Similarities Building a Network between Researchers based on the Curriculum Lattes Platform

Sérgio Antonio de Andrade Freitas[1], Edna Dias Canedo[2], Edgard Costa Oliveira[3] and Dionlan Alves de Jesus[1]

[1]*Faculty of Gama (FGA), University of Brasília (UnB), Área Especial de Indústria, Projeção A – P.O. Box 8114, Brasília-DF, CEP 72.444-240, Brazil*

[2]*Department of Computer Science, Edifício CIC/EST, Campus Darcy Ribeiro, Asa Norte - University of Brasília (UnB), P.O. Box 4466, Brasília-DF, CEP 70910-900, Brazil*

[3]*Department of Production Engineering, Technology College, University of Brasília (UnB),*

Keywords: Semantic Web, Curriculum Lattes, Algorithm, Similarity.

Abstract: Using inference machines is one resource used to assist the decision-making process in data processing and interpretation, which allows attributing knowledge to a set of information items. In this sense this work implements a similarity algorithm that calculates the percentage of adherence found amongst academic profiles at the University of Brasília (UnB). The domain base use to provide the data for the work is that of the Lattes platform. This platform holds data on the scientific production of registered university scholars. The calculation provides a rating of the individuals and the approximations between their academic production. This is achieved by taking into account a base profile which is compared to one or more destination profiles. To run this procedure, the data held in each Curriculum Lattes is extracted, and an ontology of concepts is created that holds the data on the production to supply the information needed by the comparison task. These comparisons are made in each term of the name, for all the bibliographical production for both profiles compared. Each term can have a set of synonyms that are also taken into consideration in the comparison. And at the end the results are compiled and presented in a spreadsheet that holds the summaries for all adherence percentages that were compared. Applying the algorithm determines which people in a set have more or less proximity and a semantic link with the academic output when compared to other individuals. And that produces a similarity percentage.

1 INTRODUCTION

The Web as an entity that is in constant development can increasingly attribute sense to information, through machines and software agents, to organize knowledge in many areas, with the use of standardized terminologies, as a means to structure information to produce knowledge. The data is available in many formats, namely: Web pages, files, repositories, and the Curriculum Lattes database, amongst others.

The Lattes platform http://lattes.cnpq.br/, used a the source of data in this work, provides the academic background data on the execution of scientific work and the academic output of students, lecturers and professionals involved in science and technology.

This source of data establishes, percentage-wise, how much a member is connected and in adherence, or similar, according to one's bibliographical production when compared to another member, or to a group of members of the Brazilian academic community. This work aims at developing a tool that takes the information held in the Curriculum Lattes base into account, on a researcher, to extract the data and create an ontological model. And, beyond that, to set up a mathematical model that calculates the similarity percentage that exists between the individuals/researchers. To that end, an algorithm was developed that automatizes this ontological model and that points, in a quantitative fashion, what the percentage of adherence is, as found amongst the publications of the members compared.

According to the bibliography surveyed (Shadbolt et al., 2006), (Hendler et al., 2002), (Berners-Lee et al., 2001), (Sudeepthi et al., 2012) and (de Oliveira, 2011), the semantic Web, with the use of its tools, languages, and frameworks, allows the development

Freitas, S., Dias Canedo, E., Costa Oliveira, E. and Alves de Jesus, D.
Similarities Building a Network between Researchers based on the Curriculum Lattes Platform.
DOI: 10.5220/0006664102030214
In *Proceedings of the 20th International Conference on Enterprise Information Systems (ICEIS 2018)*, pages 203-214
ISBN: 978-989-758-298-1

of the ontology of concepts and provides input items for this domain to be mapped and towards obtaining the data that is found in each Curriculum Lattes of the researchers studied.

In the development of the application, already-existing functionalities were re-used, to extract the curriculum from the Lattes platform and to create the ontology. The ontological file is read and consulted in order to obtain the bibliographical works considered in the calculation of the similarity, amongst the individuals selected. The result is exported on a spreadsheet that lists the profiles under comparison, and displays a percentage of adherence as found amongst them.

Tests were carried out to verify and validate the figures obtained. The set of tests was controlled with the knowledge of the set of profiles that would be compared. The goal was to attest whether the figures actually obtained matched that which were observed in the real world.

In order to better explain the point of this work, it is structured as follows. Section 2 presents the proposed model, with the basis of the proposal, the proposal in itself, the mathematical formulation, and the implementation of the algorithm. Section 3 describes the implementation and the tests that were run. It also presents the tables, as originated in the execution of the application, along with the final result and the percentages of similarity for the comparisons. Finally, Section 4 concludes the work.

2 MODEL FOR SIMILARITY AMONGST RESEARCHERS

The model proposed is based on the use of the platform found in the Lattes academic record database as its data source to obtain information on the academic careers of people, aimed at making comparisons via an algorithm. In possession of such data, one can consolidate inferences as found amongst the individuals found in the domain that the Lattes database is about.

In the academic community, both the teaching staff as the students corpus can have their Curriculum Lattes, as a way to build a portfolio on one's academic path, for different ends. The Curriculum Lattes base, as found in the Web, holds varied information on the academic career of any of the individuals registered. Based on this data a strategy was devised to create an ontology that would represent this model, as expressed in the Lattes database (Galego and Renata, 2013). This strategy also included a manner for extracting data that would allow a relationship amongst the individuals found in the database. Such a drive stemmed from a few questions we wanted to see answered:

1. Is it possible to establish a link between different individuals that have not met, based on the life they lead in the Academy?
2. Is it possible to establish a quantitative approach through calculation of how much similarity there is in a comparison made of individuals hitherto unknown to each other?
3. Is it possible to make this line of thought automatic with the use of an algorithm? That is, there is a possibility for drawing aspects that are similar amongst individuals and to have the process for that running on a machine, to obtain an index that allows assessing whether a person is similar or not to another one, considering the aspects of one's trajectory in the Academy.

In the analysis of the Lattes database it is possible to see that several fields separate the individuals and characterize them according to a predominance in a given area of knowledge.

Some fields found in blocks of items can be considered for the purposes of individualization, of the characteristic of each individual.

Amongst these items we can cite: academic background and titles, supplementary qualifications, professional history, research areas pursued, research projects, extension projects, development projects, reviewing work for periodicals, areas of activity, awards and titles received, work that contains the bibliographical production, articles published in periodicals, books published/organized, or editions, chapters of books published, text published in journals/magazines as news, full work published n conference proceedings, expanded abstracts published in conference proceedings, abstracts published in conference proceedings, presentations of technical production work with information on assistance and consultancy, technical work, interviews, round tables, programs and comments in the media and in the item that covers other types of technical production, and other work.

The blocks of information have other elements that hold information that is semantically relevant and that can be used. The following elements can be mentioned as examples: advice in work for degree course final project, advice and supervision completed, advice for MSC theses, end-of-course work in refresher/specialization courses, end-of-course work in degree courses, scientific introduction courses, advice of other kinds, development of teaching or instruction material, interviews, round tables, programs and comments in the media, organization of events, conferences, exhibitions and fairs, examination boards in judges committees in public tests, and other participations.

Faced with this, one can see that the wealth of information is quite impressive. A constantly updated Curriculum Lattes of an individual is considered as a very valuable source of data, for the purpose of surveying and assessing one's results. And, to materialize the proposal, the presentation of a percentage of similarity is done through a number of comparisons with other members registered in the Lattes database.

The goal is not to exhaust all the attributes found in the Lattes database, but to demonstrate that it is entirely possible to establish a manner of comparison between individuals, based on the set of information items that is laid out in each profile.

As a proposal, we set up a plan to consider the macro item Production that entails the bibliographical production, technical output, and other artistic and cultural productions. The data found in these characteristics is considered as having great semantic relevance when thinking of making comparisons between people. Once the inference – or comparison – is made of this macro item with an acceptable satisfaction margin and taking into account the computational analysis versus human analysis, it is possible to consider relating the others with the same purpose.

At first, and considering the block of elements contained in the Production element, the comparison will be made through considering the name of each title, for each bibliographical production, with the goal of achieving key terms that have a high semantic load, to allow more efficient comparisons with other terms, as found in the publications, amongst the individuals surveyed. The terms will be selected and submitted to consultations through synonyms, to encompass a larger possibility and to render the result amongst comparisons more concise.

Thus, starting from these considerations of the analysis of the data, as found in the items of the Lattes database, and knowing that, from there, it is possible to make the comparisons, one can formalize the process of comparison in a structured algorithmic form.

2.1 Calculation of the Index of Similarity

Consider a closed set of individuals identified by their Curriculum Lattes (equation 1):

$$CP = \{I_1, I_2, ...I_i, ..., I_n\} \tag{1}$$

Each individual I_1 is represented through a set LI_i where each one of its elements B_ik is the list of all the activities of I_1 in one of the large blocks j that form the Lattes database (e.g., concentration area, publications in periodicals, publications in proceedings, advisory work, etc.):

$$LI_i = \{B_i^1, B_i^2, ...B_i^k, ..., B_i^m\} \tag{2}$$

Consider now one of the individuals in named Base Individual I_{base}:

$$\exists I_{base} | I_{base} \in CP \tag{3}$$

And another individual I_{target} in CP different from I_{base}:

$$\exists I_{target}, I_{base} | I_{target} \in CP \wedge I_{base} \in CP \wedge I_{target} \neq I_{base} \tag{4}$$

The comparison between individuals I_{base} and I_{target} is possible through correlating the activities represented in their Curriculum Lattes and the comparison of the elements of one same block B^k as found both in LI_{base} as in LI_{target}. Consider an activity E_b as belonging to the group of elements that form LI_{base} of the Lattes of LI_{base} and semantically equivalent to an activity E_d as belonging to the set of activities of I_{target}:

$$\exists E_b, E_d | E_b \in B_{base}^k \wedge E_d \in B_{target}^k \wedge E_b \equiv E_d \tag{5}$$

Considering a block of B^k any kind, the set of all the elements E_b that verify equation 5 represents the set of similarities found amongst individuals I_{base} and I_{target} for a given block k:

$$Sem_k = A | E_i \in B_{base}^k \wedge E_j \in B_{target}^k \wedge E_i \equiv E_j \tag{6}$$

Where A in equation 6 represents:

$$A = \bigcup_{i=0}^{max(B_{base}^k)} \wedge \bigcup_{j=0}^{max(B_{target}^k)} \exists E_i, E_j \tag{7}$$

In applying equation 6 for each one of the blocks found in the Lattes database for two individuals and by merging the results, it is possible to obtain a similarity set between I_{base} and I_{target}:

$$CS = Sem_1 \bigcup Sem_2 ... Sem_k \bigcup ... Sem_t \tag{8}$$

Using the cardinality of CS it is possible to establish an similarity index IS between I_{base} and I_{target}:

$$IS_{base}^{target} = |CS| \tag{9}$$

In fixing individual I_{base} and doing the calculation for equation 9, for each one of the individuals of CP a set is obtained for the similarity indexes CIS_{base} between the base individual and the remainder of the individuals considered:

$$CIS_{base} = \{|IS_{base}^1|, |IS_{base}^2|, ..., |IS_{base}^n|\} \tag{10}$$

The ordained set of CIS_{base} establishes the similarity degree between any base individual and the rest of the set of individuals.

2.2 Applying the Algorithm

This section presents the step-by-step process for the similarity algorithm. The steps followed start with the reading of the data on the individuals informed through the identifier element found in the Lattes database. Following the reading of all curriculum in the Lattes database a procedure is carried out that processes the names in each publication found for both the base and target individuals. This function removes the insignificant terms and standardizes the quantity of terms that will be compared. After that, for each word selected a consultation is made to obtain the respective synonyms, should they exist.

The comparison starts when the base individual is set and all of his/her publications are compared with all the publications of all the target individuals. That is, the first term in the first publication of the base profile is compared with the first term of the bibliographical output of target individual number 1. This process is iterated until the number of terms and synonyms is exhausted, followed by the volume of one's bibliographical production and, lastly, the quantity of the target individuals themselves.

Equal occurrences that are found are accounted for and added into the percentage calculation. For the purposes of calculation, the figure for the product in the quantity of the quantities of bibliographical production of the individual are needed, with each target individual, which will be the division number for each percentage amount.

This figure is multiplied by 5, as to work with equal figures which, in the case of the term or word itself, number 5 represents the quantity of terms randomly chosen from the name of the bibliographical output. As dividend, we have the number of equal occurrences, which is incremented as every new equal term is found between the base and the target individuals.

Lastly, in order to have the amount as a percentage, it is multiplied by 100 and a classification is done of the percentages of adherence for the comparisons that have been calculated. Similarity algorithm between the bibliographical outputs for selected members:

- Reads base individual data;
- Reads target individuals' data;
- Processes the names of the publications;
- Searches synonyms in English and in Portuguese for each term found in the bibliographical output items that have been loaded;
- Fixes the base individual and, for each element of every target: (1). Compares the first term of the base element with that of the target individual; (2) If they are equal, then: Counts one equal occurrence; (3) If not: Does not count anything and moves to the next term and goes back to the previous step; (4) Moves to the next target individual and repeats procedure until the last individual.
- Calculates the similarity percentage;
- Divides the quantity of equal occurrences by the number of comparison possibilities found amongst the individuals multiplied by 5;
- Multiplies the total by 100. Classifies the percentage of the comparisons.

3 IMPLEMENTING THE PROPOSED MODEL

The infrastructure environment that was configured to support the development, execution and testing of the system took some software products as well as non-functional requirements of portability, implementation, and integration into account. As a requirement for integration, the software should have a link to the Dynamic Lattes application http://www.github.com/efgalego/DynamicLattes to obtain the ontology file and thus manage to process with the algorithm that calculates the similarity percentage (Luna et al., 2013) and (da Costa and Yamate, 2009).

Functional requirements of the application:

- Extracting data from the Lattes database;
- Creating the ontology that models the Lattes database;
- Calculating the similarity percentage;
- Taking the title of each publication into account;
- Obtaining synonyms for each English and Portuguese term;
- Making comparisons between the terms of each individual;
- Classifying the adherence percentage.

The architecture of the software constructed is built by the collaboration from other tools that assist the calculation of the similarity. The Dynamic Lattes open-source tool, apart from helping the extraction process, is an element of the Script Lattes, Onto Lattes, and Semantic Lattes tools (da Costa and Yamate, 2009).

Dynamic Lattes does the extraction, creates the ontology, and builds an OWL format file (Siddiqui and Alam, 2011). The goal is to have the software

developed read, interpret and calculate the similarity percentage for the profiles registered with the Lattes platform database, and allow the analysis for adherence of the sets of profiles under comparison.

3.1 System Modelling

Figure 1 shows the domain of the application on a business level, which helps to understand the scope of the architecture developed for the application.

The Lattes Database represents the knowledge base of the Lattes platform, which currently holds a little over three million registered curriculum. The Dynamic Lattes application (Chalco et al., 2009) is responsible for extracting the Curriculum Lattes through an identifier element, unique to each individual that is fed as a parameter to the tool. It interoperates with the Lattes knowledge through a call for a curriculum in a XML format and does the mapping of the Lattes domain in a representation of the ontology, to consolidate it via an OWL file (da Costa and Yamate, 2009).

The OWL file holds the meta data and the populated values for the curriculum that have been fed. The application developed interprets this information and makes the conversion of the ontology into objects, using the Java programming language.

This stage is shown in Figure 1 in a package that includes the calculation of the similarity percentage. After that a graph of similarities is shown, pursuant to the criterion set in the algorithm shown in Section 2.1.

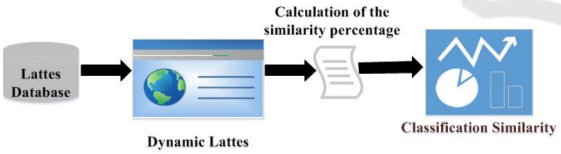

Figure 1: Domain Application.

Use case diagram shown in Figure 2 has the interaction of an user that may be the individual that wishes to find what one's similarity percentage is in relation to the other members, with the use of the software. One interacts with the Calculate Similarity Percentage use case.

The use case at first interacts with the Dynamic Lattes to be able to carry out the extraction procedures, contained in the Extract Data use case, from the Lattes Database, to Create the Ontology in the Lattes Database. After this interaction, the ontological file is created and used by the Consult Extracted Individual use case, which is when it loads the results from bibliographical output into lists for all people being fed data. From then on, other use cases are necessarily executed, as shown in Figure 2 by the <<include>> notation. Firstly, the Process Title of Each Publication use case, followed by Obtain Synonym for Each Term in English and in Portuguese and after that by the Compare Terms of Publications from Each Individual and, lastly, by the Export Output into Spreadsheet item.

Figure 4 shows The process modeling with the flow that is executed until a spreadsheet is obtained that holds the values for the comparisons of similarities as found amongst the individuals.

The process gets under way with the extraction of the data from the Lattes database that looks up the curriculum base in the Lattes platform to create the ontology of concepts (da Costa and Yamate, 2009).

The ontology is a file that will be later loaded and interpreted by the application, and also to check the existence of a record produced by the consultation. If no records are found the process is terminated and, if not, the records are loaded into lists that will be processed to extract terms regarded as non-significant from the standpoint of the application, to then be subjected to a synonym consultation. In the end, the comparisons are done and a spreadsheet is produced that holds the percentage results for each comparison, along with the quantity of terms for each individual compared, on a per-person basis.

Following the comparison procedure, the final calculation is done, added with the percentage value for similarity formulated by the mathematical relation (Chalco et al., 2009) shown in the previous chapter that determines that the similarity percentage is equal to the number of synonyms (occurrences found), divided by the number of possibilities (Cartesian product) found amongst the publications of the two profiles compared.

3.2 View of the Workflow

The process for the workflow and for information exchange amongst the elements that make the processing of the inference algorithm starts when an user accesses the Dynamic Lattes entry system, as shown in Figure 3. The user enters the list of the identifiers for all Curriculum Lattes one wishes to extract from, including the base individual and all target individuals.

The Dynamic Lattes communicates with the database of the Lattes platform by itself, advising the identifiers entered by the user to locate the respective Curriculum Lattes. The extraction then happens, when the files for each curriculum are found, and with the reply for the consultation for processing sent by

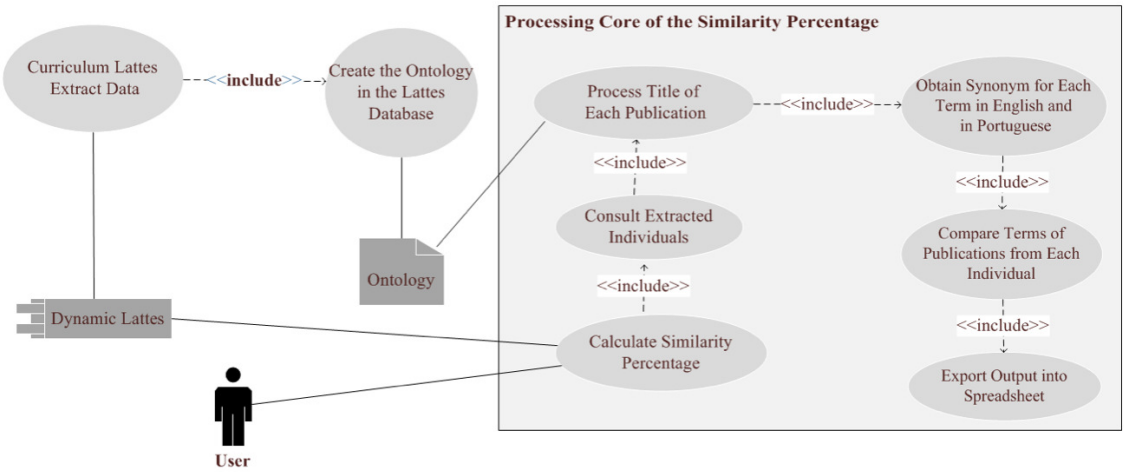

Figure 2: Use Case Diagram.

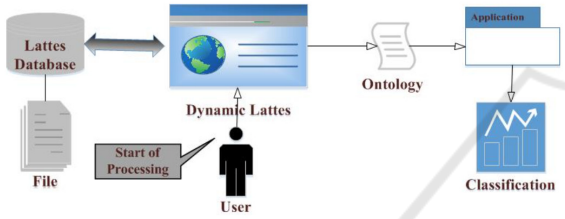

Figure 3: Process Workflow.

the Dynamic Lattes. Each file represents a curriculum in the Lattes database.

The internal process of the Dynamic Lattes should run through the following stages:

1. Obtaining the data from the Lattes Platform (XML);
2. Managing a list of researchers (called members);
3. Storing data and making it available for consultation, in an ontology-based database, to allow the use of inference monitors;
4. The user may ask the system for reports, consolidated as per group of researchers. These reports should use the data that is found in the database.

The Presentation layer is shown in Figure 5 and is responsible for the interaction with the user via Web pages. These Web pages are loaded with the data found in the Data layer. The Data layer persists and consults data with the importing of OWL format file that was generated on the extraction layer and the search for information is done through SPARQL (Zhao et al., 2014) consultations; the inferences on the ontologies are also carried out on this layer.

The extraction layer aims at making the data extractions from the Lattes Platform, as provided for download in XML format, and crush this data to generate the OWL file that will persist on the data layer. This layer provides the essential extraction functionalities as well as the creation of the ontology that will be used in the next processing stages.

Following the creation of the OWL ontology file, the application loads the file and starts the process of consultation of individuals. A base individual is defined whilst the remaining ones become target individuals. At this point the processing is done of the names of the bibliographical output, the synonyms in English and in Portuguese are obtained, along with the comparisons and the calculation of the percentage for similarity.

Finally, a compilation is run of the results from the comparison into a list, with their classification in percentage terms. With it, it is possible to identify the target individuals that are most similar and those that are more aligned with the base individual.

3.3 Implementation

The implementation of the application resorted to Dynamic Lattes, which incorporates the Script Lattes, Onto Lattes, and Semantic Lattes tools (Berners-Lee et al., 2001), pursuant to the following steps, as needed to obtain the values for the knowledge base in the Lattes Platform:

3.3.1 Knowledge Extraction from the Lattes Database

The Onto Lattes rose with the need to build an ontology to represent the data domain for the members found in the Curriculum Lattes database. It is used for the extraction of information found in the Curriculum Lattes, and provided in XML format by the Lattes database.

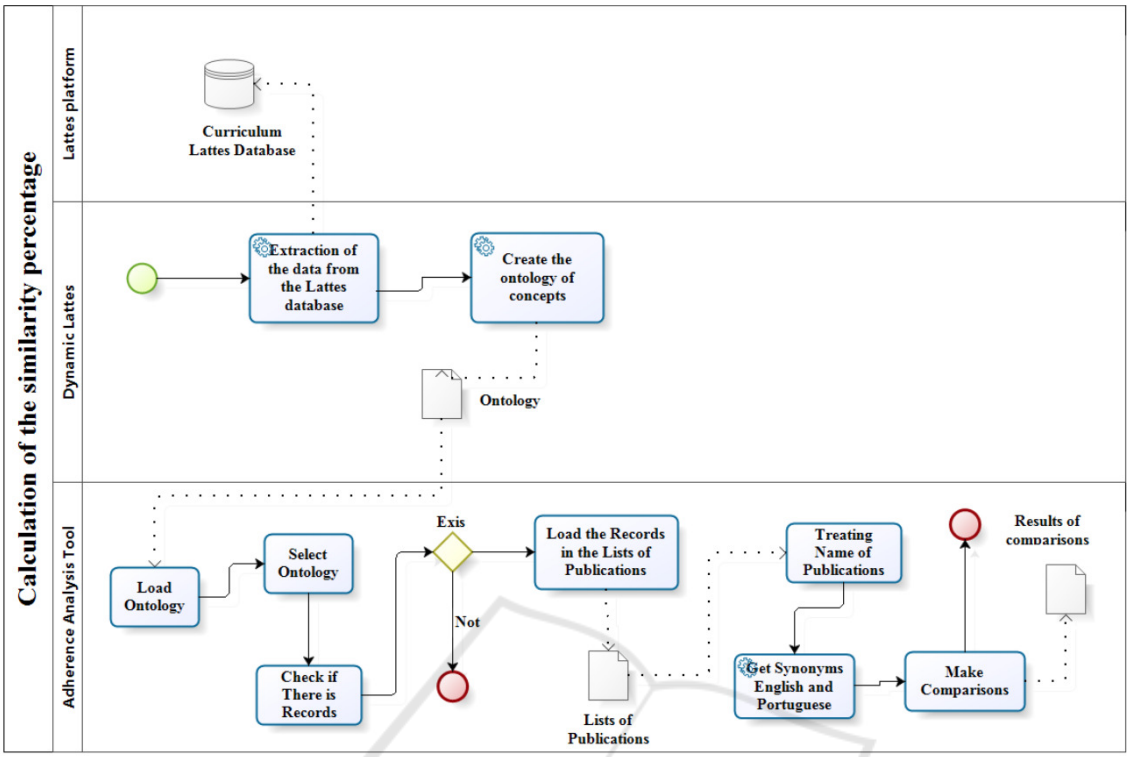

Figure 4: Process Modeling.

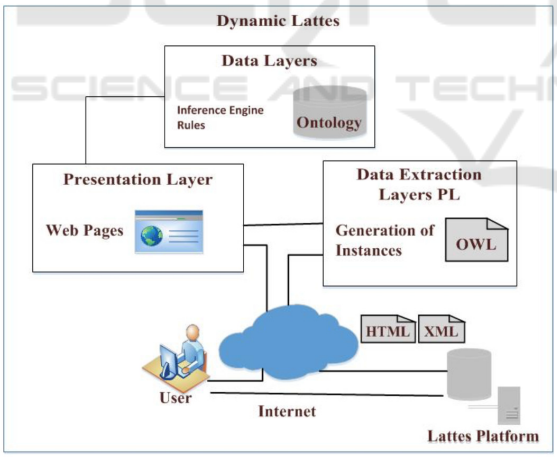

Figure 5: Architecture Dynamic Lattes.

The Script Lattes tools extracts data on the teaching staff, as registered with the Lattes database. When it is executed, it runs a download process with some specific features such as reports on academic supervisions, maps of collaboration and research maps, for members of a group.graphs (networks) of joint authorship amongst members of a group, thesis advisory work reports, reports on research projects and awards, geo-location maps, and for degrees of collaboration.

Semantic Lattes (Sudeepthi et al., 2012) aims at creating a system that can run consultations on the Curriculum Lattes. These consultations are done with natural language. The OWL ontologies were produced based on questions produced by business specialists.

Based on the functionalities of these tools, the application extracts the curriculum in a XML format from the Lattes knowledge database, creates the representational ontology for the domain of each curriculum, and creates the OWL extension file. The result of interest at such initial moment is the OWL file, which holds the ontology created and loaded to use the values in the comparison, to be the entry that triggers the file read, as well as to run the conversion and the transformation into objects in the Java programming language.

3.3.2 Importing and Reading the Ontology form the Curriculum Lattes

The OWL file is loaded into an application folder and a conversion or parsing is done of the ontological language to Java. After that, the OWL file is loaded onto the application and methods are then applied to manipulate the ontology and carry out the conversion into Java objects. The libraries will deal with the consul-

tation written in the SPARQL consultation language that is used to carry out the selection in the OWL file. That is, the OWL file holds the entire ontology and the values extracted from the Lattes database. In order to run the similarity comparisons, only a few fields are needed and, as a result, SPARQL is used to return only the values of interest to process the algorithm.

3.3.3 Treatment of Terms Loaded into a List

Treating the Name on a Publication Title. At this point the name on a publication title is processed. As each title has terms that are semantically insignificant for a comparison, such as prepositions, articles, and conjunctions, they are removes, as the stopwords that they are called. The project contains a file that holds 663 words in English, Portuguese, and Spanish. At the start of the execution, this file is opened only once for each publication name and words included in the list of "stopword" are removed from it.

Selecting Words in the Publication Title. After treating the list of words, a need was detected to standardize the number of words, to run the comparison. As it is not possible to determine how many words will come from each title, an arbitrary figure of 5 was then adopted for consideration. That is, should an article have 10 words after the treatment of the names, a random calculation is done that will select only 5 of the 10 words. Should it be under 5, the absolute value smaller than, or equal to, 5 is considered. One advantage of this approach is that it guarantees the balance between the comparisons, to avoid the possibility of, for example, comparing a 15-word article with a 3-word one.

3.3.4 Searching for Synonyms for each Term in the Publication Title

Once a list of treated and selected words is at hand, a search is made for the synonyms of each one. The manner in which it is represented entails two ways. The search for synonyms in the Internet and the search for synonyms in files found in the desktop environment.

In the latter case, the desktop environment was implemented due to issues produced by the Web domain such as, for example, a limitation of the number of requests made to Web services for synonyms, and an increase in the processing time of such requests.

This time depends on the web traffic speed, on the Internet connection and on the availability of the Web service which many times was overloaded due to the high number of requests made from a same IP address. In the test section the result from the comparisons amongst these two environments is tackled in more detail.

3.3.5 Calculating the Percentage of Adherence

The calculation is done through individual comparisons on a per-term basis. Should corresponding occurrences be found, that is, words that are equal in the domain of possibilities, a value 1 is attributed, to account for the quantity of such terms at the end, that is, the number of synonyms found.

The second element of the equation is the number of possibilities to be compared between two base individuals (base and target), as represented by a number of possibilities. When the algorithm is run, for both the base and target individuals, a list is loaded with the quantity of bibliographical outputs, and this number is multiplied amongst them, to produce the number of possibilities term.

As the algorithm defines 5 as the set number for words to be searched for with their synonyms, it is necessary to multiply this figure by the number of possibilities amongst the individuals. And at the end, in order to have a percentage value, the figure is multiplied by 100.

The result of this calculation allows the comparison with the other individuals that are being compared with the base list. It also allows the carrying out of a classification amongst them, in order to check which one has the largest similarity percentage.

3.3.6 Classification of Similarity

As the process is finalized, all comparisons amongst all the individuals having been loaded for the purposes of result organization and presentation, there is the visualization of the figure for similarity percentage between each comparison, that shows which individuals had the highest scores, being the most compliant ones with the base individual. Those that had smaller values move farther from the base individual and are less compliant.

4 ANALYSIS OF THE RESULTS

This section shows the analysis and the results obtained from the implementation, execution, and testing of the similarity algorithm as proposed for a controlled set of individuals. With such prior knowledge, one might pinpoint, in a more reliable way, the results expected, given that the comparisons between human perception and the computational perception could be better analyzed. That is, the validation of the tests

was based on the knowledge of the tester on the people about to be compared, with the result calculated by the similarity algorithm.

The analysis consists of considering aspects related to the concentration area each individual dwells in the academic domain. It was possible to see that the percentage of similarity in the comparisons had higher concentrations and increase trends amongst the individuals that belonged to the same semantic group of bibliographical outputs.

The result from the application of the algorithm to 15 target individuals and the distribution of the percentage according to one's adherence to the individual set as the base one is shown on Table 1. It is possible to see the relation of one single base individual with 15 target individuals, named Target 1 to Target 15.

The Lattes ID is the identifier element of the individual as registered with the Curriculum Lattes database, the concentration area represents the domain one is predominant, as related to one bibliographical production and academic life, based on one's degree qualification, research lines, master's degree, doctorate, teaching experience, thesis advisory work, examination boards, etc. Based on such information on the area do knowledge the individual is inserted in, the algorithm allows comparing and attributing a score to those that are the closest within a given concentration area.

The next column is the number of production items found and extracted from the Curriculum Lattes for each individual. With this number, a number for how many comparisons will be made is arrived at, as the total of comparisons is the Cartesian product found between the number of bibliographical outputs of the base individual against those of the target individual.

The number of equal terms column is filled after the execution of the process, which captures the number of equal occurrences found between the terms and their respective synonyms.

The percentage is shown for adherence or similarity for the individuals under comparison. This result comes from applying the formulation found in the section on the algorithm formula presentation.

The concentration area of the base individual is within the scope of the Software are, that is, it entails a domain related to Information Technology (IT) in general, to include Computer Science, Information Science, Software Engineering, Computer Engineering, or any other areas that may relate semantically with information, with software, and systems, amongst others.

Table 1 has nine of the fifteen individuals in this area. In analyzing this, evidence is found that approximate the other individuals characterized in their Software knowledge area, as shown in the Table, with the base individual also belonging to this domain.

This configuration produces a result that tends to higher values for the respective individuals. This predominance is explained by the fact that they share information that covers one same sampling realm, whether the exclusive term comes from the name of the bibliographical output or from a synonym derived from such original term.

Target individuals 7, 8 and 13 had the highest percentages for similarities. As it was to be expected, these individuals are from the Software domain, as much as the base individual, and share common ideas; the algorithm pointed that there is a higher concentration in similarity than with the other individuals compared.

The remaining target individuals from other concentration areas had their percentage within a range below those individuals from the software area. The explanation for this lies in the fact that they not have information in the outputs that are similar or that deal with the same academic domain set with some degree of connection between them.

It should be noted here that there is a likelihood where individuals apparently are part of a completely opposite area, and have a high similarity percentage. For example, one case of a research psychologist who searches for software capable of assisting with the treatment and monitoring of patients. Or still, of a librarian who seeks to innovate on the automation of assets and aims at obtaining or proposing applications to such an end.

The evidence for these cases is seen should outputs be published and should they be added to the Curriculum Lattes database of such members.

This analysis can be replicated to people or to groups of people who are unknown to each other, aimed at finding out and measuring similarities based on the bibliographical production.

It is possible to see that the evidence is satisfactory to allow concluding that: given a set of people, there is a degree of adherence amongst them, calculated from their bibliographical production. And such an adherence may vary according to the area of knowledge an individual is inserted in.

This test was validated, based on the personal knowledge found amongst the individuals compared and, by attesting that the values found match and are aligned with the reality observed.

Given that the tests pointed to such conclusion, it can be replicated to a larger number of elements of the set of individuals, which will display the same result trend, for people that share the scope of a simi-

Table 1: Result from the execution of the algorithm to calculate the similarity percentage amongst individuals (1st Scenario).

Individuals	Lattes ID	Concentration Area	No. of Production Items	No. of Equal Terms	% Similarity
Base Individual					
Base	0468265522433921	SOFTWARE	20		
Target Individuals					
Target 1	0395549254894676	SOFTWARE	19	85	4,47
Target 2	8424412648258970	CIVIL	22	4	0,18
Target 3	6753551743147880	ELECTRONICS	18	6	0,33
Target 4	1196380808351110	SOFTWARE	37	75	2,03
Target 5	2105379147123450	SOFTWARE	12	40	3,33
Target 6	7610669796869660	SOFTWARE	53	238	4,49
Target 7	0580891429319047	SOFTWARE	17	191	11,24
Target 8	9950213660160160	SOFTWARE	18	342	19,00
Target 9	1700216932505000	SOFTWARE	11	47	4,27
Target 10	0255998976169051	MEDICINE	99	4	0,04
Target 11	2443108673822680	ELECTRICAL	23	8	0,35
Target 12	7844006017790570	SOFTWARE	13	3	0,23
Target 13	2187680174312042	SOFTWARE	111	1649	14,86
Target 14	0571960641751286	POWER	38	17	0,45
Target 15	8075435338067780	PHYSICS	11	15	1,36

lar knowledge. With the goal of certifying and making the results reliable, other tests were carried out, seeking to diversify the control scenario between the domain of knowledge between the academic profiles at the University of Brasília (UnB).

Table 2 shows another situation, namely the base individual randomly chosen amongst the lecturers of the Software Engineering school of the UnB, at the Gama campus. The other members were also randomly chosen for the purposes of comparison, also at the Gama campus and at other University of Brasília departments.

They included 5 profiles in the concentration area of software, 2 from Portuguese language, 2 from the School of Economics, 2 from Electrical/Electronics Engineering, 2 from Civil Engineering, 1 from the Scenic Arts, and 1 from the School of Sociology, as shown on the said table, totaling 15 target individuals.

Algorithm processing showed that the base individual has his best approximation in terms of similarity with target Individual 8, also from the Course of Software Engineering. Without even going into the details of one's bibliographical production, it is possible to see a link, even in the name of the course, but the algorithm does not only take that into account but the context of the academic career between compared profiles, which can yield higher similarity indexes.

As regards the individual who drew the closest to the base individual, it is possible to pinpoint the largest concentration of bibliographic production items included in the same scope of production, of development, of study, and of research for both profiles.

The second target profile in high similarity with the base individual was number 10. It is also possible to see that the said profile is in the sane large knowledge area of Software and has bibliographical production items that share the same semantics exchanged between them in conversations.

It is possible to see that they deal with different subjects in their concentration areas and production items, but the algorithm points that there is a semantics between them. Even if it is just for some terms, when comparing in the generals sense of all publications, a higher similarity trend is attributed.

It is important to consider, for the purposes of calculation, that there is a variation, or margin, above or under the figure presented. It may vary, according to the tests, at approximately 3%. This variation is explained by the random selection of the terms that will be analyzed with the algorithm. Every execution can use terms different from the previous one.

Thus, there was the adherence expected amongst the people with similar concentration areas. It is worth mentioning that the result allows concluding that the individuals with a higher percentage have affinities in the bibliographical production, being in a similar context, though not necessarily equal.

Table 2: Result from the execution of the algorithm to calculate the similarity percentage amongst individuals (2nd Scenario).

Individuals	Lattes ID	Concentration Area	No. of Production Items	No. of Equal Terms	% Similarity
Base Individual					
Base	5685720614944773	SOFTWARE	11		
Target Individuals					
Target 1	5176634535321377	SOCIOLOGY	31	34	1,99
Target 2	3594383262391290	PORTUGUESE	54	6	0,20
Target 3	7201356664034110	PORTUGUESE	7	10	2,60
Target 4	4731226594888669	ECONOMICS	7	14	3,64
Target 5	1409988766720310	ECONOMICS	18	0	0,00
Target 6	2831991076751450	SOFTWARE	5	1	0,36
Target 7	9554285834432090	SOFTWARE	21	81	7,01
Target 8	5685720614944773	SOFTWARE	6	135	40,01
Target 9	4739013535126460	ELECTRICAL	35	9	0,47
Target 10	2193972715230641	SOFTWARE	37	478	23,49
Target 11	0716559775355685	SOFTWARE	26	32	2,24
Target 12	1386396456867680	ELECTRONICS	9	19	3,84
Target 13	3770883410480180	CIVIL	73	86	2,14
Target 14	0980291033230862	CIVIL	51	138	4,92
Target 15	2723749173803350	SCENIC ARTS	35	100	5,19

5 CONCLUSION

Based on the foregoing and on the starting hypothesis that it was possible to obtain some percentage of adherence amongst the profiles of the academic community at the University of Brasília, a process was devised that could prove and validate the initial questions on the existence of a connection between profiles that were unknown and randomly selected, as well as demonstrate the quantification, comparison and definition of how similar they are.

The result was obtained from consultations in the ontology created, based on the structure found in the Curriculum Lattes database. The ontology is a significant step towards the implementation of the Semantic Web, allowing the making of inferences, adding value to machine reasoning, to allow them to differentiate people who talk about subjects that are not exactly equal, but that are similar as they belong to one same concept domain, as postulated by the Semantic Web.

We took the set of individuals registered with the knowledge base of the Curriculum Lattes Platform into consideration, as it contains the blocks of elements on the bibliographical production and also justify the origin of the data to be compared. The algorithm developed allows inferring the existence of a certain degree of relationship and similarity amongst the bibliographical production items registered by such members.

This degree can vary, according to the level of similarity found in the individual comparison between a profile set as base individual whilst the others are targets. The results lead to the conclusion that the highest values between two profiles are shown to be the most adherent as regards the lowest values.

Intermediate values are aligned with a relationship of production in the middle range, containing some similarities, but no conclusion can be made as to whether they are from the same area or not or if they have big affinities. The lowest values in the scale show that the academic connections are increasingly farther and that is directly proportional to the number of similar terms found, making them more distant from the percentage of similarity.

The higher the number of blocks of elements found in the Curriculum Lattes database that are considered in the comparisons, which produce equal occurrences, the greater the potential is for the percentage of similarity to increase, allowing one to infer that the profiles compared are aligned in the University.

Having analyzed the test blocks and compared them with the Curriculum Lattes of the individuals registered with the Lattes platform, it is possible to prove that the values pointed by the algorithm provide a reliability margin between the result presented and the database consulted, thus validating the initial perspective of the proposal.

As a result, this work has been important for the academic community as it presents the implementation of a solution that calculates the similarity percentage amongst individuals, according to one's career in the Academy. It contributes to the identification of lecturers who have the highest similarity with one another and gives margin for the knowledge that exists between them, even allowing their cooperation in the same knowledge area.

REFERENCES

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*, pages 35–41.

Chalco, M., Pascual, J., Junior, C., and Marcondes, R. (2009). Scriptlattes: an open-source knowledge extraction system from the lattes platform. *Journal of the Brazilian Computer Society*, 15(4):31–39.

da Costa, A. P. and Yamate, F. S. (2009). *Semantic Lattes: Uma Ferramenta de Consulta de Informaes Acadmicas da Base Lattes Baseada em Ontologias*. Undergraduate thesis, Escola Politcnica da Universidade de So Paulo.

de Oliveira, A. V. (2011). Introduo a web semntica, ontologia e mquinas de busca. *Revista Tecnologias em Projeto*, 2(1):7–10.

Galego, F. E. and Renata, W. (2013). *Extrao e Consulta de Informaes do Currculo Lattes Baseadas em Ontologias*. Master thesis, Universidade de So Paulo (USP).

Hendler, J., Berners-Lee, T., and Miller, E. (2002). Integrating applications on the semantic web. *Journal of the Institute of Electrical Engineers of Japan*, 122(10):676–68.

Luna, J., Revoredo, K., and Cozman, F. (2013). Link prediction using a probabilistic description logic. *Journal of the Brazilian Computer Society*, 19(4):397–409.

Shadbolt, N., Hall, W., and Berners-Lee, T. (2006). The semantic web revisited. *IEEE Intelligent Systems Journal*, 21(3):96–101.

Siddiqui, F. and Alam, M. A. (2011). Web ontology language design and related tools: A survey. *Journal of Emerging Technologies in Web Intelligence*, 3(1):47–59.

Sudeepthi, G., Anuradha, G., and Babu, M. S. P. (2012). A survey on semantic web search engine. *IJCSI International Journal of Computer Science Issues*, 9(1).

Zhao, Y., Si, H., and Lang, Q. (2014). Knowledge sharing in virtual community based on rdf triple publication and retrieving to process spqrql query. *Journal of Software*, 9(7):1941–1951.