Visual representation of bibliographic production data from Lattes Platform

Patrícia Salles Escarassatti
University of São Paulo
Helton H. Bíscaro (heltonhb@usp.br)
University of São Paulo

Research Article

Keywords: curriculum Lattes, multidimensional projection, data visualization

Posted Date: May 24th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1665862/v1

Visual representation of bibliographic production data from Lattes Platform

Patrícia Salles Escarassatti[1] and Dr Helton H. Bíscaro[1]

[1]School of Arts, Sciences and Humanities, University of São Paulo, Arlindo Béttio, 1000, São Paulo, 03828-000, São Paulo, Brazil.
*Corresponding author: Dr Helton H. Bíscaro.

Contributing authors: patricia.salles@usp.br; heltonhb@usp.br;

Abstract

The Brazilian scientific community has available a curricular information system called Lattes. The analysis in the area of scientific production is rapidly evolving as we have an increase in the number of publications combined with increasingly interdisciplinary sources of information. The Brazilian scientific community has available a database with curricular information called Lattes which contain information about academic and scientific achievements. There is a demand to carry out the search and retrieval of information from the Lattes curriculum in order to reduce the difficulties in exploring bibliometric data and to aid in the analysis of this information. This article uses textual data mining techniques, multidimensional projection techniques and visualization based on analysis of bibliometric data extracted from the Lattes platform. The results obtained were clusters of research groups that are related to each other and clusters of research groups that work with very different themes. These results can become a tool for visually exploring bibliometric data from the Lattes platform.

Keywords: curriculum Lattes, multidimensional projection, data visualization

1 Introduction

With the rapid development of technology and more accessible devices, electronic documents such as news and scientific articles have continuously spread.

This explosion of electronic documents has made it difficult for a user to extract useful information from them (Aliguliyev, 2009).

Moreover, contemporary advances in the technological area have raised the need for management of organizational knowledge dispersed in various information sources. A key challenge for knowledge management systems is the effective discovery and use of content stored in these information sources (Abbas et al, 2014).

The monitoring of the areas of research, especially when there are multiple sources of interdisciplinary information, requires a substantial effort by researchers and research programs to carry out the analysis in the area of scientific production. Present-day scholars and scientists devote substantial effort to keeping up with advances in their fields of activity. The growing number of publications, combined with increasingly interdisciplinary sources, makes it challenging to keep up with emerging research fronts and identify key works. It's even harder to start exploring a new field without an initial reference (Dunne et al, 2012).

It is necessary to analyze production data to understand what they represent, bibliometry is the process responsible for extracting and analyzing this data. Essentially, bibliometry is a set of techniques that extracts data from publications and analyzes that data in many ways to answer the research questions that these publications represent (Belter, 2015).

In this regard, bibliometry contributes to the progress of science because it allows us to discover information in many different ways: evaluating the progress to be made, identifying the most reliable sources of scientific publication, establishing the academic basis for evaluating new ventures, identifying key scientific actors, developing bibliometric indices to evaluate academic output and so on (M. Gutiérrez-Salcedo, 2018).

There are three main types of bibliometric indicators: quantitative indicators that measure the productivity of a given researcher or research group. Quality indicators that measure the quality of a journal, researcher journal, researcher or research group. Structural indicators that measure connections between publications, authors or research groups (Durieux and Gevenois, 2010).

The quantitative indicators are intended to measure the productivity of a researcher or a research group. The simplest method of accounting is to count the number of articles published by a particular author or research group over a certain period of time. However, the number of publications does not reflect the productivity of the author or research group, and does not reflect the quality of the articles (Durieux and Gevenois, 2010).

Performance indicators help to identify the level of quality of the work of an author or research group and can be used to assess the impact of research on the scientific community. The frequency that an article, author, or journal is cited by others is an example of a performance indicator (Durieux and Gevenois, 2010).

Structural indicators deal with understanding social, intellectual and conceptual structure by means of bibliographic networks. Bibliometric mapping is a spatial representation of how disciplines, fields, specialties and documents, authors or research groups are related to each other (M. Gutiérrez-Salcedo, 2018).

The Brazilian scientific community has available a curricular information system called Lattes managed by CNPq. For this reason, the "Lattes Curriculum" is considered a national standard of information about scientific and academic achievements of students, professors, researchers and professionals involved in science and technology in general (Mena-Chalco and Junior, 2009).

The Lattes curriculum is a rich and powerful database that presents numerous potential applications (scientific, technological, economic, etc.) The Lattes curriculum displays information only in an individual way, in other words, the registered information is individually associated with each person. This characteristic does not easily provide a way to discover the bibliographic, technical or artistic productions of a given group, such as a research group, professors from an academic department or members of a Brazilian institution (Mena-Chalco and Junior, 2009).

The curriculum data are structured in a hierarchical manner. There are two modules available for filling in data (General Data module and Production module) that are used for filling in the researcher's data. The General Data module is divided into personal data, professional data, and other relevant information. In the Production module, the user registers or updates his bibliographical, technical and artistic/cultural production, as well as his completed research (Amorin, 2003).

Lattes is a tool that can be used to generate information about the research groups, institutions and authors who publish the most on a given subject. Most Brazilian academic institutions usually explore the Lattes curricula in order to produce reports on scientific productions, supervisions and projects of research groups related to these institutions. The reports are usually created by manual analysis of the Lattes data of each group member, in order to obtain a complete summary of all the 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, and highly susceptible to errors caused by manual processing (Mena-Chalco and Junior, 2009).

In specific domains, such as scientific articles and patents, there is a demand for efficient methods of searching for and retrieving information. Researchers have devoted efforts to proposing tools to deal with document analysis using visual techniques and thus creating a visualization field known as visual text mining (Eler and Garcia, 2013).

Tools for the rapid exploration of literature can help reduce these difficulties, providing concise overviews tailored to researchers needs and assisting in the generation of research. Digital libraries and search engines are useful for finding specific documents that match a search string, but do not provide the additional analysis tools needed to quickly summarize a field. Users unfamiliar

with the field of research often find it challenging to look for influential people or innovative articles, authors and journals (Dunne et al, 2012).

Thus, text mining, a technique of deriving high-quality knowledge from text, has recently attracted great attention in the research community. Search text mining topics include language identification, document grouping, summarization, text indexing, and visualization. In particular, text visualization refers to the technology that displays text data or mining results in a logical layout (for example, color graphics) so that you can view and analyze documents easily and intuitively. This presents a direct way to look at documents as well as understand the relationship between them (Chen et al, 2009).

Visual exploration of multidimensional data has become a common task in recent years and is needed to address the complexities of interpretation of large multidimensional datasets. In order to make visual exploration feasible, different approaches to visualizing information have been developed to deal with multidimensionality. These approaches are known as multidimensional data visualization techniques (Tejada et al, 2003).

Visual analysis can be performed by using information visualization techniques. Multidimensional projections are examples of these techniques, in which the original dimensions are projected into a lower dimensional (usually two-dimensional) space, and the instances are then displayed in scatter plots. This mapping process can lead to information loss, and different strategies can be applied to create the projection, but they preserve certain properties of the data distribution (Etemadpour et al, 2014a).

This work intends to use visualization techniques based on the multidimensional analysis of the data that will be extracted from the Lattes platform. In order to verify whether multidimensional projection visualization techniques can help analysis of bibliometric data of the Lattes platform, verifying the existence of patterns and the general distribution of the data.

2 Background

This section is a background about preprocessing text and visualizing multidimensional data.

2.1 Preprocessing

The preprocessing step should be performed with the aim of reducing the number of terms in the text, selecting the relevant words and structuring the data to facilitate the processing of the algorithms and generating a good visual representation of the document collection. The list below presents the preprocessing operations that are typically applied for creating a vector space model:

- **Tokenization:** The tokenization process consists of representing each word of the text in distinct units, called tokens. This way each word in the document will be represented as a token (Abasi et al, 2020).

- **Removal of stopwords:** There are words considered irrelevant to the language and that have a high frequency. It is extremely important to remove these words due to the high volume that negatively affects the clustering of textual documents, as well as making the clustering of documents more time-consuming (Abasi et al, 2020).
- **Stemming:** is a technique applied to remove prefixes and suffixes from words, this way these words will be represented by their root, for example, eating, eats, eaten, the root will be the term "eat", which will be used as a feature of the document collection (Abasi et al, 2020).
- **Zipf's Law:** Zipf's law is based on the frequency of terms that occur in many documents and therefore do not help distinguish them. The Zipf curve is designed by considering the ordering of the frequency of terms in a decreasing way and from it we can define a threshold to exclude these less significant features. In the graph of Figure 1 the horizontal axis, named $Words$, represents the terms in decreasing frequency order and the vertical axis, here called $Frequency$, represents the frequency of these terms.
- **Luhn's Cut:** From the Zipf curve, Luhn specified two thresholds to remove unrepresentative terms. The C line in Figure 1 represents one of these cuts, words on the left would be considered inappropriate, because they are the most common terms to appear in any type of document. Since the degree of frequency was proposed as a criterion, a lower bound, line D, would also be established and the terms below that cut would be considered rare and thus would not indicate significance in discriminating different documents (Luhn, 1958). To define these thresholds, Goffman established a procedure to eliminate the least relevant terms from the document. Thus the transition zone is given by the point where the word count has a frequency close to one. Therefore, finding the Goffman's point given by n would serve us as the transition point (Sequera et al, 2009). The calculation of n is done as follows:

$$n = \frac{-1 + \sqrt{1 + 8 \times I_1}}{2} \tag{1}$$

where:
I_i - number of words with frequency equal to 1.
The goal is to obtain a transition area that provides the greatest number of successes from Goffman's point.

2.2 Vector space model

In a vector space model, each document is represented by a word vector. The vector space model constructs a document vector d_j consisting of all distinct terms within the document collection.

For each term, represented as k_i in the document vector, the weights of that term will be calculated. The calculation of these weights considers two basic components, the first is the occurrence number of a term in a given document d_j and the second is the number of documents that contain this term (Kalmukov, 2020).

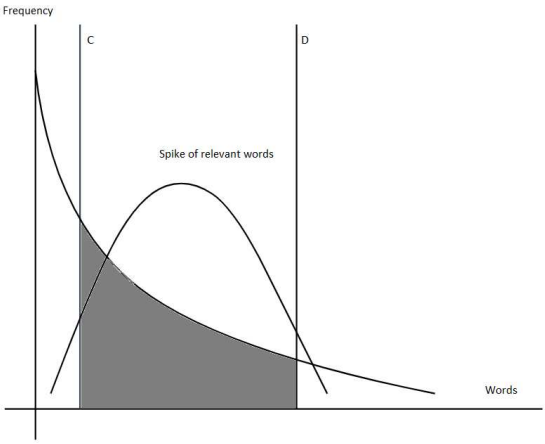


Fig. 1: Luhn's Cut

The weight $w_{j,i}$ is a positive real value associated with the pair (k_i, d_j). In the vector space model, the frequency of the term k_i within the document d_j is given as tf and provides a measure of how important this term is to the document. The factor df represents the number of documents that contain k_i, the more documents contain a term, the more common and less informative it will be. This is an inverse measure of informativity. But the term weighting models consider not the frequency of document df, but the inverse of the document frequency, idf - the fewer documents contain a term, the more informative it will be. Idf is calculated as follows (Abasi et al, 2020):

$$idf_i = log\frac{d}{df_i} \tag{2}$$

where:

idf_i - inverse document frequency of term k_i.

d - number of documents in the entire collection.

df_i - number of documents that contain the term k_i.

Intuitively, the most basic weighting scheme is the combination of these two terms:

$$w_{j,i} = tf_{j,i} \times idf_i = tf_{j,i} \times log\frac{d}{df_i} \tag{3}$$

Thus, the vector of a document d_j is represented by $\vec{d_j} = (\mathrm{w}_{j,1}, w_{j,2}, ..., w_{j,n})$. The vector space model represents the documents as a matrix mxn as follows:

$$\begin{pmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,(n-1)} & w_{1,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \dots & \dots & \dots & \dots & \dots \\ w_{(m-1),1} & \dots & \dots & \dots & w_{(m-1),n} \\ w_{m,1} & w_{m,2} & \dots & \dots & w_{m,n} \end{pmatrix}$$

2.3 Multidimensional projection techniques

Information visualization techniques generate a graphical representation in a low-dimensional space, generating a very good representation of multidimensional data and can be produced efficiently (Chalmers, 1996).

Below we will introduce different multidimensional projection techniques that help make data analysis as visual as possible.

2.3.1 Multidimensional scaling (MDS)

Multidimensional scaling (MDS) is one of the dimensionality reduction techniques that converts multidimensional data into a lower dimensional space, keeping the intrinsic information of the data. The main reason to use MDS is to get a graphical display for the data provided, so that it is much easier to understand. The only assumption of the MDS is that the number of dimensions must be one less than the number of points, which also means that at least three variables must be inserted into the model and at least two dimensions must be specified (Saeed et al, 2018).

MDS is the approach that maps the original high-dimensional data (m dimensions) into a lower dimensional space (d dimensions). It addresses the problem of constructing a configuration between the n points from a $k\,x\,k$ matrix $\mathbf{D}$, which is called an affinity matrix. MDS finds n data points $y_1, ...y_n$ from a matrix of distance $\mathbf{D}$ in a space of dimension d, so if $\hat{d_{ij}}$ is the Euclidean distance between y_i and y_j, then $\hat{\mathbf{D}}$ is similar to $\mathbf{D}$. The Multidimensional scaling can be considered as (Saeed et al, 2018):

$$\min_{Y} \sum_{i=1}^{k} \sum_{i=1}^{k} (d_{ij}^{X} - d_{ij}^{Y})^2 \tag{4}$$

where $d_{ij}^{X} = x_i - x_j{}^2$ and $d_{ij}^{Y} = y_i - y_j{}^2$.

Several different MDS techniques have been proposed. Most of them try to represent the coordinates of the observed dissimilarities in m-dimensional space. The dissimilarities are mapped in a way that tries to match the Euclidean distances. Thus, the dissimilarity ρ_{ij} between points i and j is mapped into their distance d_{ij} with minimal loss of information. The dissimilarities are related to the Euclidean distance by the function $f : \rho_{ij} \rightarrow d_{ij}(\mathbf{X})$, where $d_{ij}(\mathbf{X})$ implies that the distance d_{ij} depends on the unknown coordinates of $\mathbf{X}$. Where $\mathbf{X}$ is a matrix $n\,x\,m$, where n is the number of points and m defines the dimensional space.

There are different models of MDS that can be defined based on this mapping of dissimilarities for the distances that will be presented below.

2.3.2 Classical MDS (CMDS)

CMDS (Saeed et al, 2018) attempts to create n projections, x, of the high-dimensional points in a d-dimensional linear space, trying to organize the projections in such a way that Euclidean distance between their pairs, d_{ij}, resemble the dissimilarities between the high-dimensional points. In short, CMDS tries to minimize:

$$\chi = \sum_{i \neq 1} (p_{ij} - d_{ij})^2 \tag{5}$$

where p_{ij} is the dissimilarity between point X_i and point X_j, and d_{ij} is the distance between the projection of X_i, x_i and the projection of X_j, x_j.

2.3.3 Weighted MDS (WMDS)

In the WMDS algorithm an extra parameter must be calculated to fit the points and their corresponding dissimilarities. Applications of WMDS include dealing with missing data, weighting the data based on its reliability, and normalizing the data (France and Carroll, 2011). Once these weights are estimated, the rest of the procedure is similar to CMDS. There are different formulations proposed by authors that use the idea of weighting, one of the best known is called Sammon's Mapping.

Sammon's Mapping is popular in pattern recognition and uses the idea of weight, where the weight is $w_{ij} = p_{ij}^{-1}$, so small dissimilarities will be given greater weight than large ones. The error function for *Sammon's Mapping* is defined as:

$$\xi_{sam} = \sum_{j=i+1}^{n} \left(1 - \frac{d_{ij}(\mathbf{X})}{p_{ij}}\right)^2 = \sum_{i<j} p_{ij}^{-1} (p_{ij} - d_{ij}(X))^2 \tag{6}$$

The weights w_{ij} are specified based on some formal considerations. One way to choose w_{ij} is to equalize it with the reliability of the proximity information, which means that more reliable proximities get more weight, while unreliable proximities get less weight (Saeed et al, 2018).

2.3.4 Isometric feature mapping (Isomap)

Isomap is also a dimensionality reduction technique that maps high-dimensional structures into a low-dimensional space. Isomap uses geodesic distances instead of the Euclidean distance between each pair of points in its distribution (Saeed et al, 2018).

A possible loss function for Isomap is defined as:

$$\xi_i = \sum_{i \neq j} (g_{ij} - d_{ij})^2 \tag{7}$$

The Isomap algorithm is based on CMDS for large-scale Multidimensional scaling problems. In particular, it focuses on nonlinear varieties and higher dimensions. Since CMDS cannot handle missing dissimilarity values, they are replaced by the shortest path in the graph. This forms a fully populated matrix of pseudosimilarities on which CMDS runs (Groenen and Borg, 2013).

2.4 Graph-based visualization techniques

A graph $G = (V, E)$ is a set V of vertices and a set E of edges, in which an edge joins a pair of vertices. Usually, graphs are represented with their vertices as points in a plane and their edges as line segments or curves connecting these points. There are algorithms that produce two-dimensional images of graphs, as will be presented in the following (Fruchterman and Reingold, 1991).

2.4.1 Force-directed placement (FDP)

The basic idea of the Force-directed placement algorithm (Fruchterman and Reingold, 1991) is based on graphs, replacing vertices with steel rings and replacing each edge with a spring to form a mechanical system. The vertices are placed in an initial layout randomly and the spring forces on the rings move until the system reaches a minimum energy state.

This model of a graph represented as a physical system of rings and springs is based on the work of Eades (Eades, 1984). Eades made some remarks about the forces exerted by the spring, for example, repulsive forces are calculated between each pair of vertices, but the attractive forces are calculated only between neighbors. This reduces the time complexity because calculating the attractive forces between neighbors is $\Theta(\mathsf{E})$, while calculating the repulsive force is $\Theta(\mathsf{V}^2)$, where V represents vertices and E represents edges.

The initial configuration of the algorithm can be completely or partially specified, but usually the vertices are placed randomly in the layout. Different functions may have been chosen to calculate the repulsion and attraction force. Basically the effect of the attractive forces on each vertex and the effect of the repulsive forces are calculated. We would like the vertices to be uniformly distributed in the layout. Intuitively, the farther apart two vertices are, or the closer they are, the more strongly the correction should be considered (Eades, 1984). If f_a and f_r are the attractive and repulsive forces, respectively, with d the distance between the two vertices, then:

$$f_a(d) = \frac{d^2}{k} \tag{8}$$

$$f_r(d) = -\frac{k^2}{d} \tag{9}$$

Since each particle is subject to the forces of all the other particles and it is necessary to calculate $n(n-1)$ the force in each iteration, the calculation

of the forces is $O(n^2)$, where n is the number of particles in the system. To produce a layout it takes n iterations, the resulting algorithm will be $O(n^3)$. Therefore, although this technique generates layouts accurately, its application is limited to small datasets (Morrison et al, 2011).

2.4.2 Chalmers's algorithm

Chalmers (Chalmers, 1996) made some considerations in the standard Force-directed placement technique to make the computational cost in each linear iteration relative to n. Instead of doing all force calculations $n(n-1)$ in each iteration, force calculations will be made between each object i and the members of two sets whose size is bounded by a constant. In this way, we maintain a computational cost for each iteration that is linear over n.

2.4.3 Force scheme

The approach is also based on the concepts of force of attraction and repulsion and uses the fact that the ratio between the distances in both the original and the projected space must be constant for each pair of data points (x'_i, x'_j). The idea is to separate instances projected too close together and to bring instances projected too far apart (Tejada et al, 2003).

This force-based projection enhancement technique was used to improve point placement by recovering some of the information lost during the projection process. The main difference here is that since the points were already projected, using fast techniques, and with the effort to preserve distance, the number of iterations required to converge is very small (Minghim et al, 2006).

2.5 Least square projection (LSP)

Given a set of points $S = p_1, ..., p_n$ in $\mathbb{R}^m$, the algorithm aims to represent the points of S in a smaller dimensional space $\mathbb{R}^p$, where $p \leq m$, preserving the neighborhood relationship between the points. There are two steps involved in this projection process. First, a subset of points in S, called "control points", are projected in $\mathbb{R}^p$ by the MDS method. Making use of the neighborhood relation of the points in $\mathbb{R}^m$ and the Cartesian coordinates of the control points in $\mathbb{R}^p$, it is possible to construct a linear system whose solutions are Cartesian coordinates of the points p_i in $\mathbb{R}^p$. Least square projection seeks to preserve neighborhood relationships between multidimensional objects in the projected space, unlike other conventional projection techniques that attempt to preserve distance relationships (Paulovich et al, 2008).

2.6 Assessment measures

Some objective quality measures can be applied to evaluate the results of the different visualization techniques used in the study. The measures used were the Silhouette Coefficient and the Neighborhood Hit.

The Silhouette Coefficient is a measure used in cluster analysis and to evaluate the quality of clusters generated by projection techniques. It is a

measure calculated for each instance, so we calculate the average coefficient of all instances as the Silhouette Coefficient of the projection. This coefficient ranges from -1 to 1, where the best results are close to 1 (Eler and Garcia, 2013).

The Neighborhood Hit is a measure that represents the percentage of nearest neighbors that belong to the same class as a certain instance. The strategy aims to analyze the ability of the visualization to preserve classes in the same neighborhood, favoring visual perception. Applying this metric to projections, the calculation is performed as a function of a distance between points in the projection plane. The more separated and grouped the points are the greater the accuracy (San Roman et al, 2013).

Precision determines the fraction of records that are actually positive in the group that the classifier has declared as a positive class. Precision is the ratio $tp/(tp + fp)$ where tp is the number of true positives and fp the number of false positives. The higher the precision, the lower the number of false positive errors made by the classifier. The best accuracy value is 1 and the worst value is 0 (Tan et al, 2005).

3 Related works

Most of the multidimensional projection techniques applied in the studies create visual representations highlighting the relationship between documents from textual information (Eler and Garcia, 2013). For this purpose, a vector space model is calculated using document content and visualization techniques deal with such models to establish relationship between documents. Visual text mining tools apply dimensionality reduction techniques to the space vector model to represent document collections in visual space (2D space). Preprocessing for creating the vector space model is important for obtaining information that characterizes the data and generates visual representations.

With regard to preprocessing techniques, we can highlight that the articles (Gomez-Nieto et al, 2014), (Alencar et al, 2012), (Eler and Garcia, 2013), (Giannakopoulos et al, 2013) and (Butka and Pócsová, 2013) used tokenization techniques, stopwords stemming and the term frequency calculation technique, known as tf-idf, to preprocess the input dataset and produce a matrix representing the documents (a so-called representation based on a vector model). In this vector model each document is represented by a feature frequency vector.

In (Chen et al, 2010), (Oesterling et al, 2010), (Sherkat et al, 2018), (Thai et al, 2012) and (Dias et al, 2019) were used the tf-idf metric to help get an overview of the content of the documents being studied.

Multidimensional projections have been employed to generate global visualizations of high-dimensional datasets according to (Alencar et al, 2012). A mapping of high-dimensional data into a low-dimensional visual space, typically 2D, is performed while similar points are placed close to each other. It has

been shown that these techniques applied to document collections can generate insightful document maps that are suitable for visualization and intuitive exploration of collection content.

The study (Eler and Garcia, 2013) applied the Multidimensional Scaling technique. The experiments used datasets with few documents, as it was not considered reasonable to use Multidimensional Scaling on larger datasets due to its computational complexity. Regarding the (Etemadpour et al, 2014b) study, the Multidimensional Scaling technique was used as it is an alternative capable of handling non-linear datasets. The Multidimensional Scaling projection had a tendency to create more rounded clusters of data, this technique was evaluated in this study using eye trackers which analyzes projected multidimensional data, looking for relationship, behavior comparison and pattern identification.

Regarding multidimensional projection techniques, the (de Antonio et al, 2013), (Sherkat et al, 2018) and (Muhr et al, 2010) studies used the Force-Direct Placement technique to perform the visual representation of document collections. This technique was used as it tries to improve the closeness of similar data points and increase the separation for different data points. Documents projection based on the Force-Direct Placement approach places documents with similar group labels together, while projecting the isolated nodes away from the center of the cluster.

The Force-Direct Placement technique simulates forces of attraction and repulsion between documents depending on their similarity measures. Although this approach has great scalability, it also has many desirable properties for the purpose of this study: good quality of the resulting layout, real-time iterative positioning process and ability to be extended including other factors in the positioning process.

Another multidimensional projection technique that was applied in the selected studies is the Least Square Projection. Several studies applied this technique and compared it with other projections. For example, the study (Gomez-Nieto et al, 2014) applied the Least Square Projection technique because of its good accuracy in terms of distance preservation and its low computational cost. The projection preserves much of the original neighborhood structure of the data, ensuring that similar instances are placed close to each other in visual space. The (Andreotti et al, 2018) and (Paulovich et al, 2012) studies employed the Least Square Projection projection technique that handles large datasets but low computational cost to project document collections in 2D space.

The authors (San Roman et al, 2013) applied Least Square Projection which generates a layout that preserves neighborhood groupings in the characteristic space. First, a subsample of the data is obtained, called control points, which represents the spatial distribution of the analyzed documents. Next, the neighborhoods for these control points are calculated. These control points are then designed, showing a global view that represents groups of texts with similar content.

The Least Square Projection technique was also applied in the study (Alencar et al, 2012) for the purpose of creating a projection with temporal perception. The study generated a sequence of similarity-based maps that transmits the evolution of a collection of documents over time.

The next section will present the use of visualization techniques from the multidimensional analysis of the Lattes platform data.

4 Materials and methods

This section describes the dataset used and its characteristics and also describes the methodological procedures used to achieve at the results of this work.

4.1 Materials

The dataset of this work is composed of information from the Lattes platform of the researchers and students of the School of Arts, Sciences and Humanities of the University of São Paulo (EACH USP). The bibliometric data that were analyzed are inserted within the Production module of the Lattes platform, specifically the topic of "published articles" was analyzed.

The dataset is separated by research groups belonging to EACH USP and for each research group there is a dataset regarding bibliographic production. The items present within the bibliographic production that the researcher enters information on are articles accepted for publication, published articles, other types of bibliographic production, books and chapters, texts in newspapers or magazines, and papers in events.

In this study the item of published articles was analyzed, the main information that is registered in this item and that are present in the data set are the title of the article, year of publication of the article, language, country of publication, title of the journal or magazine of publication, authors' names, among others. The multidimensional data analysis performed in this work is composed of the information obtained from the titles of the articles published for each research group analyzed belonging to EACH USP.

The research groups analyzed in this study were the groups belonging to the Information Systems graduate program, the Textile and Fashion graduate program, and the Astrophysics group. The Information Systems research groups contain 863 published articles belonging to five research groups: Management of Information Systems (GESI), Group on Public Policies for Access to Information (GPOPAI), Artificial Intelligence Research Group (GrIA), Laboratory for Health Informatics Applications (LApIS) and the Research Group on Modeling of Complex Systems (GRIFE). The Textile and Fashion research group contains 99 published articles and the Astrophysics research group contains 76 published articles.

4.2 Methods

The algorithm for the visual representation of the bibliometric data from the Lattes Platform can be summarized in five main steps: obtaining the curricula from the Lattes Platform, extraction of the module of published articles, textual preprocessing, generation of the document matrix, and visualization of the multidimensional projection of the data.

The knowledge used to build this approach corresponds to a survey of several approaches used in the related literature on multidimensional projection of data. Figure 2 graphically represents the five main steps of this work.

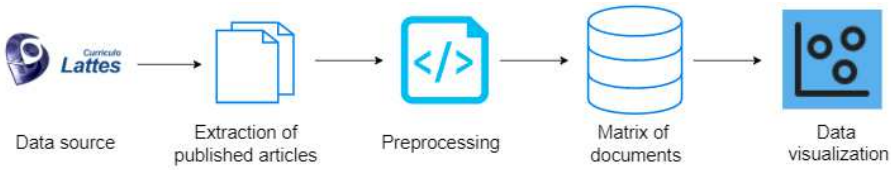


Fig. 2: Algorithm summary

The algorithm begins with the application of the preprocessing steps, the textual analysis of the articles will occur from the titles of the published articles. The first step to be performed is to exclude the titles of duplicate articles present in the same research group. For the textual analysis to be performed properly, it is necessary to collect the articles published in the same language, in the case of this study the language chosen is English, because most of the published articles are in English.

The second preprocessing step is to perform the removal of stopwords which are words considered irrelevant to the language and that have a high frequency. The goal of removing these words is to reduce their high volume that negatively affects the clustering of textual documents.

Next, the third step is the stemming process, which consists in removing prefixes and suffixes from words, so that these words are represented by their root. The goal is to represent words that are derived from each other by their root, in order to map a group of words by the same root.

The fourth stage of textual preprocessing is to transform words into tokens, this process consists in representing each word in the text in distinct units, thus each word in the document will be represented as a token. The goal of this process is to summarize the frequency of these words in each article to later generate a histogram of the most frequent words contained in the collection of published article titles.

By constructing the histogram of the frequency of terms it is possible to obtain the Zipf curve. The Zipf curve is drawn considering the ordering of the frequency of words in a descending order, and from the curve we can define a threshold to exclude the least significant terms. These thresholds are defined

using the Goffman's point, which established a procedure to eliminate less relevant terms from the documentary base.

After choosing the most significant words in the documents, a selection of articles containing these words is made. With the selection of these articles and after all the textual preprocessing performed on these documents is performed the step of creating the vector model of documents, where these documents will be represented as a matrix. The weights of each word contained in this collection of documents are calculated using the tf-idf method. Thus, the vector of a document d_j is represented by $\vec{d_j} = (\mathrm{w}_{j,1}, w_{j,2}, ..., w_{j,n})$.

Once the preprocessing stage is complete and the document matrix has been created, the process for creating the data visualization begins. The data visualization begins with the calculation of the distance between the documents that will generate a distance matrix mxm. The obtained distance matrix will be processed by multidimensional projection algorithms with the purpose of converting the multidimensional data into a lower dimensional space, keeping the intrinsic characteristics of this data.

The multidimensional projection strategy uses different types of algorithms and combinations. The multidimensional projection strategy uses different types of algorithms and combinations. The objective of performing the multidimensional projections is to verify if there is separation of clusters of research groups that work with very distinct themes and if research groups with related themes are not distinguished in the multidimensional projection visualization. In the next section the results obtained by each of the approaches will be presented.

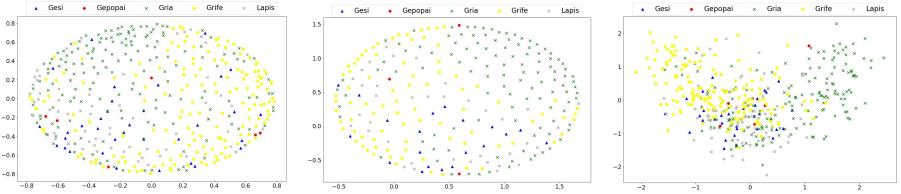
5 Results and discussion

For visual analysis of multidimensional data it is common to use dimensionality reduction techniques that project the multidimensional points into a low-dimensional visual space, and usually the projected points are displayed in the form of scatter plots in two dimensions. The projection method should preserve the distributions of the multidimensional data as much as possible in order to obtain information about this data (Etemadpour et al, 2014b). This section will present some results of techniques for projecting data onto two-dimensional visual spaces.

The Multidimensional scaling projection techniques were applied to the published articles from the research groups of the Information Systems graduate program, the Textile and Fashion graduate program, and the Astrophysics group obtained from the information entered in the Lattes platform.

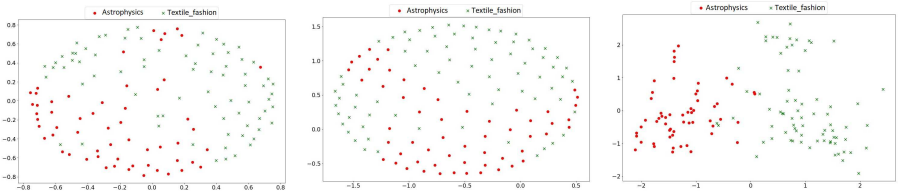
The results of the projection of the multidimensional data of the articles from the research groups of the graduate program in Information Systems are presented in Figure 3. Three scatter plots were obtained for each of the Multidimensional scaling techniques, the first plot represents the result of the Classical MDS technique, the second plot represents the Weighted MDS algorithm, and the third plot is the Isomap technique. Each color of the points in

the scatter plots represents an Information Systems research group and each point in the scatter plot represents a published article. It is observed that the research groups are not segregated in the visualizations, which can be an indication that these groups study topics that are related to each other. Another point observed is that the Isomap technique presents a scatter plot different from the other two techniques.



(a) Classical MDS (b) Weighted MDS (c) Isomap

Fig. 3: Information Systems Research Groups



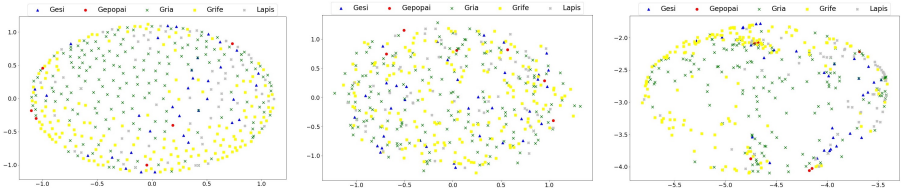
(a) Classical MDS (b) Weighted MDS (c) Isomap

Fig. 4: Textile and Fashion and Astrophysics Research Groups

Another analysis performed was to jointly compare the articles from the Textile and Fashion research groups and the Astrophysics research group. Theoretically these research groups study topics from research fields that are not very closely related to each other. The results obtained are presented in Figure 4, the green color of the points in the scatter plots represents the published articles from the Textile and Fashion research group and the red color represents the published articles from the Astrophysics research group. It can be seen that, unlike the Information Systems research groups, the published articles from Textile and Fashion and Astrophysics tend to separate more easily in the data visualization which may confirm our initial hypothesis that these research groups study topics that are not related to each other.

The Force-directed placement algorithm was applied to the same document set that was studied previously with the Multidimensional scaling technique. The results obtained with this projection technique are presented in Figure 5 and Figure 6. When the Force-directed placement algorithm and its variations

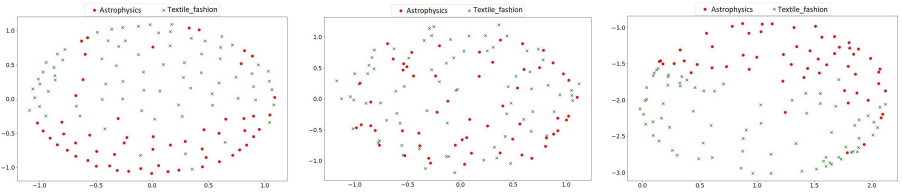
were applied to the published articles from the Information Systems research groups, no visual separation of these groups is observed in the projection obtained as shown in Figure 5.



(a) Classical Force-directed (b) Chalmers (c) Force scheme

Fig. 5: Information Systems Research Groups

Regarding the application of Force-directed placement techniques in the published articles of the Textile and Fashion research groups and the Astrophysics research group a better visual segregation of these groups is observed as shown in Figure 6.



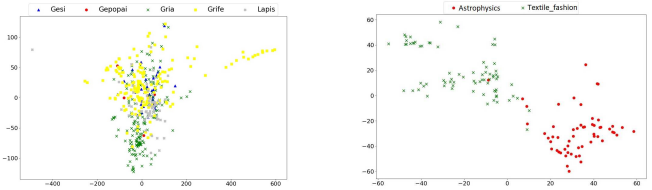
(a) Classical Force-directed (b) Chalmers (c) Force scheme

Fig. 6: Textile and Fashion and Astrophysics Research Groups

However, the Chalmers algorithm presented in item b in Figure 6 could not separate the Textile and Fashion and Astrophysics groups, possibly because it is an approach that reduces the complexity of iterations by using data samples to determine which instances are linked to each other, this sampling can cause loss of information about the articles of the research groups (Chalmers, 1996).

Finally, the Least Square Projection algorithm was applied in the published articles of the research groups of the Information Systems graduate program and the Textile and Fashion graduate program and of the Astrophysics research group, the results obtained are presented in Figure 7.

Again it is observed that there is visual separation only from the research groups of Textile and Fashion and the research group of Astrophysics. Visually, the Least Square Projection technique is the one that best separates data from these research groups when compared to the other algorithms presented previously.



(a) Information Systems Research Groups

(b) Textile and Fashion and Astrophysics Research Groups

Fig. 7: Least Square Projection

To finalize the analysis of multidimensional projection techniques, Neighborhood Hit and Silhouette Coefficient were used to compare the different layouts produced by the techniques of Multidimensional Scaling, Force-directed placement and Least Square Projection. Figure 8 presents the results of the Neighborhood Hit evaluation and Table 1 and Table 2 presents the results of the Silhouette Coefficient.

Table 1: Comparative evaluation between projection techniques using the Silhouette Coefficient in Information Systems Research Groups

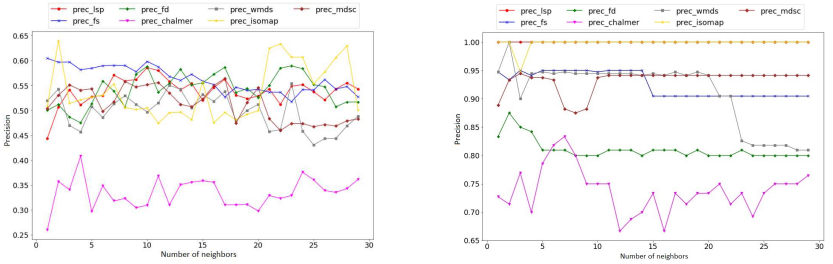
Projection techniques	Silhouette Coefficient
Multidimensional Scaling	-0,037
Weighted Multidimensional Scaling	-0,044
Force-directed Placement - Force Scheme	-0,065
Force-directed Placement	-0,077
Multidimensional Scaling - Isomap	-0,086
Force-directed Placement - Algoritmo de Chalmers	-0,105
Least Square Projection	-0,178

Table 2: Comparative evaluation between projection techniques using the Silhouette Coefficient in the Textile and Fashion Research Groups and the Astrophysics Research Group

Projection techniques	Silhouette Coefficient
Least Square Projection	0,587
Multidimensional Scaling - Isomap	0,424
Force-directed Placement - Force Scheme	0,232
Multidimensional Scaling	0,192
Force-directed Placement	0,164
Weighted Multidimensional Scaling	0,138
Force-directed Placement - Algoritmo de Chalmers	0,009

The comparative evaluation of the projection techniques for the Information Systems research groups does not show satisfactory results according to the Silhouette Coefficient values presented in Table 1. Silhouette Coefficient close to 0 means that there is a low quality in the formation of groups generated by the projection techniques. Looking at the result of the Neighborhood Hit that assesses the projections generated for the Information Systems Research Groups, none of the techniques can have a precision result greater than 70%, which suggests that the Information Systems studies are not easily separable and that possibly these groups study subjects that are related to each other.

When the evaluation is carried out comparing the projection techniques for the published articles of the research groups of Textile and Fashion and of the research group of Astrophysics, more satisfactory results are observed. The Silhouette Coefficient results presented in Table 2 indicate that the projection techniques Least Square Projection and Multidimensional Scaling - Isomap have the best results, closer to the value 1, which indicates that there is a higher quality in the separation and formation of the groups and can be confirmed with the visual observation presented previously. The best results presented by the Neighborhood Hit approach also belong to the Least Square Projection and Multidimensional Scaling - Isomap projection techniques. For the two evaluation approaches Chalmers' projection algorithm presents the worst results when it comes to separating the Textile and Fashion research groups from the Astrophysics research group, which can be confirmed by looking at the visual result presented in Figure 6.



(a) Information Systems Research Groups

(b) Textile and Fashion and Astrophysics Research Groups

Fig. 8: Comparative evaluation between projection techniques using the Silhouette Coefficient in the Information Systems Research Groups and in the Textile and Fashion Research Groups and the Astrophysics Research Group

6 Conclusion and future works

The study demonstrated that the projection techniques Least Square Projection and Multidimensional Scaling - Isomap had the best results when applied

to the articles published in the Textile and Fashion research groups and the Astrophysics research group that which extracted from the Lattes platform.

When the Information Systems research groups were evaluated it was evident that there is no visual separation of these groups, which indicates that the published articles belonging to this group are related to each other.

Our approach can be used to evaluate and analyze bibliometric data, identify how research groups may be related to each other, and assist in reporting on the scientific productions and projects of different research groups present in the Lattes curriculum.

The limitation found in this work is related to the scope of the research, where it was defined that some research groups belonging to the School of Arts, Sciences and Humanities of the University of São Paulo would be evaluated. In addition, there is a qualitative limitation, as the information entered in the Lattes platform is the sole responsibility of the user and there is no verification of this data.

Future work includes using more sophisticated multidimensional projection techniques and expanding the analysis to other research groups within the Lattes curriculum.

7 Declarations

7.1 Ethical Approval and Consent to participate

Not Applicable

7.2 Consent for publication

The authors agree to the publication of the article, if accepted.

7.3 Availability of supporting data

Not Applicable

7.4 Competing interests

The authors declare that the authors have no competing interests as defined by Springer, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

7.5 Funding

Not Applicable.

7.6 Authors' contributions

The authors developed the research in collaboration, and the author Helton Hideraldo Bíscaro is the supervisor of the author Patrícia Salles Escarassatti. Both collaborated in writing the text and designing the experiments. The

author Patrícia Salles Escarassatti carried out the implementations and data compilation.

7.7 Acknowledgments

The authors thank researcher José de Jesús Pérez Alcazar for providing the data.

References

Abasi AK, Khader AT, Al-Betar MA, et al (2020) Link-based multi-verse optimizer for text documents clustering. Applied Soft Computing Journal 87(106002)

Abbas A, Zhang L, Khan SU (2014) A literature review on the state-of-the-art in patent analysis. World Patent Information 37:3–13

Alencar AB, Paulovich FV, Börner K, et al (2012) Time-aware visualization of document collections. Proceedings of the ACM Symposium on Applied Computing pp 997–1004

Aliguliyev RM (2009) Clustering of document collection – a weighting approach. Expert Systems with Applications 36:7904–7916

Amorin CV (2003) Organização do currículo: plataforma lattes. Pesquisa Odontológica Brasileira 17:18–22

Andreotti ALD, Silva LF, Eler DM (2018) Hybrid visualization approach to show documents similarity and content in a single view. Information (Switzerland) 9(129)

de Antonio A, Moral C, Klepel D, et al (2013) 3d gesture-based exploration and search in document collectionsn. 17th International Conference on Electronic Publishing, ELPUB 2013 pp 13–22

Belter CW (2015) Bibliometric indicators: opportunities and limits. Journal of the Medical Library Association : JMLA 103,4:219–221

Butka P, Pócsová J (2013) Hybrid approach for visualization of documents clusters using ghsom and sammon projection. 8th IEEE International Symposium on Applied Computational Intelligence and Informatics, SACI 2013 (6608994):337–342

Chalmers M (1996) A linear iteration time layout algorithm for visualising high-dimensional data. Proceedings of Seventh Annual IEEE Visualization Conference (5456773)

Chen C, Ibekwe-SanJuan F, Hou J (2010) The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. Journal of the American Society for Information Science and Technology 61:1386–1409

Chen Y, Wang L, Dong M, et al (2009) Exemplar-based visualization of large document corpus. IEEE Transactions on Visualization and Computer Graphics 15(6):1161–1168

Dias AG, Milios EE, de Oliveira MCF (2019) Trivir: A visualization system to support document retrieval with high recall. Proceedings of the ACM Symposium on Document Engineering, DocEng 2019 (3345401)

Dunne C, Shneiderman B, Gove R, et al (2012) Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. Journal of the American Society for Information Science and Technology 63:2351–2369

Durieux V, Gevenois PA (2010) Bibliometric indicators: quality measurements of scientific publication. Radiology 255,2:342–351

Eades P (1984) A heuristic for graph drawing. In: Congressus Numerantium, pp 149–160

Eler DM, Garcia RE (2013) Using otsu's threshold selection method for eliminating terms in vector space model computation. 17th International Conference on Information Visualisation, IV 2013 (6676566):220–226

Etemadpour R, da Motta RC, de Souza Paiva JG, et al (2014a) Role of human perception in cluster-based visual analysis of multidimensional data projections. 5th International Conference on Information Visualization Theory and Applications, IVAPP 2014 pp 276–283

Etemadpour R, Olk B, Linsen L (2014b) Eye-tracking investigation during visual analysis of projected multidimensional data with 2d scatterplots. 5th International Conference on Information Visualization Theory and Applications, IVAPP 2014 pp 233–246

France SL, Carroll D (2011) Two-way multidimensional scaling: A review. Systems, Man, and Cybernetics, Part C: Applications and Reviews 41:644–661

Fruchterman TMJ, Reingold EM (1991) Graph drawing by force-directed placement. Software: Practice and Experience 21:1129–1164

Giannakopoulos T, Stamatogiannakis E, Foufoulas I, et al (2013) Content visualization of scientific corpora using an extensible relational database implementation. 17th International Conference on Theory and Practice of

Digital Libraries, TPDL 2013 416:101–112

Gomez-Nieto E, Roman FS, Pagliosa P, et al (2014) Similarity preserving snippet-based visualization of web search results. IEEE Transactions on Visualization and Computer Graphics 20(6629989):457–470

Groenen PJ, Borg I (2013) The past, present, and future of multidimensional scaling. Econometric Institute Report

Kalmukov Y (2020) Automatic assignment of reviewers to papers based on vector space text analysis model. ACM International Conference Proceeding Series pp 229–235

Luhn HP (1958) The automatic creation of literature abstracts. IBM Journal of Research and Development 2:159–165

M. Gutiérrez-Salcedo JAMMEHVMCM. Ángeles Martínez (2018) Some bibliometric procedures for analyzing and evaluating research fields. Applied Intelligence 48:1275–1287

Mena-Chalco JP, Junior RMC (2009) Scriptlattes: an open-source knowledge extraction system from the lattes platform. Journal of the Brazilian Computer Society 15:31–39

Minghim R, Paulovich FV, de Andrade Lopes A (2006) Content-based text mapping using multi-dimensional projections for exploration of document collections. Proceedings of SPIE - The International Society for Optical Engineering 6060(60600S)

Morrison A, Ross G, Chalmers M (2011) Combining and comparing clustering and layout algorithms. Department of Computing Science, University of Glasgow

Muhr M, Sabol V, Granitzer M (2010) Scalable recursive top-down hierarchical clustering approach with implicit model selection for textual data sets. 21st International Workshop on Database and Expert Systems Applications, DEXA 2010 (5591979):15–19

Oesterling P, Scheuermann G, Teresniak S, et al (2010) Two-stage framework for a topology-based projection and visualization of classified document collections. 1st IEEE Conference on Visual Analytics Science and Technology, VAST 10 (5652940):91–98

Paulovich FV, Nonato LG, Minghim R, et al (2008) Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. IEEE Transactions on Visualization and Computer Graphics 14(4378370):564–575

Paulovich FV, Toledo FMB, Telles GP, et al (2012) Semantic wordification of document collections. Computer Graphics Forum 31:1145–1153

Saeed N, Nam H, Haq M, et al (2018) A survey on multidimensional scaling. ACM Computing Surveys 51(1)

San Roman F, De Pinho R, Minghim R, et al (2013) A study on the role of similarity measures in visual text analytics. International Conference on Computer Graphics Theory and Applications, GRAPP 2013 pp 429–438

Sequera JLC, Castillo JRFD, Sotos LG (2009) Cluster of reuters 21578 collections using genetic algorithms and nzipf method. IADIS European Conference Data Mining 2009 pp 174–176

Sherkat E, Nourashrafeddin S, Milios EE, et al (2018) Interactive document clustering revisited: A visual analytics approach. 23rd ACM International Conference on Intelligent User Interfaces, IUI 2018 pp 281–292

Tan PN, Steinbach M, Kumar V (2005) Introduction to data mining, 1st edn. Addison Wesley

Tejada E, Nonato LG, Minghim R (2003) On improved projection techniques to support visual exploration of multi-dimensional data sets. Information Visualization 2:218–231

Thai V, Rouille PY, Handschuh S (2012) Visual abstraction and ordering in faceted browsing of text collections. ACM Transactions on Intelligent Systems and Technology 3(21)