Web 2.0 And Collaborative Tagging

Daniel Mican
Dept. of Business Information Systems
Babeş-Bolyai University
Cluj-Napoca, Romania
e-mail: daniel.mican@econ.ubbcluj.ro

Nicolae Tomai
Dept. of Business Information Systems
Babeş-Bolyai University
Cluj-Napoca, Romania
e-mail: nicolae.tomai@econ.ubbcluj.ro

***Abstract*—The aim of the present paper is to analyze the user's behavior in Web 2.0 through comparing statistical data from a Web 1.0 site and an online community from Web 2.0. In order to emphasize the difference between the two concepts, we have taken into consideration the following: new vs. returning visitors, visitor loyalty, length as well as the depth of the visit to the site. Analyzing the data, one can observe that the visitor loyalty, their lengthy visits to the site and the pages they view are much higher for the site built on the principles of Web 2.0, compared to the classical sites. Likewise, we developed and implemented a collaborative tagging system that allows every registered user to publish, save and label the content with the help of tags. Therefore, this can help in building a model that not only offers customized content but also adapts the site to user needs by providing a recommendation system and aggregate collective intelligence.**

Keywords-component; web 2.0; web analytics; folksonomies; tags; collective intelligence

I. INTRODUCTION

New technologies have created a new dynamic web, housed in a constant evolution, that adapts itself everyday more and more to users' needs and becomes an increasing part of our lives. Social networking sites, such as wiki and blogs, along with the free flow of information, web services, RSS, are designed to link people from all over the world together depending on their common interests. Therefore, a large number of companies have focused their resources on the development of certain applications that allow users to publish, organize and share information depending on their needs.

In order to understand the users' navigation behavior and to verify the pulse of various sections within a web application, it is necessary to analyze the web traffic. The analysis of the web traffic results in a series of statistical information that allows us to get a comprehensive picture of the application and to identify what works and what does not, from the visitor point of view. The use of Web Analytics is essential in order to improve the quality of the content, the various sections and the way of navigation. The literature affords but little information about Web Analytics and the analysis of web traffic. In [3], has been investigated whether the advanced web metrics that are calculated using Google Analytics could be used to evaluate the general usability of e-commerce sites and to identify the sections containing possible usability problems. Thus, a set of metrics resulted that are useful in order to identify not only the general usability problems, but also the specific ones. As to conclude, we consider appropriate to use more than one metrics in the web traffic analysis, in order to observe the existent differences between a read-only site and a read-write web application.

The new generation of web applications allows the users to control and to move from consumers to producers of information. Therefore, an increasing concern for collaborative tagging appeared [1, 2, 11, 12], that is defined as the process by which many users add metadata in the form of keywords to the content they publish or save. Due to the popularity of the collaborative tagging system we developed and implemented such a system on Intelepciune.ro and we analyzed collective intelligence aiming to discover new relations between tags.

The present paper is organized as follows: Section 2 makes a brief foray into the world of Web 2.0 and collective intelligence. Section 3 describes the phenomenon of collaborative tagging with regard to a personalized classification of the content, using tags as well as state of the art. In Section 4, we analyzed in detail the web traffic on two sites, namely one from the Web 1.0 generation and the other one from Web 2.0, using metrics such as: new vs. returning visitors, visitor loyalty, length of the visit as well as depth of the visit. Finally, Section 5 contains the experiment in which we analyzed data from the collaborative tagging system that has been implemented on Intelepciune.ro, in order to discover the existence and the intensity of certain relationships between tags. We conclude the paper by presenting our future plans.

II. WEB 2.0

Lately, people have been talking a lot about Web 2.0, concept that has become very popular among web users and developers. Its popularity has grown exponentially, due to the fact that the users have suffered a change of state, namely from consumers to producers of information. Not only their needs have changed radically but also the technology has adapted to their increasingly dynamic needs. The applications built up on the typology of Web 2.0 enables users to control the data, to communicate and to collaborate with the community members they belong to.

The concept of Web 2.0 rejoins the new generation of web applications and it has been introduced for the first time during a conference between O'Reilly and MediaLive

International [8]. The term makes a difference between them and the traditional websites belonging to Web 1.0. According to [6], Web 2.0 sheds new light on the way in which the processes are automated within these web applications and the principles underlying their development. As stated by [5], Web 2.0 means using technologies and standardized interfaces in order to generate the content in a dynamic way and to personalize it according to user needs. In [6], Web 2.0 is defined as the philosophy that maximizes the collective intelligence and the value added by each participant in a dynamic and formalized system of creation and exchange of information.

Collective intelligence is composed of all participants knowledge, that is distributed in a group, being at the same time a dynamic result that adapts itself to the constantly changing environment [4]. Likewise, collective intelligence implies collecting data from a group or even separate groups of users, combining as well as analyzing them. Collecting and analyzing the responses within a group, allows the extraction of statistical findings that reflect the collective ideas, the behavior and the preferences of an online community [7, 9].

In recent years, the rapid development of Web 2.0 web applications has generated great interest in collaborative tagging, which is also known as "folksonomy". The word "folksonomy" is formed from "folk" and "taxonomy" and is used to describe the social phenomenon of classification [1, 10, 11]. Taking into consideration user-specific data, the information retrieval systems and search engines can extend their functionalities by understanding users' wishes. When a search is carried out by analyzing collected data, these systems can generate personalized searches. Due to this fact among others, the personalized search represents one of the most promising directions heading towards the search paradigm [11].

III. FOLKSONOMIES AND COLLABORATIVE TAGGING

Collaborative tagging is described as the process through which many users add metadata in the form of keywords to the content which they publish or save. This concept is treated extensively in [1].

Tag-collaborative annotation [11], allows personalized description of the content and is a good indicator of user interests. Tag classification is an ad-hoc, user-centered classification of resources which can be shared within the community by the user. Tags are defined as data about data attached to an object (metadata).

The analysis of one object shows that it can acquire different meanings for different users. Therefore, it is very important for the user to have the possibility to classify the desired objects at their own discretion. It is well known that an object can be characterized by multiple attributes and that each user will differently rank the attributes that help him in identifying the object. Let's take, for example, Mr. Smith and the connections that various people have with him. Mr. Smith will be labeled by a subordinate as a colleague or the boss, by his wife as her husband, by his grandchildren as their grandfather, and so on. Consequently, each person that comes into contact with Mr. Smith assigns one or more tags to classify him. The same happens within the context of web application. Each person perceives the objects differently and classifies them according to their background knowledge.

Another essential aspect is the grouping of people that perceive different objects in the same way and the evolution of the connections between them. These relations vary with time over shorter or longer intervals. For instance, the subordinates that tag Mr. Smith as their boss may have the status of subordinates just for a short period of time, compared to his grandchildren. Likewise, the connections that exist among different tags, the intensity as well as their evolution in time also have to be taken into consideration. The frequency of certain links between tags is used to create tag classes that can be exploited in order to anticipate certain user behavior patterns.

Moreover, on the basis of tags that users apply to content, a personalized search and information retrieval system can be created. The goal of this system is to provide personalized content, adapted to user needs, which should take into consideration not only the taxonomy of a file but also the social information collected from the users.

Adding tags to content is a practice increasingly used in order to organize content and enable better navigation and subsequent easy searching. An increasingly large number of websites have successfully implemented systems that allow collaborative tagging and site users to freely attach the desired tags to their content.

In what follows we will present some related work concerning folksonomies and collaborative tagging systems. Reference [2] has used data from the social bookmarking site del.icio.us in order to analyze the dynamics of the collaborative tagging systems. The analysis of the distribution of the most common tags has shown the way in which the significance of certain tags results from their relationship with other tags. Thereby, a new generative model of collaborative tagging has resulted that helps the understanding of the dynamics behind the tagging phenomenon and the way in which a stabilized power law distribution of tags has come into being. As stated in [1], the structure of the del.icio.us tagging collaborative system has been analyzed in order to discover regularities in users' activities, tag frequency, the type of tags used, the increase in the popularity and their stability for a certain URL. Likewise, a new dynamic model of collaborative tagging is presented, which allows one to predict stable models based on common knowledge. In [11] a framework for automatic folksonomy evaluation has been proposed that is based on data from del.icio.us and Dogear. Moreover, [12] has proposed the TBCF, which is a new collaborative filtering approach based on the semantic distance between the associated tags by different users in order to improve the efficiency of selecting neighbors. The experimental results have shown that the TBCF brings significant improvements compared to the traditional cosine-based recommendation method.

IV. WEB 2.0 VS. WEB 1.0 CASE STUDY

In order to highlight the differences between the two generations, in this section we analyzed in detail the web

traffic on two sites: one belonging to Web 1.0 and the other to Web 2.0. Therefore, we used metrics such as: new visitors vs. returning visitors, visitor loyalty, length of visit, depth of visit.

The first metric used in our study refers to new vs. returning visitors. The new vs. returning visitors report reveals information with regard to the new visitors compared to the returning visitors on the site. The behavior of a new visitor can be quite different from that of a returning visitor. It is very important to understand these two types of behavior in order to adapt the site to visitor needs. Thus, a large number of new visitors could suggest the fact that the owner is successful in bringing new visitors to the site, whereas a large number of returning visitors could suggest that there is sufficiently attractive content to motivate the visitor to return.

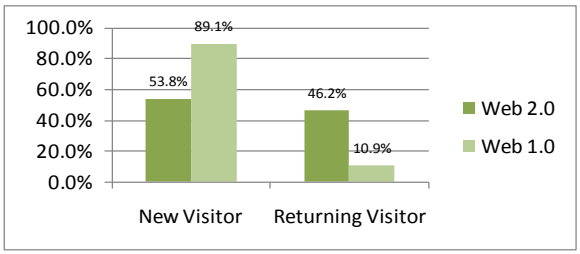

Figure 1. New vs. returning visitors

In Figure 1 we observe that in the case of Web 2.0 community more than 46.2% of the visitors return to the site, compared to just 10.9% for the Web 1.0 site. By dividing the two values we observe that Web 2.0 site is 4.2 more popular than Web 1.0 site.

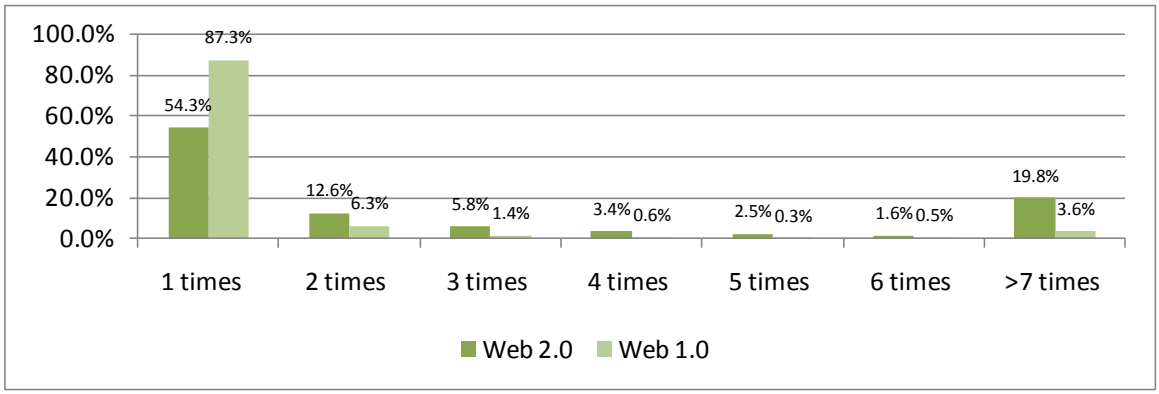

Figure 2. Visitor Loyalty

In what follows we will study the visitor behavior taking into consideration the visitor loyalty. Loyalty report helps us study the visitor loyalty observing how many times they returned to the site. Figure 2 illustrates that the visitors develop a much higher loyalty for the Web 2.0 compared to the Web 1.0 site. Therefore, we can observe that in the case of Web 2.0 more than 19,8% of the visitors return to the site more than seven times compared to just 3.6% for the Web 1.0 site. Regarding the percentage of the visitors that return to the site more than three times, we obtain, totaling the percentage, 33.1% for the Web 2.0 and 6.4% for the Web 1.0 site.

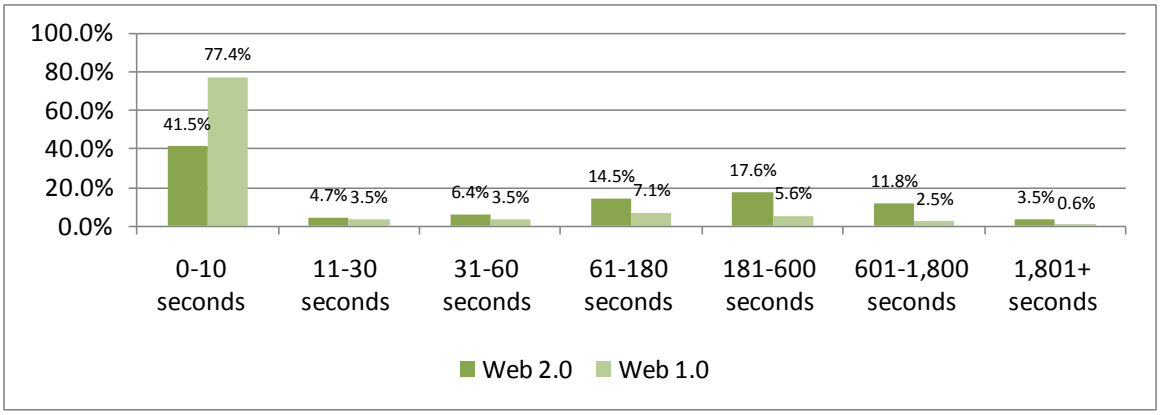

Figure 3. Length of visit

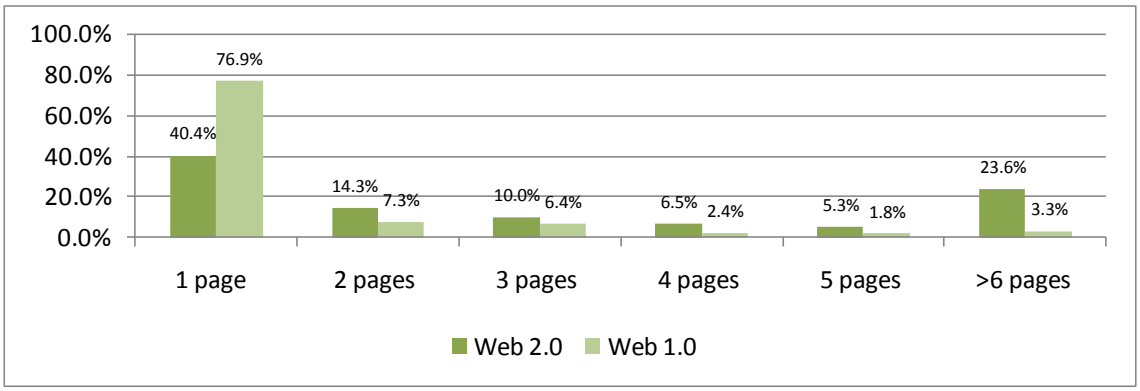

Figure 4. Depth of visit

An important criterion that has to be taken into consideration is the length of visit to the site. The length of visit gives an indication of the quality of the visit to the site. A large number of lengthy visits, suggests that there is high quality content and great functionality on the site, which leads visitors to interact more with it. Analyzing Figure 3, one can observe that approximately 80.8% of the visitors to the Web 1.0 site do not spend more than 30 seconds on it, compared to just 46.2% for the Web 2.0 site. The total number of visitors that spend more than sixty seconds on the site can be further split according to site type: 32.9% for the Web 2.0 site and 8.6% for the Web 1.0 site respectively, the former being 3.82 times greater than the latter.

In order to do a comparative analysis for the two sites, we consider analyzing the depth of the visit to be very useful. The depth of the visit is a measure that gives information about the quality of the visit to the site. A large number of pages viewed per visit suggest high usability, high functionality as well as quality content.

Therefore, Figure 4 shows that 76.9% of the users on the Web 1.0 site visit only one page compared to 40.4% for the Web 2.0 site. However, the visitors that find high usability and functionality, visiting more than 4 pages, amount to 35.4% on the Web 2.0 site in contrast to just 7.5% for the Web 1.0 site, the former being 4.75 times greater than the latter.

For a better understanding of visitor behavior and to support the comparative analysis regarding the two site types, one can also use the following indicators: average page views, time on the site, bounce rate, recency as well as other reports provided by web analytics tools.

Designing a website which can provide its visitors with additional value, has become an increasingly difficult task nowadays. Therefore, the designer must anticipate the diversity in user needs which differ from one user group to the next and which change over time. The understanding of user needs requires careful and rigorous analysis of the navigation behavior and of the interaction with the web pages. This helps developers in designing a software application that is easy to customize and adapt to the needs of each category of identified users.

Figure 5 shows information concerning the age distribution of the active members registered on Intelepciune.ro website. The analysis of the age groups of the internet users and the members of the site are very important, because the level of knowledge in using computer can differ considerably from one age category to the next. Certain user categories can adapt very quickly to the new forms of technology and interfaces, unlike others that show reticence towards major changes in functionalities and the way the information is organized.

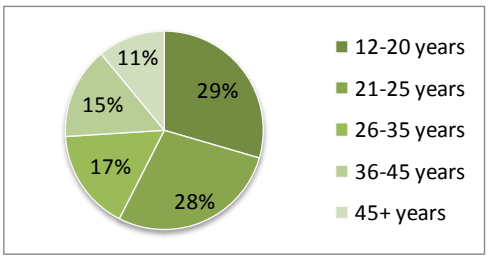

Figure 5. Age distribution

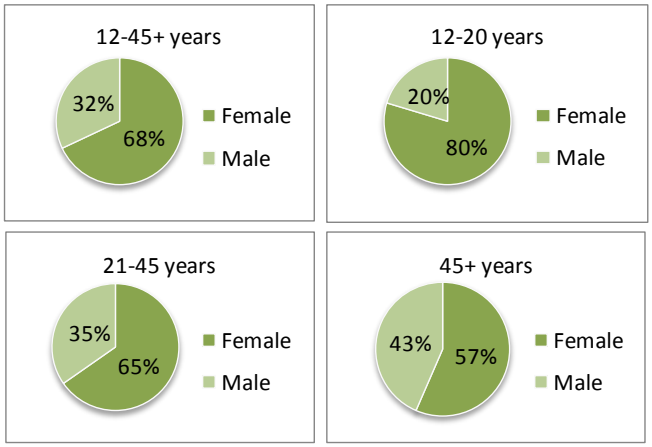

Figure 6. Gender distribution

Analyzing Figure 6 we observe a very interesting aspect namely that for the age category 12-20 years, the percentage of female members is 80, whereas the percentage of male members is only 20. This distribution is changing a great deal in the sector of users over 45 years old.

V. TAG CORRELATION ANALYSIS

Due to the capacity of collecting user data, collaborative tagging systems have the potential to become infrastructural technology that supports knowledge management activities in an organization or society, and at the same time have become a challenge for researchers in the field [10].

The many challenges posed by collaborative tagging as well as the desire to deliver personalized content and adapt the site to user needs motivated us to build and implement a collaborative tagging system on Intelepciune.ro, that is the biggest Romaian cultural site, built on Web 2.0 principles and that has more than 250.000 unique visitors per month.

The implemented system allows every registered user to publish, save and label the content with the help of tags. It also permits the extraction of information about the members of the community, their preferences for certain subjects of interest and the way they apply different tags to the existing content on the site. By collecting and analyzing the users data we observed that some tags combinations repeat themselves with a certain frequency thus, resulting various similarity between them. In order to measure these similarities between tags we used, as a coefficient of correlation, the frequency with which these were used with the others by a user to tag a web document.

This study involved a process of classification as well as extracting the first 30 tags from the social tagging system according to their popularity among users. In order to determine the maximum number of possible connections that may exist between tags, we calculated combinations of 30 taken 2 at a time.

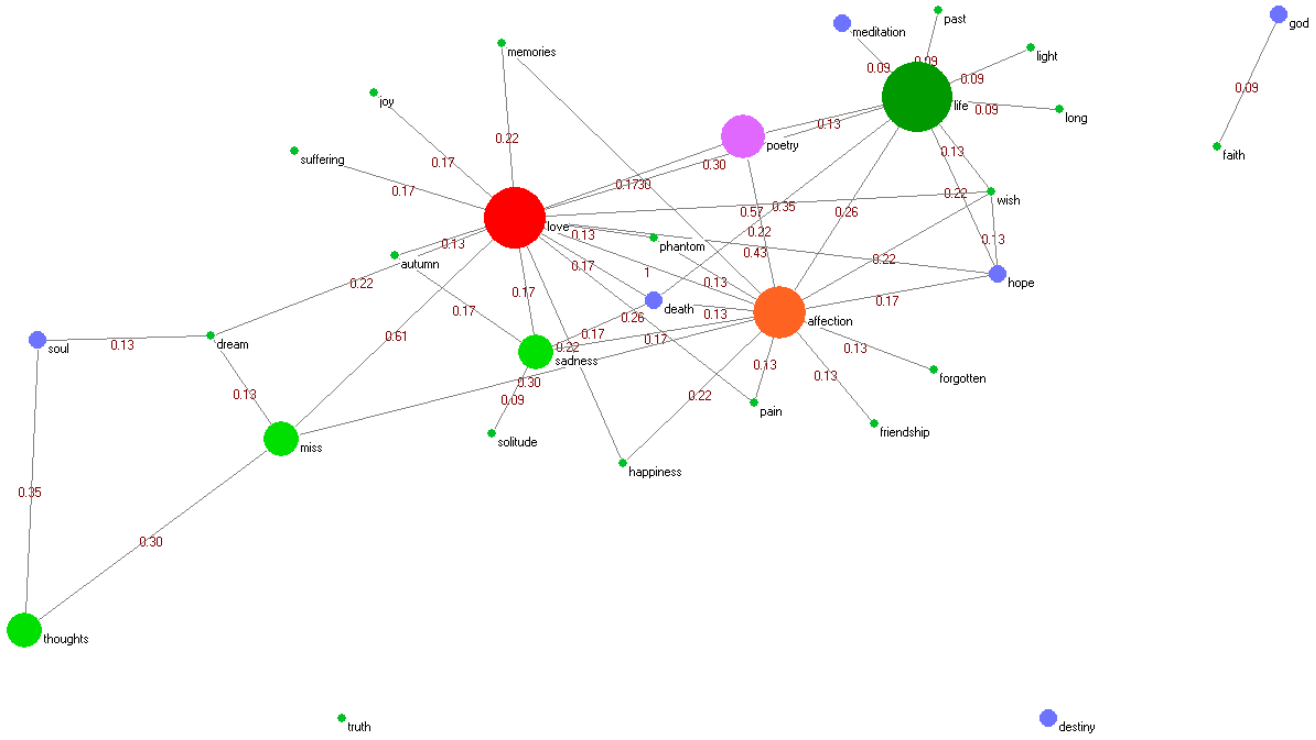

Figure 7. Tag correlation analysis

After the quantification of the frequency with which these connections appear in user preferences, we observed that from 435 possible connections only 46 have been carried out. In order to compare the results obtained we used the normalization equation, presented in [9]. Consequently, for the existing connections we obtained values included within the [0, 1] interval. The description of the value hierarchy will be made in terms of how close they are to the best results, which will always take the 1 value. Furthermore, on the basis of the quantification results we built an undirected graph in which one can observe the popularity of the tags and the connections between them. Figure 7 shows the resulted graph, in which each tag is represented by a node and the connections among them are represented by the edges between the nodes.

VI. CONCLUSIONS AND FUTURE WORK

The research we have carried out reveals the fact that Web 2.0 represents the present of web applications, becoming increasingly popular among the developers and web users. The great popularity is due to the fact that the applications built on the principles of Web 2.0 allow users to control the data, to interact not only with the application but also with other users, and to communicate and collaborate with the members of the community they are part of. Additionally, Web 2.0 offers them the opportunity to express themselves freely, to create and choose which content to read or publish, and with whom to interact. All the advantages offered by the Web 2.0 sites is reflected in the user's behavior, that is to say, they become more loyal to the site, spend more time online, and the number of viewed pages is much higher compared to the classical sites.

The analysis of the collective information about the members of the community and the way they apply different tags brought us to the conclusion that connections are being created over time between users, tags and content. Therefore, the collective classification with tags can help building a model, which to provide personalized content and adapt the site to user needs. Likewise, the collaborative classification has the advantage that it can function as a recommendation system due to the connections that result between tags, and the fact that aggregates collective intelligence.

Based upon the similarities discovered between users, tags and content, one can make a series of recommendations enhancing the capacity of adaptation and personalization of web applications. Adaptive web applications can do this only after analyzing data resulted from the users' current and former interaction with the system by aggregating collective intelligence. Due to the fact that the preliminary results are promising we plan to focus our future research on Web 3.0 and semantic web by developing a web personalization system that can be used in adaptive web applications. The system will provide the web application with the capacity to offer personalized content and to adapt by anticipating users' needs and to provide them with the information and content they need.

ACKNOWLEDGMENT

This work is supported by the Romanian Authority for Scientific Research under project IDEI_573.

REFERENCES

[1] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," Journal of Information Science, vol. 32, no. 2, pp. 198-208, April 2006, doi: 10.1177/0165551506062337.

[2] H. Halpin, V. Robu, and H. Shepherd, "The complex dynamics of collaborative tagging," in WWW '07: Proceedings of the 16th international conference on World Wide Web. New York, NY, USA: ACM, 2007, pp. 211-220, doi: 10.1145/1242572.1242602.

[3] L. Hasan, A. Morris, and S. Probets, "Using Google Analytics to Evaluate the Usability of E-Commerce Sites," in Proceedings of the 1st international Conference on Human Centered Design: Held As Part of HCI international 2009 (San Diego, CA, July 19 - 24, 2009). M. Kurosu, Ed. Lecture Notes In Computer Science, vol. 5619. Springer-Verlag, Berlin, Heidelberg, pp. 697-706, doi: 10.1007/978-3-642-02806-9_81.

[4] R. Högg, M. Meckel, K. Stanoevska-Slabeva, and R. Martignoni, "Overview of business models for Web 2.0 communities," in Proceedings of GeNeMe 2006, Dresden, pp. 23-37.

[5] IEEE Professional Communication Society Newsletter, ISSN 1539-3593, Volume 51, Number 2, February 2007.

[6] K. Maruyama, M. Matsushita, and S. Yamamoto, "Japanese Workshop on Leveraging Web2.0 Technologies in Software Development Environments (WebSDE)," in Proceedings of the 21st IEEE/ACM international Conference on Automated Software Engineering, September 18 - 22, 2006, pp. 377.

[7] D. Mican and N. Tomai, "Web 2.0 as a New Vision of Web-Based Applications," Studia Universitatis Babes-Bolyai. Informatica, Special Issue 2009, pp. 39 - 42.

[8] T. Oreilly, "What is web 2.0: Design patterns and business models for the next generation of software," Social Science Research Network Working Paper Series, August 2007, doi: http://ssrn.com/abstract=1008839.

[9] T. Segaran, Programming Collective Intelligence: Building Smart Web 2.0 Applications, 1st ed. O'Reilly Media, August 2007.

[10] H. Wu, M. Zubair, and K. Maly, "Harvesting social knowledge from folksonomies," in HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia. New York, NY, USA: ACM, 2006, pp. 111-114, doi: 10.1145/1149941.1149962.

[11] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu, "Exploring folksonomy for personalized search," in SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 2008, pp. 155-162, doi: 10.1145/1390334.1390363.

[12] S. Zhao, N. Du, A. Nauerz, X. Zhang, Q. Yuan, and R. Fu, "Improved recommendation based on collaborative tagging behaviors," in IUI '08: Proceedings of the 13th international conference on Intelligent user interfaces. New York, NY, USA: ACM, 2008, pp. 413-416, doi: 10.1145/1378773.1378843.