iadis international conference

IADIS MULTI CONFERENCE ON COMPUTER SCIENCE AND INFORMATION SYSTEMS

MCCSIS 2009

21 - 23 June
Algarve, Portugal

Proceedings of ICT, SOCIETY AND HUMAN BEINGS 2009 and WEB BASED COMMUNITIES 2009

Edited by:
Gunilla Bradley
Piet Kommers

iadis
international association for development of the information society

IADIS INTERNATIONAL CONFERENCE

ICT, SOCIETY AND HUMAN BEINGS 2009

and

IADIS INTERNATIONAL CONFERENCE

WEB BASED COMMUNITIES 2009

part of the

IADIS MULTI CONFERENCE ON COMPUTER SCIENCE AND

INFORMATION SYSTEMS 2009

PROCEEDINGS OF THE

SECTION I

IADIS INTERNATIONAL CONFERENCE
ICT, SOCIETY AND HUMAN BEINGS 2009

SECTION II

IADIS INTERNATIONAL CONFERENCE
WEB BASED COMMUNITIES 2009

part of the

IADIS MULTI CONFERENCE ON COMPUTER SCIENCE AND INFORMATION SYSTEMS 2009

Algarve, Portugal

JUNE 21 - 23, 2009

Organised by
IADIS
International Association for Development of the Information Society

Volume Editors:
Gunilla Bradley and Piet Kommers

Computer Science and Information Systems Series Editors:
Piet Kommers, Pedro Isaías and Nian-Shing Chen

Associate Editors: Luís Rodrigues and Patrícia Barbosa

ISBN: 978-972-8924-82-9

SUPPORTED BY

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA E DO ENSINO SUPERIOR Portugal

SECTION II

IADIS INTERNATIONAL CONFERENCE

WEB BASED COMMUNITIES 2009

part of the

IADIS MULTI CONFERENCE ON COMPUTER SCIENCE AND INFORMATION SYSTEMS 2009

SHORT PAPERS

REFLECTION PAPERS

KNOWLEDGE SHARING THROUGH ONLINE COMMUNITIES OF PRACTICE: THE CASE OF LINUX PORTS TO EMBEDDED PROCESSORS

S. L. Toral, F. Barrero
*E. S. Ingenieros, University of Seville **
Avda. Camino de los Descubrimientos s/n, 41092, Seville, Spain

M. R. Martínez-Torres
E. U. Estudios Empresariales, University of Seville
Avda. San Francisco Javier, s/n 41018 Sevilla (Spain)

ABSTRACT

The paper explores knowledge sharing processes in online communities analyzing their explicit and stored messages. Their content is processed using latent semantic analysis techniques, obtaining several indicators related to knowledge sharing activities. A factor analysis is then applied to obtain the main dimensions affecting knowledge sharing. The obtained results provide new insights in the underlying knowledge processes in online communities of practice.

KEYWORDS

Knowledge sharing, online communities, communities of practice, factor analysis.

1. INTRODUCTION

The notion of community has been at the heart of the Internet since its inception (Lesser et al., 2000). Initially, Internet was used by scientist to share knowledge, collaborate on research and exchange messages, and today, million of Internet users worldwide communicate themselves using electronic tools. The advent of Web 2.0 and social software have propitiated the organization of users around communities of interest. The distinctive feature of these online communities is the intensive use of electronic media for people getting in contact. The theoretical background behind online communities has been treated by numerous authors. Some of them (Preece, 2001) highlight the connection of online communities with the social learning theory and communities of practice developed by Wenger (1998), while others are focused on their relation with knowledge sharing, knowledge creation, and innovation models (Lee and Cole, 2003; Kuk, 2006).

Of particular importance are the online communities supporting OSS (Open Source Software) projects, as they are changing the way in which software is produced. Several case studies can be found on the literature. Mockus et al. (2002) raise some questions about OSS development and analyze two case examples based on Apache and Mozilla projects. Lee and Cole (2003) focused on the most well known OSS project: Linux.

In this context, knowledge sharing means transforming individual knowledge into collective, organizational knowledge (Liebowitz, 2001). That is precisely one of the main objectives of a community of practice. Knowledge sharing is not stimulated by imposing structures and tools but by rich social interaction and its immersion in practice (van den Hooff & Huysman, in press). Consequently, knowledge sharing can be considered one of the most important processes involved in the development of an online community. The purpose of this paper consists of analyzing the main dimensions affecting knowledge sharing activities in online communities, using their stored information as a staring point. This information will be processed using latent semantic analysis techniques, extracting several indicators related to knowledge sharing, and these indicators will be statistically treated to identify the critical dimensions related to knowledge sharing.

2. KNOWLEDGE SHARING

In the early days of knowledge management, knowledge was often considered an object, which could be captured, codified and stored with the aid of information technologies. Knowledge management was mainly focused on optimizing these three processes. The problem of this approach is that the stored knowledge did not reflect real practices (Wenger et al., 2002). The insight that knowledge is not simply an aggregate of information that could be de-coupled from its context was then introduced and attention shifted to the idea that knowledge is socially embedded in the context where it takes shape (van Hooff & Huysman, in press).
The online community can be defined as a social relationship aggregation, facilitated by Internet-based technology, in which users communicate and build personal relationships (Rheingold 1993). Individuals engage in knowledge sharing, problem solving, and learning through posting and responding to questions on professional advice, storytelling of personal experiences, and debate on issues relevant to the network. Examples of online communities can be found on fields like education, software development or consumer behaviour (Toral et al., 2005; Shang et al., 2006).

Online communities have been frequently connected with communities of practice (Lin and Lee, 2006), in the sense that communities develop their own routines, formal and informal ‘‘rules’’, and practices evolve as a result of learning. The concept of Communities of Practice (CoP) was developed by Lave & Wenger (1991). This concept refers to the process of social learning that occurs when people who have a common interest in some subject or problem collaborate over an extended period to share ideas, find solutions, and build innovations. CoPs are not formal structures, such as departments or project teams. Instead, they are informal entities, which exist in the mind of their members, and are glued together by the connections the members have with each other, and by their specific shared problems or areas of interest (Wenger and Snyder, 2000). Intangible, tacit knowledge embedded in an organization's members is an asset that is not easy to capture. CoPs, however, offer a practical mechanism to help their members share and internalize tacit knowledge (Wang et al., 2008). Several researchers have noted that CoPs appear to be a more effective tool for dealing with unstructured problems and knowledge sharing/ creation than traditional and formal ways of structuring interaction in organizations. Understanding the processes and mechanisms that enable members to share knowledge in CoPs is very important for knowledge sharing within and between such communities (Pan and Leidner, 2003). A methodology based on semantic analysis technoques will be used for the identification of the main dimension with an influence in knowledge sharing.

Modern semantic analysis techniques are based vector space model, in which documents are summarized and represented by vectors of words (term vectors). However, the main problem of such representation is the high dimensionality of the feature space (one dimension for each unique word). 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 traditional clustering algorithms can be then applied. For instance, clustering using Latent Semantic Indexing (Zha et al., 2001) is one of the most well-known techniques. Different to LSI, in this paper we have applied the topic model, which is a statistical language model that relates words and documents through topics. It is based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words (Blei et al., 2003; Hofmann, 2001). Representing the content of words and documents with probabilistic topics has one distinct advantage over the purely spatial representation of LSI. Each topic is individually interpretable, providing a probability distribution over a word that picks out a coherent cluster of correlated terms. Hofmann (2001) introduced the probabilistic topic approach to document modelling in his Probabilistic Latent Semantic Indexing method (pLSI). Blei et al. (2003) extended this model by introducing a Dirichlet prior, calling the resulting generative model Latent Dirichlet Allocation (LDA). In both cases, the basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Given T topics, the probability of the ith word in a given document can be written as

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j) \qquad (1)$$

where z_i is a latent variable indicating the topic from which the ith word was drawn and $P(w_i|z_i=j)$ is the probability of the word w_i under the jth topic. $P(z_i=j)$ gives the probability of choosing a word from topics j in the current document, which will vary across different documents. Intuitively, $P(w|z)$ indicates which words are important to a topic, whereas P(z) is the prevalence of those topics within a document. Latent Dirichlet Allocation combines Eq. (1) with a prior probability distribution to provide a complete generative model for documents. The Latent Dirichlet Allocation algorithm involves the selection of several parameters to achieve a topic description of the domain under study.

Table 1. Dimensions involved in LDA algorithm.

Parameter	Description
D	Number of documents in corpus
W	Number of words in vocabulary
N	Total number of words in corpus
L	Average length of document in words (L = N/D)
T	Number of topics
ITER	Number of iterations

Table 2. Selected indicators

Indicator	Description
I1	Number of topics
I2	Polisemy
I3	Rated Polisemy
I4	Average messages size (characters)
I5	Average number of messages per topic
I6	Messages per topic distribution
I7	Average messages size (words)
I8	Size of threads
I9	Average number of threads per topic
I10	Threads per topic distribution

Table 1 details the key dimensions or size parameters describing a corpus (D, W, N and L) and a topic model run (T and ITER). In our study, we will analyze several online communities, extracting several indicators related to knowledge sharing by applying the topic model for each year. Selected indicators are detailed in **Table 2**. The first indicator I1 is the number of topics. This is a parameter of the topic model. Consequently, it has been chosen attending to the perplexity criterion. Perplexity is a standard measure of performance for statistical models of natural language (Manning & Schutze, 1999), and it is defined as:

$$pplex = \exp\left(-\frac{1}{W} \sum_{n=1}^{W} \log P(w_n \mid d_n) \right) \qquad (2)$$

Perplexity varies from 1 to W; lower perplexity is better, and the maximum perplexity of W is reached when all words in the vocabulary are equally likely. In our case, LDA algorithm has been run for a number of topics varying between 1 and 50, selecting the number of topics leading to a minimum perplexity value.

Indicators I2 and I3 are related to polysemy. Generative models have the ability of capturing polysemy, where the same word has multiple meanings. This is because they do not impose restrictions about mutual exclusivity that restricts words to be part of one topic. I2 measures the polysemy as the number of times a word w_i appears more than once in different topics while I3 consider this previous value rated with the probability $P(w_i|z_i=j)$ of the word w_i under the jth topic. The next four indicators are related to messages. They consider the average size of messages in characters and words, the average number of messages per topic and their distribution over topics. Finally, the last three indicators are related to threads of discussion. It is usual that online communities are organized by threads of discussion. Threads are groups of messages sharing the same subject. A thread is initiated by someone who posts a message asking for help, suggesting some improvements, or just considering some new idea. Then people start answering this initial message, posting possible solutions, sources of information or just extending posted considerations. The indicator I8 relative to the size of threads considers those threads with al least one answer. Indicators I9 and I10 are the same than I5 and I6 but using threads instead of topics as the unit of analysis.

3. CASE STUDY

The case study is based on Linux ports to embedded processors. Linux is a PC-based operating system that has been developed as Open Source Software along the structure of the UNIX operating system, and it is one of the most prominent examples of OSS projects. The same than Windows is the most prominent operating system released under a proprietary software license, Linux is the most prominent operating system released under a free license like GPL. Although Linux started as a hobby in 1991, it represents today a serious threat to Microsoft Windows's market dominance in operating systems. Nevertheless, the proposed case study will be focused on Linux ports to other processor architectures not intended for desktop or personal computer market. There are several reasons for this choice. First, Linux is firmly in first place as the operating system of choice for smart gadgets and embedded systems. Second, in contrast to other typical open source projects or even desktop Linux project, most contributions in this field do not come from volunteers or hobbyists, but from commercial firms, many of which are dedicated embedded Linux firms. Third, there are a lot of communities supporting each one of these Linux ports, and this is an excellent opportunity for analyzing a big group of more or less “homogeneous” communities.

Among the Linux distributions, Debian is perhaps one of the most well-known distributions. The Debian Project is an association of individuals who have made common cause to create a free operating system called Debian GNU/Linux, or simply Debian for short. Twelve Debian Linux ports communities have been considered for this study (**Table 3**). Each community will be analyzed during the period of time detailed in the description column of **Table 3**. Basically, the initial year is the one in which the community had a certain activity every month. For each year and community, indicators summarized in **Table 2** were obtained. Consequently, 110 case studies were considered.

Table 3. Virtual communities considered.

Community	Description	Community	Description
Debian port to m68k (D-68k)	Motorola 68k port of Debian GNU/Linux (98-08).	Debian port to Hurd (D-HURD)	The GNU Hurd is a new operating system being put together by GNU group (99-08).
Debian port to Alpha (D-Alpha)	the Alpha family of processors port the Debian GNU/Linux (98-08).	Debian port to IA64 (D-IA64)	Debian port to Intel IA-64 (01-08).
Debian port to AMD64 (D-AMD64)	The port consists of a kernel for all AMD 64bit CPUs with AMD64 extension and all Intel CPUs with EM64T extension (04-08).	Debian port to MIPS (D-MIPS)	MIPS port of Debian GNU/Linux, able to run at both endiannesses (99-08).
Debian port to ARM (D-ARM)	ARM port for Debian GNU/Linux. Debian fully supports a port to little-endian ARM (99-08).	Debian port to PowerPC (D-PPC)	PowerPC port of Debian GNU/Linux (99-08).
Debian port to BSD (D-BSD)	This is a port of the Debian operating system, complete with apt, dpkg, and GNU userland, to the NetBSD kernel (01-08).	Debian port to S390 (D-S390)	Debian port to IBM S/390 (01-08)
Debian port to HPPA (D-HPPA)	This is a port to Hewlett-Packard's PA-RISC architecture (01-08).	Debian port to SPARC (D-SPARC)	This port runs on the Sun SPARCstation series of workstations, as well as some of their successors in the sun4 architectures (99-08).

A factor analysis will be applied to extract the main dimensions related to knowledge sharing in virtual communities. Factor analysis attempts to identify underlying variables or factors, which can explain the pattern of correlations within a set of observed variables. Factor Analysis is a way to fit a model to multivariate data, estimating their interdependence. It addresses the problem of analyzing the structure of interrelationships among a number of variables by defining a set of common underlying dimensions, the factors, which are not directly observable, segmenting a sample into relatively homogeneous segments (Rencher, 2002). Factor analysis has been performed using the principal component method.

The eigenvalues of the sample covariance matrix are shown in **Table 4**. In factor analysis it is usual to consider a number of factors equal to the number of eigenvalues higher than 1 or able to account for more than 70% of the total sample variance. In our case, study, this value is achieved with three factors.

Using the associated eigenvectors, factor loadings can be estimated. Sometimes, it is difficult to perform the right interpretation of factors using the estimated loadings. Fortunately, factor loading can be rotated through the multiplication by an orthogonal matrix, preserving the essential properties of the original loadings. Varimax method is an orthogonal rotation method that minimizes the number of variables that have high loadings on each factor. This method simplifies the interpretation of the factors. **Table 5** reports the rotated factor loadings with varimax rotation for each one of the economical areas analyzed.

Table 4. Total variance explained.

Component	Eigen-values	% of Variance	Cumulative %
1	4,852	48,524	48,524
2	2,821	28,214	76,738
3	1,861	18,610	95,347
4	,157	1,569	96,916
5	,099	,985	97,901
6	,081	,812	98,713
7	,051	,513	99,227
8	,036	,364	99,591
9	,024	,236	99,828
10	,017	,172	100,000

Table 5. Rotated Component matrix with Varimax rotation.

	F1	F2	F3
I1	-,093	,959	-,070
I2	,209	,951	,066
I3	-,069	,982	,033
I4	-,117	-,016	,973
I5	,984	-,052	-,052
I6	,945	,066	-,011
I7	-,065	,036	,980
I8	,976	,120	-,078
I9	,977	-,081	-,069
I10	,962	-,008	-,124

To extract the meaning of each factor, we move horizontally through **Table 5**, from left to right, across the three estimated loadings of each variable, identifying the highest loading and the corresponding factor. To assess significance of factor loadings, a threshold value of 0,7 was considered (Rencher, 2002). The association between variables and factors is highlighted in grey in **Table 5**. The resulting aggregation of variables leads to the following latent factors or dimensions:

- The first factor refers to topic activity. The activity around topics is highlighted by the high value of I5, I8 and I9 indicators which account the number of messages and threads associated to topics. However, the high values of the standard deviation in the messages and threads distributions per topics suggest that all the topics are not treated the same way.
- The second factor is related to knowledge creation and reuse. The number of topics and polysemy are a measure of knowledge creation and reuse. This is precisely one of the main abilities of communities of practice. Topics are continuously evolving and previous knowledge is mixed and combined to generate new knowledge.
- The third factor refers to the information provided. I4 and I7 indicators refer to the average size of messages. The availability and depth treatments of topics are also a determinant factor for a successful development of the underlying community.

The obtained results demonstrate the necessity of guiding the evolution of the virtual community. Although virtual communities are based on the volunteer collaboration of community member, people posting a message hope to find an answer to their question or an alternative solution. Consequently, virtual communities are not only a question of social participation, but also a question of the quality of the provided information. Obviously, it is very difficult to asses individually the quality of each answer, as there are thousands of them, or to evaluate when a new knowledge is created or reused. For this reason, it is necessary to set a group of indirect indicators able to measure activity around the extracted topics, or the knowledge created through the evolution of the community.

4. CONCLUSION

This paper analyzes knowledge sharing in virtual communities of practice. An automatic tool based on semantic analysis is proposed to avoid the implicit difficulties of analyzing individually thousands of messages to extract some measures about the community knowledge processes and activity. In particular, the Latent Dirichlet Allocation has been chosen as the specific algorithm to extract the topics in which the community is involved. As a case study, several communities related to Debian-Linux ports to embedded processors have been analyzed, measuring a set of predefined indicators. The application of a statistical technique like factor analysis allows the extraction of the main dimensions related to community knowledge sharing processes and activity. One of the main conclusions of this work is that community portal managers may use simple measures for evaluating actual virtual community stimulation.

ACKNOWLEDGEMENT

This work has been supported by the Spanish Ministry of Education and Science (reference DPI2007-60128) and the Consejería de Innovación, Ciencia y Empresa (reference P07-TIC-02621)

REFERENCES

Blei, D. M., Ng, A. Y., and Jordan, M. I., 2003. Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022.

Cai, D., He, X., and Han, J., 2005. Document Clustering Using Locality Preserving Indexing, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, no. 12, pp. 1624-1637.

Hofmann, T., 2001. Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning Journal*, Vol. 42, no. 1, pp. 177-196.

Kuk G. 2006, Strategic Interaction and Knowledge Sharing in the KDE Developer Mailing List, *Management Science*, Vol. 52, no. 7, pp.1031–1042.

Lave J. and Wenger E., 1991. *Situated learning: Legitimate peripheral participation*. Cambridge University Press.

Lee, G.K., Cole, R.E. 2003, From a Firm-Based to a Community-Based Model of Knowledge Creation: The Case of the Linux Kernel Development, *Organization Science*, Vol. 14, no. 6, pp. 633-649.

Lesser, E. L., Fontaine, M. A., Slusher, J. A. 2000. *Knowledge and Communities*, Butterworth-Heinemann, pp. 86-98.

Liebowitz, J. 2001. Knowledge management and its link to artificial intelligence, *Expert Systems with Applications*, Vol. 20, Iss. 1, 2001, pp. 1-6.

Lin, H.-F., Lee, G.-G. 2006, Determinants of success for online communities: an empirical study, *Behaviour & Information Technology*, Vol. 25, No. 6, pp. 479 – 488.

Manning, C. D., & Schutze, H. 1999. *Foundations of Statistical Natural Language Processing*, MIT Press, MA.

Mockus A., Fielding R.T., Herbsleb J.D. (2002), "Two Case Studies of Open Source Software Development: Apache and Mozilla", *ACM Transactions on Software Engineering and Methodology*, Vol. 11, no. 3, pp. 309-346.

Pan, S. L. and Leidner, D. E. 2003. Bridging communities of practice with informaton technology in pursuit of global knowledge sharing. *Journal of Strategic Information Systems*, 12, pp. 71–88.

Preece, J., 2001, Sociability and usability: twenty years of chatting online, *Behaviour and Information Technology*, Vol. 20, no. 5, pp. 347–356.

Rencher, A.C., 2002. *Methods of Multivariate Analysis*. 2nd ed. Wiley Series in Probability and Statistics, J. Wiley & Sons.

Rheingold, H., 1993. *The Virtual Community: Homesteading on the Electronic Frontier*, Reading, MA: Addison-Wesley.

Shang, R.-A., Chen, Y.-C., and Liao, H.-J. 2006, The value of participation in virtual consumer communities on brand loyalty, *Internet Research*, Vol. 16, No. 4, pp. 398-418.

Toral S.L., Barrero F., Martínez-Torres M.R., Gallardo S., Lillo J. 2005, Implementation of a Web-Based Educational Tool for Digital Signal Processing Teaching Using the Technological Acceptance Model, *IEEE Transactions on Education*, vol. 48, no. 4, pp. 632-641.

van den Hooff, B., Huysman, M., in press, Managing knowledge sharing: Emergent and engineering approaches, *Information and Management*, in press.

Wang, C.-Y., Yang, H.-Y., Chou, T., 2008, Using peer-to-peer technology for knowledge sharing in communities of practices, Decision Support Systems, 45, pp. 528–540.

Wenger E. 1998, *Communities of practice: learning, meaning, and identity*, Cambridge: Cambridge University Press.

Wenger, E., McDermott, R., and Snyder, W. M. 2002. *Cultivating Communities of Practice*. Harvard Business School Press, Boston, USA.

Wenger, E.C. and Snyder, W.M., 2000, 'Communities of practice: the organizational frontier, *Harvard Business Review*, Vol. 78 No. 1, pp. 139-44.

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, pp. 1057-1064, Cambridge, Mass.: MIT Press.