Extracting Usage Patterns and the Analysis of Tag Connection Dynamics within Collaborative Tagging Systems

Daniel MICAN, Nicolae TOMAI
Business Information Systems Department,
Faculty of Economics and Business Administration
Babeş-Bolyai University, Cluj-Napoca, Romania
daniel.mican@econ.ubbcluj.ro, nicolae.tomai@econ.ubbcluj.ro

Collaborative tagging has become a very popular way of annotation, thanks to the fact that any entity may be labeled by any individual based on his own reason. In this paper we present the results of the case study carried out on the basis of data gathered at different time intervals from the social tagging system developed and implemented on Întelepciune.ro. Analyzing collective data referring to the way in which community members associate different tags, we have observed that between tags, links are formed which become increasingly stable with the passing of time. Following the application of methodology specific to network analysis, we have managed to extract information referring to tag popularity, their influence within the network and the degree to which a tag depends upon another. As such, we have succeeded in determining different semantic structures within the collective tagging system and see their evolution at different stages in time. Furthermore, we have pictured the way in which tag recommendations can be executed and that they can be integrated within recommendation systems. Thus, we will be able to identify experts and trustworthy content based on different categories of interest.

Keywords: *Tags, Collaborative tagging, Network Analysis, Tag Recommendation, Collective Intelligence*

1 Introduction

Collaborative tagging has become a very popular method of online resource annotation within Web 2.0. Due to this fact, an increasing preoccupation towards collaborative tagging has arisen within the academic community [3, 5, 15, 16]. This is defined as being the process through which several users add metadata in the form of keywords to the content they are publishing or saving. Thus, users can attribute tags to photos, clips, sites, books, e-mails, people or basically any entity that can produce meaning as they please. Thanks to the ability to collect data from users, collaborative tagging systems, in Wu's vision [14] have the potential to become an infrastructure tech in support of knowledge management activities in any organization or society, thus becoming a challenge for the researchers in its field.

The present paper constitutes a sequel to the research presented in [11]. In the aforementioned study, it has been observed that following the analysis of collective data regarding the manner in which community members associate different tags that connections form between tags. In the following we will look into the extent to which tag connections are maintained and become more stable with time. In accomplishing this goal we have composed a case study using data gathered at different time intervals from the social tagging system developed and implemented on Întelepciune.ro.

In further study, we will represent tags and the connections between them in the form of a graph and will apply the specific methodology for network analysis. This will be done in order to extract information referring to the way in which users attribute tags and give birth to different semantic structures within the collaborative tagging network. In this respect, we will study different tag attributes such as centrality, market share and market share by centrality. These attributes help us identify and gain a global view on tag popularity, their influence within the network and the extent to which a tag depends upon an-

other. Moreover, we will also study the evolution of these attributes at different stages in time. Also within this study we will analyze the opportunity of offering a contextual dimension to recommender systems by means of tags. Thus we have simulated for exemplification five recommendations for four of the most popular tags extracted from the experiment. The lists of recommendations were made based on data produced as a result of applying the three presented formulas for calculating similarities.

Below, the study is structured as follows: in Chapter 2 will treat theoretical aspects referring to collaborative tagging systems. In chapter 3 we present the case study's details and in the end we will lay out the conclusions.

2 Collaborative Tagging Systems

"Folksonomy" is a term derived from joining the words "folk" and "taxonomy" and is used to describe the social phenomenon of classification [3, 14, 15]. Collaborative tagging became popular in the same time as sites such as Flickr, del.icio.us or Technorati which implemented this concept. They allow tagging different resources, photos, web pages or blog posts with a set of key-words, called tags, chosen freely by each individual user.

Collaborative tagging can be described as the process through which users add metadata in the form of key-words to the content they are publishing or saving. Collaborative annotation by means of tags [15] allows for a personalized description of the content and represents a good indicator of users' interests. Classification by means of tags is done on-the-spot and spontaneously, it is user-centered and can be shared with the community. Tags can also be defined as data attached to an object (metadata). Collaborative tagging [13] is placed between the area of traditional representation methods and that of information retrieval (IR), being part of the new generation of technologies used in retrieving, representing and producing information.

The main advantages and characteristics [6] of collaborative social tagging systems are described as follows:

- the resource can be tagged with a multitude of key-words;
- users may use their own words to give meaning to the tagged resource;
- tags can be shared to create knowledge through aggregation;
- enable the development of communities based on similar interests;
- users can quickly and easily tag resources without the need of prior knowledge of classification or indexation.

In the conceptual model for social tagging systems presented by Marlow [9], tags are considered links between users and resources, and based on those links connections between resources and between users are formed. Links can be deduced by representing data as a social network and analyzing its structure towards identifying resource and user communities respectively. Users who employ the same tags in tagging a resource can be classified in a single community within the network. As such, we can claim that tags can identify an explicit connection between resources through the users who tag them and between users through the resources that they tag.

Peters [13] develops and extends Marlow's theories so that if the resources are labeled with the same tag they form a thematic group. These form a bibliographic group if they are labeled with the same tag or by the same user. If users apply the same tag, they are tied to a thematic group and if they describe the same resource they become part of the same bibliographic group. In Figure 1, which is a sum of Marlow and Peter's theories, document 1 and document 2 are linked both thematically and bibliographically, due to the fact that they are labeled with the tag 1. Users 1 and 2 are linked thematically due to the fact that they have labeled document 2 with tag 1. Users 1, 2 and 3 form a bibliographic group due to the fact that they described the same resource: document.

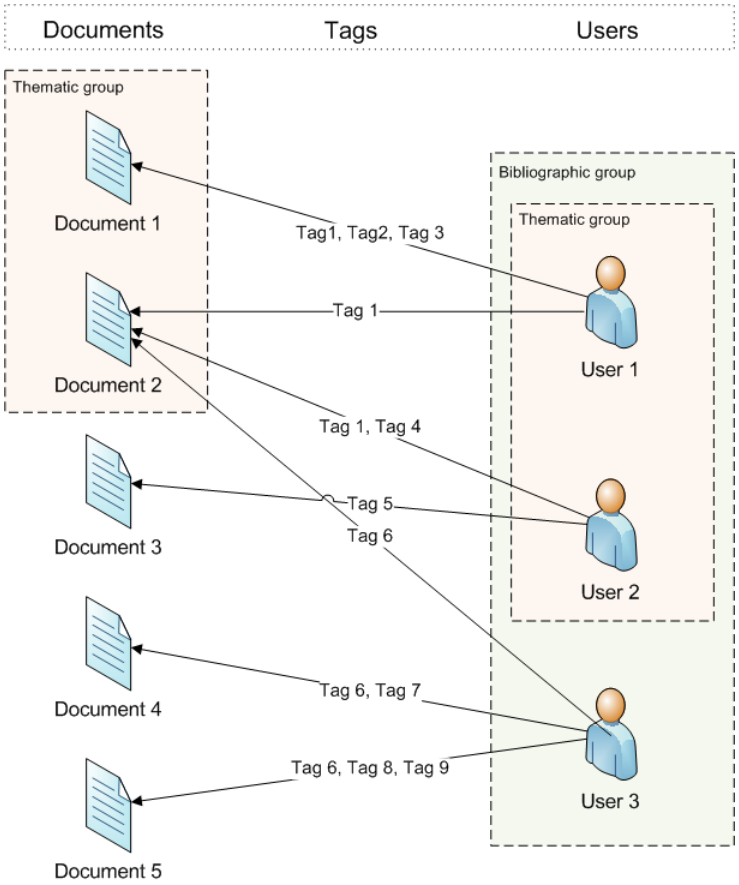

Fig. 1. The conceptual model of a collaborative tagging system

Based on the tripartite document, composed of resources, tags and users presented in Figure 1, we can extract three networks which can be shaped using the graphs theory [13]. Therefore, we can extract a network made up of resources which are linked together by tags, one composed of users connected via tags and one comprised of resources and users. In the case of the latter, nodes can be represented by resources as well as users. The purpose of analyzing these types of networks can be that of creating clusters of resources and users towards creating recommender systems.

Collaborative tagging generates a mass of valuable information [6] regarding the tags used, the tagged resources and the people that tag them. Thusly:

- information regarding tags will indicate the terms which are useful to users, new terms for existing concepts and new concepts conjunctively with the associated terms;
- information regarding the tagged resources will highlight the resources deemed valuable by users thanks to their labeling, therefore becoming of interest;
- information regarding users who tag resources will point to the tags and resources which a person has used in creating a profile that contains that person's preferences.

For storing data referring to the tags associated by users to a resource, three approaches have stood out [7], [8]. These are: MySQLicious, Scuttle and Toxi, all of which are highlighted in Figure 2. The schematic of the database for the MySQLicious approach contains only one table, which is denormalized. Tags are stored in a single field, delimited by space, presuming that each tag is composed of only one term.

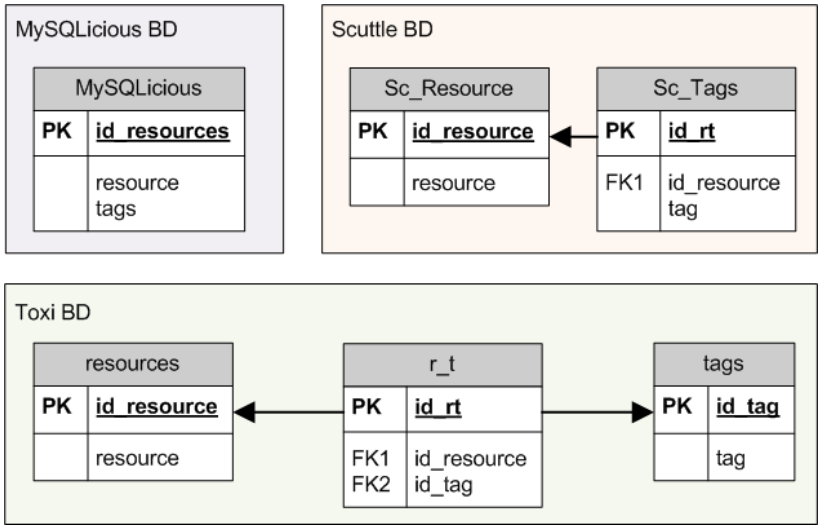

Fig. 2. Existing approaches towards collaborative tagging

The schematic for the database of the Scuttle approach contains the resources and the tags tables. The advantage of this approach above the former is that a resource can be assigned a tag that is composed of multiple terms. The third approach is the Toxi and it contains three normalized tables: resources, tags and resources_tags. The Toxi approach has an advantage over the first two thanks to the fact that tags are stored only once, making it much more scalable.

In our opinion, the greatest issue with the above approaches is that they contain no information regarding the users who tag resources. As we can observe in Figure 1, a collaborative tagging system's conceptual model contains three entities: resources, tags and users. From this point of view, we consider the presence of a model which can store information regarding users who tag resources to be highly suitable. This model can be observed in Figure 3. If building a personalized recommender system for recommending resources, tags or users is desired, we suggest this approach.

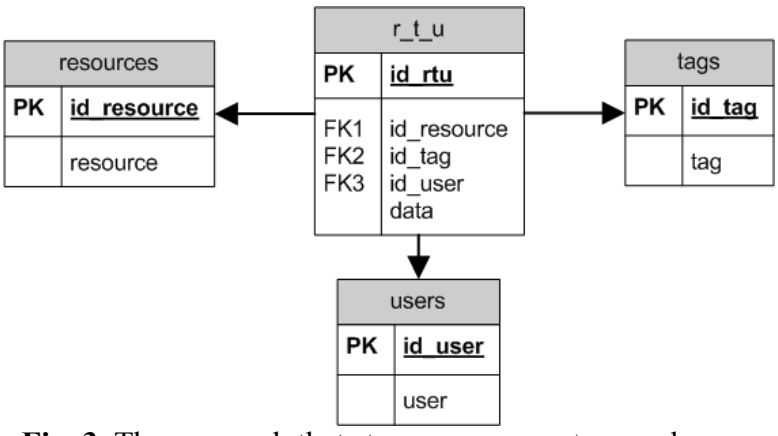

Fig. 3. The approach that stores resources, tags and users

Collaborative tagging systems, due to their social characteristic and the fact that users can label resources without using a limited vocabulary, may be employed in search engines, reputation systems, indexing, storing, personalizing and data mining systems. Taking into account user-specific data, information retrieval systems and search engines may extend their functionality by understanding the user's intent. Once a search is executed, they can produce a personalized search [15], which is one of the most promising directions the traditional search paradigm is heading towards.

3 Case Study

In order to analyze the usage patterns and the dynamics of connections between tags within collaborative tagging systems, we have conducted a case study using data collected from the collaborative tagging system implemented on *intelepciune.ro*. This offers data referring to the way users label text by means of tags, allowing for the extraction of preferences, connections and their popularity. The purpose of this case study was studying the opportunity to develop and offer a new contextual dimension, by means of tags, to the WSNRS system [10]. With the help of tags, by adding context, we aim to add a new dimension to recommended resources, users and groups of interest. In the following we will describe the collaborative tagging context, the degree of knowledge and will detail the completed experiment.

3.1 The Context of the Case Study and State of the Art

Following the analysis of an object we can see that it may hold different meanings to different people. Because of this, it is very important that each user have the possibility of tagging and classifying by their own reasoning the desired objects. It is well known that an object can be characterized through multiple attributes, and each user will rank the attributes by which they will identify an object differently. We can exemplify by using Mr. Smith and the connections that different people have with him. Mr. Smith will be labeled by an employee as boss or coworker, by his wife as husband, by his nephews as uncle and so on. We see that every person he interacts with gives one or more labels to classify Mr. Smith. The same happens in a collaborative tagging system. Each person perceives objects differently and classifies them depending on the degree of knowledge or by diverse personal motivations [11].

An essential element in collaborative tagging is identifying the existing connections between tags, users, objects and their evolution in time. These connections may vary in time, in longer or shorter intervals. For example, the employees who tag Mr. Smith as "boss" could only end up being his employees for a limited amount of time, unlike his nephews. Also, the number of connections that exist between entities, as well as their evolution in time need to be taken into consideration. The frequency with which certain connections appear between tags can be used in calculating similarities which may be used to anticipate certain user behaviors.

Adding tags is an increasingly used practice for organizing content and searching it easily. An increasing number of web applications have successfully implemented systems which allow collaborative tagging and offer users the possibility of attaching tags freely to content. A brief glance at context and knowledge reveals the fact that a series of relevant papers and researches have already appeared. Halpin [5] used data from the social bookmarking site del.icio.us to examine the dynamics of collaborative tagging systems. By analyzing the distribution of the most commonly encountered tags it has been shown how the significance of certain tags is defined by their relation to other tags. From this was born a generative model for collective tagging which helps in understanding the dynamics behind the tagging phenomenon and the way in which a stable distribution, based on a law of strength between tags may be created.

The TBCF algorithm put forward by Zhao [16] aim to represent a new approach in collaborative filtering based on the semantic distance between tags used by different users to enhance the efficiency of neighbor selection. Experimental results have shown that TBCF brings significant improvements to the traditional method of recommendation based on the cosines. The structure of the del.icio.us collaborative tagging system has been analyzed by Golder [3] to discover irregularities in user activities, tag frequency, the types of tags used, increases in popularity and their stability towards a certain URL. A dynamic collaborative tagging model which offers the possibility of predicting stable models based on common knowledge is also presented. Xu [15] suggested an automated rating frame-

work based on the folksonomies derived from del.icio.us and Dogear. Experimentation has shown it to be capable of improving search quality significantly.

3.2 Experiment Description

Taking into account the many challenges presented by collaborative tagging systems, we have built and implemented such a system to Întelepciune.ro. The implemented system allows any registered user to publish and label content with up to five tags. The system also offers the possibility of extracting information regarding the members of the community, their preferences on certain subjects of interest and the manner in which they associate tags to the published content. One of the collected data analysis' goals was to observe tag dynamics and if connections between tags become apparent, as well as their stability in time.

The data within the system were collected in two separate time intervals, following the first and the second year of implementation. According to the collected data, the implemented system was used in the tagging of 5.374 resources with a total 11.223 tags, of which 3.924 were unique. Based on system usage statistics, it is shown that 81.73% of users used at least one tag to label the published resources.

In order to discover the tags which are most frequently used together in tagging the same resource and the similarities between them, we have conducted an experiment using the collected data. We extracted the first 50 tags in order of popularity from the collaborative tagging system. In order to establish the maximum number of connections that can exist between two tags, we calculated combinations of 50, taken in pairs. As such, we generated a set of 1225 possible connections between the 50 tags and analyzed the frequency with which they were associated by users in published texts. Users could attach up to 5 tags to a text and the processed data were extracted on all five positions. The point was not to discover if 5 tags can appear together in the labeling of a text, but the frequency with which two tags are used together to label the same text.

The experiment had two phases. In the first, the data collected within a year was analyzed and in the second the data collected after two years. Initially, following the quantification of the frequency with which these connections appear in text labeling we observed that of 1225 possible connections, 218 materialized, of which only 93 with a frequency greater than 2. In the second phase, 358 connections materialized, 203 having a frequency greater than 2 and 125 greater than 3.

In order to measure the similarity between two tags, we used three formulas for measuring the degree of correlation between two pairs of tags. The first formula calculates the similitude between two t_i and t_j tags thusly:

$$similarity(t_i, t_j) = \frac{N(t_i, t_j)}{max\ (N(t_i, t_j))}$$

In the above formula, $N(t_i, t_j)$ represents the frequency with which the two tags have been jointly used to label the same text, and $max\ (N(t_i, t_j))$ represents the greatest frequency with which two tags have been jointly used to label a text. As such, we have produced values between [0, 1] for the existing values. The strongest connection will always have the value 1, and in case there is no connection between two tags, it will have the value 0. The closer a connection's value is to 1, the stronger it is.

The second formula is represented by the support used in data mining to generate association rules. An association rule's support is defined as being a fraction of a transaction that satisfies the union of items from a rule's antecedent and consequence. The support function adapted for our case is:

$$support(t_i \cup t_j) = \frac{N(t_i, t_j)}{Ntt}$$

In the above formula we have $support(t_i \cup t_j)$ for an association rule between two tags, t_i and t_j, which is calculated as a report between $N(t_i, t_j)$, which represents the frequency with which the two tags have been jointly used in tagging the same text and Ntt, which

represents the total number of text labelled with tags.

The third function measures cosine distance [5] that captures the extent tags' joint appearance and which can be interpreted as a metric of similarity. These are calculated by using the formula:

$$cosine(t_i, t_j) = \frac{N(t_i, t_j)}{\sqrt{N(t_i) * N(t_j)}}$$

In the above formula, $cosine(t_i, t_j)$ represents the similarity between t_i and t_j. $N(t_i)$ and $N(t_j)$ respectively, represent the frequency with which each of these tags have been used in labeling texts, and $N(t_i, t_j)$ represents the frequency with which the two tags have been jointly used to label the same text.

Table 1. Similarities between tags and their evolution between the two phases

		$similarity(t_i, t_j)$			$support(t_i \cup t_j)$			$cosine(t_i, t_j)$		
Tag A	**Tag B**	**Ph.1**	**Ph. 2**	**Evol.**	**Phase 1**	**Phase 2**	**Evol.**	**Ph. 1**	**Ph. 2**	**Ev.**
love	affection	1	1	0	0.01071	0.01303	0.00232	0.136	0.161	0.025
love	miss	0.61	0.77	0.16	0.00652	0.01005	0.00353	0.151	0.192	0.041
affection	miss	0.30	0.49	0.19	0.00326	0.00633	0.00307	0.081	0.131	0.050
miss	thoughts	0.30	0.33	0.03	0.00326	0.00428	0.00102	0.147	0.168	0.021
love	hope	0.43	0.3	-0.13	0.00466	0.00391	-0.00075	0.123	0.104	-0.019
love	pain	0.26	0.27	0.01	0.00279	0.00354	0.00075	0.089	0.094	0.005
love	desire	0.57	0.27	-0.30	0.00605	0.00354	-0.00251	0.189	0.114	-0.075
love	life	0.30	0.24	-0.06	0.00326	0.00316	-0.00010	0.036	0.041	0.005
love	dream	0.22	0.21	-0.01	0.00233	0.00279	0.00046	0.071	0.074	0.003
love	happiness	0.22	0.21	-0.01	0.00233	0.00279	0.00046	0.072	0.082	0.010
life	death	0.35	0.21	-0.14	0.00372	0.00279	-0.00093	0.081	0.086	0.005
affection	desire	0.22	0.20	-0.02	0.00233	0.00261	0.00028	0.078	0.092	0.014
life	hope	0.22	0.19	-0.03	0.00233	0.00242	0.00009	0.058	0.074	0.016
love	sadness	0.17	0.17	0	0.00186	0.00223	0.00037	0.042	0.053	0.011
thoughts	dreams	0.26	0.17	-0.09	0.00279	0.00223	-0.00056	0.240	0.177	-0.063
affection	life	0.26	0.16	-0.10	0.00279	0.00205	-0.00074	0.033	0.029	-0.004
thoughts	soul	0.35	0.16	-0.19	0.00372	0.00205	-0.00167	0.176	0.100	-0.076
love	autumn	0.13	0.14	0.01	0.00140	0.00186	0.00046	0.043	0.060	0.017
love	suffering	0.17	0.14	-0.03	0.00186	0.00186	0	0.074	0.065	-0.009
life	time	0.14	0.14	0	0.00186	0.00186	0	0.072	0.072	0
miss	memories	0.14	0.14	0	0.00186	0.00186	0	0.103	0.103	0
affection	dream	0.13	0.13	0	0.00167	0.00167	0	0.048	0.048	0
affection	friendship	0.13	0.13	0	0.00167	0.00167	0	0.059	0.059	0
miss	dreams	0.13	0.13	0	0.00167	0.00167	0	0.108	0.108	0
love	soul	0.11	0.11	0	0.00149	0.00149	0	0.035	0.035	0

The results gathered following the application of the three formulas to the two time spans, as well as the evolution between them can be observed in *Table 1*. The connections are ranked by degree of $similarity(t_i, t_j)$, calculated on the basis of data extracted in phase 2.

By observing the data in Table 1, we can affirm that the connections that form between tags generally maintain their stability in time. This was deducted considering that the value of the evolution is insignificant compared to the values of similitude resulted in the two phases of analysis. Still, we must take into account the fact that there are certain exceptions from this rule and dealing with these exceptions is a subject we will attend to in future researches.

Following the quantification of the results from phase 1, we constructed a network, represented graphically by an undirected graph. In the case of the graph, the nodes are repre-

sented by tags and their dimensions are directly proportionate to the frequency with which they appear in the labeling of texts within the system. The connections between tags are represented by the edges labeled with the value produced by the calculation of similarity, $similarity(t_i, t_j)$ between t_i and t_j.

```
*Vertices 30
1 "love" x_fact 8 y_fact 8 ic OliveGreen bc OliveGreen
2 "affection" x_fact 7 y_fact 7 ic Red bc Red
3 "life" x_fact 6 y_fact 6 ic Orange bc Orange
4 "miss" x_fact 5 y_fact 5 ic Thistle bc Thistle
5 "thoughts" x_fact 4 y_fact 4 ic ForestGreen bc ForestGreen
6 "sadness" x_fact 4 y_fact 4 ic ForestGreen bc ForestGreen
7 "soul" x_fact 4 y_fact 4 ic ForestGreen bc ForestGreen
8 "hope" x_fact 2 y_fact 2 ic Periwinkle bc Periwinkle
9 "dream" x_fact 2 y_fact 2 ic Periwinkle bc Periwinkle
10 "pain" x_fact 2 y_fact 2 ic Periwinkle bc Periwinkle
11 "death" x_fact 2 y_fact 2 ic Periwinkle bc Periwinkle
12 "meditation" x_fact 2 y_fact 2 ic Periwinkle bc Periwinkle
13 "happiness" x_fact 2 y_fact 2 ic Periwinkle bc Periwinkle
14 "god" x_fact 1 y_fact 1 ic PineGreen bc PineGreen
15 "destiny" x_fact 1 y_fact 1 ic PineGreen bc PineGreen
16 "desire" x_fact 1 y_fact 1 ic PineGreen bc PineGreen
17 "loneliness" x_fact 1 y_fact 1 ic PineGreen bc PineGreen
18 "friendship" x_fact 1 y_fact 1 ic PineGreen bc PineGreen
19 "autumn" x_fact 1 y_fact 1 ic PineGreen bc PineGreen
20 "memories" x_fact 1 y_fact 1 ic PineGreen bc PineGreen
21 "time" x_fact 1 y_fact 1 ic PineGreen bc PineGreen
22 "separation" x_fact 1 y_fact 1 ic PineGreen bc PineGreen
23 "suffering" x_fact 1 y_fact 1 ic PineGreen bc PineGreen
24 "memory" x_fact 1 y_fact 1 ic PineGreen bc PineGreen
25 "wisdom" x_fact 1 y_fact 1 ic PineGreen bc PineGreen
26 "dreams" x_fact 1 y_fact 1 ic PineGreen bc PineGreen
27 "light" x_fact 1 y_fact 1 ic PineGreen bc PineGreen
28 "faith" x_fact 1 y_fact 1 ic PineGreen bc PineGreen
29 "weeping" x_fact 1 y_fact 1 ic PineGreen bc PineGreen
30 "theater" x_fact 1 y_fact 1 ic PineGreen bc PineGreen
*Edges
1 2 1
1 4 0.77
2 4 0.49
4 5 0.33
1 8 0.3
1 10 0.27
1 16 0.27
1 3 0.24
1 9 0.21
1 13 0.21
3 11 0.21
2 16 0.2
3 8 0.19
1 6 0.17
5 26 0.17
2 3 0.16
5 7 0.16
1 19 0.14
1 23 0.14
3 21 0.14
4 20 0.14
2 9 0.13
2 18 0.13
4 26 0.13
1 7 0.11
1 20 0.11
2 13 0.11
1 5 0.1
1 11 0.1
1 24 0.1
2 6 0.1
2 20 0.1
3 7 0.1
4 9 0.1
4 23 0.1
1 26 0.09
2 10 0.09
4 10 0.09
4 24 0.09
5 20 0.09
6 19 0.09
10 23 0.09
22 29 0.09
14 28 0.09
```

Fig. 4. Representing the graph in *.paj format

For the connections' graphic representation, we have created an export module within the system implemented on *intelepciune.ro* that extracts useful data and generates a *.paj file. This is a special format which contains the graph's representation and which can be imported to an analysis program for social networks. The structure of the generated file can be observed in Figure 4.

In the following we have put together the graph's representation based on the Kamada-Kawai algorythm, employing the option to draw the disconnected components. Other possible representations based on the

Kamada-Kawai algorithm could have been made by setting the centroid in the middle or fixing the first and last nodes. As an alternative to the Kamada-Kawai algorithm we have the Fruchterman-Reingold algorithm, with a 2D or 3D representation option.

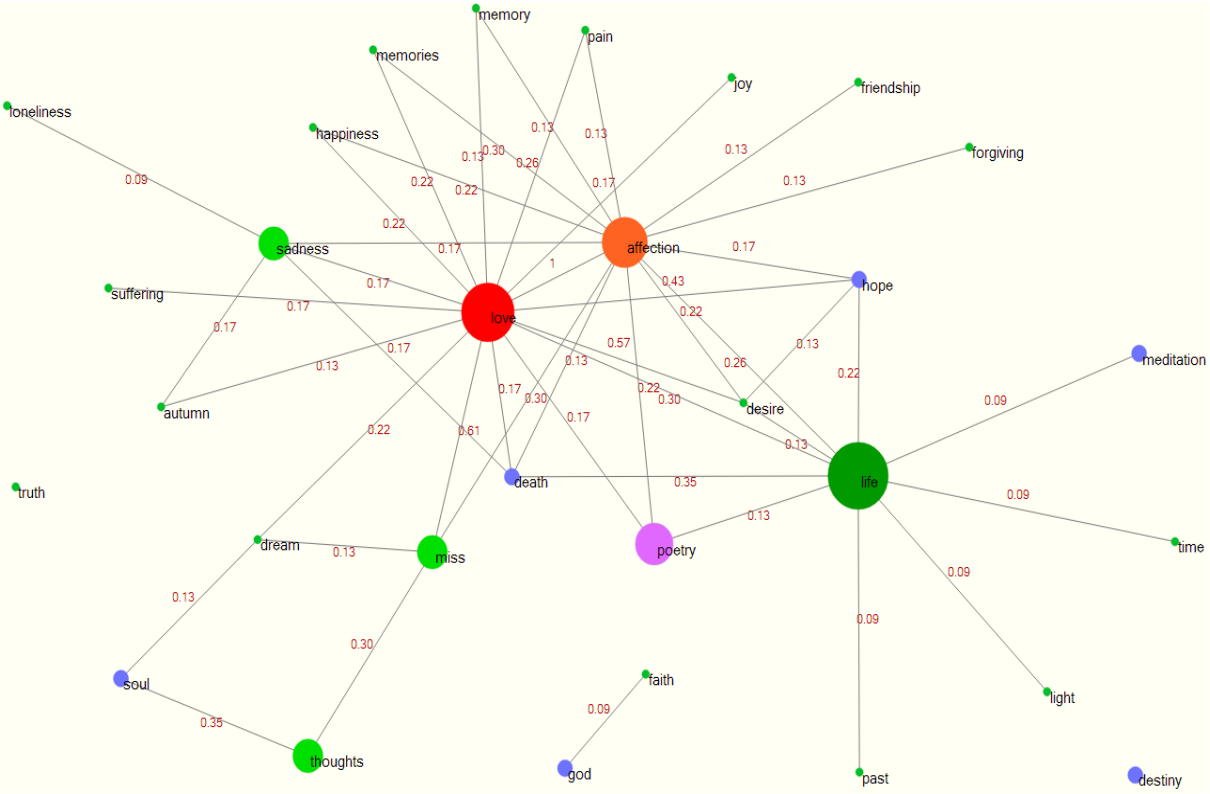

Fig. 5. Visualizing the relevant connections between tags, phase 1

In Figure 5 we can observe that the most popular tag is *life*, followed by *love* and the strongest connection is between affection and love. We can also see that the *love* tag is the central node in this network.

The graph in Figure 6 is the result of the quantification of the results from phase 2. In putting together this graph the same methodology used in the drawing of the graph from phase one was employed. As such, it can be seen that the most popular tag is *love*, followed by *affection* and *life*, and the strongest connection remains that between *love* and *affection*. Furthermore, we observe that the *love* tag remains the central node in this network as well. One other thing that can be observed is the fact that together with the system's maturing, a part of the existing connections between the most popular tags lose their intensity.

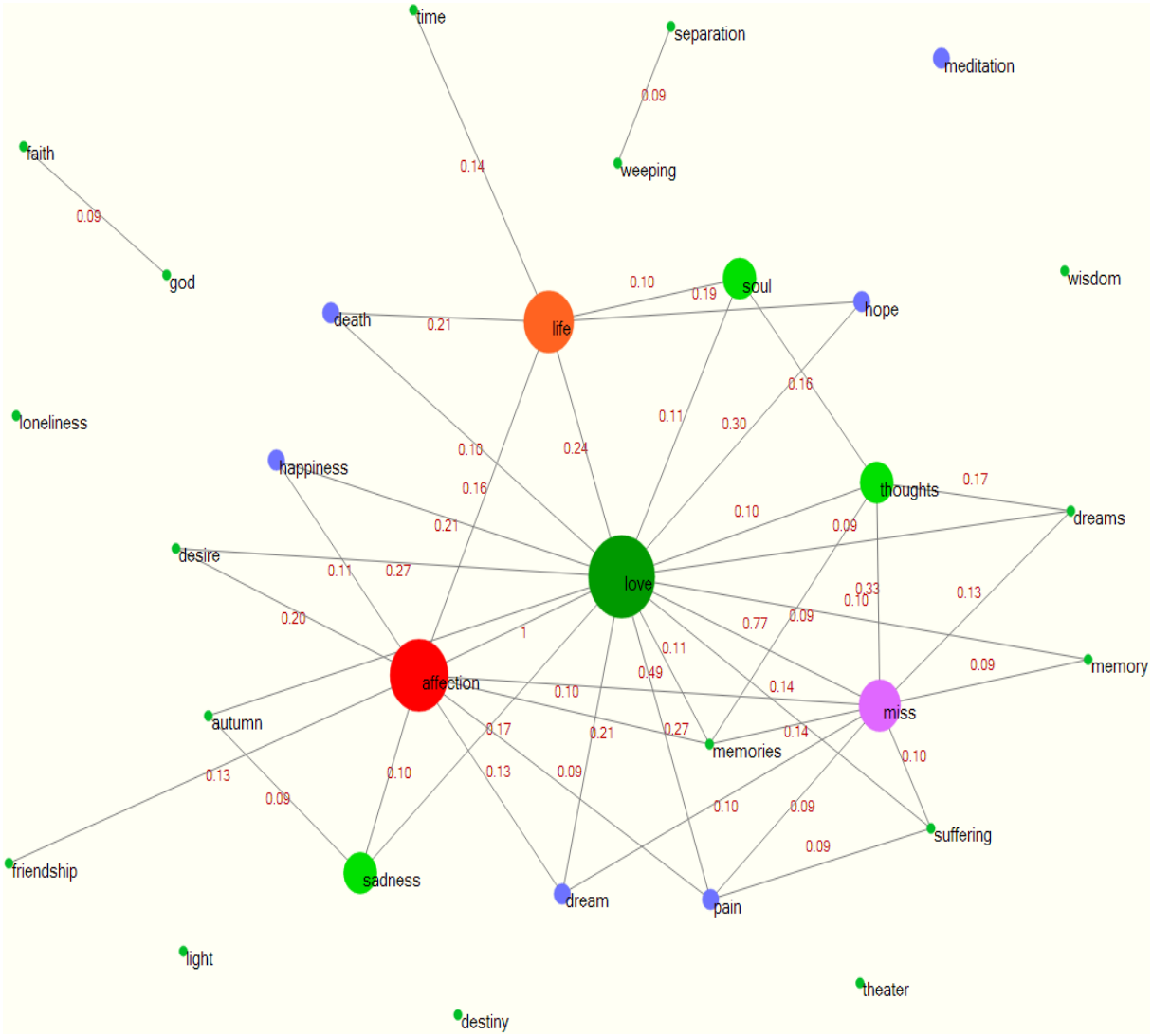

Fig. 6. Visualizing the relevant connections between tags, phase 2

Representing tags and the connections between them in the form of a graph and applying the specific methodology in analyzing networks can provide important aspects referring to the way in which users attribute tags, as well as the way in which certain semantic structures within the collaborative tagging systems appear. In his paper, Freeman [2] stresses the idea that everyone agrees with: centrality is an especially important structural attribute of social networks. One of the given measures for calculating centrality is the degree of normalized centrality $C'_D(t_i)$ of a node that has its formula presented onwards and which we have adapted to our case:

$$C'_D(t_i) = \frac{\sum_{i=1}^{n} a(t_i, t_j)}{n-1}$$

In this formula, $a(t_i, t_j)$ represents the number of tags adjacent to the t_i tag and n represents the number of tags present within the graph.

The evolution of the degree of centrality and of the market share for the most popular tags can be seen in Table 2. These measures help us in identifying the dominant tags and having a global view of their popularity and influence within the system. The market share in this case represents the proportion of the total tags held by a certain tag. In the table, CpC represents the market share calculated according to centrality and CpP represents the market share calculated according to the popularity of tags amongst users. The resulted values are calculated for the entire domain of thirty tags, but for exemplification purposes, only the first ten have been shown in the

table. The recordings in the table are organized by the tag's degree of centrality calculated according to the data extracted in the second phase.

Table 2. The evolution of the degree of centrality and the market share for tags

	$C'_D(t_i)$			CpC			CpP		
Tag	**Ph. 1**	**Ph. 2**	**Evol.**	**Ph. 1**	**Ph. 2**	**Evol.**	**Ph. 1**	**Ph. 2**	**Evol.**
love	0.552	0.586	0.034	0.182	0.193	0.011	0.129	0.155	0.026
affection	0.483	0.345	-0.138	0.159	0.114	-0.045	0.113	0.131	0.018
miss	0.138	0.310	0.172	0.045	0.102	0.057	0.034	0.055	0.021
life	0.345	0.207	-0.138	0.114	0.068	-0.045	0.146	0.115	-0.031
thoughts	0.069	0.172	0.103	0.023	0.057	0.034	0.034	0.037	0.003
pain	0.069	0.138	0.069	0.023	0.045	0.023	0.018	0.028	0.010
memories	0.069	0.138	0.069	0.023	0.045	0.023	0.016	0.018	0.002
soul	0.069	0.103	0.034	0.023	0.034	0.011	0.031	0.036	0.005
dreams	0.103	0.103	0.000	0.034	0.034	0.000	0.019	0.028	0.009
sadness	0.172	0.103	-0.069	0.057	0.034	-0.023	0.034	0.036	0.001

It is interesting to observe that the result of the market share differs depending on the entry data used. For example, in the cases of the tags *miss* and *life* the difference between the two market shares differs substantially. In the case of the *miss* tag, we have high values for phase 2 in $C'_D(t_i)$ = 0.310 and CpC = 0.102 and a greatly reduced value for CpP = 0.0055. This indicates that this is a tag with a strong influence on the other tags, but with a reduced popularity in what concerns its use in text labeling. Unlike the tag *miss*, *life* is an especially popular tag in labeling texts, but with a reduced influence on other tags, a claim sustained by the CpC value of 0.068, much lower than the CpP value of 0.115.

Collaborative tagging and tags offer semantic information on labeled resources, on user preferences and their evolution in time. Due to the fact that they sum up user associations, they can be used in constructing recommendation systems for tags, resources and users.

One example can be the case in which a user executes a search following a tag and the system can supply him with a list of the most similar tags for the respective search. Likewise, at the moment in which a user wishes to publish a text and finishes filling out the textbox for the first tag, the system can automatically fill in the other four textboxes with tags similar to the imputed one. This can lead to a notable increase in usability of the form used for text insertion and brings with itself a series of advantages. Among them, we would mention the easing of the manual labeling process, eliminating spelling mistakes which may arise and increasing usability. The recommendation based on extracting similarities between tags which have been used together by the community works very well regarding users, as well as new resources in the system.

In Table 3 we have simulated, for exemplification purposes, a number of five recommendations for four of the most popular tags extracted within the experiment. The recommendation lists were made based on the data produced as a result of applying the three formulas for calculating similarity presented above: $similarity(t_i, t_j)$, $support(t_i \cup t_j)$ and $cosine(t_i, t_j)$. We can see that the recommendations resulted from applying $similarity(t_i, t_j)$ and $support(t_i \cup t_j)$ are equivalent, but differ from the other results following the use of the similarity formula $cosine(t_i, t_j)$.

Table 3. Simulating some recommendations for the most popular tags

$similarity(t_i, t_j)$	$support(t_i \cup t_j)$	$cosine(t_i, t_j)$
love		
affection - 1.000	affection - 0.01303	miss - 0.192
miss - 0.770	miss - 0.01005	affection - 0.161
hope - 0.300	hope - 0.00391	desire - 0.114
pain - 0.270	pain - 0.00354	hope - 0.104
desire - 0.270	desire - 0.00354	pain - 0.094
life		
love - 0.240	love - 0.00316	death - 0.086
death - 0.210	death - 0.00279	human - 0.082
hope - 0.190	hope - 0.00242	hope - 0.074
affection - 0.160	affection - 0.00205	time - 0.072
time - 0.140	time - 0.00186	love - 0.041
affection		
love - 1.000	love - 0.01303	love - 0.161
miss - 0.490	miss - 0.00633	miss - 0.131
desire - 0.200	desire - 0.00261	desire - 0.092
life - 0.160	life - 0.00205	friendship - 0.059
dream - 0.130	dream - 0.00167	dream - 0.048
miss		
love - 0.770	love - 0.01005	love - 0.192
affection - 0.490	affection - 0.00633	thoughts - 0.168
thoughts - 0.330	thoughts - 0.00428	affection - 0.131
memories - 0.140	memories - 0.00186	dreams - 0.108
dreams - 0.130	dreams - 0.00167	memories - 0.103

Based on the tags users associate to the content, a personalized search and content recommendation system can be built. One such system can supply personalized content taking into account a document's taxonomy as well as the social information collected from users.

5 Conclusions

The research carried out in this article is based on the data collected within the social tagging system developed and implemented on Întelepciune.ro. Following the analysis of collective intelligence concerning the way in which community members associate different tags, we have observed that with time, connections between tags are outlined. Moreover, connections between tags are maintained and become more stable with the passage of time. The results obtained have a high degree of similarity towards all calculation formulas that were identified and used. Therefore, we can assert that the calculated degree of similarity between tags can be used in building recommender systems. These will be qualified to recommend tags, resources and users. Likewise, we consider that social annotation by means of tags can offer a contextual extension to any recommender system.

As a result of the representation of tags and the connections between them as a graph, we have applied the network analysis specific methodology. By studying different measures such as centrality, the market share and the market share depending on centrality, we managed to have a global vision on the tagging network. We thus managed to extract information referring to tag popularity, their influence within the network and the extent to which a tag depends upon another. By analyzing the way in which users attribute tags we managed to determine different semantic structures within the social tagging network

and see their evolution at different times.
For the future, we aim to continue the research from [10] towards improving the model through integration of a recommender system that uses tags. This way, we will be able to identify experts and trustworthy content by different categories of interest. We also wish to bring to fruition a hybridization through aggregation with the system, based on extracting association rules from navigation sessions, as proposed in [12]. We will thereby orient our research towards propositioning a hybrid system whose aim will be to solve the problems present in classic recommender systems.

References

[1] R. Agrawal, T. Imieliński, A. Swami, "Mining association rules between sets of items in large databases", in *Proc. of the 1993 ACM SIGMOD international Conference on Management of Data*, Washington, D.C., 1993, pp. 207-216

[2] L. C. Freeman, "Centrality in social networks conceptual clarification", *Social Networks*, vol. 1, no. 3, pp. 215-239, 1979

[3] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems", *Journal of Information Science*, vol. 32, no. 2, pp. 198-208, 2006

[4] J. T. Hackos, *Content Management for Dynamic Web Delivery*, Wiley, 1st edition, Indianapolis, Indiana, 2002

[5] H. Halpin, V. Robu, H. Shepherd, "The complex dynamics of collaborative tagging", in *Proceedings of the 16th international conference on World Wide Web: ACM, WWW '07*, New York, NY, USA, 2007, pp. 211-220

[6] S. Hayman and N. Lothian, "Taxonomy Directed Folksonomy", in *Proceedings of the World Library and Information Congress: 73rd IFLA General Conference and Council*, Durban, South Africa, 2007

[7] P. Keller, (2005), Tags: Database schemas. [Online]. Available: http://www.pui.ch/phred/archives/2005/04/tags-database-schemas.html

[8] T. Krohn, M. C. Kindsmüller, M. Herczeg, "myPIM: A Graphical Information Management System for Web Resources", In *Proceedings of the 3rd International Conference on the Pragmatic Web: Innovating the Interactive Society (ICPW '08), ACM*, New York, NY, USA, 2008, pp. 3-12

[9] C. Marlow, M. Naaman, D. Boyd, M. Davis, "Position paper, tagging, taxonomy, flickr, article, toread", in *Collaborative Web Tagging Workshop, at WWW2006*, Edinburgh, Scotland, 2006

[10] D. Mican, L. Mocean, N. Tomai, "Building a Social Recommender System by Harvesting Social Relationships and Trust Scores between Users", In *Business Information Systems Workshops*, LNBIP, Volume 127, Part 1, Springer, Berlin Heidelberg, 2012, pp. 1-12

[11] D. Mican and N. Tomai, "Web 2.0 and Collaborative Tagging", in *Proceedings of the 2010 Fifth International Conference on Internet and Web Applications and Services, ICIW 2010*, IEEE Press, Barcelona, Spain, 2010, pp. 519-524

[12] D. Mican N. Tomai, "Association-rules-based recommender system for personalization in adaptive web-based applications", in *Proceedings of the 10th international conference on Current trends in web engineering (ICWE'10)*, Springer-Verlag, Berlin Heidelberg, 2010, pp. 85-90

[13] I. Peters, *Folksonomies. Indexing and Retrieval in Web 2.0 (Knowledge and Information)*, De Gruyter; 1 edition, Berlin, Germany, 2009

[14] H. Wu, M. Zubair, K. Maly, "Harvesting social knowledge from folksonomies", in *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia*. *Hypertext '06*. ACM, New York, NY, 2006, pp. 111-114

[15] S. Xu, S. Bao, B. Fei, Z. Su, Y. Yu, "Exploring folksonomy for personalized search", in *Proceedings of the 31st Annual international ACM SIGIR Conference on Research and Development in infor-*

mation Retrieval. SIGIR '08, ACM, New York, NY, 2008, pp. 155-162

[16] S. Zhao, N. Du, A. Nauerz, X. Zhang, Q. Yuan, R. Fu, "Improved recommendation based on collaborative tagging behaviors", in *Proceedings of the 13th international conference on Intelligent user interfaces*, ACM, New York, NY, USA, 2008, pp. 413-416

Daniel MICAN is teaching assistant and webmaster at Faculty of Economics and Business Administration, Babeş-Bolyai University of Cluj-Napoca. He holds a PhD diploma from 2013 and his main research areas are focuses on: web applications, content management systems, collective intelligence, recommendation systems, search engine optimization, online marketing.

Nicolae TOMAI is full Professor at Faculty of Economics and Business Administration, "Babeş-Bolyai" University of Cluj-Napoca. His main research areas are: Fundamentals of Computer Science, C#, Computer Networks and Distributed Systems, E-business, Mobile systems His work includes 19 books, 71 scientific papers published, 1 patent, innovation patent 5, active member in 18 research contracts.