Hierarchical Clustering Based on Reports Generated by Scriptlattes

Wonder Alves, Saulo D. Santos, Pedro Schimit

Hierarchical Clustering Based on Reports Generated by Scriptlattes

Wonder A. L. Alves, Saulo D. Santos and Pedro H. T. Schimit

Universidade Nove de Julho, São Paulo, Brazil
{wonder,saulods,schimit}@uni9.pro.br

Abstract. Scriptlattes has been used as an important tool to analyze a curriculum database of Brazilian researchers (Lattes curriculum). Such analysis enables a user generating reports for specific research field, knowledge area, graduate courses, and so forth. However, when users need a report of a Graduate Program, for instance, it is necessary to create subsets of information in order to run the script. Since each report needs a subset, in this paper we propose a hierarchical clustering method to categorize reports generated by Scriptlattes. Finally, experimental results show hierarchical clustering of a higher education institution, and approaches to stress such clustering.

Keywords. Lattes curriculum· Sucupira platform· Scriptlattes· Clustering· Hierarchical clustering.

1 Introduction

The growth of scientific community in the last decades created new issues to quantify the production, quality and innovation of research results. Besides the evaluation of researchers, institutes and countries concerning the research that has been done, keep a robust database of the activities and production is becoming crucial to manage the research properly. In this direction, several computational tools specific to the scientific community have been developed, including: social networks [1], institutional [2] and patents [3] repositories, curriculum platforms [4] and others [5] [6].

In Brazil, it is practically mandatory that all researchers keep their Lattes curriculum updated. Created with effort from the National Counsel of Technological and Scientific Development (CNPq - *Conselho Nacional de Desenvolvimento Científico e Tecnológico*), Lattes platform allows that every curriculum is available online to public consultation [7,8]. Therefore, if a researcher wants to apply to a position in an University, sure his Lattes will be consulted previously by the University commission. According to CNPq, the Lattes Platform exceeds 3 million Lattes curriculum, being 6.35% of PhDs, 10.85% of masters 27.69% of graduates, 16.18% of specialists, 35.54% other levels and 3.39% of Lattes curriculum that do not have the information [9].

Although the public information is individually available for each researcher, automatic compilation of data to generate reports concerning scientific production of a certain research group is not an easy task, and it is not provided by

Lattes plataform. Regarding this problem, ScriptLattes has been created as an open source tool to make the information extraction easier. Given a group of researchers registered in the Lattes platform, ScriptLattes downloads the curriculums in html format, extracts the information of interest, eliminates redundancies and create reports [8]. It is important to remember that the Scriptlattes has been widely used by various studies in Brazil and the results so far have been of great value as can seen in [10,11,12,13,14,15,16,17,18].

Informations that can be collected by using ScriptLattes are, for instance, statistic descriptions and collaboration graphs. This collection of information of curriculums is performed globally for all dataset and then the Scriptlattes displays the report produced by HTML pages containing links between its pages. Further, Scriptlattes user can be interested in getting reports with different views of dataset, such as: specific research line, course, knowledge area and so forth. Each report needs a subset of the dataset (eg, segmenting the curriculum by some criteria) and a execution of the subset.

From this limitation of Scriptlattes, this paper proposes a method that use Scriptlattes functionality to categorize reports generated by Scriptlattes. Precisely, this categorization is modeled by elements of graph theory in which is constructed a partially ordered set that induces a hierarchy based on a categorization of the data. Thus, the nodes represent subsets of curriculum grouped by some criteria that associates them. Finally, running a Scriptlattes for each node, user has different views of the dataset with its own settings (subsets of curriculum and customizations in general).

This paper is organized as follows: Section **??** has definitions and properties about hierarchical clustering. In Section **??**, the proposed method is described. Results and applications of the hierarchical clustering of a higher education institution are in Section **??**. Finally, Section 5 has conclusions and future works.

2 Theoretical Background

Consider Cv a set of Lattes curriculum. In machine learning, we say that $\mathbb{P}$ is a clustering (partition) on the dataset Cv if and only if $\mathbb{P}$ containing n groups (subsets) $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_n$ of Cv such that the following conditions hold:

1. The clustering $\mathbb{P}$ does not contain the empty set, i.e.,

$$\emptyset \notin \mathbb{P}. \tag{1}$$

2. The union of the groups in $\mathbb{P}$ is equal to Cv, i.e.,

$$\mathrm{Cv} = \bigcup_{\mathcal{S}_i \in \mathbb{P}} \mathcal{S}_i. \tag{2}$$

3. The intersection of any two distinct groups in $\mathbb{P}$ is empty.

$$\forall \mathcal{S}_i, \mathcal{S}_j \in \mathbb{P}, \mathcal{S}_i \neq \mathcal{S}_j \Rightarrow \mathcal{S}_i \cap \mathcal{S}_j = \emptyset. \tag{3}$$

In this paper, we denote that $\mathbb{P}(C)$ is the group $\mathcal{S}_i \in \mathbb{P}$ containing the Lattes curriculum C, i.e., $\mathbb{P}(C) = \mathcal{S}_i$ if and only if $C \in \mathcal{S}_i$. Thus, we define a binary relation *finer than* on the set of clusterings of Cv, as follows: for any two clustering $\mathbb{P}_i$ and $\mathbb{P}_j$ on Cv, we say that $\mathbb{P}_i$ is finer than $\mathbb{P}_j$ (and that $\mathbb{P}_j$ is coarser than $\mathbb{P}_i$) if and only if every group of $\mathbb{P}_i$ is a subset of some group of $\mathbb{P}_j$, i.e.,

$$\forall C \in \mathrm{Cv}, \mathbb{P}_i(C) \subseteq \mathbb{P}_j(C). \tag{4}$$

In this case, we say that $\mathbb{P}_i$ is a refinement of a clustering $\mathbb{P}_j$ and written as $\mathbb{P}_i \preceq \mathbb{P}_j$. More details see in [19,20].

This relation $\preceq$ on a subset $\mathcal{T}$ of clusterings of Cv constitutes a partially ordered set (poset) and thus, we can run a scriptlattes for each clustering $\mathcal{T}_i \in \mathcal{T}$ and presents them in a manner categorized by a hierarchy based on the Hasse diagram of poset $(\mathcal{T}, \preceq)$. To illustrate this idea, consider the following example: let $\mathbb{P}_{\text{Area}}$, $\mathbb{P}_{\text{Course}}$ and $\mathbb{P}_{\text{ResearchLine}}$ partitions of Cv clustered by knowledge area, course and research line, respectively, in such a way to satisfy: $\mathrm{Cv} \preceq \mathbb{P}_{\text{ResearchLine}} \preceq \mathbb{P}_{\text{Course}} \preceq \mathbb{P}_{\text{Area}} \preceq \{\mathrm{Cv}\}$. Thus, we have:

- $\mathcal{S}_i \in \mathbb{P}_{\text{Area}}$ is the group of Cv belonging the knowledge area i;
- $\mathcal{S}_{ij} \in \mathbb{P}_{\text{Course}}$ is the group of $\mathcal{S}_i \subseteq \mathrm{Cv}$ belonging the course j of the knowledge area i;
- $\mathcal{S}_{ijk} \in \mathbb{P}_{\text{ResearchLine}}$ is the group of $\mathcal{S}_{ij} \subseteq \mathcal{S}_i \subseteq \mathrm{Cv}$ belonging the research line k of the course j of the knowledge area i.

Therefore, $\mathcal{T} = \mathbb{P}_{\text{Area}} \cup \mathbb{P}_{\text{Course}} \cup \mathbb{P}_{\text{ResearchLine}} \cup \mathrm{Cv}$. Fig. 1 shows part of a branch of the Hasse diagram $(\mathcal{T}, \preceq)$.

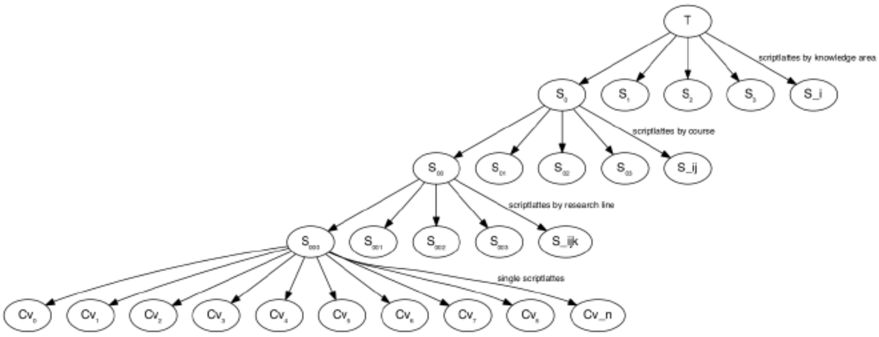

Fig. 1: Example of part of a branch of the Hasse diagram $(\mathcal{T}, \preceq)$.

To build the hierarchy of clusterings $\mathcal{T}$ is employed the technique traditional of machine learning, called hierarchical clustering [21,22], described in Section 2.1.

2.1 Hierarchical Clustering

Hierarchy of clustering can offer more information about the structure of the curriculums in the dataset. With a hierarchy, the cluster of curriculums can be seen at different levels, i.e., from the bottom level where each curriculum forms an independent cluster (singleton clusters) to the top level with only one cluster containing all the curriculums. The hierarchical clustering involves creating clusterings that have a predetermined ordering from top to bottom. There are two traditional approach for construction hierarchical clustering [21,22], divisive and agglomerative:

- *Divisive*: In this method is assigned all of the curriculums to a single cluster and then partition the cluster to two least similar clusters. Finally, we proceed recursively on each cluster until there is one cluster for each curriculums.
- *Agglomerative*: Initially, the set of all curriculum is considered to be a single cluster and then, recursively, is calculated the similarity (e.g., distance) between each of the clusterings and join the two most similar clusterings.

In order to build the Hierarchy of clustering, it is required to determine the (dis)similarity between each pair of clusterings using a distance function. The main methods used to measure the distance between the clusters are [21,22]:

1. *Single Linkage*: the distance between two clusters is defined as the shortest distance between two curriculum in each cluster, i.e.,

$$\forall \mathcal{T}_i, \mathcal{T}_j \in \mathcal{T}, D_{min}(\mathcal{T}_i, \mathcal{T}_j) = \min\{dist(C_i, C_j) : C_i \in \mathcal{T}_i, C_j \in \mathcal{T}_j\}. \quad (5)$$

2. *Complete Linkage*: the distance between two clusters is defined as the longest distance between two curriculum in each cluster, i.e.,

$$\forall \mathcal{T}_i, \mathcal{T}_j \in \mathcal{T}, D_{max}(\mathcal{T}_i, \mathcal{T}_j) = \max\{dist(C_i, C_j) : C_i \in \mathcal{T}_i, C_j \in \mathcal{T}_j\}. \quad (6)$$

3. *Average Linkage*: the distance between two clusters is defined as the average distance between each curriculum in one cluster to every curriculums in the other cluster.

$$\forall \mathcal{T}_i, \mathcal{T}_j \in \mathcal{T}, D_{avg}(\mathcal{T}_i, \mathcal{T}_j) = \frac{1}{|\mathcal{T}_i| \times |\mathcal{T}_j|} \sum_{C_i \in \mathcal{T}_i} \sum_{C_j \in \mathcal{T}_j} dist(C_i, C_j). \quad (7)$$

Note that, the distance functions between clusterings need of a distance function between curriculums that is defined in Section **??**.

3 Proposed Method

Hierarchical clustering allows the creation of a cluster tree, as mentioned previously. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. This allows adjustment of the grouping level most suitable for our pretensions. Therefore, it is necessary to build a distance function between the curriculum, that is used in the kernel of hierarchical clustering algorithms.

Consider $\mathrm{Inf}_{\mathrm{Field}}(C)$ the function that represents the value associated in a given field on the Lattes curriculum $C \in \mathrm{Cv}$. For example, $\mathrm{Inf}_{\mathrm{Area}}(C)$ is the value associated the knowledge area of curriculum C. Thus, consider $\mathcal{F}$ a list of associated functions on the Lattes curriculums. Note that, $\mathcal{F}$ are informations that are associated in each row of dataset, which is built by using a file where each row contains: personal ID, name, list of associated informations. Table 1 has the structure of the dataset.

Table 1: Structure of the dataset

Dataset				
Personal ID	**Name**	**Inf$_{\mathbf{Area}}$**	**Inf$_{\mathbf{Course}}$**	**Inf$_{...}$**
Lattes ID of the person 1	person 1	ID of area 1	ID of course 1	...
Lattes ID of the person 2	person 2	ID of area 1	ID of course 1	...
Lattes ID of the person 3	person 3	ID of area 1	ID of course 2	...
Lattes ID of the person 4	person 4	ID of area 1	ID of course 2	...
Lattes ID of the person 5	person 5	ID of area 2	ID of course 3	...
Lattes ID of the person 6	person 6	ID of area 2	ID of course 4	...
Lattes ID of the person 7	person 7	ID of area 2	ID of course 4	...
Lattes ID of the person 8	person 8	ID of area 2	ID of course 4	...
Lattes ID of the person ...	person ...	ID of area ...	ID of course ...	...

The distance function between curriculum is defined as following:

$$\forall C_i, C_j \in \mathrm{Cv}, dist(C_i, C_j) = \sum_{k \in \mathcal{F}} \omega_k |\mathrm{Inf}_{\mathrm{k}}(C_i) - \mathrm{Inf}_{\mathrm{k}}(C_j)|, \tag{8}$$

where ω_k is a weight associated the function $\mathrm{Inf}_{\mathrm{k}} \in \mathcal{F}$.

For example, the hierarchy of clusterings shown in Fig.1 can be built using single linkage as the distance function between clusterings and the functions associated $\mathrm{Inf}_{\mathrm{Area}}$, $\mathrm{Inf}_{\mathrm{Course}}$ and $\mathrm{Inf}_{\mathrm{ResearchLine}}$ also used to define the following distance function between curriculums:

$$\begin{aligned} \forall C_i, C_j \in \mathrm{Cv}, dist(C_i, C_j) = {} & \omega_{\mathrm{ResearchLine}} |\mathrm{Inf}_{\mathrm{ResearchLine}}(C_i) - \mathrm{Inf}_{\mathrm{ResearchLine}}(C_j)| + \\ & \omega_{\mathrm{Course}} |\mathrm{Inf}_{\mathrm{Course}}(C_i) - \mathrm{Inf}_{\mathrm{Course}}(C_j)| + \\ & \omega_{\mathrm{Area}} |\mathrm{Inf}_{\mathrm{Area}}(C_i) - \mathrm{Inf}_{\mathrm{Area}}(C_j)|, \end{aligned} \tag{9}$$

where $\omega_{\mathrm{ResearchLine}} = 1$, $\omega_{\mathrm{Course}} = \omega_{\mathrm{ResearchLine}} + \max\{\mathrm{Inf}_{\mathrm{ResearchLine}}(C) : C \in \mathrm{Cv}\}$ and $\omega_{\mathrm{Area}} = \omega_{\mathrm{Course}} + \max\{\mathrm{Inf}_{\mathrm{Course}}(C) : C \in \mathrm{Cv}\}$.

Once the hierarchy of clusterings $\mathcal{T}$ is build, we can visit the clusters $\mathcal{T}_i \in \mathcal{T}$ (see Fig. 2) and associate each cluster $\mathcal{S}_k \in \mathcal{T}_i$ the result of Scriptlattes. Moreover, depending on the cluster height in the hierarchy we can provide different settings for Scriptlattes, allowing different treatment to clusterings by knowledge areas, courses, research lines and single curriculum.

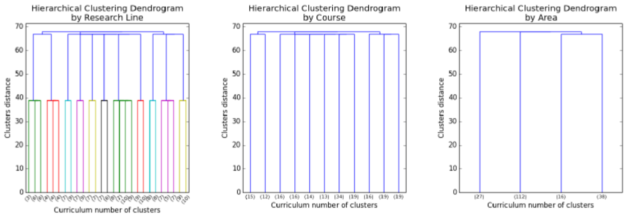

Fig. 2: Hierarchical clustering dendrograms

A implementation in Python language for constructed the hierarchy of clusterings $\mathcal{T}$ can be acessed at link: https://goo.gl/9v6Wvw

4 Application: Data Extraction for Sucupira Platform

The Sucupira platform is a management tool used by the Coordination for the Improvement of Higher Education Personnel (CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). By using this platform, programs report their performance to the coordination annually and, at the end of each four-year period, while CAPES verify if they show minimum quality required for maintenance. Data is imported directly from the Lattes platform, and it refers to the scientific production of each of the professors belonging to the assessed course. After the import, we need to perform a manual checking, since many information regarding the registration of the professors data as well as the scientific productivity may still be incomplete [13].

For this purpose, we construct a sequence of hierarchies ($\mathcal{T}^{2013}$, $\mathcal{T}^{2014}$, $\mathcal{T}^{2015}$, $\mathcal{T}^{2016}$, $\mathcal{T}^{2013-2016}$), where each element of this sequence contains a hierarchy of clusterings configured to execute scriptlattes for a specific period. Moreover, the clusters formed by curriculums of a same course are configured with the Qualis of knowledge area. Once the hierarchies of clusterings ($\mathcal{T}^{2013}$, $\mathcal{T}^{2014}$, $\mathcal{T}^{2015}$, $\mathcal{T}^{2016}$, $\mathcal{T}^{2013-2016}$) are build, we can visit each hierarchy of clusterings $\mathcal{T}^{\text{Period}}$ and run a Scriptlattes for each level of them in $\mathcal{T}^{\text{Period}}$. Finally, the reports produced by Scriptlattes throughout hierarchies of clusterings are organized into a web page that can be accessed at link: https://db.tt/5JnRPH7e

Note that it is possible to make a thorough checking of the information to be submitted to Sucupira platform by using the hierarchies of clusterings based on reports generated by scriptlattes. It is worth remembering that the preliminary check of scientific productivity of each professor permits the identification of those who may have not achieved the goals set by the program, determined by

CAPES. With this result, managers can improve research strategies in order to improve their production during the period of quadrennial, avoiding any unfortunate surprises when assessing the course. A further hierarchy of clustering without period setting can be seen in this link: https://db.tt/ZOxTXHiP

5 Conclusion and Future Works

In this paper, we proposed a method that makes use of the Scriptlattes to build a hierarchy of clusterings based on reports generated by scriptlattes. In this hierarchy, the nodes represent subsets of curriculum grouped by a criteria that associates them and Scriptlattes is executed for each node. The hierarchy of clusterings built provides more information about the structure of the curriculums in the dataset. Thereby, the cluster of curriculums can be seen at different levels, i.e., from the bottom level where each curriculum forms an independent cluster (singleton clusters) to the top level with only one cluster containing all the curriculums. It allows the Scriptlattes user to access different views on the reports generated from the dataset. Moreover, it has been presented a example of application for data extraction for Sucupira platform, which used the proposed hierarchy approach. As a future work, we intend to study the similarities between clustering on different hierarchies, and analyze the evolution of collaboration graph along the hierarchy of clusterings.

References

1. Kadriu, A.: Discovering Value in Academic Social Networks: A Case Study in ResearchGate. In: Proceedings of the ITI 2013 35th International Conference on Information Technology Interfaces, IEEE (2013) 57–62
2. Lynch, C.A.: Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. Libraries and the Academy **3** (2003) 327–336
3. Ernst, H.: Patent Information for Strategic Technology Management. World Patent Information **25** (2003) 233 – 242
4. Fernández-Breis, J.T., Castellanos-Nieves, D., Hernández-Franco, J., Soler-Segovia, C., del Carmen Robles-Redondo, M., González-Martínez, R., Prendes-Espinosa, M.P.: A Semantic Platform for the Management of the Educative Curriculum. Expert Systems with Applications **39** (2012) 6011–6019
5. Burns, C.S., Lana, A., Budd, J.: Institutional Repositories: Exploration of Costs and Value. D-Lib Magazine **19** (2013) 1
6. Edgar, B.D., Willinsky, J.: A Survey of Scholarly Journals using Open Journal Systems. Scholarly and Research Communication **1** (2010)
7. Lane, J.: Let's Make Science Metrics more Scientific. Nature **464** (2010) 488–489
8. Mena-Chalco, J.P., Junior, R.M.C.: ScriptLattes: An Open-source Knowledge Extraction System from the Lattes Platform. Journal of the Brazilian Computer Society **15** (2009) 31–39
9. (CNPq)
10. Ferraz, R.R.N., Quoniam, L.: A Utilização da Ferramenta Computacional Scriptlattes para Avaliação das Competências em Pesquisa no Brasil. Revista Prisma. Com (2014)

11. Ferraz, R.R.N., Quoniam, L.M., Maccari, E.A.: The use of Scriptlattes Tool for Extraction and On-line Availability of Academic Production from a Department of Stricto Sensu in Management. In: 11th International Conference on Information Systems and Technology Management–CONTECSI. Volume 17. (2014)
12. Giordano, D.M., Bruning, E., Bordin, A.S.: Uso do Scriptlattes e Gephi na Análise da Colaboração Científica. Anais do Computer on the Beach (2015) 239–248
13. Nigro, C.A., Ferraz, R.R.N., Quoniam, L., Alves, W.A.L.: Strategic Management of Research Productivity from Graduate Medicine Program by the use of Scriptsucupira Computational Tool. In: Proceedings of the 13th International Conference on Information Systems and Technology Management. (2016)
14. Mena-Chalco, J., Cesar-Jr, R.: Bibliometria e Cientometria: Reflexões Teóricas e Interfaces. Pedro & João Editores, São Carlos (2013)
15. Mena-Chalco, J.P., Digiampietri, L.A., Lopes, F.M., Cesar, R.M.: Brazilian Bibliometric Coauthorship Networks. Journal of the Association for Information Science and Technology **65** (2014) 1424–1445
16. Mena-Chalco, J.P., Junior, R.M.C.: Towards Automatic Discovery of co-authorship Networks in the Brazilian Academic Areas. In: IEEE Seventh International Conference on e-Science Workshops (eScienceW). (2011) 53–60
17. Perez-Cervantes, E., Mena-Chalco, J.P., Cesar-Jr, R.M.: Towards a Quantitative Academic Internationalization Assessment of Brazilian Research Groups. In: E-Science (e-Science), 2012 IEEE 8th International Conference on. (2012) 1–8
18. Perez-Cervantes, E., Mena-Chalco, J.P., Oliveira, M.C.F.D., Cesar, R.M.: Using link prediction to estimate the collaborative influence of researchers. In: 2013 IEEE 9th International Conference on eScience (eScience). (2013) 293–300
19. Brualdi, R.: Introductory Combinatorics. Pearson Education. Pearson Education International (2012)
20. Newman, M.H.A.: Elements of the Topology of Plane sets of Points. Dover Publications (1992)
21. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons (2012)
22. Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning. Volume 1. Springer series in statistics Springer, Berlin (2001)