19	http://es.wikipedia.org/wiki/Jos%C3%A9_Luis_Sampedro	http://www.clubcultura.com/clubliteratura/clubescritores/sampedro/home.htm
20	http://es.wikipedia.org/wiki/Fred_Scholz	
21	http://es.wikipedia.org/wiki/Joseph_Stiglitz	http://www.josephstiglitz.com/
22	http://es.wikipedia.org/wiki/Vandana_Shiva	http://www.navdanya.org/
23	http://es.wikipedia.org/wiki/Carlos_Taibo	http://www.carlostaibo.com/
24	http://es.wikipedia.org/wiki/Esther_Vivas	http://esthervivas.com/
25	http://es.wikipedia.org/wiki/John_Zerzan	http://www.johnzerzan.net/
26	http://es.wikipedia.org/wiki/Jean_Ziegler	

Source: Authors