Methodological issues for virtual communities analysis in the context of Big Data

Sergio L. Toral, Mª Rocio Martínez-Torres
University of Seville
Seville, Spain
storal@us.es, rmtorres@us.es

Nicoletta Fornara
Università della Svizzera Italiana, USI
Lugano, Switzerland
nicoletta.fornara@usi.ch

***Abstract*—Virtual communities represent today en emergent phenomenon through which users get together to create ideas, to obtain help from one another, or just to casually engage in discussions. Their increasing popularity as well as their utility as a source of business value and marketing strategies justify the necessity of defining some specific methodologies for analyzing them. The aim of this paper is providing new insights into virtual communities from a methodological viewpoint, highlighting the main trends and challenges.**

Keywords-Virtual communities; Big Data; Social network analysis; Semantic analysis; Non-reactive data collection..

I. Introduction

Virtual communities were born as places on the Web where people can find and then electronically 'talk' to others with similar interests. However, very quickly it became clear that virtual communities can also generate a business value (Chen et al., 2012). For instance, virtual communities can be leveraged to provide access to consumers and consumer data (Spaulding, 2010), to create new innovations (Chesbrough, 2006) or to support the generation of new developments (Martinez-Torres, 2014). The emergence of customer-generated Web 2.0 content on various forums like newsgroups, social media platforms and crowd-sourcing systems have propitiated new opportunities for researchers and practitioners, but it is also demanding new methodologies. The main challenge faced in this topic is dealing with the huge amount of information available. Moreover, this information is spread over websites, and it is non-structured, which means that the information is not structured in a database. On the contrary, data must extracted from users interactions and shared content.

In this context, the purpose of this paper is providing an in-depth analysis of the research flow chart in virtual communities analysis, Figure 1.

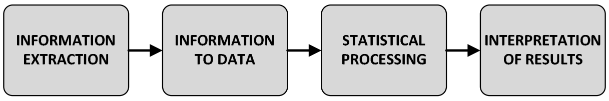

Figure 1. Research flow chart in virtual communities.

This chart includes the extraction of information, the transformation of this information into data, and the subsequent statistical analysis of data. It can be noticed the multidisciplinary scope of virtual communities analysis. The extraction of information consists not only on the software tools to access this information, but also choosing the relevant information in the context of virtual communities, above all, participation and shared content. It should also consider the ethical and legal rules for data collection and the policy for accessing and using the data published on Web sites stated by the data provider. The transformation of information into data refers to techniques like social network analysis and natural language processing as the most suitable techniques for extracting data from participation and shared content. Finally, the statistical processing includes the sets of statistical techniques for data mining.

will be explained using several case examples of virtual communities analysis.

The rest of the paper is structured as follows. Section II details the information extraction block of the research flow chart in virtual communities. Section III deals with the transformation of information into data, and Section IV illustrates several case examples including the statistical processing. Section V describes the future trends and challenges and finally, Section VI concludes the paper.

II. Information Extraction

Information extraction is mainly related to computer science and information systems disciplines. It involves computer science techniques for accessing the source code of web pages and extracting the relevant information, that is, information about users and shared content. A crawler is a software program that can follow the hyperlink structure of websites for accessing the desired information. However, the main challenge in this point is that websites have very different styles and they are programmed in a wide variety of formats. As a result, the crawler must be customized for each particular website. The crawler can be easily programmed using the most popular programming languages like Python, Matlab, Java or R. The best alternative consists of using a programming language supporting regular expressions, which facilitate the string characters processing.

Another key point for data extraction is deciding which is considered a relevant information. Basically, three types of data can be distinguished:

- Participation: users in virtual communities are usually registered with an alias or email. Using this identification, they can interact with other users of the community. For instance, they can post messages, ideas, reviews, innovation and these posted messages can receive answers, comments or even evaluations from other community users. Participation information refers to all the possible forms of interactions among users within the community.

- Content: the title and body of shared messages constitute another piece of information that can be analyzed in virtual communities. Additional elements of information are the content of answers and comments as well as the tags or keywords in which sometime messages are required to be classified.
- Other data: virtual communities can also incorporate additional sources of information that can be extracted. For instance, the number of readings a message have received, the reputation of users within the community, or their trust index.

All this information belongs to the so called non reactive data. It is not based on questionnaires but on the objective data and tracing users leave when they are part of a community. The main advantage of working with non reactive data is that the sample is actually the whole population. Once the crawler extract the desired information, it can actually reach all the users of the community. The disadvantage is that non reactive data in the context of Web 2.0 generate a huge volume of information that requires specific methodologies to be transformed into data, which is the topic of the next section.

III. Transforming Information into Data

A. *Participation*

The natural way of dealing with participation features is Social Network Analysis (SNA). SNA consists of modeling a community as a graph where nodes represent users indentified by their email or alias, and arcs represent the different types of interactions among users. A triple level of analysis can be done with social networks. The first level is the local one defined by the local topological properties of nodes as part of the network. The second level is the global one given by the global properties of the network as a whole. Finally, the third level refers to assimilating networks to complex network models such as random, free scale or small world networks.

The local level considers the interaction of a given node with its closer neighborhood, usually its one hop neighborhood. The in-out-all degree of a node is given by the number of in-out-all arcs incident on this node, and represents the number of other users interacting with him. It is a measure of participation intensity. Sometimes it is also interesting considering not only the participation intensity but the position of the user within the whole network. This value is given by centrality. However, centrality can be measured using several criteria. Closeness centrality is based on the distance between a given node and the rest of the community. It is a measure of the ability to reach other nodes following the shortest path. Betweenness centrality is focused instead on the role of a node as a mediator among the rest of nodes. Finally, eigenvector centrality considers the eigenvector corresponding to the dominant eigenvector of the graph's adjacency matrix. Each measure of centrality captures a different meaning of centrality and their values can be even quite different, depending on the topology of the networks. The most appropriate type of centrality depends on the issues the social network is modeling and the final application. Finally, clustering coefficient is a measure of local cohesion, that is, to which extent one hop neighbors are connected among them. Several other local properties can be derived from the position of nodes within the network. In the context of virtual communities, local properties are interesting to find specific profile of users. For instance, local properties can reveal specific group of users like the core of the community o those key users that facilitate the diffusion of messages or information through the community.

As a difference to the local level, the global level considers the community as a whole. Parameters like size, density, the average shortest path or the diameter of the network can be used to compare networks and to determine their optimal structure. Several of the local properties of nodes like degree or centrality can also be averaged to calculate a global parameter of the network. One important issue that combines the local and global level is the detection of sub communities within the network. This detection is based on local properties of nodes, like cores, cliques or p-cliques. Sub communities within virtual communities can reveal structural patterns. For instance, if there is a giant component where all the nodes are connected, or if the network is divided in several unconnected sub communities, etc.

The last level of analysis consists of the analysis of networks from the perspective of complex network models. Simple networks can be described as random networks, which exhibit a high similarity regardless of what part is examined. As a difference, a complex network is a network that has certain significant topological features that do not occur in simple networks, like a heavy tail in the degree distribution, a high clustering coefficient or a hierarchical structure. This is the case of virtual communities and many other existing networks, with topological structures very different from random networks. The two famous models of complex networks are the scale-free networks model (Barabási and Albert, 1999), and small-world networks model (Watts and Strogatz, 1998). In scale free networks, the degree distribution of nodes follows a power law distribution. That means there is a small percentage of nodes concentrating the majority of interactions. In small-world networks, most of the nodes can be reached from every other by a small number of hops or steps. Both phenomenon can be observed in existing virtual communities (Chau & Xu, 2007). For instance, the Web has been found to have both small-world and scale-free properties (Albert and Barabási, 2002). Virtual communities can also be studied by determining to what extent they can be approached by complex network models.

B. Shared content

Natural language processing (NLP) is a set of techniques from a subspecialty of computer science and linguistics that uses computer algorithms to analyze human (natural) language. The vast amount of data on the Web and social media has made possible new applications. The most frequent applications utilizing NLP include among others information retrieval, information extraction, language modeling, spelling correction, question answering, text classification, sentiment analysis, etc.

In the case of virtual communities, most common techniques include information retrieval, language modeling, text classification and sentiment analysis.

Information retrieval consists of finding material of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored in computers). The simplest approach to deal with text analysis consists of obtaining the term-document incidence matrix, where each cell contains the number of times each word appears in each document. One of the most commonly used models of information retrieval is vector space model (Salto and McGill, 1983). This model considers a V dimensional vector space where words are the axes of the space and documents are points or vectors in this space, being V the number of words of the vocabulary. Obviously, for big collections this still is a very high dimensional space. However, vectors are very sparse vectors since most entries are zero. Using this matrix, similarities among documents or similarities with queries can be evaluated as the proximity of vectors in this V dimensional space, for instance using the cosine of the angle between them. However, the high dimensionality of the feature space us a problem when working with big collection of documents. Therefore, it is desirable to first project the documents into a lower-dimensional subspace in which the semantic structure of the document space becomes clear (Cai et al., 2005). In the low-dimensional semantic space, the similarity measures or clustering algorithms can be then applied. To this end, spectral clustering (Shi and Malik, 2000; Ng et al., 2001), clustering using Latent Semantic Indexing (Zha et al., 2001), and clustering based on non negative matrix factorization (Xu and Gong, 2004) are the most well-known techniques. Particularly, Latent Semantic Indexing (LSI) decomposes a term document matrix using a technique called singular value decomposition (SVD) to construct new features as combinations of the original features, significantly reducing the high-dimensionality problem of the feature space (Deerwester et al., 1990).

Language modeling is another important topic for virtual communities. The goal of language modeling is to assign a probability to a sentence. Although language modeling is mainly used in applications related to machine translation, spell correction or speech recognition, it can also be applied to identify the main topics of discussion in a collection of documents. Usually, the parameters of the language model are trained using a training set and then they are validated using a test set. Comparisons between two different models can be performed using an evaluation metric over the test set when solving a specific task, like a spelling corrector or a speech recognizer. An alternative consists of using an intrinsic evaluation, such as the perplexity. Perplexity is the probability of the test set, normalized by the number of words.

Text classification is another wide topic in natural language modeling. Basically, text classification of the task of assigning any kind of topic category to any piece of text. The input is a document and a set of classes, and the goal given this document is to predict a particular class from that set of classes. This task can be done in several ways. The simplest possible text classification method is to use hand written rules. However, building and maintaining these rules is expensive. That is way supervised machine learning is typically used to perform the classification. In this case, the input is a document, a set of classes and a training set of hand-labeled documents, and the goal of machine learning is to produce a classifier that maps each document to a class. There are lots of machine learning classifiers like naïve bayes, logistic regression, k-nearest neighbors, etc. They can evaluated using the typical measure of precision, recall and F measures.

A particular interesting part of text classification is sentiment analysis. Sentiment analysis is the detection of attitudes and dispositions towards objects or persons using documents. It can include the detection of the holder or source of the attitude, the target or aspect of the attitude, the type of attitude from a set of classes (like, love, hate, desire,...) of just using a simple weighted polarity (positive or negative)

IV. CASE EXAMPLES

This section illustrates several case examples of virtual communities where participation and shared content data were statistically processed to obtain some conclusions about their structure, behavior or users' features.

A. Open Source communities

Open source communities (OSS) emerged as a new paradigm of software creation opposed to traditional proprietary software schemes. The main resource of open source software projects are their subjacent community. Within the community, hundreds or thousands of individuals spread over the world shared their knowledge and propose new developments and ideas that drive the evolution of the target software (Martinez-Torres & Diaz-Fernandez, 2014).

Open source communities have been studied from the perspective of SNA. The local analysis has studied a phenomenon like participation inequality, typical in virtual communities (Kuk, 2006). For instance, the Gini coefficient was used to provide a measure of the level of participation based on the numbers of postings made by individual developers within a mailing list (Martinez-

Torres et al., 2009). This analysis has also focused on specific profile of users such as the core of the community, responsible of the majority of contributions, or the so called brokers of knowledge, which behave as intermediaries between expert software developers and peripheral users (Toral et al., 2010). These profile of users can be distinguished using local properties of nodes within the network, like the degree or the brokerage role. A brokerage role happens when a given node is the middle one in a directed triad, defined a set of three vertices and the lines among them (Toral et al., 2010).

As a difference to the local level, the global network analysis considers a set of networks or the evolution of the same network over time. Global parameters of networks are then measured and statistically processed. For instance, structural equation modeling was used to determine the main antecedents of online communities' success, quantifying the strength of the relation through the standardized path coefficients (Toral et al., 2009a). Factor analysis is another multivariate statistical technique used to identify the patterns followed by a given community over time (Toral et al., 2009b). Finally, a stepwise regression analysis was used to validate several hypotheses regarding the structure of the community and its incidence over its final activity and participation (Toral et al., 2009c).

Regarding the last level of participation analysis, OSS networks have been assimilated to scale free networks in which interactions among nodes follow a power law distribution (Valverde et al., 2006). This distribution reveals a social network's hierarchical organization with a core group on top of the hierarchy.

Shared content in open source communities has also been studied in several works. They are focused on open source repositories. For instance, Kawaguchi et al. (2006) proposed a tool called MUDABlue for the automatic categorization of software systems, relying only on the source code. This tool is based on LSI and it is able to properly categorize software systems based not only on usage, but also on architectures and libraries used. Topics discovery is another application of language modeling to OSS repositories. For instance, the work from Martinez-Torres et al. (2013) is based on the latent Dirichlet allocation algorithm developed by Blei et al. (2003), and it is focused on mailing lists repositories. In addition to the topic extraction, this study also applies a factor analysis to distinguish the patterns of knowledge sharing within OSS communities.

B. Open Innovation communities

Open innovation represents an effective strategy to provide organizations with access to a wider range of ideas in the worldwide market, reducing the costs associated with R&D (Chesbrough, 2006). One of the most popular alternatives for open innovation implementation is open innovation communities, which promote the generation of new ideas, the interactions among users as well as the interactions among the development team and customers (Di Gangi and Wasko, 2009).

Participation is a key mechanism for developing ideas, as interactions among users enable them to build on one another's knowledge and experiences. Typically, community members can participate sharing innovations, but also commenting and scoring other shared innovations. As a difference to other communities, shared ideas are evaluated by the company promoting the community, so those ideas which are selected to be adopted are publicly shown in the website.

One of the challenges of open innovation communities is that they tend to generate a huge volume of ideas, hindering the process of ideas evaluation. That is why many studies are focused on the identification of a special group of users called lead users (Von Hippel, 1986), with the ability of anticipating innovations earlier than the rest of the community. Some previous studies have been focused on identifying this group of users using several of their topological features within the community social network (Martinez-Torres, 2013). This local level of participation analysis considers the specific properties of lead users as stated in von Hippel's previous works (Von Hippel, 1986; 1988). Other local analyses have considered intermediaries in innovation networks, that have been proved to be facilitators of the innovation process. This role is developed by innovation brokers, which also collaborate in the diffusion of ideas (Winch & Courtney, 2007).

The global perspective can be used to explore idea providers' network connectivity. The study of Björk and Magnusson (2009) concludes that there is a clear interrelationship between the network connectivity and the quality of the innovation ideas created. In this case, authors use the idea of group degree centrality as an extension of node centrality to analyze and compare subnetworks.

Finally, content analysis have also been used in the context of open innovation. Again, the content approach is used to deal with the huge amount of collected information. This fact can bring some difficulties in finding the desired information. Finding potential solvers for a given problem can be solved by means of concept recommendation, which consists of assisting users to choose the right tag or to improve their search experience (Damljanovic et al., 2012). Content analysis can also be used to discriminate between adopted and non adopted ideas. For instance, to obtain the different perceptions of users belonging to the open innovation community and the company sustaining the community and evaluating shared ideas (Rufo et al., 2013).

C. E-word of mounth communities

With the advancement of Internet technologies, informal communication between consumers over particular products or services has become widely available and popular on the Web. Through eWOM (electronic word of mouth), customer can share their thoughts, opinions and feelings about products and services (Jeong and Jang, 2011). As a difference to traditional WOM, eWOM is directed at multiple individuals, is anonymous and is available at any time. Many studies on eWOM focus on the influence that product reviews could have on consumption decisions and sales in different sectors. The quality of the reviews and the reputation of the reviewers are specifically considered to be important factors affecting purchase decisions. This led to the problem of identifying a specific group of users, called influencers, which tend to be early adopters in markets, they are trusted by others, and have a large social network (Kiss & Bichler, 2008). The same than in previous case examples, influencers can also be distinguished attending to their local properties as part of a community network. For instance, Ku et al. (2012) proposes the identification of these users through their trust networks. Trust relationships in an opinion-sharing community are likely influenced predominantly by the reviews and preferences of trusted members.

Regarding the global network analysis, heterogeneities among different product categories were analyzed following several criteria like density, distance, degree, cohesion and centrality (Wang et al., 2011).

Diffusion is another issue that has been studied from the perspective of complex network models. For instance, scale free networks facilitate the diffusion of information through the whole network because they tend to contain centrally located and extensively high degree “hubs”. This is the focus of viral marketing, which refers to marketing techniques that use social networks to produce increases in brand awareness by ”viral” diffusion processes, analogous to the spread of pathological and computer viruses. This techniques works better if they are centered on influencers, that constitute the hubs of the network (Kiss & Bichler, 2008).

Content analysis techniques have been applied with different objectives. For instance, influencers can also be distinguished analyzing the shared content. The quality of shared opinions are highly dependent on the author's level of expertise (Huang et al., 2010), and the level of specialization can be obtained from the tags in which users are required to categorize their posted ideas (Martinez-Torres 2013). A different approach to content analysis consists of considering the quantity of emotional expression in shared ideas (Li et al., 2010). trustable reviewer should write relatively fair comments on the products, highlighting the merits but also the defects of the product. On the contrary, those users with very extreme evaluations both positive or negative are less trustworthiness. Sentiment analysis techniques can be applied not only for the identification of influencers but also to monitor the emotions of users about specific products or even a brand (Mostafa, 2013).

V. FUTURE TRENDS AND CHALLENGES

Virtual communities can offers numerous business potentials for companies. They provide a framework for organizing activities around a collective aim, taking advantage of connecting people spread over the world. They can also be useful creating new knowledge or discovering new knowledge, for example, with regard to customer habits, churn prediction, or new product trends (Heidemann et al., 2012). A big amount of data is today easily accessible though these communities, and with the emergence of new data collection technologies and advanced data mining and analytical tools, the analysis of big data is becoming a keystone of competitive advantage. The world's volume of data doubles every eighteen months, for example, and enterprise data are predicted to increase by about 650% over the next few years (Chang et al., 2014).

However, the use of virtual communities in the business context also goes along with some challenges and risks. The first challenge is selecting in which areas virtual communities can be leveraged reasonably. Before deciding about using virtual communities, companies must first analyze the goal of the community and the business functions or areas where they can create value. Moreover, communities must be also organized and they some guide and structure in order to successfully achieve their objectives. A special profile like the community manager is necessary for this purpose. Structure is another important aspect of brand communities, as members usually exhibit different levels of engagement. In general, a wide variety of structures can be found depending on the specific characteristics of virtual communities, varying more flat to more hierarchical structures. There isn't an optimal structure, although the objective is achieving a good level of engagement of community members.

Another major risk for companies adopting virtual communities scheme is the loss of control over shared information. This is the case of open innovation communities, where possible innovation are also visible for competitors. Too much openness can negatively impact companies' long-term innovation success, due to this loss of control, but a closed innovation approach does not serve the increasing demands of shorter innovation cycles and reduced time to market (Enkel et al., 2009). The optimum lies in a good balance between both approaches, although many companies are not prepared for such a cultural change.

Another critical point of virtual communities refers to the privacy risks. Public exposition of personal information, ownership of data provided within virtual communities or fake profiles able to distort shared information are open issues and still a challenge for companies.

Despite all those risk and challenges, there is a general consensus about the importance of virtual communities and the possibilities for new business potentials in the short term.

VI. CONCLUSIONS

This paper summarizes several methodological issues well suited for virtual communities in the context of big data. A research flow chart is first proposed and then detailed in the subsequent sections. Finally, several case examples are presented to visualize practical applications of the proposed methodology.

However, the topic of virtual communities is still a very large, interdisciplinary and emergent area of research, which requires further studies to complete previous addressed issues.

ACKNOWLEDGEMENTS

This work has been supported by the Consejería de Economía, Innovación, Ciencia y Empleo (Research Project with reference P12-SEJ-328), the ISCH COST Action IS1004 (www.webdatanet.eu), and the 5th Research Programme of the University of Seville.

REFERENCES

[1] Albert, R., Barabási, A.-L. (2002). Statistical mechanics of complex networks. Reviews of Modern Physics, Vol. 74, Iss. 1, pp. 47–97.

[2] Barabási A. L. & Albert R. (1999) Emergence of Sealing in Random Networks. Science, Vol. 286, pp. 509-512.

[3] Björk, J., and Magnusson, M. (2009). Where Do Good Innovation Ideas Come From? Exploring the Influence of Network Connectivity on Innovation Idea Quality, Journal of Product Innovation Management, Vol. 26, Iss. 6, pp. 662-670.

[4] Blei, D. M., Ng, A. Y., & Jordan, M. I., (2003). Latent Dirichlet Allocation, Journal of Machine Learning Research, 3, 993-1022.

[5] Cai, D., He, X., and Han, J. (2005). Document Clustering Using Locality Preserving Indexing, IEEE Transactions on Knowledge and Data Engineering, Vol. 17, Iss. 12, pp. 1624-1637.

[6] Chang, R. M., Kauffman, R. J., Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data, Decision Support Systems, doi: 10.1016/j.dss.2013.08.008

[7] Chau, M., Xu, J. (2007). Mining communities and their relationships in blogs: A study of online hate groups, Int. J. Human-Computer Studies, Vol. 65, Iss. 1, pp. 57–70.

[8] Chen, H., Chiang, R. H., Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact, MIS Quarterly, Vol. 36, no. 4, pp. 1165-1188.

[9] Chesbrough, H. (2006). Open innovation-the new imperative for creating and profiting from technology. Boston: Harvard Business School Press, 2006.

[10] Damljanovic, D., Stankovic, M., Laublet, P. (2012). Linked Data-Based Concept Recommendation: Comparison of Different Methods in Open Innovation Scenario, The Semantic Web: Research and Applications, Lecture Notes in Computer Science Vol. 7295, pp 24-38.

[11] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis, Journal of the American Society of Information Science, Vol. 41, Iss. 6, pp. 391–407.

[12] Di Gangi, P. M., and Wasko, M. (2009). Steal my idea! Organizational adoption of user innovations from a user innovation community: A case study of Dell IdeaStorm, Decision Support Systems, Vol. 48, Iss. 1, pp. 303–312.

[13] Enkel, E., Gassmann, O., Chesbrough, H. (2009). Open R&D and open innovation: exploring the phenomenon, R&D Management, Vol. 39, Iss. 4, pp. 311–316

[14] Heidemann, J., Klier, M., Probst, F. (2012). Online social networks: A survey of a global phenomenon, Computer Networks, Vol. 56, Iss. 18, pp. 3866-3878.

[15] Huang, Y., Bessis, N., Norrington, P., Kuonen, P., Hirsbrunner, B. (2013) Exploring decentralized dynamic scheduling for grids and clouds using the Community-Aware Scheduling Algorithm. Future Generation Computer Systems, Vol. 29, Iss. 1, pp. 402-415.

[16] Jeong, E., Jang, S. (2011). Restaurant experiences triggering positive electronic word-of-mouth (eWOM) motivations, International Journal of Hospitality Management, Vol. 30, Iss. 2, pp. 356-366.

[17] Kiss, C. and Bichler, M. (2008) Identification of influencers — Measuring influence in customer networks, Decision Support Systems, Vol. 46, pp. 233–253.

[18] Kuk, G. (2006). Strategic interaction and knowledge sharing in the KDE developer mailing list, Management Science, Vol. 52, Iss. 7, pp. 1031–104.

[19] Li, Y. M., Lin, C. H., Lai, C. Y. (2010). Identifying influential reviewers for word-of-mouth marketing, Electronic Commerce Research and Applications, Vol. 9, pp. 294–304.

[20] Martínez-Torres, M.R., Toral, S.L., Barrero, F., Cortes, F. (2009). The role of Internet in the development of Future Software Projects, Internet Research, Vol. 20, Iss. 1, pp. 72-86.

[21] Martínez-Torres, M.R., Toral, S. L., Barrero, F., Gregor, D. (2013). A text categorisation tool for open source communities based on semantic analysis, Behaviour & Information Technology, Vol. 32, Iss. 6, pp. 532-544.

[22] Martinez-Torres, M. R. (2013). Application of evolutionary computation techniques for the identification of innovators in open innovation communities, Expert Systems with Applications, Vol. 40, Iss. 7, pp. 2503-2510.

[23] Martínez-Torres, M.R. & Diaz-Fernandez, M. C. (2014). Current issues and research trends on open-source software communities, Technology Analysis & Strategic Management, Vol. 26, Iss. 1, pp. 55-68.

[24] Martinez-Torres, M. R. (2014). Analysis of open innovation communities from the perspective of social network analysis, Technology Analysis & Strategic Management, doi: 10.1080/09537325.2013.851378.

[25] Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments, Expert Systems with Applications, Vol. 40, Iss. 10, pp. 4241-4251.

[26] Ng, A. Y., Jordan, M., and Weiss, Y. (2001) On Spectral Clustering: Analysis and an Algorithm, Advances in Neural Information Processing Systems, 14, Cambridge, Mass.: MIT Press, pp. 849-856.

[27] Rufo, I., Martínez Torres, M. R., Toral, S. L. (2013). Open Innovation Implementation through Open Innovation Communities: the Case of Starbucks, Proc. Annual Conference of the Academy of Innovation and Entrepreneurship, Oxford, UK.

[28] Salto, G., and McGill, M. J. (1983). An Introduction to Modern Information Retrieval, McGraw-Hill, New York.

[29] Shi, J., and Malik, J. (2000). Normalized Cuts and Image Segmentation, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22, Iss. 8, pp. 888-905.

[30] Spaulding, T. J. (2010). How can virtual communities create value for business?, Electronic Commerce Research and Applications, Vol. 9, pp. 38–49

[31] Toral, S.L., Martínez-Torres, M.R., Barrero, F. (2010). Analysis of Virtual Communities supporting OSS Projects using Social Network Analysis, Information and Software Technology, Vol. 52, Iss. 3, pp. 296-303.

[32] Toral, S.L., Martínez-Torres, M.R., Barrero, F., Cortes, F. (2009a). An empirical study of the driving forces behind online communities, Internet Research, Vol. 19, Iss. 4, pp. 378-392.

[33] Toral, S.L., Martínez-Torres, M.R., Barrero, F. (2009b) Modelling mailing list behaviour in open source projects: the case of ARM embedded Linux, Journal of Universal Computer Science, Vol. 15, no. 3, pp. 648–664.

[34] Toral, S.L., Martínez-Torres, M.R., Barrero, F. (2009c). Virtual Communities as a resource for the development of OSS projects: the case of Linux ports to embedded processors, Behaviour and Information Technology, Vol. 28, Iss. 5, pp. 405-4119.

[35] Valverde, S., Theraulaz, G., Gautrais, J., Fourcassie, V., Sole, R.V. (2006). Self-organization patterns in wasp and open source communities, IEEE Intelligent Systems, Vol. 21, Iss. 2, pp. 36-40.

[36] Von Hippel, E. (1986) Lead users: a source of novel product concepts, Management Science, Vol. 32, Iss. 7, pp. 791-805.

[37] Von Hippel, E. (1988). The Sources of Innovation, New York, Oxford University Press.

[38] Wang, K.-Y., Thongpapanl, N., Wu, H.-J., Ting, I-H. (2011) Identifying Structural Heterogeneities between Online Social Networks for Effective Word-of-Mouth Marketing, 2011 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), , pp. 418-422.

[39] Watts, D. J. & Strogatz, S. H. (1998) Collective Dynamics of 'Small World' Networks. Nature, Vol. 393, Iss. 6684, pp. 440-442.

[40] Winch, G. M., Courtney, R. (2007). The Organization of Innovation Brokers: An International Review, Technology Analysis & Strategic Management, Vol. 19, Iss. 6, pp. 747-7633.

[41] Xu, W., and Gong, Y. (2004). Document Clustering by Concept Factorization, Proc. Intl. Conf. Research and Development in Information Retrieval, pp. 202-209.

[42] Zha, H., Ding, C., Gu, M., He, X., and Simon, H. (2001). Spectral Relaxation for k-Means Clustering, Advances in Neural Information Processing Systems, 14, Cambridge, Mass.: MIT Press, pp. 1057-1064.