Content analysis of open innovation communities using latent semantic indexing

M. R. Martínez-Torres

Facultad de Turismo y Finanzas, University of Seville, Avda. San Francisco Javier, s/n, 41018 Sevilla, Spain

Open innovation (OI) represents an emergent paradigm by which customers and users are involved as part of the innovation processes of organisations. One of its most popular implementation schemes is OI communities, which have been popularised by the use of social software. Through these communities, users are free to post, share, comment and evaluate other users' ideas, and they can interact with other users as well as with the innovation department and experts of the organisation. One of the challenges of OI communities is distinguishing the most innovative ideas, as they receive hundreds or even thousands of ideas. This paper proposes a novel approach for this task consisting of analysing the content of shared ideas. Through this analysis, several conclusions about the decision processes of the organisation can be inferred. The obtained results can help OI managers to improve ideas evaluation processes.

Keywords: open innovation; innovation communities; latent semantic indexing; ultrametric trees; innovation policies

1. Introduction

Open innovation (OI) has emerged as a new paradigm for organisations to gain opportunities and advantages using internal and external resources to drive their innovation processes (Chesbrough 2003; Huizingh 2011). The proliferation of information and communication technologies and the popularity of social software have propitiated the use of the Internet as the primary communication channel for customer integration (Mortara and Minshall 2011; Arenas-Marquez, Martinez-Torres, and Toral 2014). Therefore, virtual communities constitute an important external source of innovation for those firms that can implement a constructive relationship with them (Jeppesen and Molin 2003; Dahlander, Frederiksen, and Rullani 2008). These communities can be managed by one company, or they can be organised as online contests, which are intended as competitions among users in order to reach the best idea/proposal and the winner is rewarded (Harland and Nienaber 2014). Through these communities, users can share ideas with the rest of the community and the staff of the organisation, and they can also comment and score other previously posted ideas. Recent studies address that innovation in online communities also takes place through communication among participants and thus through the recombination of

*Email: rmtorres@us.es

ideas, leading to a common perception of what is valuable (Di Maria and Finotto 2008; Toral, Martínez-Torres, and Barrero 2009). One important advantage of these communities is that they encourage interactions among people. The feeling of being part of a community helps users to learn from the experiences of other users. In fact, social learning has been highlighted as an important antecedent for knowledge sharing and can also be an important driver for innovations (Toral, Martínez Torres, and Barrero 2010).

Research in the field of OI communities has been mainly focused on the identification of those best ideas that can be potentially applicable by the organisation. However, the main problem of OI communities is that they generate a huge volume of information that can saturate the absorptive capacity of organisations (Martinez-Torres 2013). Shared ideas should be individually assessed by the innovation department of the organisation or by some specific expert to decide if they can be applied or not. Therefore, the assessment of ideas is a costly task in terms of time and human resources. Practical implementations of OI communities demonstrate that only a small fraction of contributions are really attractive for the company and finally implemented (Di Gangi and Wasko 2009). Another problem is that decisions about the adoption of ideas can be influenced by the strategic policies of the company, and decision makers can be averse to those ideas that can lead to a radical change in the strategy of the company. A possible alternative consists of including a scoring system in the OI website, allowing the community to score shared ideas. However, the problem of this scheme is just the opposite to the evaluation guided by the strategic innovation policy of the company. The proposed ideas can be excellent for users but non-affordable or prohibitive for the company. Moreover, it has been pointed out that these non-affordable ideas are precisely those that receive a better score by community users (Martínez-Torres 2013). Different approaches are those based on distinguishing best ideas using the patterns of behaviour of community users who have posted them. In this line, the lead user theory (Von Hippel 1988) states that user innovation is more likely to emerge among the so-called lead users, who face needs and demands much earlier than the rest of customers. The behaviour of lead users has been described in the literature by their creativity and the active engagement in problem-solving activities (Amabile et al. 2005). These approaches based on the participation features of users have been treated in the literature from the perspective of social network analysis, modelling the community and their interactions as social networks (Ganley and Lampe 2009; Martínez-Torres 2012).

This paper follows a different approach focusing on the content of ideas and comments rather than on the activity of users. The main hypothesis is that content analysis techniques such as Latent Semantic Indexing (LSI) can be used to understand the decision processes and to check to what extent the decision-making of the company is focused on certain topics at the expense of some others. This analysis requires first to know the decision about each shared idea. This information can be easily obtained from OI communities, as the company makes public the decision about each idea. In this analysis, three different categories of ideas have been considered: implemented, partially implemented and acknowledged (but not implemented) ideas. Extracted topics per category reveal company's preferences, while comments and scores show the customer preferences. The obtained results can help decision makers to better approach to customers' needs and preferences.

This work differs from previous literature review in several ways. First, the focus is not on community users but on the content of exchanged information. Second, text analysis techniques allow the analysis of a great amount of information without the limitation of performing a manual identification, evaluation and coding of relevant text, leading to a richer scope and scale of the analysis.

The remainder of this paper is structured as follows. Section 2 reviews the methodologies and applications of text analysis techniques. Section 3 describes the proposed methodology based on the combination of LSI and ultrametric trees. Section 4 introduces the case study and Section 5 shows the obtained results. Discussion and implications are detailed in Section 6. The paper concludes in Section 7.

2. Related work

The increasing amount of data available on the Web provides a huge amount of useful information that can be processed to discover useful knowledge from the Web (Roussinov and Zhao 2003). This is the case of virtual communities, where the information is publicly available. Text analysis tools provide a set of techniques for automatically analysing documents without the limitation of manually reading, understanding, annotating and interpreting pieces of relevant text (Martinez-Torres et al. 2013).

Several previous studies have exploited the automated summarisation of documents using text analysis techniques, for instance, by representing a set of documents with a list of the most representative topics (Toral et al. 2010), using concept maps or clustering messages into semantically homogeneous groups (Anjewierden et al. 2011). All those approaches inherently rely on the algorithms representing the content of the text messages or documents through a vector space model (Zhai 2009), in which each document is represented by a vector of words (keywords or index terms). Text analysis algorithms compute the similarities between the documents based on their vector representations. Using similarity values, they extract the underlying topics or classify documents in categories. The main problem of the vector space model is the high dimensionality of the feature space (one dimension per keyword). Subspace-based algorithms reduce the high dimensionality by projecting the documents into a lower dimensional subspace in which the semantic structure of the document space becomes clearer (Cai, He, and Han 2005). LSI is one of these subspace techniques widely used in several research domains (Oudshoff et al. 2003). LSI decomposes a keyword-document matrix using a technique called singular value decomposition (SVD) to construct new features as combinations of the original ones, reducing significantly the high-dimensionality problem of the feature space (Deerwester et al. 1990; Lee and Yang 2009). LSI is based on three basic claims: (1) semantic information can be derived from a word-document co-occurrence matrix; (2) dimensionality reduction is an essential part of this derivation and (3) words and documents can be represented as points in Euclidean space. Once the dimensionality is reduced, documents can be easily managed in this lower dimensional space for classification, categorisation, association or summarisation of documents (Oudshoff et al. 2003).

Textual analysis has been widely used for processing exchanged information in different kinds of virtual communities. Open source software communities constitute one of the clearest examples of collective intelligence, where users post ideas, solutions, source code, bugs and interact with other users in order to improve the underlying software. The tool CATOSEM proposed by Martínez-Torres et al. (2013) applies textual algorithms to extract the main topics of discussion within Linux distribution lists. This tool categorises messages attending to their affinity, facilitating the search of related information within a huge number of messages usually sorted by date or threads of discussion. Content analysis has also been used to analyse virtual communities' dynamics. Koh and Kim (2004) studied how the level of community knowledge sharing activity is able to improve virtual community outcomes. Other issues such as knowledge creation, users' motivation and participants' performance have been studied in discussion forums and

online health communities (Ginossar 2008). Automatic knowledge discovery from texts allows enhancing communication in virtual communities, facilitating the evaluation of the community performance. This paper extends these previous works to the case of OI communities, where a previous categorisation of ideas based on their suitability is given by the company. Performance in the case of OI communities heavily depends on the assessment of ideas. In this line, this paper provides methodology to monitor the assessment of ideas and to identify the similarities and differences between users and company preferences when deciding about potential innovations.

3. Methodology

Figure 1 details the methodology followed in this paper. The first step is the selection of the keywords which should be representative enough of the target domain to be studied. Once the set of keywords is chosen, the proposed procedure is similar for each subset of ideas.

In the context of open communities, the keywords can be easily obtained using the tags provided by the OI website. When users decide to post a new idea, they must categorise the idea using a set of tags provided by the organisation. These tags represent the different areas in which users are expected to share their innovations. The number and scope of tags are high enough to cover all the areas of the organisation. The rest of steps of the block diagram are detailed in the following subsections.

3.1. *Data collection*

OI websites typically categorise ideas as implemented, partially implemented or acknowledged. This last status means that the idea has been evaluated but it has not been considered as implementable.

The first step for data collection consists of separating the different kinds of ideas according to their status. Once the ideas are discriminated, the text associated with each idea is captured and stored for the subsequent text processing and analysis. The text associated with each idea includes the original text submitted by the author as well as all the comments posted by the community.

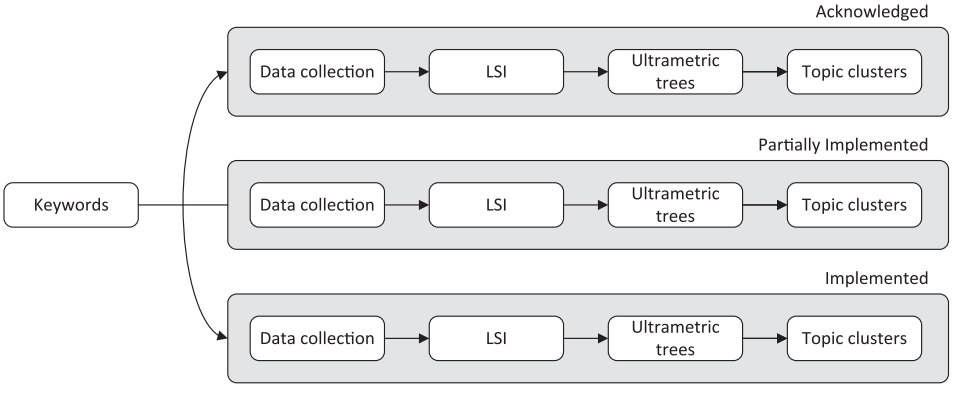

Figure 1. Block diagram of the methodology.

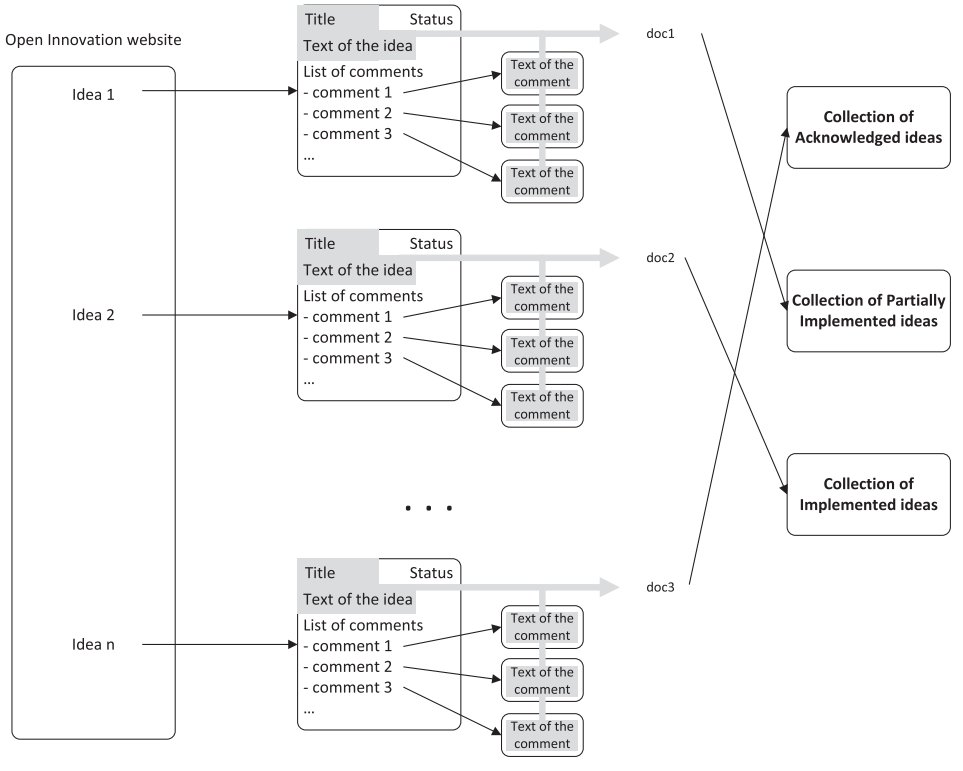

Figure 2. Crawler operation.

A specific crawler has been designed for discriminating and extracting the text associated with shared ideas. It processes html web pages following their hyperlink structure. First, ideas are classified attending to the content of the status field of ideas. Then, the link to each shared idea is followed to access the title and the body of the idea. Finally, the links of the received comments are explored to add their text to the body of the original idea. As a result, three collections of separate documents were obtained. Figure 2 details the crawler operation.

3.2. *LSI*

Given n documents that collectively contain m keywords, the term-document matrix A is an $m \times n$ matrix such that a_{ij} is the number of times that word i occurs in document j, being $n = m$. Matrix A defines an m-dimensional space in which each document corresponds to a point. LSI projects the row and column vectors of A into a lower dimensional space, one in which comparisons of these vectors are less susceptible to the effects of variance in word selection. This is accomplished by first computing the SVD of A. The SVD decomposes an $m \times n$ matrix A into the product of three other matrices

$$A = U\Sigma V^{\mathrm{T}}, \tag{1}$$

where U is an $m \times n$ orthogonal matrix, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ is an $n \times n$ diagonal matrix containing the singular values σ_i of A and V is an $n \times n$ orthogonal matrix. SVD is closely related

to the standard eigenvalue-eigenvector decomposition of a square symmetric matrix. In fact, U is the matrix of eigenvectors of the square symmetric matrix AA^{T}, while V is the matrix of eigenvectors of $A^{\mathrm{T}}A$. Σ^2 is the matrix of eigenvalues for both AA^{T} and $A^{\mathrm{T}}A$ (Deerwester et al. 1990). Although A can be reconstructed according to Equation (1), a rank-k approximation of A can be calculated by setting all but the highest k singular values in Σ to 0, yielding Σ_k. This effectively truncates U to be an $m \times k$ matrix and V^{T} to be a $k \times n$ matrix because when computing $U\Sigma_k V^{\mathrm{T}}$ only the first k columns of U and the first k rows of V^{T} enter into the computation. This approximation, A_k, is the closest rank-k approximation to A (Golub and Van Loan 1989).

The truncated SVD of A yields an embedding of words and documents in a k-dimensional space where, typically, $k = n$. Just as each row of A corresponds to a word and each column of A corresponds to a document, each row of U corresponds to a word and each column of V^{T} corresponds to a document. The similarity of two words in the k-dimensional space is determined by comparing the first k elements of the corresponding rows of the U matrix. Likewise, the similarity of two documents in this new space is determined by comparing the first k elements of the corresponding columns of V^{T}. Consequently, by using SVD, it is possible to reduce the number of dimensions in a term-document space. By using only k dimensions to reconstruct a term-document space, LSI no longer recalculates the exact number of occurrences of terms in documents. Instead, LSI estimates the number of occurrences based on the dimensions that have been retained. In the reduced dimensional reconstruction of the term-document space, the meaning of individual words is inferred from the context in which they occur. This means that LSI largely avoids problems of synonymy, since synonymous words are usually used in the same context (de Boer and van Vliet 2008).

3.3. *Ultrametic trees*

Once LSI provides the low-dimensional semantic space, a clustering algorithm is used to obtain the main topics by aggregating the selected keywords using the similarities or distances in this reduced semantic space. The resulting dendrogram is a rooted tree that represents the result of a hierarchical clustering. Leaves represent data objects, while internal nodes represent clusters at various levels. The final classification of the clustering algorithm heavily depends on the dendrogram distance definition. Therefore, the selection of the most appropriate clustering criteria for the data being investigated is quite important. Each of these dendrogram distances is in fact an ultrametric distance. In this paper, we represent the obtained dendrogram as an ultrametic tree using the algorithm UPGMA (Unweighted Pair Group Method with Arithmetic Mean).

The UPGMA method is widely used to develop taxonomies with numerical data obtained from a set of taxa, which are units grouped in a hierarchical classification (Gronau and Moran 2007). UPGMA differs from other hierarchical clustering algorithms in the definition of dissimilarity. Given a matrix of taxa (subjects or objects), this method constructs the bottom-up phylogenetic tree from the leaves (set of taxa). Let us consider an ultrametric tree T and let us define height(u) as the path length from a node u to any of its descendant leaves (since T is an ultrametric tree, all paths should have the same length). Let us also consider i and j as the descendant's leaves of u in two different subtrees. To ensure that the distance from the root to both descendants i and j is the same, consider height(u) $= M_{ij}/2$, with M_{ij} being the distance from i to j. For any two clusters $C1$ and $C2$ from the T tree, we define

$$\mathrm{dist}(C_1, C_2) = \frac{\sum_{i \in C_1, j \in C_2} M_{ij}}{|C_1| \times |C_2|}.$$

Notice that $\text{dist}(C_1, C_2) = M_{ij}$ for all $i \in C_1$ and $j \in C_2$. For example, let u be the lowest common ancestor node of i and j, and $\text{dist}(C_1, C_2) = 2\,\text{height}(u)$. For any node C_x whose ancestor is not u, it is understood that

$$\text{dist}(C_1 \cup C_2, C_x) = \frac{\text{dist}(C_1, C_x) + \text{dist}(C_2, C_x)}{2}. \tag{2}$$

In the UPGMA method, distances are calculated using an arithmetic average depending on the number of elements in each cluster.

3.4. *Topic clusters*

Once the ultrametric tree is obtained, keywords can be aggregated to obtain the main topics of discussion. Clusters are defined by the branches of the built ultrametric tree. The decision about the final clusters relies on the analyst, who must consider the consistency of the subjacent meaning of the cluster.

4. Case study

IdeaStorm is an online user innovation community created by Dell, where end users freely share innovative ideas with other community members and Dell to improve its products and services (Di Gangi and Wasko 2009). The goal of the user innovation community is to improve Dell's existing products and services as well as offer ideas on new market opportunities. Users also have the option of commenting and scoring other users' posted ideas.

Users have to register with an alias to be allowed to participate in the community. This alias is unique within the community and identifies those users posting ideas, comments or ratings. Whenever a user posts an idea to the IdeaStorm website, it should be classified attending to a limited number of tags. The list of tags is provided by Dell and they cover all the areas in which innovations are expected. Ideas can also receive comments and votes from other community members.

Dell evaluates shared ideas analysing their content and the scoring received by the rest of the community. The content is actually the main factor in the decision of adopting or not a posted idea. Typically, ideas are first assessed and classified by the innovation department and then presented to a group of experts for their evaluation in random order. As a result, ideas receive a status which is publicly shown: under review, acknowledged, partially implemented and implemented. This paper is focused on the three last categories, as they represent ideas that have received a final decision.

Following the proposed methodology, the designed crawler has been used to extract the information from the IdeaStorm website. The target information of this work is the title of the idea, its body and the content of all the received comments. All this information is merged to obtain the text associated with each idea. In what follows, the text associated with each idea will be designated as documents. Additionally, the categories or tags selected by the author when posting the original idea have also been collected. Table 1 represents the possible set of tags provided by Dell.

Ideas are automatically categorised by the crawler using the status of shared ideas. As a result, three separate collections of documents were extracted: acknowledges ideas, with a total of 190 documents, partially implemented ideas, with 170 documents, and implemented ideas, with 190 documents. Each collection has been separately analysed using LSI and ultrametrics trees to

Table 1. Set of tags provided by Dell to categorise shared ideas.

New product ideas	Mobile devices	Desktops and laptops	Accessories (keyboards, etc.)	IdeaStorm
Alienware	Software	Gaming	Dimension	XPS
Monitors and displays	Advertising and marketing	Servers and storage	Inspiron	Laptop power
Vostro	Operating systems	Dell community	Sales strategies	Service and support
Dell web site	Latitude	Broadband and mobility	Studio	Netbooks
Linux	Women's interest	Printers and ink	PartnerStorm	Education
Optiplex	Precision workstations	Environment	Retail	

extract the main topic clusters. The set of keywords of Figure 1 is obtained as a combination of previously shown tags and the most frequent words in each collection of documents.

5. Results

Several features of the three separate collections of documents were analysed before proceeding with the semantic analysis. More specifically, the differences between acknowledged, partially implemented and implemented ideas have been statistically tested using four different indicators: the number of received comments, the number of received votes, the number of tags in which ideas are included and the size of documents measured by the number of characters. The average values of these indicators for every collection of documents are given in Table 2.

In order to compare the means of variables from Table 2, a Kruskal–Wallis test has been performed. The Kruskal–Wallis test is a nonparametric version of one-way analysis of variance. The assumption behind this test is that the measurements come from a continuous distribution, but not necessarily from a normal distribution. The test is based on an analysis of variance using the ranks of the data values, not the data values themselves.

The low p-value for each variable in Table 3 suggests that the mean of one of them is significantly different from the other sample means. The null hypothesis can be rejected, so it can be concluded that the average values of the considered indicators are statistically different.

These average values clearly show differences among the three collections of documents. Ideas implemented or partially implemented receive more comments and votes that acknowledged ideas. This fact means that ideas adopted by Dell are in general popular and high scored by the innovation community. The higher size in characters of implemented and partially implemented ideas can be explained by the higher number of comments these ideas receive. Regarding the

Table 2. Features of acknowledged, partially implemented and implemented ideas.

	Acknowledged	Partially implemented	Implemented
Number of comments	1.60	26.11	15.37
Votes	2.37	277.14	38.47
Number of tags	2.07	1.71	1.52
Size (characters)	1349.07	8323.61	4864.41

Table 3. Kruscal–Wallis test.

	No. comments	Votes	No. tags	Size
χ^2	324.70	218.03	49.04	101.18
Df	2	2	2	2
Asymp. sig.	.000	.000	.000	.000

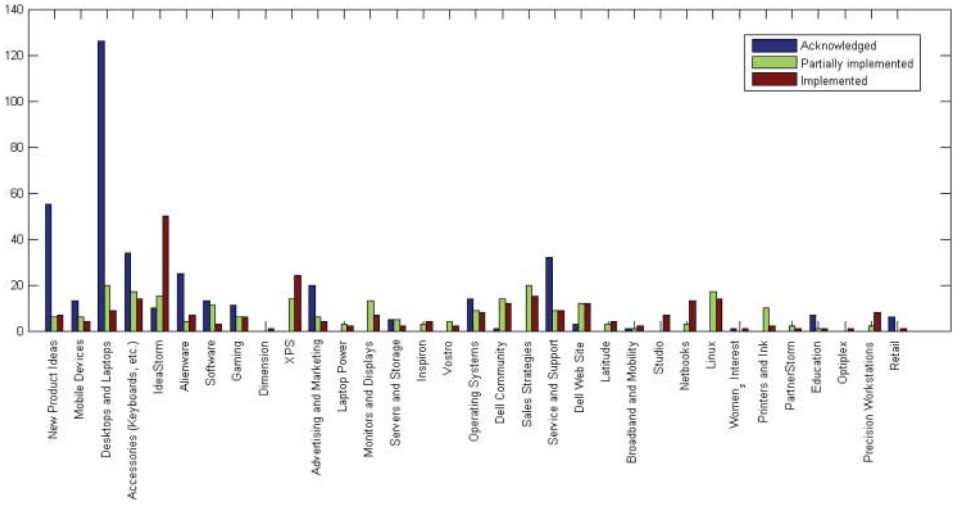

Figure 3. Number of acknowledged, partially implemented and implemented ideas per tag. (a) Acknowledged ideas, (b) partially implemented ideas and (c) implemented ideas.

number of tags, implemented and partially implemented ideas are focused on less categories than acknowledged ideas, which can be interpreted as a higher level of concretion.

If we focus on the last two groups of Table 2, partially implemented and implemented ideas, it can be noticed that the number of comments and, above all, the number of votes are much higher in the case of partially implemented ideas. Actually, this is the collection of ideas preferred by the community. This result suggests that the organisation is sensible to the preferences of the community and decides implementing these ideas but not in their whole extent. Probably, the cost of implementing the most scored and popular ideas is prohibitive. An intermediate solution consists of only implementing the affordable part of this set of ideas.

The distribution of the number of ideas per tag is detailed in Figure 3. This figure shows that acknowledged ideas are focused on more categories, especially those related to Desktops and Laptops, New Product Ideas and Service and Support. Partially implemented and implemented ideas are more uniformly distributed in the different categories considered by Dell.

The differences among the different topics treated in each collection of ideas can be further analysed through semantic analysis. For this purpose, the set of keywords of Table 1 and related words frequently used through the set of documents were selected, leading to a final list of 94 keywords for the LSI algorithm. Using this list, three keyword-document matrices were built, one for each collection of documents. An SVD is then applied to each matrix, reducing the dimensionality of the problem by applying a percentage threshold of 0.95 to their corresponding eigenvalues. The value of 0.95 represents a good balance between preserving the variance of the

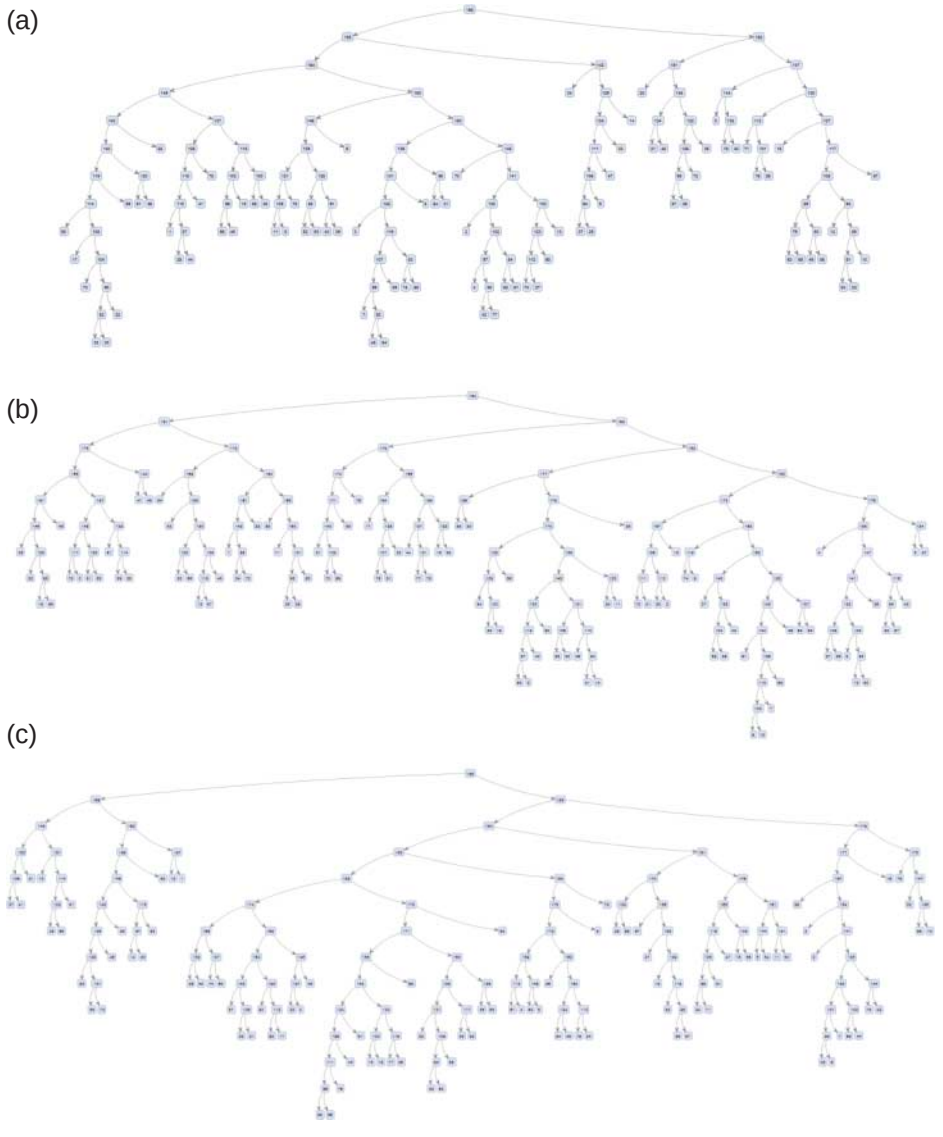

Figure 4. Ultrametric trees for the three collections of documents.

original data and the desired reduction of dimensionality (Toral et al. 2006). A hierarchical clustering algorithm is next applied to this low-dimensional space using the average link method and the Dice distance as the numerical distance metric (Toral et al. 2007). The resulting dendrogram is represented as an ultrametric tree for the three collections of documents (Figure 4). Terminal nodes at the end of the branches represent the original set of keywords, while intermediate nodes represent possible aggregations of these keywords.

Ultrametric trees help to visualise and interpret the main topics of interest of the three collections of documents. Figure 5 details the main topic for the collection of acknowledged ideas. The

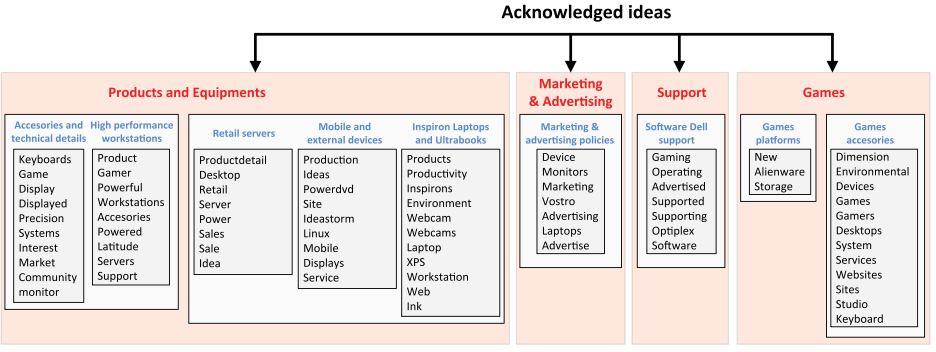

Figure 5. Topics of acknowledged ideas.

four big blocks correspond to the four main branches that can be distinguished in the third level of the ultrametric tree of Figure 4(a).

Products and equipments are the first block and cover two main categories: one related to high-performance equipments, such as workstations and accessories, and the other one focused on Laptops, Desktops, Servers and External Devices. This topic corresponds to the denser branch of the tree, which is consistent with the results of Figure 3, where ‘Desktops and Laptops’ and ‘New Product Ideas’ represent the two most covered tags by acknowledged ideas. The rest of blocks refer to Marketing and Advertising, Support provided by Dell and Gaming. In this last block, there are two main categories: one refers to Game Platforms, such as Alienware computers, which are suited for gaming, and the other is focused on Games Accessories.

The main topics of interest in the case of partially implemented ideas are detailed in Figure 6. Several differences with respect to the case of acknowledged ideas can be distinguished. Products and Accessories also appear as a separate block, but it corresponds to a considerable less dense branch of the tree compared to the case of acknowledged ideas. Instead, high-performance equipments block appears now as a separate block (in the case of acknowledged ideas was a subcategory of Products and Accessories). Gaming & Software block has a wider scope than in the previous case, including not only Game Platforms but also Software in general. Finally, Sales, Marketing and Management includes three categories: Community Ideas and Support, Laptop Advertising and Technical Details (both of them were a separate category for acknowledged ideas) and Sales Policy, which is a new category with respect to the previous case.

The topics of the implemented ideas are shown in Figure 7, where four main blocks can be distinguished. The block named Products and Accessories in the case of acknowledged ideas is now split into Products and Accessories and high-performance equipments in the case of partially implemented ideas. But in the case of implemented ideas, as shown in Figure 7, it can be observed that there are three blocks covering the main product lines of Dell: high-performance equipments, Desktops and Laptops. These blocks include Support and Accessories as categories within each block.

However, the denser part of the tree corresponds to the block named as Sales, Marketing and Management, which covers on one hand the community support and Ideastorm website organisation and on the other hand, the Sales and Market policies. It is interesting to notice that the collection of implemented ideas is the only one that includes a specific category related to the Ideastorm website. That means that Dell is really interested in improving its website and promoting the relationships with customers and users. Some other categories especially relevant in the

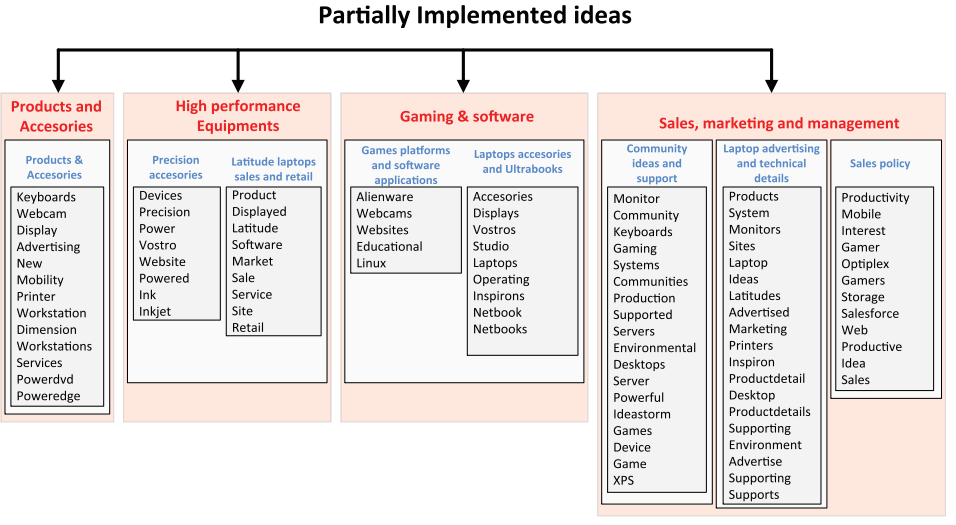

Figure 6. Topics of partially implemented ideas.

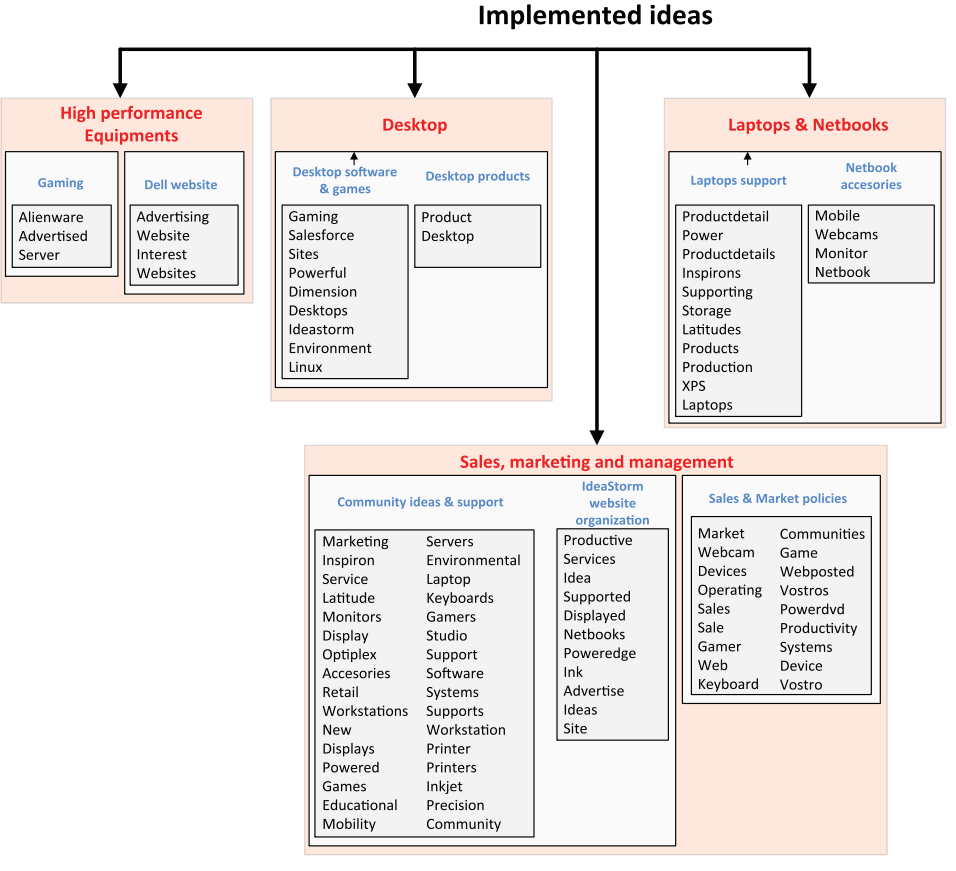

Figure 7. Topics of implemented ideas.

Table 4. Cophenetic correlation coefficient using different similarity measures.

	Cophenetic correlation coefficient		
Similarity	Acknowledged	Partially implemented	Implemented
Euclidean	0.828	0.917	0.906
Jaccard	0.358	0.975	0.525
Dice	0.849	0.927	0.950

collection of implemented ideas but not in the collection of acknowledged ideas are Netbooks and XPS Laptops.

The accuracy of the obtained clusters can be evaluated using the cophenetic correlation, which has been widely used as a criterion for evaluating the efficiency of various clustering techniques (Rashedi and Mirzaei 2013). Basically, the cophenetic correlation coefficient is a measure of how faithfully a dendrogram preserves the pairwise distances between the original unmodelled data points. Table 4 details the value of the cophenetic correlation coefficient for each collection of documents considering several measures of similarity. Best results correspond to the Dice distance, which is the one selected in our previous results.

6. Discussion and implications

One of the most important challenges of OI communities is the decision about which ideas should be implemented by the organisation. Shared ideas are evaluated by the innovation department of the organisation with the collaboration of specific experts in certain matters. They should decide about the viability and suitability of shared ideas, their cost and how these ideas can affect current products and services. They perform a technical and economical evaluation of shared ideas, considering also the strategic innovation policies of the organisation. However, some ideas can be interesting but they can suppose an important change in the production lines or even in the way the organisation is facing some necessities. In this case, ideas must overcome the resistance to change in the staff evaluating those ideas (Keupp and Gassmann 2009). To avoid losing fresh and groundbreaking ideas, OI communities develop their own evaluation system by which the community can score ideas, highlighting those ideas considered most interesting. This scoring system exerts pressure on the organisation in the sense that it is compelled to listen to the community for implementing the most relevant shared ideas.

This paper analyses shared ideas as separate collections of documents grouping acknowledged, implemented and partially implemented ideas. Figure 3 shows that the majority of received ideas belong to the category of ‘Desktops & Laptops’ and ‘New Product ideas’. However, only a small percentage of them are finally partially or totally implemented. The comparison of the obtained results through Figures 5–7 reveals that ideas related to these tags appear grouped together in the case of acknowledged ideas, but split into several more detailed blocks in the case of partially or totally implemented ideas. More specifically, high-performance equipments is a separate block for partially implemented ideas, and high-performance equipments, Desktops and Laptops & Notebooks are clearly distinguished blocks in the case of implemented ideas. This result suggests that ideas have more chances of being implemented when they are more specific to a certain product line, and that Dell tends to implement more easily ideas related to specific and more professional products rather than those focused on products oriented to the general clients. This is

the case of ideas related to Latitude Laptops, Netbooks or Workstations, which are mainly targeted for business use. In these three cases, Figure 3 shows that almost all the shared ideas about these topics were finally implemented. The position of Dell in this point can be explained by two considerations. First, users of high-performance equipments are professional users and they are aware of their real needs. Therefore, when they post ideas, they are thinking in problems and necessities related to their daily work. As a difference, non-professional users post ideas related to mid- or low-level computer systems and probably with lower accuracy than professional users (sometimes, posted ideas are only individual preferences rather than innovations). The second consideration is that it is easier to modify high-performance equipments than mid- or low-level computer systems. In the former, the price is not the determinant criterion of users when choosing an equipment, while the latter is more standardised systems which a downward adjusted price level.

Gaming is an issue clearly distinguished in the set of acknowledged and partially implemented ideas, but it is not so explicit for the set of implemented ideas. Obviously, Gaming is an important issue for a big audience of Dell products, since computer systems are also used by many people for leisure and recreational purposes. In the case of acknowledged ideas, posted innovations are related to software and accessories while in the case of partially implemented ideas they are specifically linked to the Alienware, which is Dell's premier gaming brand. In this case, the company is sensitive to the general audience and decides to implement some of the posted ideas or at least a partial affordable part of these sets of ideas.

As a difference to Gaming, the Ideastorm website organisation only appears as a specific category of the implemented ideas' topics. This result means that Dell is clearly committed to the development of the OI community, and it is open and receptive to all suggestions in this line. However, it is also true that innovations related to the website organisation are quite easy to implement since they refer to software changes rather than more costly changes, for instance, in the manufacturing processes.

Finally, Sales, Marketing and Advertising, and Community Support, are explicit categories in the three collections of considered ideas.

The comparison of the three sets of ideas shows that there is a trade-off between the resilience to change and the necessity of being receptive to the posted ideas. In general, there is a trend in Dell decision-making to trust specialised users when assessing innovations related to their product lines. However, this is not the case of Community Support or Marketing and Advertising ideas, probably because they are more affordable by the company. The partially implemented ideas constitute an intermediate area where the company can be sensible to certain breaking ideas, but implementing only the affordable part of them.

This paper shows how computational methods for content analysis can help community managers and companies to extract information about user preferences and ideas evaluation processes. Today, companies put increasing attention on secondary sources of information, above all on those related with the Internet, virtual communities and social software. Much of the information about the brand, products and services is today spread through specific or general-purpose communities, and customers and users are also using these communities when making their final decisions (Martinez-Torres 2014). Therefore, companies must be aware about all the information that is publicly available and they need to establish a social media monitoring to systematically gather, analyse and manage social media data. However, the main problem is the overabundance of information, which makes unaffordable the use of manual techniques. Computational methods for text analysis can overcome this problem by automatically analysing and processing shared

information. Although this paper shows how to extract the main topics within shared information, some other analyses are also possible, such as using classifiers or sentiment analysers.

7. Conclusion

This paper has analysed three collections of shared ideas in OI communities attending to the decision of the company about their potential applicability. The aim has been to detect the criteria behind the company's decision-making and to test to what extent the criteria are influenced by the innovation community preferences. A semantic analysis has been applied to the three collections of considered ideas to extract their main topics of interest. The comparative analysis of the obtained results reveals that decision-making is affected by the pressure of community preferences. Nevertheless, the resilience to change is more evident in those ideas affecting the manufacturing process or the main product line than in the marketing policies or the community support.

Future research will focus on finding sentiments or emotions when analysing the content of shared information. Similar to topic analysis, recognition of sentiments and emotions also requires advanced analytical tools, and can complement the proposed approach by adding the attitude and personal disposition towards the shared content.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was supported by the Consejería de Economía, Innovación, Ciencia y Empleo under the Research Project with reference P12-SEJ-328, and by the Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad under the Research Project with reference ECO2013-43856-R.

Notes on contributor

M.R. Martínez-Torres is an Associate Professor in Management and Business Administration at Business Administration and Marketing Department, University of Seville. Her main research interests include intellectual capital and knowledge management, and social network analysis. She has co-authored articles in many leading academic and professional journals, including *Information and Management*, *IEEE Transactions on Education*, *Computers & Education*, and *Behaviour and Information Technology*.

References

Amabile, T. M., S. G. Barsade, J. S. Mueller, and B. M. Staw. 2005. "Affect and Creativity at Work." *Administrative Science Quarterly* 50: 367–403.

Anjewierden, A., H. Gijlers, B. Kolloffel, N. Saab, and R. de Hoog. 2011. "Examining the Relation Between Domain-related Communication and Collaborative Inquiry Learning." *Computers & Education* 57 (2): 1741–1748.

Arenas-Marquez, F. J., M. R. Martinez-Torres, and S. L. Toral. 2014. "Electronic Word-of-Mouth Communities from the Perspective of Social Network Analysis." *Technology Analysis & Strategic Management* 26 (8): 927–942.

de Boer, R. C., and H. van Vliet. 2008. “Architectural Knowledge Discovery with Latent Semantic Analysis: Constructing a Reading Guide for Software Product Audits.” *Journal Systems and Software* 81 (9): 1456–1469.

Cai, D., X. He, and J. Han. 2005. “Document Clustering Using Locality Preserving Indexing.” *IEEE Transactions on Knowledge and Data Engineering* 17 (12): 1624–1637.

Chesbrough, H. 2003. *Open Innovation: The New Imperative for Creating and Profiting from Technology*. Boston, MA: Harvard Business School Press.

Dahlander, L., L. Frederiksen, and F. Rullani. 2008. “Online Communities and Open Innovation.” *Industry and Innovation* 15 (2): 115–123.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. “Indexing by Latent Semantic Analysis.” *Journal of the American Society of Information Science* 41 (6): 391–407.

Di Gangi, P. M., and M. Wasko. 2009. “Steal My Idea! Organizational Adoption of User Innovations from a User Innovation Community: A Case Study of Dell IdeaStorm.” *Decision Support Systems* 48 (1): 303–312.

Di Maria, E., and V. Finotto. 2008. “Communities of Consumption and Made in Italy.” *Industry and Innovation* 15 (2): 179–197.

Ganley, D., and C. Lampe. 2009. “The Ties that Bind: Social Network Principles in Online Communities.” *Decision Support Systems* 47 (3): 266–274.

Ginossar, T. 2008. “Online Participation: A Content Analysis of Differences in Utilization of Two Online Cancer Communities by Men and Women, Patients and Family Members.” *Health Communication* 23: 1–12.

Golub, G. H., and C. F. Van Loan. 1989. *Matrix Computations*. Baltimore, MD: Johns Hopkins University Press.

Gronau, I., and S. Moran. 2007. “Optimal Implementations of UPGMA and Other Common Clustering Algorithms.” *Information Processing Letters* 104 (6): 205–210.

Harland, P. E., and A. M. Nienaber. 2014. “Solving the Matchmaking Dilemma Between Companies and External Idea Contributors.” *Technology Analysis & Strategic Management* 26 (6): 639–653.

Huizingh, E. K. R. E. 2011. “Open Innovation: State of the Art and Future Perspectives.” *Technovation* 31 (1): 2–9.

Jeppesen, L. B., and M. J. Molin. 2003. “Consumers as Co-developers: Learning and Innovation Outside the Firm.” *Technology Analysis & Strategic Management* 15 (3): 363–383.

Keupp, M. M., and O. Gassmann. 2009. “Determinants and Archetype Users of Open Innovation.” *R&D Management* 39 (4): 331–341.

Koh, J., and Y.-G. Kim. 2004. “Knowledge Sharing in Virtual Communities: An E-business Perspective.” *Expert Systems with Applications* 26 (2): 155–166.

Lee, C. H., and H. C. Yang. 2009. “Construction of Supervised and Unsupervised Learning Systems for Multilingual Text Categorization.” *Expert Systems with Applications* 36 (2): 2400–2410.

Martínez-Torres, M. R. 2012. “A Genetic Search of Patterns of Behaviour in OSS Communities.” *Expert Systems with Applications* 39 (18): 13182–13192.

Martínez-Torres, M. R. 2013. “Application of Evolutionary Computation Techniques for the Identification of Innovators in Open Innovation Communities.” *Expert Systems with Applications* 40 (7): 2503–2510.

Martínez-Torres, M. R. 2014. “Analysis of Open Innovation Communities from the Perspective of Social Network Analysis.” *Technology Analysis & Strategic Management* 26 (4): 435–451.

Martínez-Torres, M. R., S. L. Toral, F. Barrero, and D. Gregor. 2013. “A Text Categorisation Tool for Open Source Communities Based on Semantic Analysis.” *Behaviour and Information Technology* 32 (6): 532–544.

Mortara, L., and T. Minshall. 2011. “How Do Large Multinational Companies Implement Open Innovation?” *Technovation* 31 (10/11): 586–597.

Oudshoff, A. M., I. E. Bosloper, T. B. Klos, and L. Spaanenburg. 2003. “Knowledge Discovery in Virtual Community Texts: Clustering Virtual Communities.” *Journal of Intelligent & Fuzzy Systems* 14 (1): 13–24.

Rashedi, E., and A. Mirzaei. 2013. “A Hierarchical Clusterer Ensemble Method Based on Boosting Theory.” Knowledge-Based Systems 45: 83–93.

Roussinov, D., and J. L. Zhao. 2003. “Automatic Discovery of Similarity Relationships Through Web Mining.” *Decision Support System* 35: 149–166.

Toral Marìn, S. L., F. J. Barrero Garcìa, R. Martínez-Torres, S. Gallardo Vázquez, E. Vargas, and V. G. Ayala. 2006. “Planning a Master’s Level Curriculum According to Career Space Recommendations Using Concept Mapping Techniques.” *International Journal of Technology and Design Education* 16 (3): 237–252.

Toral, S. L., M. R. Martínez-Torres, and F. Barrero. 2009. “Modelling Mailing List Behaviour in Open Source Projects: The Case of ARM Embedded Linux.” *Journal of Universal Computer Science* 15 (3): 648–664.

Toral, S. L., M. R. Martínez Torres, and F. Barrero. 2010. “Analysis of Virtual Communities Supporting OSS Projects Using Social Network Analysis.” *Information and Software Technology* 52 (3): 296–303.

Toral, S. L., M. R. Martínez-Torres, F. Barrero, and M. R. Arahal. 2010. “Current Paradigms in Intelligent Transportation Systems.” *IET Intelligent Transport Systems* 4 (3): 201–211.

Toral, S. L., M. R. Martinez-Torres, F. Barrero, S. Gallardo, and M. J. Duran. 2007. “An Electronic Engineering Curriculum Design Based on Concept-mapping Techniques.” *International Journal of Technology and Design Education* 17 (3): 341–356.

Von Hippel, E. 1988. *The Sources of Innovation*. New York: Oxford University Press.

Zhai, C. 2009. *Statistical Language Models for Information Retrieval*. Princeton, NJ: Morgan & Claypool Publishers.