Journal of the Brazilian Computer Society, 2009; 15(4):31-39. ISSN 0104-6500

scriptLattes: an open-source knowledge extraction system from the Lattes platform

Jesús Pascual Mena-Chalco*, Roberto Marcondes Cesar Junior

Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo – USP,
Rua do Matão, 1010, 05508-090, São Paulo, SP, Brazil

Received: July 5, 2009; Accepted: December 15, 2009

Abstract: The Lattes platform is the major scientific information system maintained by the National Council for Scientific and Technological Development (CNPq). This platform allows to manage the curricular information of researchers and institutions working in Brazil based on the so called Lattes Curriculum. However, the public information is individually available for each researcher, not providing the automatic creation of reports of several scientific productions for research groups. It is thus difficult to extract and to summarize useful knowledge for medium to large size groups of researchers. This paper describes the design, implementation and experiences with scriptLattes: an open-source system to create academic reports of groups based on curricula of the Lattes Database. The scriptLattes system is composed by the following modules: (a) data selection, (b) data preprocessing, (c) redundancy treatment, (d) collaboration graph generation among group members, (e) research map generation based on geographical information, and (f) automatic report creation of bibliographical, technical and artistic production, and academic supervisions. The system has been extensively tested for a large variety of research groups of Brazilian institutions, and the generated reports have shown an alternative to easily extract knowledge from data in the context of Lattes platform. The source code, usage instructions and examples are available at http://scriptlattes.sourceforge.net/.

***Keywords**: academic production report, Lattes platform, knowledge discovery.*

1. Introduction

Knowledge extraction and visualization from large datasets is an important research topic in computer science with strong potential impact in all scientific fields[12]. The research on this topic tipically involves the treatment of large datasets which can not be processed and understood by human experts due to its volume, diversity and complexity. Techniques from datamining, summarization, visualization, network modeling and high-performance computing are often brought together in order to solve problems of this nature. The present work is focused on the summarization and visualization of scientific reports obtained from the Brazilian Lattes platform.

The *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (Brazilian National Council for Scientific and Technological Development - CNPq), makes efforts to integrate the Curricula of people associated to Brazilian scientific communities, in a curricular information system denominated Lattes[I]. For this reason, the so-called "Lattes Curriculum" is considered a national standard of information about the scientific and academic accomplishments of students, professors, researchers and professionals involved in science and technology in general.

The Lattes curriculum is used for academic evaluation because it represents the history of scientific, academic and professional activities[1]. It is hence a rich and powerful database that presents inumerous potential applications (scientific, technological, economical, etc.) The Lattes curriculum, in HTML format as available in the CNPq site, displays information only in a personal way, i.e., the registered information is individually associated to each person. This characteristic does not easily provide a way to figure out the bibliographical, technical or artistic productions of a given group, such as a research group, professors of an academic department or members of a Brazilian institution.

Currently, most of the Brazilian academic institutions usually explore the Lattes curricula in order to elaborate reports about scientific productions, supervisions, and projects of research groups related with these institutions, as well as to evaluate the graduate programs in Brazil. The reports are typically created by manually-assisted analysis of the Lattes curriculum data of each member of the group in order to obtain a complete digest of all scientific productions, supervisions and projects of the group. It is important to note that, despite having structured information, this procedure is very cumbersome and time consuming, being highly susceptible to errors caused by the manual treatment.

I. The software platform, available at http://lattes.cnpq.br/, was named Lattes in honor to Cesare M. G. Lattes, a Brazilian physicist, co-discoverer of the pion or pi meson, one of the nuclear particles.

*e-mail: jmena@vision.ime.usp.br

There are some interesting bibliometric questions that may be answered about a group just based on the respective Lattes curricula:

- How many bibliographical, technical or artistic productions were elaborated?
- What is the profile (i.e. proportion of publication) of the different types of bibliographical productions?
- How is the regularity and the evolution of the publications along the years?
- How is the collaboration/cooperation among researchers?
- How many thesis and dissertations have been concluded?
- What is the geographic distribution of the researchers?
- What is the scientific formation influence of the considered researchers?

The scriptLattes, an open-source system, was designed to provide answers to the above questions through automatically created reports. Given a group of researchers registered in the Lattes platform, the scriptLattes download their Lattes curricula from the CNPq site, extract the information of interest, eliminate the redundant scientific productions and create reports about the production, reports of academic supervisions as well as the collaboration graph and the research map from the members of the group. We believe that the introduced system is a useful tool to easily extract knowledge about the Lattes curricula of a group. This knowledge may be used to explore, identify or validate patterns of academic activities, thus bringing bibliometric information about a group of interest [22].

This open-source system runs on a PC with GNU/Linux using Perl modules and basic structures of programming languages. The scriptLattes is a project registered in the Free Software Competence Center at University of São Paulo, being hosted at SourceForge [29]. To the best of our knowledge, the system is the first to be widely used in several Brazilian academic groups, including to the University of São Paulo (USP), the State of São Paulo Research Foundation (FAPESP), and the Agency for Agro-business Technology in the State of São Paulo (APTA), for instance. The system was successfully tested with at least 300 research groups of Brazilian institutions.

Therefore, the present paper describes a system that allows scientific data summarization from a strucutred database of curricula vitae, i.e. the Lattes platform. In this context, the paper describes a new system that allows the extraction of useful summarized knowledge from large set of data, a task that would be too difficult (in many cases, impossible) to be performed manually. In order to produce such a system, different solutions and algorithms have been adopted or proposed and implemented, being described in the paper. Possible applications of the system are also discussed. The paper's contributions show how Computer Science tools (developed data-structures, algorithms, visualization techniques and software system) solve an important knowledge extraction problem from large datasets. The relevance of the paper's contributions relies in the context of knowledge extraction and visualization systems that deal with possibly large datasets, an important Computer Science research topic[10,11,12,15].

The remaining of the paper is organized as follows: Section 2 discusses some important background references followed by Section 3, which describes the modules of the proposed system. Some results illustrating the use of scriptLattes, as well as a form to explore the obtained information, are described in Section 4. Finally, the conclusions and future directions are summarized in Section 5.

2. Background

There has been intensive growing attention to the problem of extracting meaningful summarized knowledge from large volumes of data[12]. Papers under different perspectives (knowledge extraction, e-Science, data intensive paradigm, etc.) have appeared in most important Computer Science forums and discussed the main involved aspects, from computational tools to applications. For instance, Kouzes and collaborators[15] describe some of the typical architecture components of a knowledge extraction systems. Some of these components inspired the tools implemented in the scriptLattes. An interest resource is the Digging into Data website[11], which presents a National Science Foundation (among other funding agencies) call for applications to extract knowledge based on data-driven inquiry of large sets of books (mainly on Social Sciences and Humanities). The book[10] discusses how knowledge extraction plays a central role in modern scientific fields such as environment, health and scholarly communication. Different problems, systems and computational solutions are presented by the authors in this edited book. An important field of application for such techniques is the analysis of scholar data such as publications, supervisions, collaboration and scholar influence[10]. The analysis of academic data is carried out in different levels, from the case where students and researchers are exploring collaborations and supervisors to institutional levels of academic assessment of whole departments[16]. Many works have been devoted to the analysis of co-authorship and collaborations from papers databases. Many of the issues in this research appear from the low degree of structure of the data as well as the ambiguities often present[9,22,27]. Another important issue is the analysis of the networks themselves[14,17,20,28].

It is worth noting that co-autorship is not the only approach to create scientific networks. For instance, text-mining helps the generation of paper networks allowing clustering and hierarchical analyis of large datasets of papers based on subject. In all such cases, visualization is of utmost importance in order to produce good interfaces to allow the user to understand the summarized data[26]. Visualization of networks plays a central role in helping the user to understand and to interact with data, mainly because large volumes are tipically involved[3].

3. Modules of the System

The input of the system is composed by an ASCII list of Lattes curriculum's IDs in conjunction with the time period of each member of the group to be analyzed, i.e., the years where each member has been associated to the group (research group, institute, department, university, etc.) The ID of a Lattes curriculum is a number of 16 algarisms associated to each person registered in the Lattes platform, being easily obtained from the Lattes curriculum. Therefore, the IDs are commonly used in the request of a given curriculum. See in Figure 1 an example of the input file format.

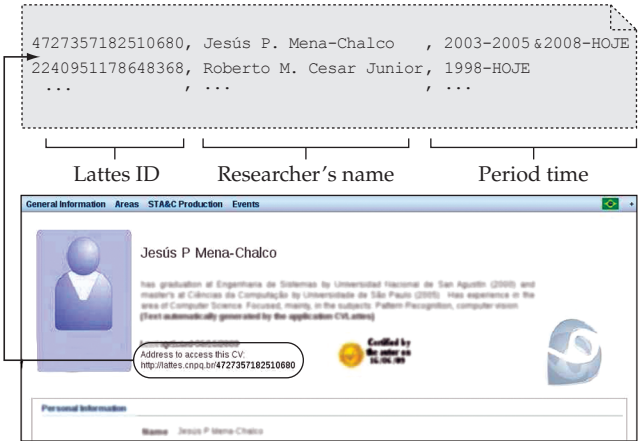

Figure 1. Example of input file (dotted lines). Each line corresponds to data from an author. The period time (third column value) is optional to each researcher. The figure shows where the Lattes ID may be obtained from the Lattes curriculum.

The system has been divided into six modules. Figure 2 shows the schematic data flow diagram of the whole system, where each module is responsible to process a given type of information of the Lattes curricula. The system output is given by several reports, in HTML format, showing the summarized information in terms of bibliographical, technical and artistic production, as well as academic supervisions, collaborations and research map among the members. The HTML format was chosen in all reports because of being a standard format to be visualized on the internet.

3.1. Data selection

This module allows to download the Lattes curricula, in HTML format, from the site of the Lattes platform. In order to facilitate the comparison between strings, the curricula were normalized to use the same characters codification.

The curricula are downloaded in HTML format because they are publicly available only as plain HTML. Public users of the Lattes platform do not have access either to the Lattes database or to data in XML format. As a result, special attention is devoted to extract the information about the scientific productions, as explained in the data preprocessing module.

3.2. Data preprocessing

In this module, a HTML parser is used to extract the information about the professional address, the list of productions, the list of ongoing and concluded supervisions, and the available photograph in the Lattes curriculum. The list of productions and the list of supervisions are limited by the

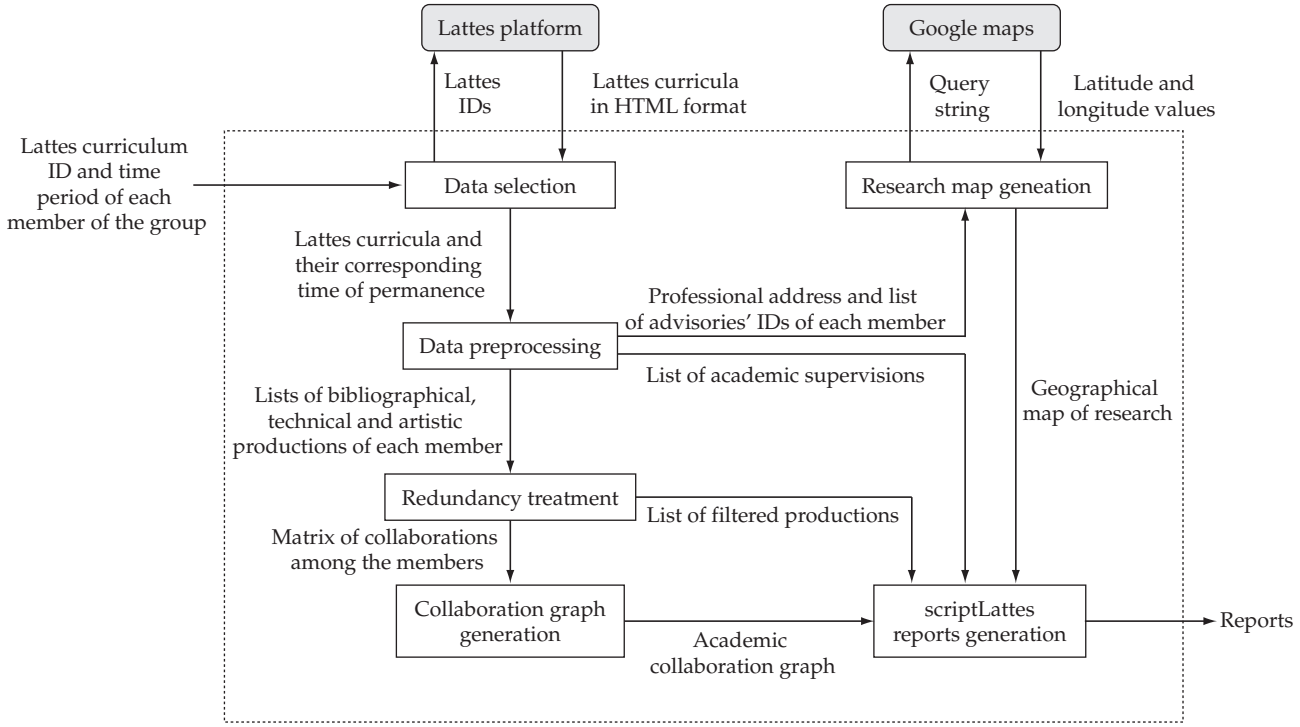

Figure 2. Schematic data flow diagram of the system (dotted lines). Each block (solid lines) represents a module while each arrow represents the information flow between modules or external platforms (gray boxes).

years indicated in the time of permanence of each member of the group.

It is important to note that the types of scientific productions considered by the scriptLattes are those registered in the Lattes platform. All available information from the Lattes curricula are considered as validated by the CNPq. See in Table 1 the types of productions and supervisions considered by the system.

3.3. Redundancy treatment

It is common that the productions are made in collaboration with one or more collaborators of the same group. Therefore, the same production (e.g. a journal paper) may appear duplicated because of being declared in each curriculum vitae of the corresponding co-authors. Thus, this module allows to detect and to eliminate the duplicated productions of the lists obtained after preprocessing the Lattes curricula. Furthermore, the duplicated productions are used to detect collaboration: two researchers are defined as collaborators in the software if they have a common production detected by this module.

Table 1. Information extracted from the Lattes curriculum.

Personal information
Name
Professional address
Bibliographical production
Articles in scientific journals
Book published/organized
Book chapter published
Articles in newspapers/magazines
Complete works published in proceedings of conferences
Expanded summary published in proceedings of conferences
Summary published in proceedings of conferences
Articles accepted for publication
Presentations of work
Other kinds of bibliographical production
Technical production
Patented or registered software
Not patented or registered software
Technological products
Techniques or process
Technical works
Other kinds of technical production
Artistic productions
Artistic/cultural production
Ongoing / concluded supervisions
Postdoctorate supervision
Ph.D. thesis
Master's thesis
Monograph of completion for improvement/specialization
Works of completion for graduation
Scientific initiation
Other academic advisory

The productions title is considered as a standard characteristic in order to compare the scientific productions elaborated by the analyzed group. This characteristic is adopted because it is always displayed in all productions presented by the Lattes curricula in HTML format. Note that we also considered other characteristics such as the authors name or DOI number. Nonetheless, these values have not yet been standardized in the Lattes platform, being more difficult to parse and to compare by automatic means in the current version.

The following algorithm summarizes the process applied by this module on a list of scientific productions:

```
FILTER-PRODUCTIONS-LIST(List, x)
1  for p_i ← 1 to |List| - 1
2      do Ap_i ← AUTHOR(List[p_i])
3         isNotSimilar ← TRUE
4         for p_j ← p_i + 1 to |List|
5             do Ap_j ← AUTHOR(List[p_j])
6                if Ap_i ≠ Ap_j and Compare(List[p_i],List[p_j])> x
7                   then ▷ p_i and p_j have x percent of similarity
8                        matrixOfCollaborations[Ap_i][Ap_j]++
9                        matrixOfCollaborations[Ap_j][Ap_i]++
10                       isNotSimilar ← FALSE
11     if isNotSimilar
12      then ▷ The production p_i is unique into the List
13           PUSH(FilteredList,List[p_i])
14 return FilteredList
```

The Filter-Productions-List procedure requires a list of productions and a x value as parameters. The x value is the desired threshold (as a percentage value) in order to identify the similar productions. The longest common subsequence algorithm[5] was used by the Compare procedure in order to calculate the percentage of similarity between the title of the scientific productions (for comparison purposes, only the size of the longest common subsequence was used as a measure of distance between productions). Thus, the publications list is filtered by removing the equal or similar publications into the *List*. Currently, in this module, the productions with similarity of 92% in their titles are identified and merged. This percentage was adopted after preliminary experiments. This value can be easily modified.

All redundant scientific productions are recorded in a bi-dimensional matrix, *matrixOfCollaborations*, that stores the number of bibliographical, technical and artistic collaborations among the members. Let n be the length of elements into *List*. Since we have compared the *i-th* production with the $(n - i)$ remaining productions, $O(n^2)$ comparisons are required to filter the complete list.

Note that the process of this module is individually applied to each type of productions. Therefore, there may exist some publications with similar titles but belonging to different types. Those are not identified by the system as being the same.

3.4. Collaboration graph generation

Commonly, a scientific collaboration graph describes research activities that have been carried out by a research group[19]. In this module, scriptLattes uses a graph to represent the collaboration among members of a group based on scientific productions. Each member is represented by a node. An edge is created between a pair of nodes whenever a common production of the corresponding researchers is detected by the redundancy treatment module. In other words, if two researchers have co-authored a common production, their respective nodes in the collaboration graph are linked by an edge. The edge weights are calculated as the number of co-authored scientific productions. Therefore, a weighted graph is generated.

The process in this module is simplified by using the *matrixOfCollaborations*, a symmetric matrix that contains the total amount of co-authored productions between members, being computed by the redundancy treatment module. *matrixOfCollaborations* is taken as an adjacency matrix that represents the graph of collaborations. In the generated graph, it is possible to observe the collaboration among members and clusters of cooperation. This graph is an instrument that helps to discover the researchers with more activity of co-authoring within the group and could be used in detailed analysis of co-authoring as[17,20,28].

3.5. Research map generation

It is often desired to know the geographic location of members of a particular group as well as of their alumni. In this context, scriptLattes generates a "research map" which represents the geographic location of the group members on the world. This module allows to create a research map using an external platform of geographic addresses.

The country name, city name, and the zip code number of each member, available in the Lattes curricula, is used to query the location: the longitude and latitude values are obtained if the query string is valid. Furthermore, the geographic location of each formed PhD by the group is also plotted in the map. Thus, the research map shows where the formed students are working, given an idea of the influence maintained by the research group. It is important to note that, to accomplish the retrieval of the latitude and longitude values of the geographic position of each member, an interface with Google Maps is used.

3.6. Reports creation

Reports of the productions, as well as for ongoing and concluded supervisions, are created in this module. These reports are separated by type and show a quantitative information disclosed per year in inverse chronological order.

Bar charts are associated to the reports, where the bar lengths are proportional to the values of scientific productions of the group. Hence, it is possible to discover if the production volume of a particular group came to baseline.

Additionally, the generated reports show web links to search engines (including Google and Scholar, among others) in order to find possible citations or similar works. Currently, reports in HTML or JSP format are generated. These formats have been chosen because of being suitable to the internet. The collaboration graph can be visualized in PNG, PostScript or using an interactive java applet[18].

4. Results

We describe two experiments that illustrate the software potential applications.

4.1. Experiment using a single group

In this section we show an example of reports generated by the system. We use the Lattes curricula of 41 professors associated to the Department of Computer Science of the Institute of Mathematics and Statistics at the University of São Paulo (DCC-IME-USP). The time of permanence of each member was indicated as an input data, whenever the information is available.

Figure 3 shows some generated reports of the group. The complete list is available in[23]. See in Figure 3c the generated collaboration graph. Figure 3d presents an example of the generated research map. Each member of the group is represented in green and each formed PhD in blue. The advisor-advisee relation is represented with a thin straight-line. The name of the researcher and the complete address are shown by clicking the location points.

The matrix of collaborations can be used in different approaches to explore and to extract knowledge. For instance, in Figure 4 it is possible to observe the temporal collaborations of a group of professors in the last five triennia (years 1994-2008). The temporal academic collaboration graphs show how the collaborations were made during each period, i.e. the evolution of collaborations among the members through the years. Note that members 1 and 2 have been collaborating in the last five triennia in a stable way. Member 6 only had collaboration with other members in the triennium 2003-2005. On the other hand, member 5 has been collaborating in the group since the triennium 2000-2002. With the temporal graph of collaborations it is possible to automatically characterize the collaboration among members. In particular, it is possible to determine the degree of temporal collaboration of the whole group using measurements such as those described in[6].

4.2. Experiment using several groups

The second experiment briefly investigates the profile (proportion of publication) of bibliographical productions from several groups with distinct characteristics.

To define the research groups, we used the Brazilian research groups database[II] maintained by the CNPq. On May of 2009, using the keyword *computação* as query in the

II. Available in http://dgp.cnpq.br/buscaoperacional/.

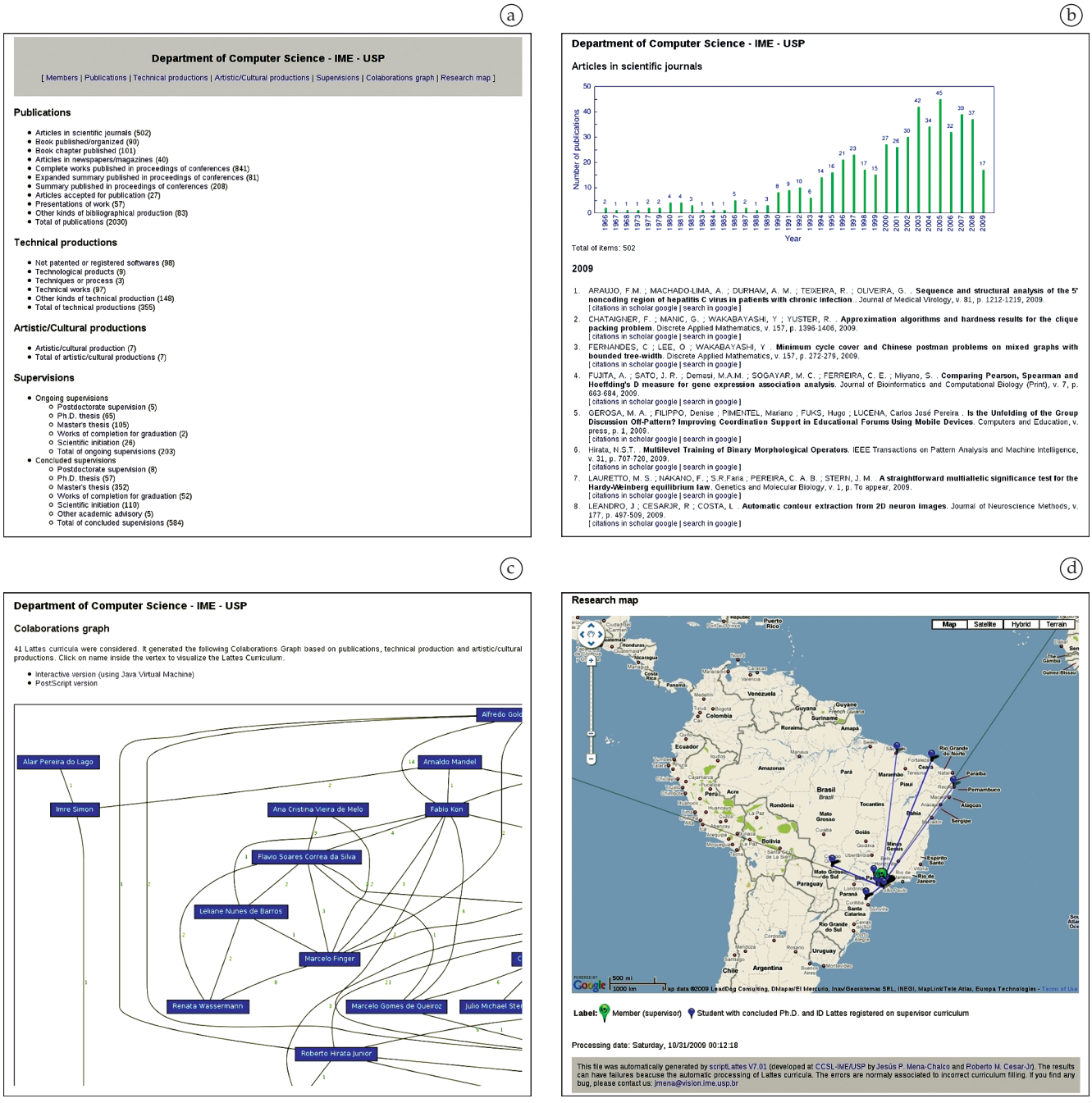

Figure 3. An example of reports obtained with scriptLattes: a) Index of generated pages; b) List of articles in scientific journals; c) Collaboration graph; (d) Research map. These results are available at http://www.vision.ime.usp.br/creativision/publications_dcc/.

database, 56 research groups in Computer Science belonging to São Paulo State were extracted. In all groups, the lists of researchers registered on the CNPq were used to create the lists of Lattes curricula. Altogether, 56 lists were created as input to the scriptLattes. In order to consider the same time periods of the scientific productions, in the remaining of this experiment the productions of the last 10 years were considered, i.e., scientific productions between 1999 and 2008.

Figure 5a shows the total amount of bibliographical productions extracted from 306 members belonging to the research groups using scriptLattes. Note that these global amounts indicate a preference to publish as follows: complete works in conferences, summary in conferences and articles in scientific journals. See the radar chart in Figure 5b. The data visualization through radar charts helps to display both the dominant type of bibliographical production for a given group, and which groups are most similar, as described below.

The reports of bibliographical productions of each group were used to explore/classify the profiles of the research groups. The scientific productions of the last 10 years of each bibliographical type were concatenated to

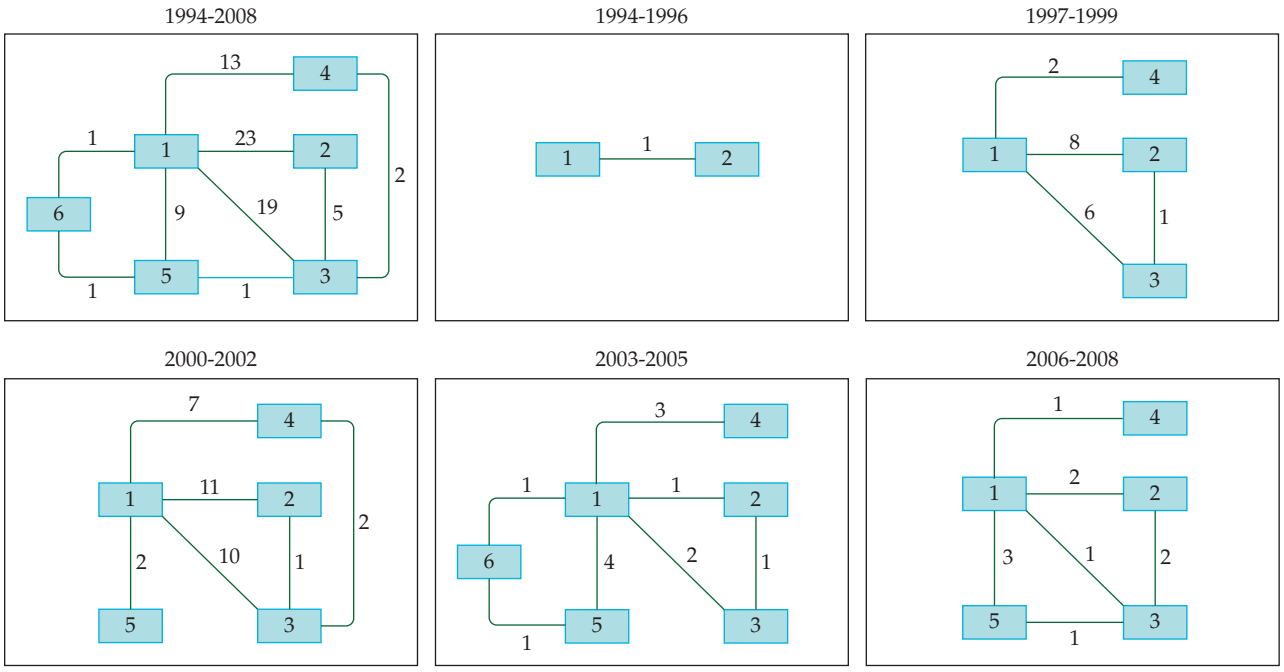

Figure 4. Temporal graph of collaborations based on the productions belonging to the Computer Vision Research Group[24] of the Department of Computer Science at University of São Paulo. The edge weight represents the number of co-autored productions.

a)

Type of bibliographical production	Total
1. Articles in scientific journals	1816
2. Book published/organized	220
3. Book chapter published	403
4. Articles in newspapers/magazines	310
5. Complete works published in proceedings of conferences	4544
6. Expanded summary published in proceedings of conferences	586
7. Summary published in proceedings of conferences	2264
8. Presentations of work	1233
9. Other Kinds of bibliographical production	472

b)

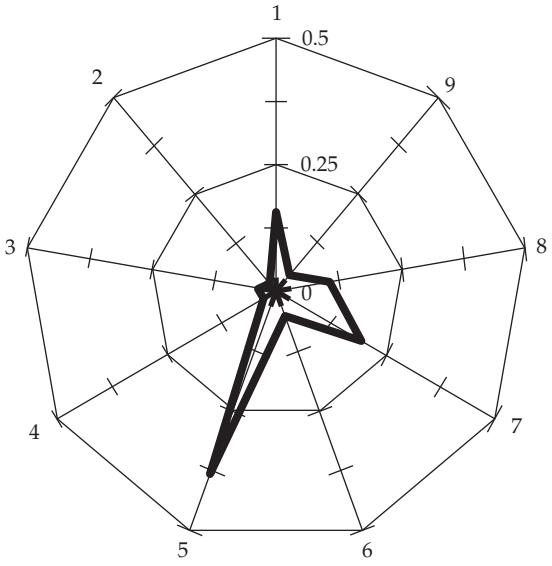

Figure 5. Bibliographical productions extracted from 56 research groups associated to Computer Science in the São Paulo State: a) Total amount of scientific productions; b) Radar chart of the productions profile (proportion of publication). The axes, represented by a number, are related to each type of production listed in a).

form a vector of 90 elements. Then, the vector was normalized by dividing its elements by the maximum value. Thus, each element stores a relative amount of annual scientific productions of the group. Principal component analysis[13] (PCA) was carried out to examine the 56 vectors of temporal scientific productions. This analysis allows to transform correlated data into a smaller number of uncorrelated principal components. In particular, the first two components were used to define four classes of research groups. The groups were clustered in classes using the centroid distance method. It was adopted the centroid distance method because it is a natural form to cluster few groups. Please refer to[8] for more detailed descriptions. The first two components take into account more than 36.5% of the total variation of the analyzed vectors.

Figure 6a shows the results of clustering the 56 Computer Science research groups using the PCA data extracted by the scriptLattes reports. Each research group is represented by a

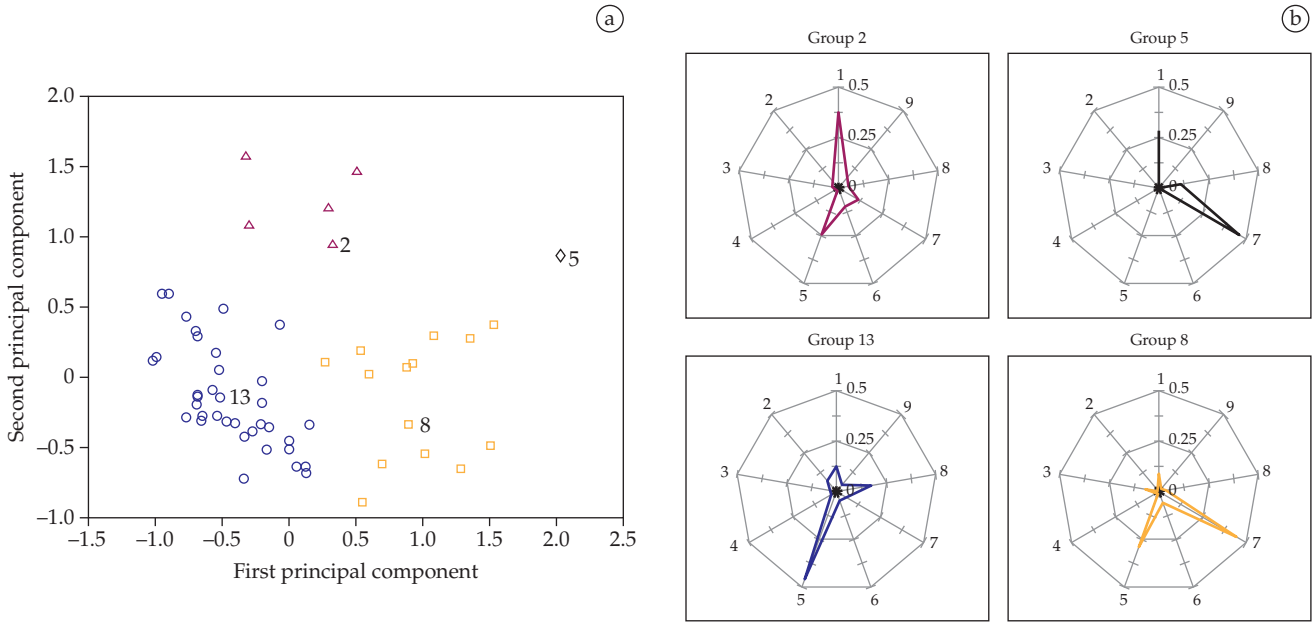

Figure 6. Analysis of the scientific productions profile: a) Clustering of 56 research groups in Computer Science associated to the São Paulo State. Note that four classes were identified. Each class is represented by a different geometric form. b) Radar charts of representative groups of the identified classes. The axes, represented by a number, are related to each type of production listed in Figure 5a.

regular shape (e.g. circle, square, triangle or diamond) and corresponds to one of four identified classes.

In Figure 6b the production profiles from four groups are shown: 2, 5, 8 and 13. These groups were selected because they are the more representative of the identified classes. Observe that the first principal component represents variation in the profile from complete works in conferences to summaries published in conferences. The second is rather flat, probably representing the variation in the profile from complete works in conferences to articles in scientific journals. This semantic interpretation is dependent upon the area of the analyzed research groups.

Similar approaches can be explored in order to analyze the overall progress of productions developed by scientific and technological Brazilian groups, i.e. to analyze the patterns of increasing or decreasing of bibliographical productions through the years. In this context, the quantitative values obtained with scriptLattes can be used to automatically estimate the different patterns of productivity and show if any pattern is significantly influenced by the area of knowledge (e.g. Social, Applied, Medical or Biological Sciences).

5. Conclusions

In this paper we have presented an open-source system for extraction and visualization of knowledge from Lattes curricula. The designed system allows to analyze scientific productions based on curricula registered in the Lattes platform. It is a simple form to obtain a survey of significant performance indicators of research groups, and to analyze the progress of scientific productions and the relevant information about their activities (e.g. see in reference[30] an analysis of the performance of research groups associated to Agronomy, Genetics and Sociology).

The scriptLattes only deal with structured information from Lattes currila. A natural path to be improved in the system is the use of information from other more general or semi-structured sources. In that sense, we can explore new methods in regard to automatically extract information[7,27].

An important issue regarding the generation of the collaboration graph is the treatment of the correct title disambiguation in author citations. Alternative methods, such as in[9], may be explored with the aim of use an unsupervised learning approach. Currently, the scriptLattes is being improved with the purpose of have a more general and efficient system. Future improvements aims at exploring the analysis about co-authoring networks[2,17,20,28], measuring and extracting proximity in networks[14], finding and evaluating community structure in networks[21], and applying ontologies[4,25]. We believe that these new approaches will allow more contributions to the extraction and management of knowledge from the Lattes platform.

While there is great potential to explore the Lattes curricula vitae there still needs to be important effort on the part of institutions which host the data to supply software interfaces, facilitating the data extraction process to the pursuit of knowledge discovery on Lattes platform. Until that time, there is an important opportunity for combine efforts to provide the bibliographical production data that can integrate the information that is already available on the internet.

Acknowledgements

The authors would like to thank Fabio Kon, Fabrício Martins Lopes and Yossi Zana for discussions and suggestions on this work and the anonymous reviewers for the critical comments and valuable suggestions. This work was supported by CAPES, CNPq and FAPESP.

References

1. Amorin CV. Curriculum vitae organization: the Lattes software platform. *Pesquisa Odontológica Brasileira* 2003; 17(1):18-22.
2. Balancieri R, Bovo AB, Kern VM, Pacheco RCS and Barcia RM. A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na Plataforma Lattes. *Ciência da Informação* 2005; 34(1):64-77.
3. Börner K, Chen CM and Boyack KW. Visualizing knowledge domains. In: Cronin, B. (Ed.). *Annual Review of Information Science and Technology* 2003; 37(1):179-255.
4. Castaño AC. *Populando ontologias através de informações em HTML*: o caso do currículo Lattes. [Master's thesis]. São Paulo: Universidade de São Paulo; 2008.
5. Cormen TH, Leiserson CE, Rivest RL and Stein C. *Introduction to algorithms*. 2 ed. Cambridge: MIT Press; 2001.
6. Costa LF, Rodrigues FA, Travieso G and Villas Boas PR. Characterization of complex networks: a survey of measurements. *Advances in Physics* 2007; 56(1):167-242.
7. Day MY, Tsai TH, Sung CL, Lee CW, Wu SH, Ong CS et al. A knowledge-based approach to citation extraction. In: Zhang D, Khoshgoftaar TM and Shyu ML. (Eds.). *Proceedings of the International Conference on Information Reuse and Integration*; 2005; Las Vegas Hilton. Las Vegas: IEEE Systems, Man, and Cybernetics Society; 2005. p. 50-55.
8. Duda RO, Hart PE and Stork DG. *Pattern classification*. 2 ed. New York: John Wiley & Sons; 2000.
9. Han H, Zha H and Giles CL. Name disambiguation in author citations using a K-way spectral clustering method. In: *Proceedings of the 5 ACM/IEEE-CS Joint Conference on Digital Libraries, Tools & techniques: identifying names of people and places*; 2005; Denver. Canada: ACM; 2005. p. 334-343.
10. Hey T, Tansley S and Tolle K. (Eds.). *The fourth paradigm*. Redmond, Washington: Microsoft Research; 2009.
11. The Digging into Data Challenge. 2009. Available from: http://www.diggingintodata.org/. Access in: 20/10/2009.
12. Communications of the ACM: Surviving the data deluge 2008; 51(12). New York, NY, USA: ACM; 2008.
13. Jolliffe IT. *Principal component analysis*. 2 ed. New York: Springer-Verlag; 2002. (Series in statistics)
14. Koren Y, North SC and Volinsky C. Measuring and extracting proximity in networks. In: Eliassi Rad T, Ungar LH, Craven M and Gunopulos D. (Eds.). *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2006; Philadelphia. Philadelphia: ACM; 2006. p. 245- 255.
15. Kouzes RT, Anderson GA, Elbert ST, Gorton I and Gracio DK. The changing paradigm of dataintensive computing. *Computer* 2009; 42(1):26-34.
16. Laender AHF, Lucena CJP, Maldonado JC, Souza e Silva E and Ziviani N. Assessing the research and education quality of the top Brazilian Computer Science graduate programs. *ACM SIGCSE Bulletin* 2008; 40(2):135-145.
17. Liu X, Bollen J, Nelson ML and Van de Sompel H. Co-authorship networks in the digital library research community. *Informations Processing and Management* 2005; 41(6):1462-1480.
18. Project Zoomable Visual Transformation Machine. 2009. Available from: http://zvtm.sourceforge.net/. Access in: 20/10/2009.
19. Maia MF and Caregnato SE. Co-autoria como indicador de redes de colaboração científica. *Perspectivas em Ciência da Informação* 2008; 13(2):18-31.
20. Nascimento MA, Sander J and Pound J. Analysis of SIGMOD's co-authorship graph. *SIGMOD Record* 2003; 32(3):8-10.
21. Newman MEJ and Girvan M. Finding and evaluating community structure in networks. *Physical Review E* 2004; 69(2):026113.
22. Nicholson S. The basis for bibliomining: frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services. *Informations Processing and Management* 2006; 42(3):785-804.
23. University of São Paulo - USP. *Publications of the Department of Computer Science*. São Paulo, 2009. Available from: http://www.vision.ime.usp.br/creativision/publications_dcc/. Access in: 20/10/2009.
24. Vision Research Group - IME - USP. *Publications of the Vision Research Group*. São Paulo: University of São Paulo, 2009. Available from: http://www.vision.ime.usp.br/creativision/publications_vision/. Access in: 20/10/2009.
25. Pacheco RCS and Kern VM. Uma ontologia comum para a integração de bases de informações e conhecimento sobre ciência e tecnologia. *Ciência da Informação* 2001; 30(3):56-63.
26. Paulovich FV, Nonato LG, Minghim R and Levkowitz H. Least square projection: a fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics* 2008; 14(3):564-575.
27. Peng F and McCallum A. Information extraction from research papers using conditional random fields. *Informations Processing and Management* 2006; 42(4):963-979.
28. Said YH, Wegman EJ, Sharabati WK and Rigsby JT. Social networks of author-coauthor relationships. *Computational Statistics & Data Analysis* 2008; 52(4):2177-2184.
29. Project scriptLattes. *scriptLattes*: uma ferramenta para extração e visualização de conhecimento a partir de Currículos Lattes. São Paulo: Universidade de São Paulo, 2009. Available from: http://scriptlattes.sourceforge.net. Access in: 20/10/2009.
30. Sobral FAF, Almeida MRC and Caixeta MVG. As lideranças científicas. *Ciências & Cognição* 2008; 13(2):179-191.