Competence mining for virtual scientific community creation

Sérgio Rodrigues and Jonice Oliveira
COPPE/UFRJ, Computer Science Department, Graduate School of Engineering, Federal University of Rio de Janeiro, PO Box 68.513, ZIP Code 21.945-970, Rio de Janeiro, Brazil
Fax: (55 21) 2562-8676 E-mail: searo@cos.ufrj.br
E-mail: jonice@cos.ufrj.br

Jano Moreira de Souza
DCC/IM/UFRJ, Computer Science Department, Institute of Mathematics, Federal University of Rio de Janeiro, PO Box 68.513, ZIP Code 21.945-970, Rio de Janeiro, Brazil
Fax: (55 21) 2562-8676 E-mail: jano@cos.ufrj.br

Abstract: A problem currently faced is the inability of an organisation to know the competences that the organisation masters, thereby bringing forth greater difficulties to the decision-making process, planning and team formation. In the scientific environment, this problem prejudices the multi-disciplinary research and communities creation. We propose a technique to create/suggest scientific web communities based on scientists’ competences, identified using their scientific publications and considering that a possible indication for a person’s participation in a community is her/his published knowledge and degree of expertise. The project also proposes an analysis structure providing an evolutionary visualisation of the virtual scientific community knowledge build-up.

Keywords: knowledge management; text mining; virtual communities; community creation; decision analysis.


Biographical notes: Sérgio Rodrigues did his undergraduate work (2004) in Computer Science at the Federal University of Rio de Janeiro (UFRJ) and he is currently an MS student at UFRJ’s Graduate School of Engineering (COPPE). His personal research interests are in knowledge management, text mining and database.

Jonice Oliveira did her undergraduate work (2001) in Computer Science at the Federal University of Rio de Janeiro (UFRJ) and received her master’s degree (2003) in Systems Engineering from the Graduate School of Engineering (COPPE) at UFRJ. She is a DSc student at the same institution and her personal research interests are in knowledge management, database, CSCL and CSCW.
Jano Moreira de Souza is a Professor of Computer Science at the Graduate School of Engineering (COPPE) and Mathematics Institute of the Federal University of Rio de Janeiro (UFRJ). His area of specialisation is databases, and he is involved in research in fields such as CSCW, decision support systems, knowledge management and GIS. Professor Souza received his bachelor’s degree in Mechanical Engineering (1974) and his master’s degree (1978) in Systems Engineering from the Federal University of Rio de Janeiro (Brazil), and his PhD in Information Systems (1986) from the University of East Anglia, England.

1 Introduction

People have been using knowledge, at least implicitly, in organisations for a long time. Company knowledge, competitive knowledge, process knowledge, business knowledge, in short, knowledge has been the force behind millions of strategic and operational decisions throughout time. However, the recognition of the fact that knowledge is a resource that needs to be managed is relatively recent. When attempting to practice knowledge management, we should first understand how knowledge is obtained, who has it in an organisation, in what form it is and what are the physical and cultural barriers that should be overcome to codify and distribute it. We should also know how information technology supports this activity when knowledge transfer in common conditions is impossible.

As information technology is a contributing part to knowledge management, providing a computational environment in this context may enable improvements in knowledge management applied at an institution. Care should be taken when the organisation is a scientific institution because it has its peculiarities like related processes and professionals and data used. The integration possibilities of these two areas, knowledge management and the scientific environment, comprise the main motivation of this research work, dealing mainly with the computational support for internal knowledge discovery. That is, the competences of a research centre and its continuous evolution through the use of the web communities.

This paper is divided into six sections. Following the introduction, some differences regarding scientific knowledge are mentioned in Chapter 2. In the next section, the text introduces the virtual community concept and the work proposal is presented in Section 4 with a competence mapping environment for the creation of a virtual scientific community through a community evaluation tool. The prototype and current state of this proposal are discussed in Section 5, and future works and the conclusion of this work are presented in Section 6.

2 Scientific knowledge

One of the first to define scientific knowledge was Socrates [1], and, for him, knowing a subject or concept consisted of ‘gathering the components of a singular thing, or of a real substance, and joining the similar, and separating the unsimilar ones, to form the concept or the definition of the singular thing’. In this way, to ‘join the similar ones’, it is
necessary for one to have principles, axioms, definitions and demonstrations, so that a concept to be defined as true. In other words, scientific knowledge is the knowledge resulting from scientific activities, and its objective is to demonstrate by argumentation, a proposed solution to a problem, relative to a certain issue [2].

But, what is the difference between business knowledge and scientific knowledge?

The first difference is related to the analysis of the data used and of the knowledge construction process. A business environment usually has as a reference, administrative and production data created by a well-known process. In scientific activities, data is usually provided by simulations, previous experiences, calculations and mathematical models. That is, while business activities work on given easily-structured data, the scientific activities work, almost all the time, with more complex and unstructured data that can be in distributed databases.

Frequently, the activities executed in a business domain are well defined, and the knowledge needed for the execution of each of those activities is well known, while the scientific activities, mainly on the demonstration phase, comprise sequences of attempts because the domain is not completely known. In other words, scientific knowledge is built gradually according to the results of a number of activities and it can be subject to constant alterations. And, for each new knowledge stage, documents such as technical reports, books chapters and papers are created to show solidified hypotheses and disseminate new knowledge. The main knowledge type used in the scientific environment is externalised knowledge, and scientific community mainly bases itself on this knowledge type.

Independent of the complexity of the manipulated data the information analysed, and of the way in which it is structured, there is another factor in scientific knowledge construction, which is, collaboration. According to Schur [3], collaboration is the essence of science because there is on account of people’s union, the possibility of knowledge exchange for common activity execution (peer-to-peer collaboration); and dissemination of an acquired knowledge (mentor-student). Researchers with diverse knowledge domains who do not share a common background, can interact by exchanging results (interdisciplinary) or just by publishing the research results achieved (producer-consumer).

Related to the level of inter-personal collaboration, collaboration in scientific environments is usually more restricted, and occurs among a small number of people working in the same group, dealing with or researching more specific items of their domain. Many researchers are not aware of other researchers who are working on similar projects, as they are based at another research centre, and the distance is a hindrance to regular contact. Noting that collaboration is the basis for science, the use of virtual communities is a way to provide better communication and researcher interaction in the same domain – independent of synchronous communication and physical presence. Then, the four types of scientific collaboration proposed by Schur [3] can be applied easily.

3 Virtual communities

Communities are formed by people (members) with some common interests and who share information for some time [4]. As per Seely Brown [5], communities are
When there is computational support for the existing communication, there is no need for real meetings and the relationships among the members can be maintained at a distance. These communities are known as virtual communities [6,7] or better still, ‘communities mediated by computer’ and ‘communication mediated by computer’, often considered synonymous [8], enabling groups to build new communities based on mutual interests rather than geography [9–11].

The term virtual community was coined by Rheingold [12] who applied the term to groups of individuals sharing not just mutual interests, but also ‘social network capital, knowledge capital, and communion’ [12]. It means electronically-mediated social contacts, the pool of knowledge and experience possessed by participants, and the willingness to provide mutual support [13]. Applying virtual communities in science, the value of networking to facilitate joint research is well established [14–19], but a broader concept of community - such as that of Rheingold [12] – applied specifically to science, has appeared only recently [20].

The formation of virtual communities presupposes the existence of an organising agent whose function is to join the members and relevant resources for the community. Moreover, it should be able to collect the descriptive data about the use by the community, its evolution and the transactions made.

The term community in this article is used as a virtual community synonym, and all the interactions, either asynchronous or synchronous, are made on the web, allowing the community members to contribute, to discuss and to acquire new knowledge with one another, as proposed for Ram et al. in [21]. Remember that the communities mentioned in this work have as their main purpose the acquisition, exchange and dissemination of scientific knowledge in a certain research domain, and fostering location-independent scientific collaboration.

4 Competence recognition and community creation

Scientific organisations are institutions whose main product is knowledge. Due to the growth of research centers, the physical distances between them and the new knowledge areas that have arisen in science, the universities and scientific institutions face a problem. The same problem is faced by the business world: the need to know what they know (internal competences), and who the owners of this knowledge are.

As stated in section two, the scientific documents by being the main source for information dissemination, reveal a lot of things: work in progress, results obtained, ideas, confirmed hypotheses and associations made. Thus, we can extract a great deal of information from scientific publications, such as researcher’s competence.

This work aims at mapping researcher’s competence in his/her publications, using methods and techniques applied to the area of knowledge discovery from texts or, as commonly entitled, text mining. This discovery is essential for a scientific organisation to be able to discover which areas of knowledge have active professionals, as well as how internal knowledge is divided. Our approach attempts to connect researchers with common interests, suggesting virtual communities (and supporting their
infrastructure and evolution), assuming that people who work in a specific area and have similar competences can share information and knowledge.

Next, we shall describe how personal competences are discovered in scientific publications by using text mining, as well as community suggestions in which the researcher can become a member. We shall also show mechanisms to describe and analyse the knowledge exchanged and acquired, and to evaluate the importance of a community.

4.1 Competence extraction architecture

In our approach, each researcher has a directory where all his/her publications are placed. The process can be started from there with a significant base of texts. For each text, an identification key is created and stored in the database. This identification key will render future searches possible without running the algorithm once more for this text (unless extraction parameters are altered).

Figure 1 illustrates the text extraction, storage and analysis stages. As a result, the algorithm that will be presented will supply a list of relevant words that can indicate the competencies in the text.

Figure 1 Vision of the textual information recovery system

In parallel, the text is submitted to the token generation algorithm. Tokenising consists of a rule of word identification (the tokens). This technique suggests that the tokens are defined as string of alphanumeric characters without spaces, and could include hyphens and accentuated letters. Therefore, the white space is the resource used to separate words. Apparently, this algorithm application looks simple; however, a number of problems can be found when applying this technique:

- Punctuation – the words can be attached to punctuation like commas, semicolons and full stops. The recognition of this punctuation may seem easy; however, in the cases of abbreviations, for instance, the full stop is problematic.
- White space – sometimes the white space is not a semantic separation among words. For instance, in the group of words ‘Rio de Janeiro’ (a Brazilian city), the algorithm should point that this is only a word (Rio de Janeiro), and should not present as two relevant words (‘Rio’ and ‘Janeiro’), which mean nothing in the context.

After breaking the text into words (tokens), the algorithm removes words not relevant – stop words (i.e, prepositions, adverbs, and others). The group of stop words removed from the texts is called stop list. This list of irrelevant words is deeply dependent of the language and on the context used. The ideal application happens when
the tool implementing the algorithm allows a stop word choice, as well as the creation of possible words that can become irrelevant in the mined text context, as shown in Figure 2.

**Figure 2** Tokenising – starting from a specific text (I), the words are obtained (II) and, then, the stop words are removed (III)

When stop words are removed, the remaining words are considered filtered and should enter a new selection process. In this phase, the next procedure comprises the creation of weights for each word type. An easier artifice is to indicate that all words have the same weight, thus, the relevance degree of each token is given from the frequency it appears in the text. The most significant alternative suggests the creation of a list of words and their respective weights. In this case, the algorithm counts token frequency and also analyses whether the recovered words have relevance in the context by the weight defined. It is interesting to observe the fact that a word that has a high frequency, does not indicate for sure that it is significant in the context.

In our approach, we do not use weights; we use the stemming technique to measure the relevance of a term by removing suffixes in an automatic operation. Ignoring the issue of precisely where the words originate, we can say that a document is represented by a vector of words, or terms. Terms with a common stem will usually have similar meanings, for example: CONNECT, CONNECTED, CONNECTING, CONNECTION, CONNECTIONS. Then, after we have the words, we apply the stemming to count the relevance of a term. Terms are related to a person’s competence, and these competences and the expertise degree (using the measure of relevance of a term) are stored in the database.

### 4.2 Suggesting appropriate communities to the researcher’s profile

After the researchers’ competence identification phase, our environment: (i) whether there are any communities about a topic, the environment searches a number of people with similar interests and proposes creation a new community (Figure 3), or (ii) suggests existing communities that match the profile of a future member (Figure 4). Depending on the topic, a researcher can belong to more than one community.

As the communities are only based on competencies than on historical relationships, it is natural that all community members are not equally important in terms of their contributions. For this, the community must have mechanisms to allow knowledge exchange and the sharing of it, motivating members to interact with one another. We can mention as knowledge to be shared: definitions of scientific processes (as experiment definitions), definitions of models, documents, raw data, class diaries, training material, calls for papers and a number of ideas.
Figure 3  The environment suggests a new community creation based on the competencies of researches

<table>
<thead>
<tr>
<th>Researcher A</th>
<th>Researcher B</th>
<th>Researcher C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database Systems</td>
<td>GIS</td>
<td>Component Design</td>
</tr>
<tr>
<td>Object Design</td>
<td>Decision Support Systems</td>
<td></td>
</tr>
<tr>
<td>CAD</td>
<td>Database Systems</td>
<td></td>
</tr>
</tbody>
</table>

New Community

Database Community

- Researcher A
- Researcher B
- Researcher C

Figure 4  A researcher is suggested as a new member of an existing community

Researcher N  

Suggestion  

Database Community

- Researcher A
- Researcher B
- Researcher C

To minimise the heterogeneity, our approach allows for the use of synchronous and asynchronous collaboration tools in a community, so that the knowledge can be better disseminated. Each community has the following tools:

- Discussion list – for synchronous and asynchronous communication and knowledge exchange
- Electronic meeting (chat) – allows synchronous interaction and discussions between the members of the community. Electronic meetings tools permit online interviews, so that a researcher with knowledge about an issue can be consulted synchronously.
- Video conference – bears the same functionality of an electronic meeting, allowing visualisation of the members when they are connected.
- Forum – some themes belonging to a researchers’ domain of a community can be discussed separately.
- Surveys – some topics are taken for voting and ranking.
- News – for the dissemination of news, events and conference deadlines.
- Documents upload/download and links. – allows the sharing of documents and suggestions through web pages.
4.3 Evaluation of community development

What makes communities successful over time is their ability to generate enough excitement, relevance and value to attract and engage members [22]. One factor for the success of a community is constant management to provide the community’s own internal direction and dynamism. To reach it, it is necessary to observe the community’s natural evolution and sometimes to encourage members’ participation.

As the community grows, new members bring new interests and may pull the focus of the community in different directions. Changes in the core science or technology of a community constantly reshape it, often bringing in professionals from neighbouring disciplines or introducing technological advances that change their way of working [22].

We propose some reports for observing the evolution of a community. This kind of analysis is especially useful in the following ways:

- The intellectual capital of the institution not be associated exclusively to people who own critical knowledge, but can be distributed among the members of a research team;
- Make the identification of knowledge areas possible even with a shortage of professionals, and then plan a way to acquire this knowledge, by training or external researcher recruiting;
- Make the regular appraisal of each researcher’s knowledge level possible.

Our proposal uses an online analytical processing (OLAP) structure to provide more solid and evolutionary visions about the researchers’ knowledge acquisition process and knowledge flow in a community. With this tool, it is possible to perform queries, such as:

- How were researcher’s competences developed during his/her academic life? Dimensions: Competence x Researcher x Time.
- Which competences are more broached in a community? Dimensions: Competence x Community.
- How has the knowledge been added in the communities in the last years? Dimensions: Competence x Community x Time
- Which competences were mentioned in the authors’ last papers more often? Dimensions: Competence x Researcher x Time
- Which are the most important researchers in a community? Dimensions: Researcher x Competence x Community
- What knowledge area is losing value, in other words, is not being discussed nor it has publications about it? Dimensions: Time x Competence
- Which are the weaker areas of a university, in other words, competences with few owners? Dimensions: Researcher x Competence x Community
- Which are the knowledge areas that have the largest number of publications? Dimensions: Competence x Publication
This OLAP structure is in the design phase, but the idea is to use the output data from publication mining, in order to create the OLAP model about competences, researchers and communities during time, and to display these reports to each community.

5 The prototype

The environment is implemented using an Active Server Pages (ASP) and MS-SQL Server like database system. The text mining engine is implemented in C++ and used as a CGI program.

The prototype input is the set of texts to be mined. Then, a filter is applied that removes the insignificant terms (stop words). With only the possible relevant terms of the scientific text, an analysis to obtain the frequency of each term in the text is carried out. Next, after the frequency analysis, words with similar semantics, such as synonyms, will be searched and these terms divided into clusters. Afterwards, we have as an output list with the keywords that obey a minimum limit of frequency.

The system employs a database that enables the user to choose and increase when necessary, the lists of stop words that should be filtered. Moreover, the user can define a threshold, indicating the initial frequency when a word becomes relevant.

Another function in the environment is the pre-definition on the context presented. The user can define which words should be considered relevant and the system will combine this indication with the degree of frequency that these words appear in the text. An idea (still in implementation phase) is to automatically define this pre-defined list, starting from the publication mining history, and based on this list, the system can suggest more suitable weights.

Still in the removal phase of stop words, the method called stemming is applied. In this stage, the suffixes of the words are removed and the accountancy of the relevant words is made, starting from their radicals. This makes it possible for words like CONNECT, CONNECTION, CONNECTIONS, CONNECTED, and CONNECTING to be just computed as CONNECT.

The example below shows, in simple steps, the algorithm presented being used in the developed prototype.

Example: Text mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what we are familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently is not relevant to our needs, in order to find the relevant information. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down.
<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>aside</td>
<td>1,8519%</td>
</tr>
<tr>
<td>automatically</td>
<td>1,8519%</td>
</tr>
<tr>
<td>computer</td>
<td>1,8519%</td>
</tr>
<tr>
<td>conventional</td>
<td>1,8519%</td>
</tr>
<tr>
<td>currently</td>
<td>1,8519%</td>
</tr>
<tr>
<td>different</td>
<td>3,7037%</td>
</tr>
<tr>
<td>discovery</td>
<td>1,8519%</td>
</tr>
<tr>
<td>element</td>
<td>1,8519%</td>
</tr>
<tr>
<td>experimentation</td>
<td>1,8519%</td>
</tr>
<tr>
<td>explored</td>
<td>1,8519%</td>
</tr>
<tr>
<td>extract</td>
<td>1,8519%</td>
</tr>
<tr>
<td>extracted</td>
<td>1,8519%</td>
</tr>
<tr>
<td>extracting</td>
<td>1,8519%</td>
</tr>
<tr>
<td>facts</td>
<td>1,8519%</td>
</tr>
<tr>
<td>familiar</td>
<td>1,8519%</td>
</tr>
<tr>
<td>find</td>
<td>1,8519%</td>
</tr>
<tr>
<td>goal</td>
<td>1,8519%</td>
</tr>
<tr>
<td>heretofore</td>
<td>1,8519%</td>
</tr>
<tr>
<td>hypotheses</td>
<td>1,8519%</td>
</tr>
<tr>
<td>information</td>
<td>9,2593%</td>
</tr>
<tr>
<td>known</td>
<td>1,8519%</td>
</tr>
<tr>
<td>knows</td>
<td>1,8519%</td>
</tr>
<tr>
<td>linking</td>
<td>1,8519%</td>
</tr>
<tr>
<td>looking</td>
<td>1,8519%</td>
</tr>
<tr>
<td>material</td>
<td>1,8519%</td>
</tr>
<tr>
<td>means</td>
<td>1,8519%</td>
</tr>
<tr>
<td>needs</td>
<td>1,8519%</td>
</tr>
<tr>
<td>new</td>
<td>5,5556%</td>
</tr>
<tr>
<td>order</td>
<td>1,8519%</td>
</tr>
<tr>
<td>problem</td>
<td>1,8519%</td>
</tr>
<tr>
<td>pushing</td>
<td>1,8519%</td>
</tr>
<tr>
<td>relevant</td>
<td>3,7037%</td>
</tr>
<tr>
<td>resources</td>
<td>1,8519%</td>
</tr>
<tr>
<td>search</td>
<td>3,7037%</td>
</tr>
<tr>
<td>textmining</td>
<td>5,5556%</td>
</tr>
<tr>
<td>typically</td>
<td>1,8519%</td>
</tr>
<tr>
<td>unknown</td>
<td>3,7037%</td>
</tr>
<tr>
<td>user</td>
<td>1,8519%</td>
</tr>
<tr>
<td>web</td>
<td>1,8519%</td>
</tr>
<tr>
<td>written</td>
<td>5,5556%</td>
</tr>
</tbody>
</table>
It is important to emphasise the amount of irrelevant tokens that are perfectly comprehensible because of the infinity of terms and conjunctions that are not interesting in this context. Below, in Table 2, the output based on the frequency of the words is considered relevant.

Table 2  List of relevant words (applying the stemming)

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Extract</td>
<td>3</td>
<td>5,5556%</td>
</tr>
<tr>
<td>Information</td>
<td>5</td>
<td>9,2593%</td>
</tr>
<tr>
<td>New</td>
<td>3</td>
<td>5,5556%</td>
</tr>
<tr>
<td>Textmining</td>
<td>3</td>
<td>5,5556%</td>
</tr>
<tr>
<td>Write</td>
<td>3</td>
<td>5,5556%</td>
</tr>
</tbody>
</table>

In this example, we use the same weights for all words and the threshold equal to three. Due to this, the word ‘search’ which in a certain way is significant in this small text, is not so significant in the system output. For this word to be considered, the user should indicate its appropriate weight.

It is important to detach the presence of the word ‘extract’ as a relevant word. In Table 1, the words ‘extract, ‘extracted’ and ‘extracting’ are presented with just one repetition. However, through the stemming, it was possible to compare the radicals (in the case, ‘extract’) and to compute them as three repetitions. The same happens with the word ‘written, which had its nomenclature altered to ‘write’.

Another functionality, still under development, is the dictionary of synonyms (comparing the synonyms of the words). The use of this dictionary will make it possible for words such as ‘texts’ and ‘publications’ to be counted as only one.

Nowadays, the prototype just deals with texts written in only one format archive format (.txt), it having been extended to analyse texts in other file formats, such as files HTML, MS-Word (.doc) and Acrobat (.pdf).

Finishing the keyword extraction process, the system saves this analysis, the identification of each publication, as well as of its authors. This information will be vital in the process of community formation.

Each community has a set of competences previously mapped and, with text extraction, the environment presents a set of communities more related with the discovered knowledge. Moreover, the environment enables the addition of competences to the community and to the author after each publication analysed, and the suggestion of the creation of a new community if necessary.

The analysis of a community’s evolution using OLAP is still in the design phase.

6 Conclusion and future works

The presented work is a proposal for internal competence identification in scientific organisations, starting from research publications. As a way of knowledge creation, acquisition and dissemination, we have virtual communities, which in this context, are proposed and created automatically by the identification of professionals with similar
competences. For the evolution analysis of the acquired knowledge, at a personal and organisational level, this environment uses an OLAP tool and some types of pre-definition queries. The environment is web-based, providing better integration among the researchers because it allows synchronous and asynchronous communication among the community members and not demanding the physical presence for it.

This work is still in implementation phase, and one of the future steps is its completion and putting it in place at the university where this research is being carried out. In the next step, we have the analysis of the benefits brought by environment adoption in a research centre, as well as the difficulties faced and the criticism received.

As this work develops, we plan other types of document analysis which are also significant for the identification of internal competences of a research institution, such as class diaries, course proposal, researchers’ e-mails and their attachments. The proposal of automatic dissemination of documents and content to a community is still at the study phase, using for such purposes, text mining techniques to discover publications of interest to a community.

References


