SOCIAL NETWORK ANALYSIS AND TEXT MINING TO IDENTIFY INNOVATIVE RESEARCH THEMES: SUPPORTING THE BRAZILIAN NATIONAL AGENDA FOR RESEARCH WITHIN GRADUATE EDUCATION

Alessandra Brandão

Center for Strategic Studies and Management (CGEE) – abrandao@cgee.org.br

Adriana Villela

Center for Strategic Studies and Management (CGEE) – avillela@cgee.org.br

Carlson Oliveira

Center for Strategic Studies and Management (CGEE) – carlson@cgee.org.br

André Brasil

Brazilian Federal Agency for Support and Assessment of Graduate Education (CAPES) – andre.brasil@capes.gov.br

Alexandre Santos

Center for Strategic Studies and Management (CGEE) – asantos@cgee.org.br

Abstract

This paper describes the results of a new methodological application on Future-Oriented Technology Analysis (FTA), which was based on social network analysis. This methodology combined the results of a survey designed to identify significant innovative research themes, with the scientific production from faculty members and the student body of graduate courses. The final result was a relevant contribution to the continuous improvement of the research agenda associated with the Brazilian System of Graduate Courses (SNPG).

In fact, such agenda is a pivotal part of the Brazilian National Graduate Plan (PNPG), a decennial program designed by the Brazilian Federal Agency for Support and Assessment of Graduate Education (CAPES). The plan's primary goals are to examine the current situation of the SNPG, forecast its expected growth and guide the design of public policy to promote scientific and technological production in Brazil.

To contribute to the making of the agenda mentioned above, our survey consulted the Graduate Program Directors (GPD) in the country, inquiring about their ideas on strategic and innovative research themes for the next decade. Since most of the Brazilian research takes place within graduate courses, their input was of utmost importance, providing over 8,000 theme suggestions.

Since it was expected that not all of these themes would be genuinely innovative, sometimes reflecting the current research from graduate courses, a cross-checking network analysis was conducted, and gave rise

to 214 semantic and co-authorship networks derived from data obtained from the scientific production conducted by both faculty members and graduate students.

Considering the volume of data available and the number of networks generated, exploration would be challenging for decision-makers. For that reason, this study also explored new ways of visualizing treated data to help communicate findings using network representations.

Finally, the network analysis performed was able to identify which of the suggested themes reflects research already under development, and which ones were indeed looking into the future. It was also possible to get an essential insight into the scientific and technological competencies in Brazil, as well as support the decision-making processes associated with the programs coordinated by CAPES.

The next step of the study will include not only data obtained from Brazilian databases but, rather, those obtained from international sources in order to design better international cooperation initiatives.

Keywords: social network analysis; innovative research themes; text mining and data visualization

Introduction

The Brazilian System of Graduate Courses (SNPG) has counted with different instruments to guide its development over the years. Among them, one of the most important is a national plan, which was designed in the early 1970's to establish and support the improvement of human resources related to science and technology activities in Brazil. (Guimarães & Humann, 1995)

The first edition of this Brazilian Plan for Graduate Education (PNPG 1975-1979) aimed to stimulate the expansion of graduate courses in the country, which had been taking place in an essentially spontaneous way. Since it was believed this evolution should happen in a coordinated fashion, subject to state planning and integrated with social and economic development policies, the original PNPG became the first real strategic plan for graduate education in Brazil. (Hostins, 2006)

Many years later, the sixth edition of the PNPG aims to define new guidelines, strategies, and goals for the advance of graduate education and research within the period of 2011-2020. In its core, this plan addresses the need for a continuous evolution of the system used to evaluate the quality of the courses, the importance of interdisciplinarity in research, the need to reduce regional asymmetries, the importance to train high-quality human resources for the productive sector, as well as the need to prioritize research in strategic areas, such as health, environment, energy, agriculture, defence and so on. (BRASIL. Ministério da Educação. CAPES, 2010)

From the research priorities mapped in the most recent PNPG, the special committee in charge of the plan's monitoring started designing a National Research Agenda to be conducted within Brazilian graduate education. According to a committee's report (Comissão Especial de Acompanhamento do PNPG 2011-2020, 2014), its formulation should consider major research themes, articulating the views of different funding agencies and other relevant social actors.

The research detailed in this paper is part of such effort, as it was commissioned by the Brazilian Federal Agency for Support and Assessment of Graduate Education (CAPES) to support the committee in charge of developing the planned agenda.

This work’s initial premise was based on the decision to consult every single graduate program in Brazil to consider which themes could be part of the aforementioned agenda. Since there were over 4,000 such programs by the start of this research, and they were organized into 9 broad groups and 49 evaluation fields, it was necessary to adopt a new methodological

application on Future-Oriented Technology Analysis (FTA), based on social network analysis, to deal with the potential complexity of the results, as discussed in the following section.

1. Background

1.1 Social network analysis, text mining and FTA

The social network analysis (SNA) studies the properties of individuals considering both local and global contexts, seeking to highlight how the individual relates to the whole and how the whole affects the individual (M. NEWMAN, 2010). This set of techniques differs from the usual statistical analysis, not focusing on actors or reductions from models, but rather, on the relational information between the objects of study, allowing both qualitative and quantitative analyses. Any set of entities that have, between pairs of its elements, symmetrical or asymmetrical relations that can be explained and quantified as characteristics, meanings, activities or origins, among others, can be defined as a network.

The analysis of networks has nodes and edges as basic objects of study. The first ones represent the elements of the network, and the latter outline the relations between nodes. At the node, individual quantitative or qualitative attributes of network elements can be represented by characteristics such as color, size, or format. On the other hand, edges can express attributes such as the intensity or weight of the relationship (according to the line thickness), whether it is directed or symmetrical and what type of relationship exists between nodes, for example.

Complementing the network analysis on this study, we have text mining as an extension of data mining. This process can be defined as one of extracting unknown and useful information from textual documents written in natural language. The main techniques used to mine text are:

- Natural Language Processing: It is a method that seeks to use computers to improve the understanding of natural language. This is achieved through the use of techniques which result in quicker text processing, especially by manipulating strings from survey results.
- Information Retrieval: Uses statistical or semantic methods and measures to automatically process text from documents to find which ones have the answers to the question (but not the answer itself). Even though such techniques were already employed since 1975, this method only gained notoriety with the popularization of the Internet.
- Information Extraction: Its main objective is to search for relevant chunks of text in a document, in order to extract specific information from these parts. Although this is a more limited technique to deal with natural language comprehension, it is still widely used in data mining, especially in social network analysis.

As we can see, data and text mining are methods for automatically extracting patterns and trends from complex information sources, in order to build network analysis that would not be possible otherwise. Such kinds of study and FTA share the same philosophical basis with systems theory and complexity theory which, according to Nugroho and Saritas (2009), might be used to conduct research on:

- analysis of the data generated by an FTA exercise, containing the opinion of experts, answers to queries and related statistics. The main benefit here is that the focus of analysis may shift from the attributes of the individual to the relationships and alliances between individuals in the network.

- generation of subsidies to define the limits and scope of an FTA exercise, helping to decide which themes are relevant and how they relate.
- identification and selection of key actors by understanding their position or relative importance in their networks over time.
- capture weak signals of change that can enrich understanding of systems under analysis and help structure a common agenda to facilitate transitions of such systems in the future.
- foster interdisciplinary collaborations and actions among relevant actors to foster transitions of the system under review.

The potential combinations between FTA and the network analysis described above were previously mapped by Scapolo and Porter (in CAGNIN et al., 2008), who suggest that network analysis can support three of six families of FTA methods: monitoring, trend analysis plus simulation, and modelling (the other three families would be creative methods, expert opinion, and scenarios). In their analysis, the authors conclude that three themes deserve attention for the development of new methods for use in FTA activities:

i) use of advanced tools to deal with all kinds of data and information, including search, mining, organization, visualization, and interpretation of information resources available in electronic media;

ii) methods capable of extracting, organizing, comparing and combining a diversity of human judgments, including interests and opinions, thus giving greater robustness, quality and scientific validity to participatory approaches;

iii) network collaboration tools that enable the broad contribution of anyone, anywhere. In this context, network analysis approaches have the potential to positively impact the first two themes.

In this article, we discuss two kinds of such networks. In the first one, there is only one type of edge present, so we use SNA to exhibit the relation between nodes which were generated from a survey administered to the graduate programs in Brazil. In this case, the network shows which topics have the strongest relations. For the second kind of network, there is more than one possible relationship between nodes. In this case, the networks will have multiplex edges (MAIA et al, 2015), representing either co-authoring or contextual similarity relations.

2. Methodological approach

When it was first resolved to consult the graduate programs about research themes for the national agenda, some decisions had to be made. The first one related to the unit for consultation. From the possibility to survey every researcher to the alternative of only contacting the higher education institutions, a choice was made for the Graduate Program Directors (GPD). The idea is that they would be asked to answer a survey in a way that reflected the ideas of their whole research teams. With that, it would be possible to reach the whole system, but with a more manageable number of potential responses.

2.1 Identification of strategic themes: an electronic survey

By the time we started designing the survey which would be conducted with the Brazilian GPD, one of the basic premises employed to choose its questions was to encourage respondents to

provide a large volume of textual answers. This way it would be possible to conduct a broad semantic analysis of the results, especially through the use of text mining techniques.

So, we first started by giving respondents the opportunity to suggest five different themes they considered to be relevant for the future of research in Brazil. For each one of them, they should include name, description, and keywords. Besides that, they had the possibility to defend each answer by talking about its strategic importance from the regional, national and international points of view, the last one also relating to its relevance for the field of knowledge in general. Fig. 1 brings a graphical representation of this questionnaire's design.

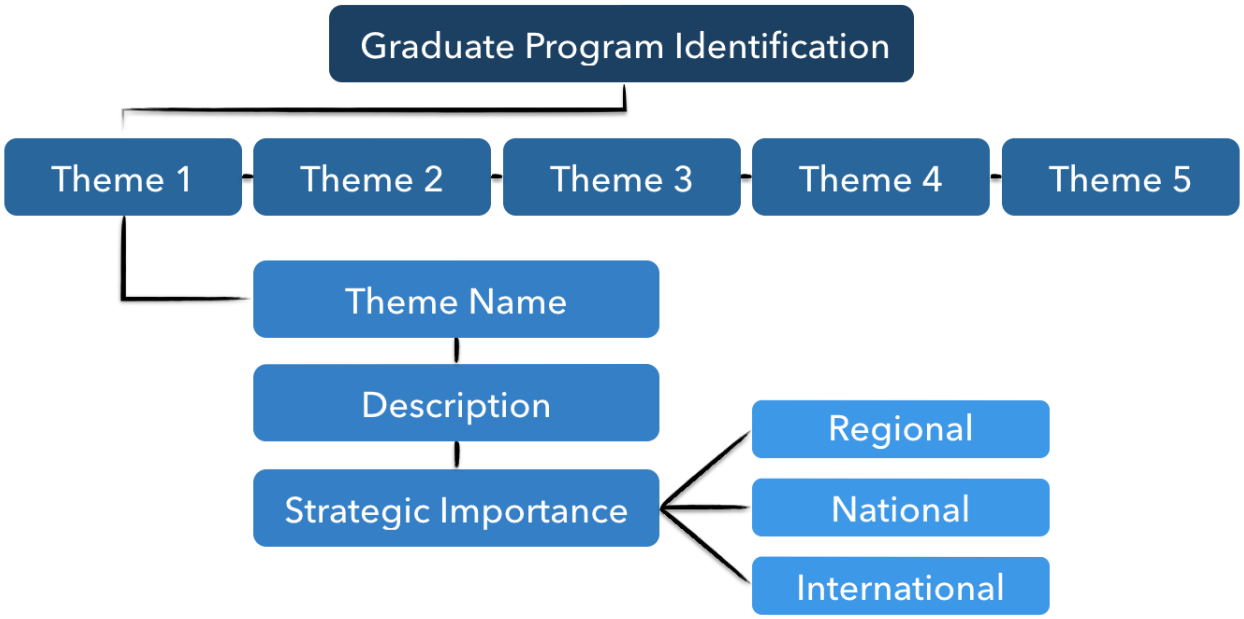

Fig. 1. Preliminary structure of the form, provided by CAPES.

Once the questionnaire was ready, the electronic survey was prepared to be carried out electronically using CGEE's insightSurvey tool (IS), which allows: a) configuration of electronic forms with several types of questions; b) managing participants; c) setting up application rounds with automatic closing; d) directing communication with the participants; e) sending invitations; f) reinforcing invitations to respondents; (g) including additional information to support respondents over the course of the survey's application; h) extracting data for integration with treatment tools and databases; i) monitoring real-time responses by basic descriptive tabulation.

In order to ensure that the comprehension of the questions was adequate, as well as to test the navigational structure of the electronic form, two pre-tests were carried out with representatives from CAPES' Technical and Scientific Committee of Higher Education (CTC-ES) and from the agency's different evaluation fields.

After some minor changes, the final questionnaire was certified so that the survey could be conducted within the universe of 4,815 GDP[1].

[1] A total of 4,447 distinct graduate programs were consulted, but some were associations between different higher education institutions, counting with more than one GPD per program. That raised the number of consultants to 4,815.

2.2 Strategic themes: generating clusters

After the survey was conducted, the final number of respondents was of 3,281, representing 68.14% of the universe consulted. Since each respondent could suggest up to five research themes, these 3,281 GPD provided a total of 8,643 themes. Because of this volume of data and the number of networks which could be generated from it, exploration would be challenging for decision-makers to extract the information they might need to contribute to the National Agenda for Research within Graduate Education (ANPPG).

For that reason, this study also explored new ways of visualizing treated data to communicate findings using network representations. This was done by grouping the datasets into networks and sub-networks of themes by evaluation field. Through the use of text analysis algorithms, nodes representing each of the themes suggested were created and clusterized by semantic similarities identified between titles and descriptions. Edges were also included to show how the nodes related among themselves.

Accompanying each node, as related metadata, was also the original text describing the theme, their corresponding strategic relevance (regional, national and international) as well as the keywords indicated by the GPD to characterize the suggestion. All this was done using CGEE's insightNet Browser, a network visualization tool accessible from Mozilla Firefox that can present clearly the results of the conducted survey.

The necessary steps to build the strategic theme networks are displayed in Fig. 2.

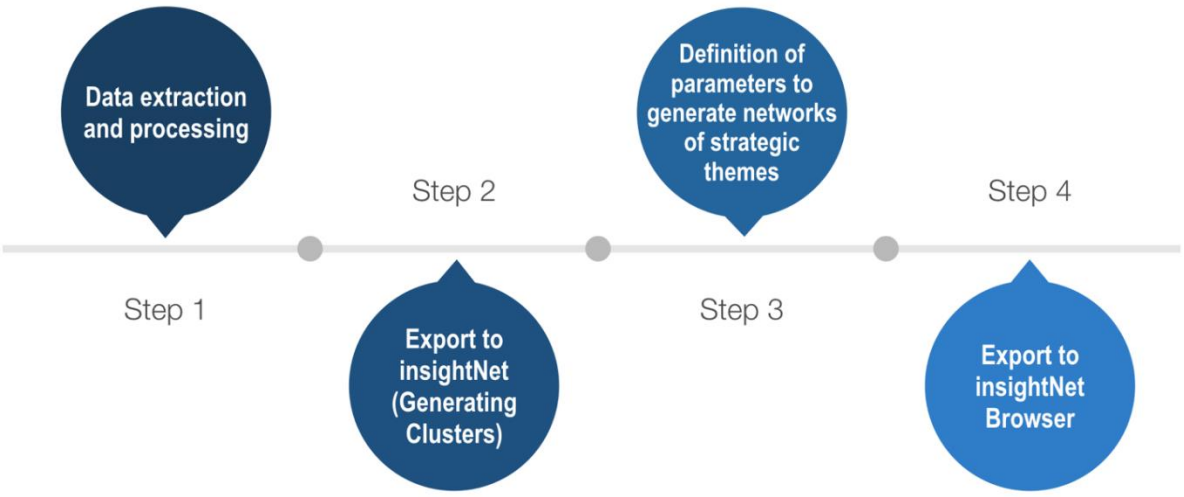

Fig. 2. Steps to build the strategic theme networks

The whole process of data extraction and processing was attended closely by the research team. Because of that, it was noticed that the evaluation fields of 'Anthropology and Archeology' and 'Education' presented results that did not follow the typical pattern of the other 47 fields. In both cases, we noticed a high incidence of duplicated or extremely similar responses, resulting in clusters with an unusual level of similarities.

This duplication of answers was exhaustively investigated in search for possible problems in the database generated by the survey. By the end of the analysis, it was confirmed that the graduate program coordinators in these fields had discussed the survey questions in advance, electing the answers which should be provided by all programs. With that, there was no nonconformity verified within the final data, and we could proceed to the next step of the analysis.

At this point, the task at hand was to export the already cleaned data to insightNet, in order to clusterize the themes and build the networks. The result from this phase included nodes and edges generated from the semantic similarity between descriptions of each theme, and it also contained all of the metadata shown in Table 1.

Table 1 - Attributes, metadata and respective descriptions used in the construction of strategic theme networks.

ATTRIBUTE	DESCRIPTION
NODE	Representation of titles texts and descriptions of themes. The diameter of each node is proportional to its number of edges
EDGE	Numerical representation of the similarity between two nodes, with values varying between 0 and 1.
KEYWORD	Set of keywords provided by GPD
DESCRIPTION	Description of the themes, as proposed by the GPD
NATIONAL	Relevance of the theme proposed to Brazil, according to its strategic importance.
REGIONAL	Regional relevance of the theme proposed, according to its strategic importance.
INTERNATIONAL	International relevance of the theme proposed, according to its strategic importance.
NUMBER OF EDGES	Number of node interactions defined by semantic similarity

From this developed data structure, the third step was to review the generated networks to identify relationships. This parameterization procedure aimed to produce groups which were semantically and hierarchically connected, according to similarities between themes. Such groupings would later be evident in visualization parameters of the networks, from node color to edge distribution.

Finally, the generated networks were exported to the visualization tool adopted for this research, insightNet Browser, a technology that enables experts to have a quick and interactive overview of the information collected, even for large sets of documents.

2.3 Co-authoring and semantic similarity networks of the scientific production

Acknowledging that the survey was conducted on behalf of CAPES, also a funding agency, it was anticipated that the GPD consulted would not only mention genuinely innovative research

themes on their answers but would also report the current research they conducted on their graduate courses. This was an expected result since some directors might think they could lose funding if their everyday research was not considered relevant or innovative.

As an outcome of this foreseen bias in the survey's results, a cross-checking network analysis was in order, combining the answers from the GPD with two more types of networks designed from the scientific production from faculty members and the student body of graduate courses.

The first kind of network associated researchers based on co-authorship algorithms, identifying those who published together and building networks based on that information. The other type was built according to semantic similarities amidst researchers' scientific publications. The similarity algorithm employed allowed the design of networks by the resemblance between their curricula, even if the faculty members or students had not published together.

The steps for designing the scientific production networks mentioned above were the same ones presented in Fig. 2, and the data necessary for this part of the research came from crossing datasets from CAPES and the National Council for Scientific and Technological Development (CNPq).

From CAPES, we obtained a comprehensive list of faculty members and students from all graduate courses involved in this research. Such list brought information from the years 2013 to 2015 (inclusive) and included ID_PROFESSOR and ID_STUDENT, with their respective Brazilian Taxpayer Registry (CPF), an official identifier of individuals which is both unique and permanent. Furthermore, CAPES' database also provided a corresponding ID_LATTES for each person, a piece of information that allowed the crossing with the second dataset.

This was a dataset obtained from the Lattes Platform, an information system, maintained by CNPq, that includes an extensive set of information from individual researchers in Brazil. From the ID_LATTES provided by CAPES, the curricula of the identified faculty members and graduate students were extracted in XML format from the Platform.

The curricula obtained was distributed among the nine broad groups and 49 evaluation fields in which the Brazilian graduate courses are divided. With that, the total number of curricula available at this phase of the study was of 87,019 for faculty members and 381,491 for enrolled students. Since some professors or event students can be involved in more than one graduate course, sometimes in distinct evaluation fields, these sets of curricula contain repetitions.

With this volume of data available, the building of the networks based on semantical similarity and co-authorships was performed through the use of CGEE's insightNet. This tool allowed for the identification of faculty members, graduate students, and the institutions to which they were associated. In addition to that, we were also able to identify keywords related to their respective fields of knowledge, according to a set of parameters adopted for the insightNet configuration to ensure the homogeneity of the information collected. These parameters included:

- Publication types selected - thirteen types of products were considered valid for this research: bibliographic production; article published in journals (complete, abstract); papers published in proceedings (complete, abstract, extended abstract); published books; published book chapters; organization of published work; patents and registrations; software; protected cultivar; industrial design; trademark; integrated circuit topography; registered cultivar.
- The value of the minimum degree of similarity between publications was established at 90.

- Publications with a degree of similarity of less than 0.15 were ignored.
- The distributions selected for the visualization of the networks were ForceAtlas2, OpenOrd or YifanHu proportional followed by Noverlap to correct the overlap of nodes.

From the parameters described above, co-authoring and semantic similarity networks were built for both faculty members and graduate students according to the programs evaluation fields and broad groups, as shown in Table 2. From that, the final networks were exported in the .gexf format, so they could be imported into the insightNet Browser.

Table 2. Number of processed curricula from faculty members and graduate students, by type of network

GROUPING	SUBJECTS	TYPE OF NETWORK	NETWORKS	PROCESSED CURRICULA
BROAD GROUPS	Faculty members	co-authorship	9	87,019
BROAD GROUPS	Faculty members	semantics similarity	9	87,019
EVALUATION FIELDS	Faculty members	co-authorship	49	87,019
EVALUATION FIELDS	Faculty members	semantics similarity	49	87,019
EVALUATION FIELDS	Faculty members + graduate students	co-authorship	49	468,960
EVALUATION FIELDS	Faculty members + graduate students	semantics similarity	49	468,960
TOTAL			**214**	**1,285,996**

Finally, it is worth to mention that the effort to build such networks provided us with two complete sets of keywords. The first one came from the suggestions of strategic themes by the GPD, and the second was a result of the analysis of current research within graduate courses. By comparing this two groups of keywords we were then able to identify which research suggested as innovative on the survey was already being conducted within graduate education in Brazil. Of course, from that, we were also able to identify the ones that were, in fact, unprecedented.

3. Results, discussion, and implications

As it was mentioned above, a total of 3,281 graduate program directors answered the survey for this research, representing 68.14% of the universe consulted. Even though the GPD could suggest up to five research themes on their answers, for the purpose of this study, everyone who suggested at least one theme was considered a respondent. With that, a total of 8,643 themes were raised by the end of the survey.

This volume of answers was considered to be very good, and the percentage of respondents exceeded 65% in eight of the nine broad groups of evaluation fields in which graduate programs are organized. For instance, the group of 'Engineering', consisting of four distinct fields of evaluation, reached a percentage of 72.77% of respondents.

From the point of view of the fields of evaluation, the highest percentages of responses were in 'Biological Sciences III' (88.89%), 'Social Work' (88.57%), 'Nutritional Science' (86.21%), 'Town Planning and Demography' (85.11%) and 'Anthropology and Archaeology' (82.76%). It is important to note, however, that there was a high percentage of responses in all fields with fifteen of them surpassing 70% of respondents; twenty-one with more than 60%; and eight over 50%.

From the themes suggested at the survey, we proceeded to the next step of the analysis, with the development of strategic theme networks.

3.1 Strategic theme networks

The 8,643 themes the respondents provided were distributed according to the 49 evaluation fields adopted by CAPES to organize graduate education in Brazil. In order to present these themes in a way that would be easy for decision-makers to visualize and navigate, this study chose to exhibit the treated data using network representations.

As an example of that, Fig. 3 brings a screen capture of the interactive strategic theme network for the 'Astronomy and Physics' evaluation field.

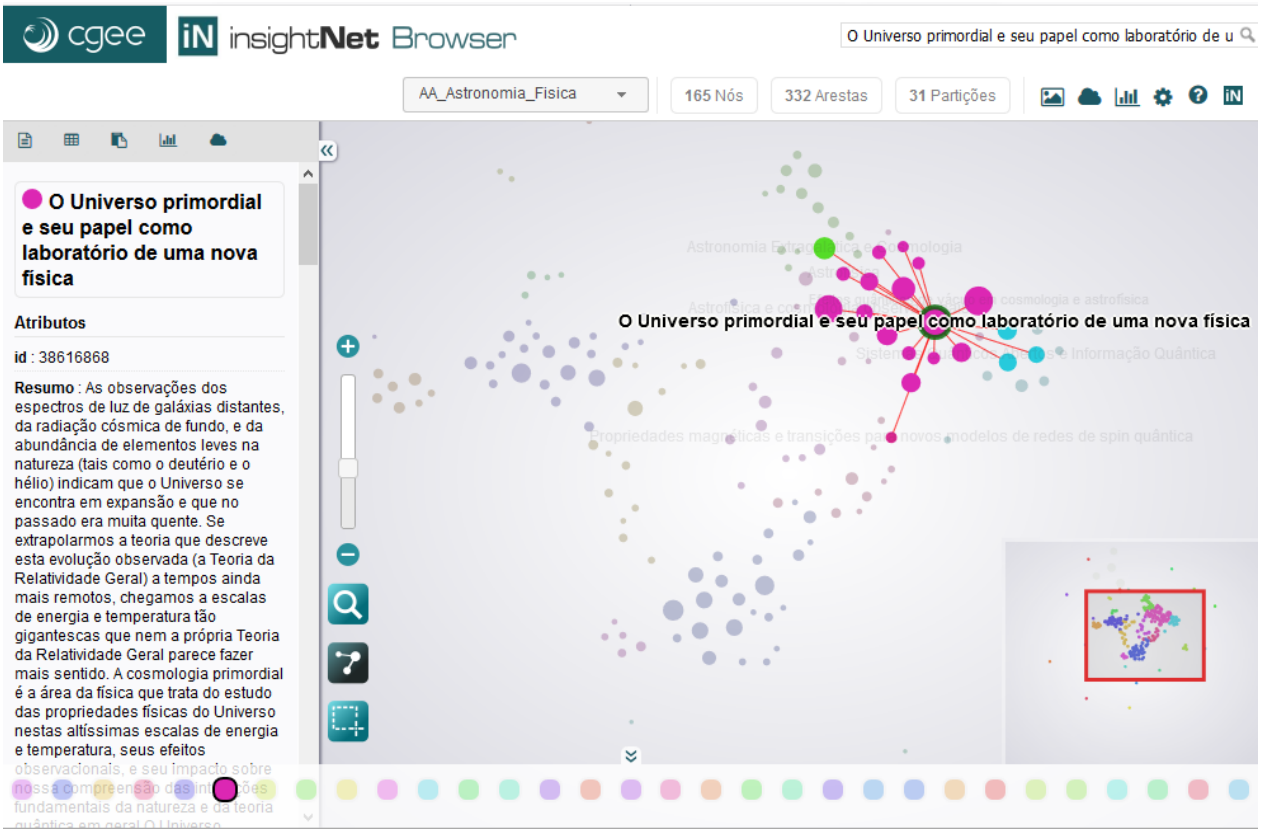

Fig. 3. Strategic theme network of the Astronomy and Physics evaluation field

On this view, the panel at the left side displays all the metadata of the selected theme, including title, name and institution of the GPD, description and strategic relevance for the theme

(national, regional and international). The map shows each strategic theme (node) and the semantic similarity (edge) between all strategic themes for each evaluation field and broad group. The clusters of similar strategic themes are identified by node colour.

For the evaluation field of 'Astronomy and Physics', the graduate program directors suggested a total of 165 strategic themes, which are clustered by semantic similarity into 31 groups, and establishes 332 distinct relationships, as represented by the edges of the network.

Through the dynamic exploration of this network, field experts can easily identify research trends, find recurrent themes from distinct graduate courses, identify potential research groups and even explore themes which do not relate to the main clusters and that can either be irrelevant to the research conducted within the field, or can be the kind of innovative proposal this research is looking for.

Another way to explore the generated networks is through keyword clouds, another function of the insightNet Browser. In Fig. 4, we see the keywords extracted from the same network exhibited in the previous image, 'Astronomy and Physics'. In this case, the word size represents the frequency of each term in the set of responses.

Fig. 4. Keyword cloud generated from the Astronomy and Physics evaluation field network

3.2 Co-authoring and semantic similarity networks of the scientific production

As described throughout the presentation of the methodological approach for this research, the second part of the study consisted on analyzing the current scientific output from Brazilian graduate courses, to compare that to what GPD stated to be strategic and innovative research themes for the next decade.

So, from a comprehensive list of faculty members and graduate students, we were able to build two types of networks, either by co-authoring relationships amongst researchers or through existing semantic similarity among their curricula. Fig. 5 and Fig. 6 presents examples of such networks for the 'Astronomy and Physics' evaluation field, where the first one is a co-authorship network and the second one is based on semantic similarity.

Each of the screen-captures come from interactive panels, where the size of the node represents the volume of scientific production for each researcher. On the dynamic version of these panels, by clicking on the nodes we may see green, red and black edges, representing co-authorship relationships, semantic similarity or both, respectively.

From a comparison between the two panels, you may notice that the similarity semantic network is denser than the one with co-authorships (13,435 vs. 3,896 edges, for the same number of 1,998 faculty members). This result demonstrates a potential for collaboration that can be explored since there are many similar curricula that could lead to future co-authorships.

It is also important to mention that these networks were built both showing relationships among faculty members, only, and also including graduate students along with the faculty. These more extensive networks provided not only a broader view of each field's scientific production but also allowed us to verify how co-authoring among students and their own supervisors is taking place within graduate courses.

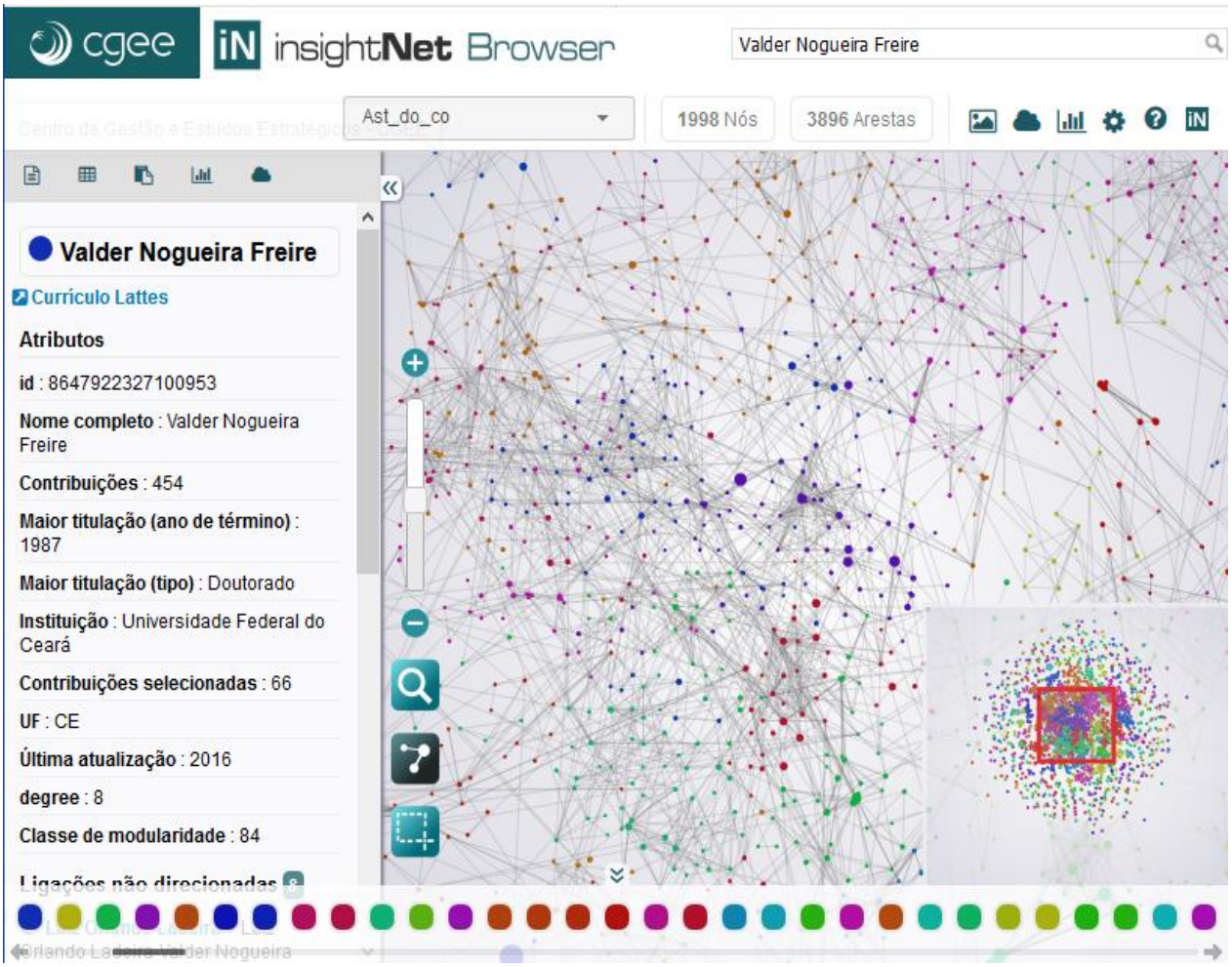

Fig. 5. Scientific production network of faculty members of graduate courses in the Astronomy and Physics evaluation field (by co-authorship)

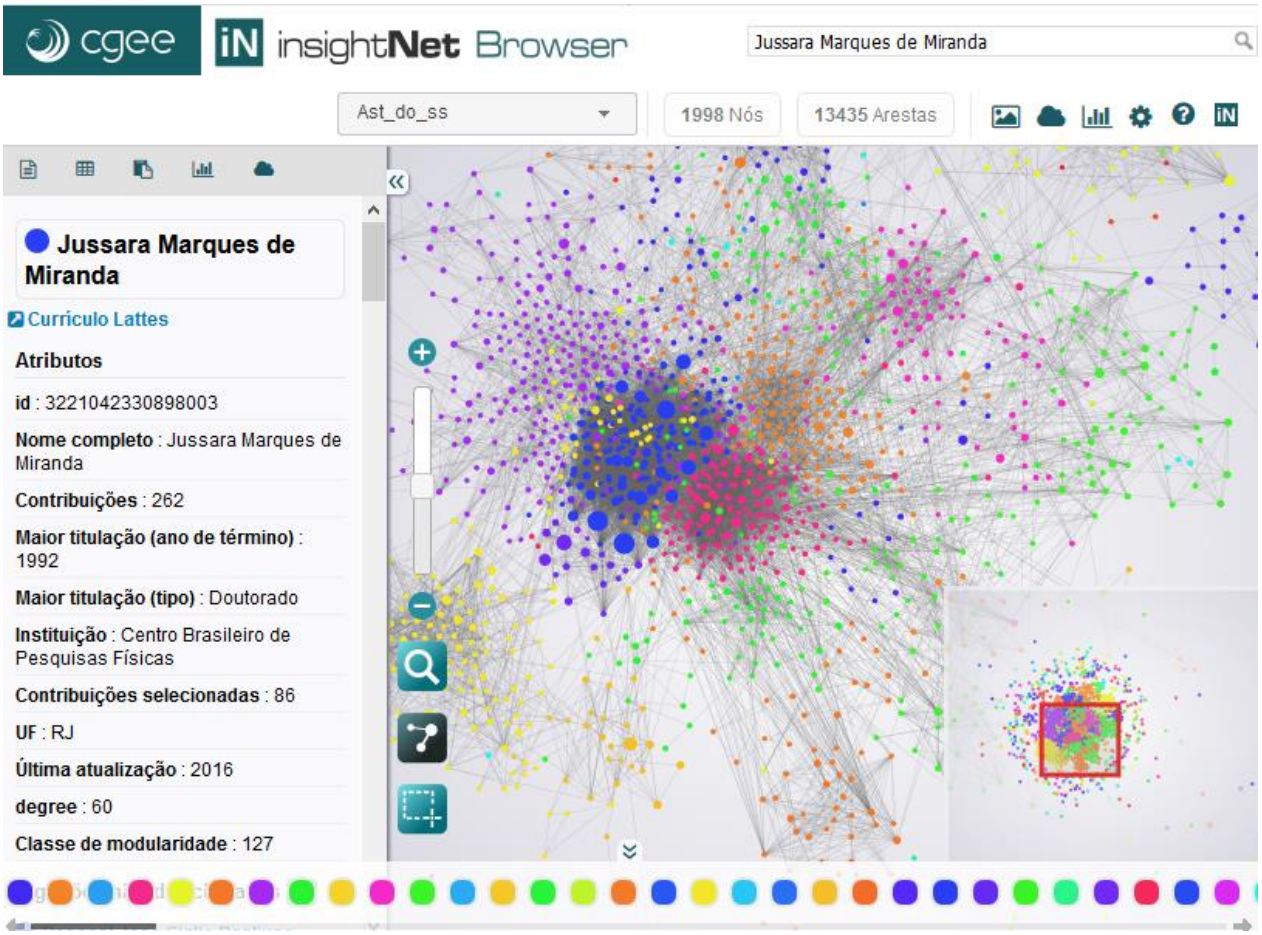

Fig. 6. Scientific production network of faculty members of graduate courses in the Astronomy and Physics evaluation field (by semantic similarity)

3.3 Comparing strategic theme networks with scientific production networks

After asking the Brazilian GPD about innovative research themes for the future of graduate education in the country, and mapping the scientific production already being conducted by their courses, it was time to cross-check this two groups of information in order to answer a pivotal question: are the themes proposed really about the future of research, or are they only a mere representation of what is already done within the courses?

To do that, this study extracted keywords from both the GPD suggestions and the scientific production of faculty members and graduate students (mapped from Lattes Platform), and compared them, considering their respective evaluation fields and broad groups.

To conduct this comparison, each keyword underwent treatment consisting of withdrawing special characters, treating the plural of simple terms, and correcting obvious typing errors, and so on. Each change was then registered with the objective of maintaining traceability.

After the identification and preparation of the keywords, it was verified which ones were a match between the two sets of data at hand. From that crossing, we got four distinct groups of information:

- Survey keywords - Number of distinct keywords, per evaluation field, which resulted from the survey with GPD;
- Lattes keywords - Number of distinct keywords identified in the curricula of faculty members and graduate students from each evaluation field;
- Found keywords - Number of survey keywords which were found in Lattes curricula, within evaluation field researchers (faculty + students);
- New keywords - Number of keywords which were proposed by GPD on the survey but were not found within researchers' curricula in the field.

This final information is very relevant for this study because it helps us find research themes that were suggested by graduate program directors as innovative and that might be, in fact, a new proposal for the future of research.

Of course, these results demand further investigation by specialists in each field, so we provided a visualization interface, built using Tableau Desktop software, to allow access to the high volume of keyword clouds generated. Fig. 7 shows the first page of the interface designed, exhibiting the four groups of information mentioned above, as well as the percentage of keywords from the survey which might represent innovation. From this page, analysts can explore the whole set of keywords per evaluation field or broad group, contributing to the analysis of the distinct sets of networks also provided.

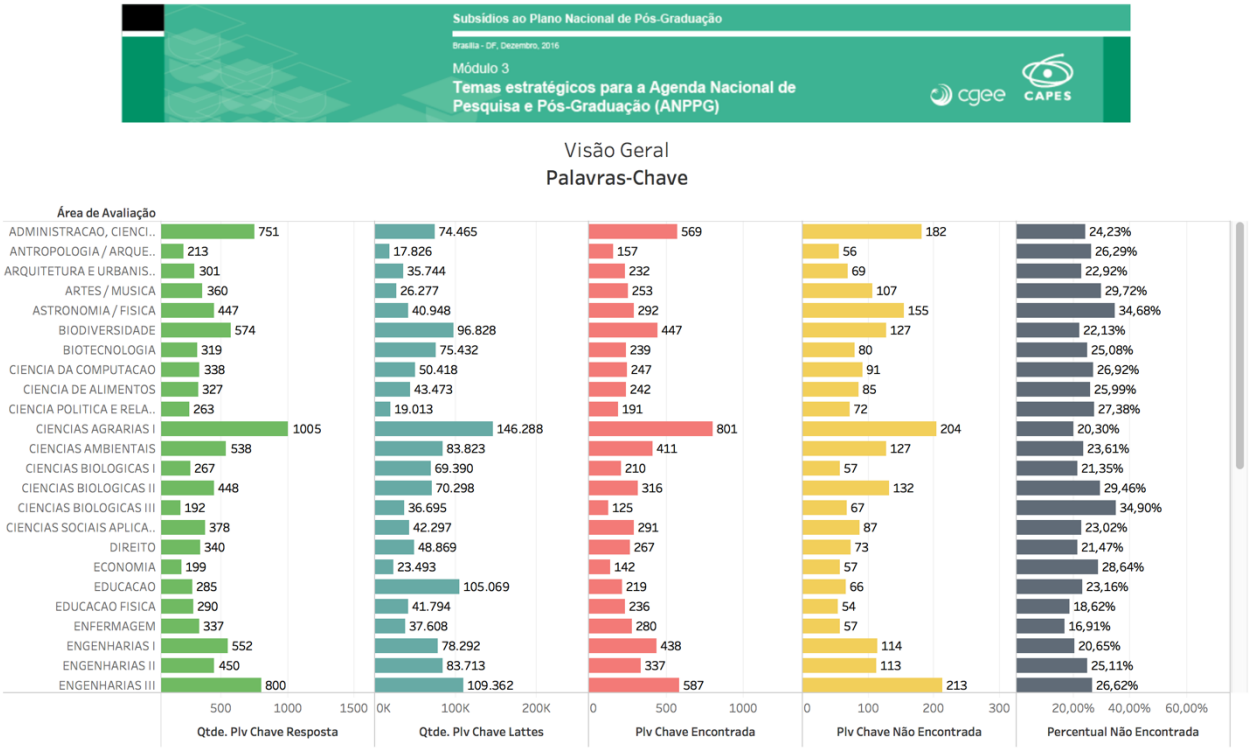

Fig. 7. Tableau dashboard presenting results from keyword cross-checking, according to evaluation field

As a reference, Table 3 shows information on keyword cross-checking for each of the 49 evaluation fields considered in this study.

Table 3. Co-occurrence of keywords extracted from the strategic theme networks and the scientific production of graduate courses[2]

EVALUATION FIELD	SURVEY KEYWORDS	LATTES KEYWORDS	FOUND KEYWORDS	NEW KEYWORDS	% OF NEW KEYWORDS
BUSINESS AND ADMINISTRATION, ACCOUNTING AND TOURISM	751	74,465	569	182	24%
ANTHROPOLOGY AND ARCHAEOLOGY	213	17,826	157	56	26%
ARCHITECTURE, INTERIOR AND INDUSTRIAL DESIGN	301	35,744	232	69	23%
ARTS	360	26,277	253	107	30%
ASTRONOMY AND PHYSICS	477	40,948	292	155	35%
BIODIVERSITY	574	96,828	447	127	22%
BIOTECHNOLOGY	319	75,432	239	80	25%
COMPUTER SCIENCE	338	50,418	247	91	27%
FOOD SCIENCE AND TECHNOLOGY	327	43,473	242	85	26%
POLITICAL SCIENCE AND INTERNATIONAL RELATIONS	263	19,013	191	72	27%
AGRICULTURAL SCIENCES	1,005	146,288	801	204	20%
ENVIRONMENTAL SCIENCES	538	83,823	411	127	24%
BIOLOGICAL SCIENCES I	267	69,390	210	57	21%
BIOLOGICAL SCIENCES II	448	70,298	316	132	29%
BIOLOGICAL SCIENCES III	192	36,695	125	67	35%
JOURNALISM AND INFORMATION	378	42,297	291	87	23%
LAW	340	48,869	267	73	21%
ECONOMICS	199	23,493	142	57	29%
EDUCATION	285	105,069	219	66	23%
PHYSICAL EDUCATION, THERAPY AND REHABILITATION	290	41,794	236	54	19%
NURSING	337	37,608	280	57	17%
ENGINEERING I	552	78,292	438	114	21%
ENGINEERING II	450	83,713	337	113	25%
ENGINEERING III	800	109,362	587	213	27%
ENGINEERING IV	581	79,745	395	186	32%
TEACHING AND LEARNING	503	63,612	376	127	25%
PHARMACY	378	56,005	269	109	29%
PHILOSOPHY AND ETHICS	165	1,522	75	90	55%
RELIGION AND THEOLOGY	120	11,645	87	33	28%
EARTH SCIENCES	393	53,471	266	127	32%

[2] Evaluation fields ordered alphabetically, according to their names in Portuguese. This way it becomes possible to compare the table with the panel exhibited in Fig. 7.

EVALUATION FIELD	SURVEY KEYWORDS	LATTES KEYWORDS	FOUND KEYWORDS	NEW KEYWORDS	% OF NEW KEYWORDS
GEOGRAPHY	265	47,957	222	43	16%
HISTORY	383	45,319	303	80	21%
INTERDISCIPLINARY	1,332	187,698	1,065	267	20%
LITERATURE AND LINGUISTICS	604	86,345	479	125	21%
MATHEMATICS AND STATISTICS	392	26,411	223	169	43%
MATERIALS SCIENCE	171	35,003	126	45	26%
MEDICINE I	371	94,277	281	90	24%
MEDICINE II	467	88,690	352	115	25%
MEDICINE III	242	37,056	183	59	24%
VETERINARY MEDICINE	394	67,500	319	75	19%
NUTRITIONAL SCIENCE	153	20,331	116	37	24%
DENTAL STUDIES	371	55,500	306	65	18%
TOWN PLANNING AND DEMOGRAPHY	285	27,512	197	88	31%
PSYCHOLOGY	421	51,901	354	67	16%
CHEMISTRY	365	87,945	268	97	27%
PUBLIC HEALTH	331	54,476	260	71	21%
SOCIAL WORK	237	15,854	173	64	27%
SOCIOLOGY	235	35,419	197	38	16%
ZOOTECHNICS AND FISHERIES	393	61,032	318	75	19%
TOTAL	**19,556**	**2,849,641**	**14,739**	**4,787**	**25%**

With that, the total number of keywords identified from the survey results came to 19,556. After analysis from the 2,849,641 obtained from faculty members and graduate students curricula, we identified that 4,787 were not found, which could mean that they represent research that is still not being conducted within Brazilian graduate education.

Another interesting thing we notice is that the results are quite different among evaluation fields. Even though an average of 25% of keywords were not identified as current research, there are fields like ‘Philosophy and Ethics’ and ‘Mathematics and Statistics’ that showed a high percentage of new keywords (55% and 42%, respectively), while ‘Sociology’ and ‘History’ presented only 16% of different keyword from those in their scientific production.

Conclusions

The main goal from this research was to provide decision-makers in charge of building the Brazilian National Agenda of Research within Graduate Education (ANPPG) with information to subsidize their work. In order to achieve that, it was decided to look for such information inside graduate programs and, with over 4,000 of them in the country, that was a great challenge.

So, after the decision to present a survey to every graduate program director in Brazil and designing a questionnaire which could result in thousands of research themes suggested to

compose the ANPPG, the hurdle was for a way to present these results to the Brazilian Agency for Support and Evaluation of Graduate Education (CAPES), which requested the study.

The decision was to make use of a new methodological application on Future-Oriented Technology Analysis (FTA), which was based on social network analysis. With that, we were able not only to present the results of the survey in an interactive and intuitive way but also include additional networks of scientific production within graduate education.

From the crossing of these two networks, and the keywords associated to them, we were also able to suggest which of the proposed research themes might actually be new, and which were already reflecting current research within the courses. But, of course, all the data that was acquired and the visualization tools with which they were presented became just a means to the final result on the horizon.

The desired conclusion is that of a new ANPPG, and it is essential for committees of experts from each of the 49 evaluation fields at CAPES to analyze the results achieved so far. Only through their expertise, it will be possible to make sense of such a valuable study, in order to find the pearls among the sand from the themes suggested by the GPD and their relation to the current scientific research that was already mapped.

This necessary step for the new agenda of research is planned to take place in the third quarter of 2018, when representatives from the 49 evaluation fields are expected to attend nine different meetings at CAPES, one for each broad group of fields. From that, we expect to obtain final reports of future research believed to be relevant and innovative for each group, and these results will aid the Special Committee in charge of following-up on the Brazilian Plan of Graduate Education to update the ANPPG for another decade of advancement in the country's research.

References

BRASIL. Ministério da Educação. CAPES. (2010). Plano Nacional de Pós-Graduação - PNPG 2011-2020 (Vol. 1). Brasília: CAPES.

Comissão Especial de Acompanhamento do PNPG 2011-2020. (2014). Elaboração da Agenda Nacional de Pesquisa. Brasília.

Guimarães, J. A., & Humann, M. C. (1995). Training of human resources in science and technology in Brazil: The importance of a vigorous post-graduate program and its impact on the development of the country. Scientometrics, 34(1), 101–119. Kluwer Academic Publishers.

Hostins, R. C. L. (2006). Os Planos Nacionais de Pós-graduação (PNPG) e suas repercussões na Pós-graduação brasileira. Perspectiva, 24(1), 136–160. Florianópolis.

Newman, M. E. J. Networks: An Introduction. USA: Oxford University Press, 2010.]

Nugroho, Y.; Saritas, O. Incorporating network perspectives in foresight: a methodological proposal, Foresight, 11(6): 21-41, 2009.

Maia, J. M. F.; Ladeira, A. V. G. C.; Cagnin, C. H.; Villela, A. B. C. Análise de redes e FTA para uma avaliação estratégica dos Institutos Nacionais de Ciência e Tecnologia. Parcerias Estratégicas. Brasília: CGEE, v. 20, n. 40, p. 101-123, jun. 2015.

Pezzini, A. Mineração de textos: conceito, processo e aplicações. 2016. Available in: <http://www.revistas.udesc.br/index.php/reavi/article/download/6715>. Access in: Feb. 10, 2018.

Cagnin, C.; Keenan, M. Johnston, R.; Scapolo, F.; Barré, R. Future-Oriented Technology Analysis – Strategic Intelligence for an Innovative Economy. Springer-Verlag Berlin Heidelberg, 2008.

Ciarli, T. & Coad, A. & Rafols, I. (2015). Quantitative analysis of technology futures: A review of techniques, uses and characteristics. Science and Public Policy. 43. scv059. 10.1093/scipol/scv059.

Mena-Chalco, J. P. & Cesar Junior, R. M. ScriptLattes: an open-source knowledge extraction system from the Lattes platform. *J. Braz. Comp. Soc.* [online]. 2009, vol.15, n.4 [cited 2018-05-12], pp.31-39. Available from: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-65002009000400004&lng=en&nrm=iso>. ISSN 0104-6500. http://dx.doi.org/10.1007/BF03194511.