

Automatic Identification of Teams based on Textual Information Retrieval

Fabrício Enembreck[‡], Edson E. Scalabrin[‡], Cesar A. Tacla^{*}, Bráulio C. Ávila[‡]

[‡]*Pontifical Catholic University of Paraná,*

R. Imaculada Conceição, 1155, 80.215-901, Curitiba, PR, Brazil.

{fabricio, scalabrin, avila}@ppgia.pucpr.br

^{*}*Universidade Tecnológica Federal do Paraná*

Av. Sete de Setembro, 3165, 80.000-000, Curitiba, PR, Brazil.

{tacla@cpgei.cefetpr.br}

Abstract

A common problem in organizations is to identify people with the right competencies to form a specialized team, in academic or industrial environments, which is capable of executing a self-managed project of Research and Development (R&D). This work presents a technique that allows for identifying people who have the most appropriated competencies to form a R&D team extracting information from their curriculum vitas (CVs). Information extraction in this work is carried out by means of textual retrieval techniques in document databases. The system was evaluated with data of real projects producing expressive results when identifying people to participate in research projects.

1. Introduction

A common problem in organizations is to identify people or suitable competencies to form a specialized self-managed research and development (R&D) team in either academic or industrial environments.

In Brazil, the Lattes database is an important source of data where one can find people in a number of knowledge areas [Lattes06]. The Lattes database contains curriculum vitas (CVs) of researchers, students, managers, professionals, and other actors of the National System of Science, Technology and Innovation. The CNPq (National Council for Scientific and Technical Development) uses such information (i) to rank candidates applying for scholarships or other kind of financial support; (ii) to select consultants, and members of committees; (iii) to assess the research and post graduation courses.

The Lattes database is representative of the Brazilian scientific population since the CNPq demands that the scholarship holders (researchers, under-graduated, master, and doctorate students), research advisors, and all the people requesting or having funds from CNPq register their CVs in the Lattes System. The register is also mandatory for all researchers and students who participate of the Directory of Groups of Research in Brazil.

In this context, given an R&D project description, our objective is to extract information that allows for identifying people having the most appropriate competencies to participate in the project using the Lattes database. In order to extract such information, we apply textual retrieval techniques.

The rest of this article is organized as follows: (i) we discuss issues on forming teams; (ii) we describe the technique to extract information from the Lattes database; (iii) we present the experiments and their results, (IV) a discussion on the results and (v) finally, we offer a conclusion.

2. Forming Teams

Considering that a whole team has more resources (e.g. information, and competencies) than its members have separately, the organization, qualification, and training of teams is useful to develop tasks that are very difficult or cannot be achieved individually. The diversity of competencies in a team makes complex tasks be possible.

In general, competency managed organizations put their strategic objectives in relation with their key business processes. Applying a systematic approach to measure the individual competencies, an organization can: (i) locate and map the knowledge

capital; (ii) use such information to analysis the employees and the organization in order to reduce the cost of training, to hire people with suitable competencies, to improve the performance of its workforce, to place more efficiently its human capital, and to aid its management to take strategic decisions (e.g. to launch new projects).

A team can be formed by people who work in the same or different areas. Nowadays, a modern approach is the formation of self-managed teams [ATTARAN99] [ELMUTI99] [SCHOLTES96] [MORAVIC98], with trained members to perform all or most part of the required tasks, without an immediate supervisor and taking decisions in a collective way, with support of coordinators. Each team must pursue its objectives trying to achieve a good performance level.

The formation and assessment of teams are not trivial tasks. The following measures may be considered in assessing the team: (i) the productivity, which consists of satisfying or exceeding criteria that refer to the quantity (e.g. number of products) and/or quality (e.g. quality of provided service); (ii) the personal satisfaction during the project, and finally, (iii) the engagement of the members in collaborative work.

To form a team presenting cohesion the following criteria may be considered: (i) to enroll members with diversified and complementary competencies; (ii) to enlist members with similar values and experiences; (iii) to keep the team as small as possible, but with a sufficient number of members to perform its functions and achieve the planned goals. Teams with low cohesion with a requirement of high performance generally do not succeed. In other words, the greatest possibilities of success are in multidisciplinary teams, where knowledge is transversal to the disciplines, and team members have complementary affinities and competencies, with capacity to add value to the project.

Forming teams is to select human resources. In the majority of cases, the best resources are not available. The management of the project must grant that the available resources satisfy the project requirements [PMBOK05]. When the management of the project has the responsibility for composing the team, it should consider the characteristics of the available human resources. It means take into consideration previous experiences - select individuals that have worked in the same or similar projects; the personal interests - whether the individuals are interested in working in the project; the personal characteristics - whether the individuals are able to work collaboratively, and finally, the

availability – whether the individuals will be available at the right moment.

Forming teams may require negotiation or previous assignments. The last case occurs when the project is the result of a proposal and the specified team was promised as part of the proposal. It may happen also in internal projects, where selected people are the ones who work in a department. In both cases, the management of the project can get external services of specific individuals or groups of individuals to carry out certain tasks of the project. So, employing external competencies is required when the organization does not have the necessary ones to satisfy the project requirements.

Members of a team work either in integral, partial, or variable time. The team starts to exist when all the members and other actors are effectively engaged in the project. Such engagement can be formal or informal, ranging from superficial to depth, depending on the project needs in relation to each person's abilities.

Forming teams involving various organizations (e.g. research laboratories, state agencies, private companies, and financial sponsors) may represent an additional difficulty in identifying and selecting members. Large projects as Genoma and Digital TV require a diversity of competencies coming from different organizations. Any tool or technique that makes easier the task of forming teams is fundamental for fast selecting the right competencies in order to answer a call for project.

3. Textual Information Retrieval

In a previous work [TACENE05], we described a textual information retrieval technique for comparing user profiles in computer supported collaborative work systems. In the present work, we use the same technique, but with a different objective: to find similarities between a description of a research project and the profiles of potential project members (candidates).

The profiles are automatically generated from knowledge items (KIs), i.e., articles, and texts written by the candidate. The KIs of a candidate are found in his/her CV saved in the Lattes database. We next describe the technique used for generating the profiles and the measures for scoring the similarity between a project description and a candidate's profile.

3.1 Generating Profiles for Candidates

In order to generate a profile, we use document classification [GOLLER00] and information retrieval techniques [YATES99] on the selected KIs.

Usually, document classification is divided into two main phases: learning and classification. In the learning phase, a collection of already classified documents (the training set) is given as input for building a representation of each class. The first step consists of selecting a set of features that are important for representing the documents. A well-known approach is to consider a number of terms to represent the documents' features. Next, the classification phase puts a new document in a particular class based on the similarity between the document and each one of the classes.

In our case, the KIs coming from a CV represent the training set. Hence, KIs are represented as vectors of relevant terms. A common measure of relevance for terms is the TF-IDF (Term Frequency - Inverse Document Frequency) measure [SAL89]. TF-IDF states that the relevance of a term in a KI is in direct proportion to its frequency in the KI, and in inverse proportion to its incidence in the whole collection D of KIs. The IDF part for the i th term is given by $\log(|D|/DF_i)$ where DF_i is the number of KIs containing the term i . TF_i designates the frequency of the i th term in a particular KI. The TF-IDF formula is given in equation (1).

$$(1) \quad TFIDF(i) = TF_i \times \log\left(\frac{|D|}{DF_i}\right)$$

A KI is considered to be a vector $\mathbf{d} = \{TF_1 * \log(N/DF_1), TF_2 * \log(N/DF_2), \dots, TF_m * \log(N/DF_m)\}$. To learn a profile based on such KIs, machine-learning algorithms (e.g. decision trees, naïve Bayes, neural networks, and nearest-neighbor) can be applied to the data [GOLLER00]. Here, we use a centroid-based approach [ENEBAR03] in which an average vector represents the profile of a candidate. Equation (2) gives the centroid vector \mathbf{c} for a collection D of KIs for a candidate.

$$(2) \quad \mathbf{c} = \frac{1}{|D|} \times \sum_{d \in D} \mathbf{d}$$

The greater the size of the KI collection, the more time consuming is the generation of the profile. So, in order to limit the number of KIs in the computation, we can set up the system to use only the KIs produced during a certain period or the last n KIs.

3.2 Text Classification

We use the centroid approach [ENEBAR03] for calculating the similarity between a project description and a candidate's profile. Classical techniques could be used such as computing the cosine or the Euclidian distance. However, we cannot use such techniques directly because the learning phase is done separately for each candidate. Their profiles have different terms and consequently different dimensions. Of course, the profiles may have common terms depending on the similarity of their KIs. For instance, candidates working on the subject probably have similar terms.

Thus, besides the similarity computation we have two additional problems: normalizing the profiles to be compared and to discover which terms best discriminate the profiles. Important terms for a candidate may not be very good for discriminating his/her profile from the other ones. For instance, the term "project" is important for most candidates, thus such a term is not a good discriminator.

In order to solve these additional problems, we propose a method for normalizing the profiles and discovering the terms that best discriminate them (Section 3.2.1). The outputs of this step are the input for the next one, i.e., scoring the similarity between a project description and a profile (Section 3.2.2).

3.2.1 Computing the Discriminating Power of Terms. In order to measure the discriminating power of the terms figuring in the profiles, we use the *Gini* index technique [SHAKAR00]. Let $\{c_1, c_2, \dots, c_m\}$ be the set of profiles computed according to Equation (2) and T_i the vector derived from the relevance of the term i in all the profiles — $T_i = \{c_{1i}, c_{2i}, \dots, c_{mi}\}$. T'_i is the vector T_i normalized with the one-norm — $T'_i = \{c_{1i} / \|T_i\|_1, c_{2i} / \|T_i\|_1, \dots, c_{mi} / \|T_i\|_1\}$ the discriminating power of i — p_i — is given by Equation (3).

$$(3) \quad p_i = \sum_{j=1}^m T'^2_{ji}$$

p_i is equal to square of the length of the T'_i vector. So p_i is always in the range $[1/m, 1]$. p_i has the lower value when $T'_{1i} = T'_{2i} = \dots = T'_{mi}$, whereas the higher value of p_i is given when only one profile has the term i .

The p_i measure acts as a normalizing parameter allowing profiles with different terms to be compared. We compute the p_i discriminating power for the terms of the whole collection of profiles. Such

p_i measures are used in the similarity computation step described in the following sub-section.

3.2.2. Comparing a profile and a project description. In order to quantify the similarity between a project description c_1 and the profile c_2 , we create a comparable vector c'_2 as follows (same size and same terms): for each term i in c_1 that exists in c_2 , c'_{2i} keeps its TF_iIDF_i value. When a term i figuring in c_1 is not found in c_2 then c'_{2i} is set to zero. Next the similarity score between c_1 and c_2 is computed using the terms' discriminating power p_i according to Equation (4). It is important to stress that the score between c_1 and c_2 is not symmetric.

(4)

$$similarity(c_1, c'_2, p) = \frac{\sum_{i=1}^{|c_1|} c_{1i} \times c'_{2i} \times p_i}{|c_1|}$$

In Equation 4, we compute the average quality of the terms inside the profiles, taking into account the discriminating power of each term. To increase the system performance we use only “good” discriminating terms, thus we have introduced an empirical threshold to avoid the influence of the poor ones. We consider terms having a p_i greater than or equal to 0.3, otherwise we arbitrarily set the p_i to zero.

4. Approach for Automatically Forming Teams

We use the technique presented in the previous section to develop a system able to identify people that may have the required competencies for a project. The inputs for the system are a project description and a set of CVs coming from the Lattes database. The output is a set of people that could form the team. Figure 1 shows the processing steps for the system.

In order to evaluate the proposed approach, we have selected CVs from a well-known Brazilian master course in Computer Science. For each CV, we selected the 20 most important KIs according to the following priority: (i) publications in scientific journals, (ii) publications in books or book chapters, and (iii) publications in recent conferences.

A typical KI contains the following information: (i) authors, (ii) article or chapter title; (iii) journal title, book title or conference name; (iv) publishing company; (v) year of publication; (vi) keywords; (vii) knowledge areas; (viii) additional references; (ix) media. Such information follows a standard format specified by the Lattes database.

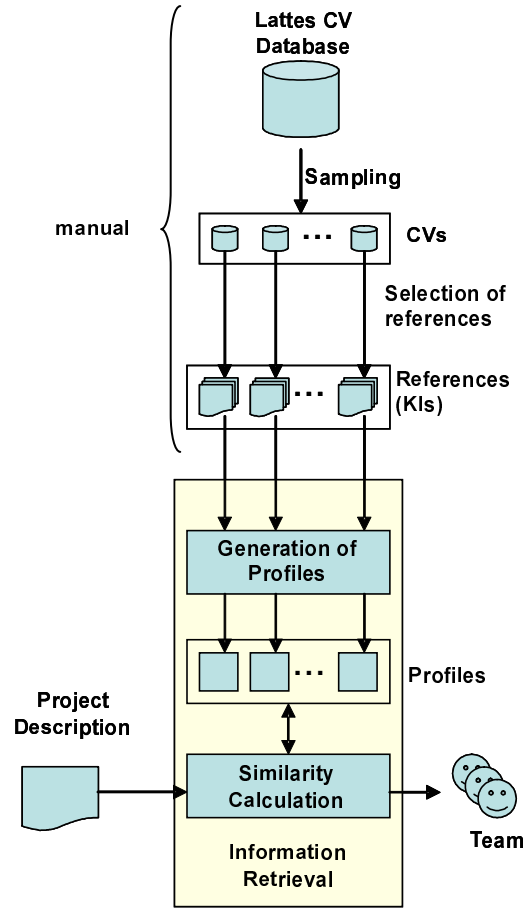


Fig 1. System Processing Steps.

The choice of a specific master course facilitates the analysis of results because the course has three well-defined groups of research, each group with its own set of professors.

The project descriptions should be realistic and the involved teams should be known in advance in order to measure the quality of the results produced by our approach. Thus, we use the last 52 projects of master's dissertations written by the professors of the course. These projects are frequently multidisciplinary and may involve people from different groups of research. The description of each project consists of a less than 500 words text. The header of the text was not considered in order to introduce an extra difficulty to the classification task.

5. Experiments

The master program on Computer Sciences has 22 professors and 3 groups of research. For each

professor 22 KIs were selected, and each KI was represented by a 500 terms vector.

The evaluation methodology compares the candidates (i.e. the potential team members) proposed by the system with the examiners of the Master's dissertation (i.e. the actual members of the project). Generally, there are three internal examiners, the advisor included, and one external. These three internal examiners give a strong indication of the professors who are interested or whose works are related with the project.

The evaluation uses two scores: FF (first-first) and AG (Advisor's Group). The FF score compares only the first candidate proposed by the system with the project manager (i.e. in our case, the advisor).

The AG score compares the team proposed by the system with the three groups of research. The AG score is higher when the system recovers people from the group of research of the advisor. Many times, people that are in the same group of research present similar profiles and even a human being would have difficulty in selecting one of them to a team. Thus, suggesting people that belong to the same group as members of a team seems a coherent alternative.

Table 1. Preliminary Results.

FF	AG
10/52 = 19,23%	39/52 = 75,00%

Our preliminary results (Table 1) show that the system is able to identify approximately 20% of the project managers (the advisors). The low value for the FF score is explained by the superposition of the research subjects, projects, and publications of people belonging to the different groups of research. People often work on the intersection of different themes what makes difficult to identify which of these subjects is more important for them.

On the other hand, the system has got a reasonable value for AG, demonstrating that the context of the project could be identified in the great majority of the projects. Despite the good preliminary results, we try to explain why the system fails for the 13 remaining cases.

Analyzing these cases, we remark that eight of them represent unusual situations that the system could not foresee adequately. These eight projects were qualified as innovation projects, and the advisor haven't had an important scientific production related to it. We can say that these projects represent new directions of research. Therefore, it would be practically impossible to the system to foresee the advisor. Considering this, we revise the preliminary results of table 1 by excluding these innovative projects. Revised results are shown in table 2.

Table 2. Revised Preliminary Results.

FF	AG
10/44 = 22,72%	39/44 = 88,63%

We have noticed a significant improvement in the results when the innovation projects had been excluded from the initial base of projects. Logically, a research institution has always some pioneering projects because it is essential for the research. Although this is not the objective of this work, the developed methodology allows these pioneering projects to be identified automatically, configuring itself as an important tool in the management of scientific politics in research programs.

6. Discussion

The Lattes database represents a rich source of information for the management of politics of incentive to the research. Despite this, few automatic mechanisms have been developed to explore this potential. Amongst the advantages of using the Lattes database is the standard format for describing the scientific production reducing the information noise. However, despite the effort of standardization of information in the Lattes system, many researchers have different interpretations for some inputs of the system. To make our approach more realistic and scalable, no pre-processing was done on the KIs extracted from the CVs.

Another possible source of noise is the use of an automatic translator from Portuguese to English used on the project abstracts. No post-processing was carried out on the text produced by the translator. Considering these sources of noise in the system, we believe that the results of our approach are satisfactory.

Although project management systems are capable of controlling temporal constraints, resources, scheduling and quality of tasks in projects, they fall in providing high level functionalities. Recently Artificial Intelligence has been successful exploited and tools based on Data Mining, Multi-Agent Systems and Knowledge Representation Approaches (Ontologies) are becoming common [ENEMBRE04]. We believe that the Information Retrieval concepts discussed in this paper provide an important contribution in such a direction.

7. Conclusions

In this paper we have presented a methodology for discovering members of project teams. The discovering of the right skills in project teams is a

challenge for any organization. A good choice for the work groups can reduce project risks, cost and guarantee a successful collaboration process.

In most cases the group formation is based on the feeling of the project manager and/or organizational structure of the institution, what do not guarantee the best team. Although a successful collaboration process depends on subjective factors like confidence, friendship and other forms of interpersonal relationship, this research regards only professional aspects. This is a requirement for organizations like the CNPq where impersonality is imperative.

The methodology has produced quite good results for the sample of CVs and projects used in the experiments. However many questions are open. We arbitrarily select 20 references for each researcher but why not use the entire CV? Using the entire CV could favor more experienced researchers but this could be a desired behavior! Most part of scientific articles is published in English, but the project descriptions were written in Portuguese, so English was chosen as target language. Manual translation could be more accurate but it is much more expensive too. Another important point is the size (the number of terms) of the candidate profiles. We have tested our approach varying the vector size from 200 to 5.000, and the 500 terms vector has produced the best balance between the processing time and the quality of results.

6. References

- [ATTARAN99] Attaran, Moshen; Nguyen, Tai T., "Self-Managed Work Team", *Industrial Management*, 41, July/August, 1999: 24.
- [ELMUTI99] Elmuti, Dean, "Sustaining High Performance Through Self-Managed Work Teams", *Industrial Management*, 41, July/August 1999: 4-8.
- [ENEBAR03] F. Enembreck, and J-P. Barthès, "Agents for Collaborative Filtering", *Cooperative Information Agents VII*, Springer-Verlag, LNAI 2782, M. Klusch, A. Omicini, S. Ossowski, H. Laamanen (eds.), Helsinki, August, 2003, pp. 184-191.
- [ENEMBRE04] Fabricio Enembreck; Indira, Thouvenin; et al., An Ontology-Based Multi-Agent Environment to Improve Collaborative Design. In: 6th International Conference on the Design of Cooperative Systems - COOP'4, Hyeres, 2004. p. 81-89
- [GOLLER00] GOLLER, C.; LÖNING J.; WILL, T.; WOLF, W., *Automatic Document Classification: A thorough Evaluation of various Methods*, IEEE Intelligent Systems, n° 14, vol 1, pp. 75-87, 2000.
- [Lattes06] <http://lattes.cnpq.br>
- [MORAVIC98] Moravic, Milan, "The Well-managed SMT", *Management Review*, 87, June 1998: 56-59.

[PMBOK05] Duncan, W. R., "A Guide to the Project Management Body of Knowledge", PMI Publishing Division, Forty Colonial Square, Sylva, North Carolina, USA. ISBN 1-880410-12-5

[SAL89] G. Salton, *Automatic Text Processing: The Transformations, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, 1989.

[SHAKAR00] S. Shankar, S., and G. Karypis, "A Feature Weight Adjustment Algorithm for Document Categorization", *KDD-2000 Workshop on Text Mining*, Boston, USA, August 2000.

[SCHOLTES96] Scholtes, Peter R.; et al, "The Team Handbook", Madison, Wisconsin: Oriel Incorporated, second edition, 1996: 1-16.

[TACENE05] TACLA, Cesar; ENEMBRECK, Fabrício. An Awareness Mechanism for Enhancing Cooperation in Design Teams. In: THE 9TH INTERNATIONAL CONFERENCE ON CSCW IN DESIGN, 2005, Coventry. The 9th International Conference on CSCW in Design. 2005.

[YATES99] R. B. Yates & B. RIBEIRO NETO. "Modern Information Retrieval". Addison-Wesley, 1999.