Uso do scriptLattes e Gephi na Análise da Colaboração Científica

Douglas M. Giordano[1], Eduardo Bruning[1], Andrea Sabedra Bordin[1]

[1]Universidade Federal do Pampa(UNIPAMPA)
97.546-550 – Alegrete – RS – Brazil
douglasmontanhagiordano@gmail.com, eduardo.bruning@hotmail.com, andreabordin@unipampa.edu.br

Abstract. *The study of scientific collaboration through Social Network Analysis (SNA) allows the discovery of results that can be input for the definition of research management policies at various institutional levels. This paper presents two tools that enable the network analysis of an R & D context: scriptLattes and Gephi and extensions implemented in scriptLattes in order to refine the analysis. The use of tools is demonstrated in a study examining the network of researchers teachers and students of a university. The results show the feasibility of the use of the tools and also the existence of a network with many fragmented with a single component member.*

Resumo. *O estudo da colaboração científica através da Análise de Redes Sociais (ARS) permite a descoberta de resultados que podem ser insumos para a definição de políticas de gestão de pesquisa em vários níveis institucionais. Este trabalho apresenta duas ferramentas que viabilizam a análise da rede de um contexto de P&D: scriptLattes e Gephi, bem como as extensões implementadas no scriptLattes com o objetivo de refinar as análises. A utilização das ferramentas é demonstrada em um estudo de análise da rede de pesquisadores professores e alunos de uma universidade. Os resultados mostram a viabilização do uso das ferramentas e também a existência de uma rede fragmentada com muitos componentes isolados com um integrante.*

1. Introdução

No âmbito acadêmico é comum que professores pesquisadores e alunos colaborem na produção de material científico. Essas colaborações podem ser analisadas e oferecem informações para a definição de políticas na gestão de pesquisa.

A Análise de Redes Sociais (ARS) é um dos métodos utilizados para analisar a colaboração científica. Com ele o grupo de pesquisadores sob análise e suas relações de coautoria são representados através de uma rede e várias métricas podem ser calculadas para analisar a referida rede.

O objetivo deste trabalho é apresentar duas ferramentas que podem ser utilizadas conjuntamente para analisar uma rede de colaboração científica: scriptLattes[1] e Gephi[2],

[1] http://scriptlattes.sourceforge.net/

[2] http://gephi.github.io/

assim como apresentar as extensões implementadas no scriptLattes para refinar o trabalho de análise.

. O scriptLattes foi escolhido por ser um sistema de código aberto que possibilita o acesso as informações dos pesquisadores cadastrados na plataforma Lattes a partir do número identificador de cada currículo. O Gephi foi escolhido por ser uma ferramenta de análise de rede free muito popular no domínio de análise de rede.

Estudos de análise de colaboração científica envolvendo a utilização dessas duas ferramentas são recentes, tais como os de Ferraz, Quoniam e Alvares (2014); Mena-Chalco et al (2014) e Ferraz e Quoniam (2013). Entretanto entende-se que este trabalho contribua neste escopo ao detalhar o uso de tais ferramentas e apresentar novas extensões implementadas no scriptLattes.

Para demonstrar a viabilidade das extensões implementadas e uso das ferramentas foi desenvolvido um estudo de análise da rede de colaboração científica de professores e alunos de um campus de uma universidade federal. Os resultados da análise desta rede buscam fornecer o perfil de colaboração do grupo de pesquisadores na referida instituição.

As seções seguintes apresentam os conceitos das técnicas e ferramentas utilizadas na pesquisa. Na seção 2 é apresentado o conceito de análise de rede social. Na terceira e quarta seção são apresentadas as ferramentas scriptLattes e Gephi utilizadas na pesquisa. Na quinta seção é apresentado o estudo de caso utilizando as ferramentas, assim como as extensões implementadas no scriptLattes. Por fim são apresentados os resultados da pesquisa, a discussão e a conclusão.

2. Análise de Rede Social

A modelagem de sistemas em rede vem sendo aplicada em áreas diversas como epidemiologia (MOORE; NEWMAN, 2000) e colaboração científica (NEWMAN, 2004). Uma rede pode ser representada por um grafo G = (V, E) formado por vértices (V) e arestas (E). Cada vértice ou nodo representa um ator e cada aresta representa a relação existente entre dois atores integrantes da rede.

Dados modelados em rede podem ser analisados através de métricas de análise de rede social (ARS). Segundo Wasserman e Faust (1994) a área de análise de rede social (*social network analysis*) tem atraído muito interesse nas últimas décadas. Através das métricas de ARS é possível identificar aspectos, tais como: a) padrões de relacionamento entre os atores de uma rede; b) a conectividade entre os mesmos; c) a formação de clusters; d) a evolução da rede ao longo do tempo e, e) o fluxo de comunicação, informação e conhecimento dentro da rede.

Para analisar a rede segundo uma perspectiva de estrutura geral deve-se utilizar a medida de densidade. Segundo Scott (2000), a densidade descreve o nível geral de ligações entre os pontos de um grafo. Um grafo "completo" é aquele em que todos os pontos são adjacentes um ao outro, ou seja, cada nodo é ligado diretamente a todos os outros pontos.

Na perspectiva individual, existem algumas métricas de centralidade que procuram descrever as propriedades de localização de um ator na rede. Os atores mais importantes ou mais proeminentes estão normalmente localizados em posições

estratégicas dentro da rede (WASSERMAN; FAUST, 1994). A centralidade de grau (*centrality degree*) de um ator corresponde ao número de arestas incidentes ou ao número de vértices adjacentes a ele. Segundo Freeman (1979) a centralidade de grau reflete a posição e o papel do ator em termos de popularidade e atividade. Em redes valoradas, onde a aresta possui um peso, a centralidade de grau pode levar em conta o valor ou peso da aresta. Em redes de coautoria essa medida determina o grau de colaboração de um ator.

Em uma perspectiva de grupo, é possível a identificação de componentes ou subredes dentro de uma rede. Componentes são subredes cujos nodos estão conectados por algum caminho dentro da rede, mas desconectados entre as demais subredes. Se uma rede contém um ou mais pontos "isolados", esses pontos também são chamados de componentes. Componente gigante é o nome dado a subrede que contém o maior número de nodos conectados. Numa rede de coautoria a presença de mais de um componente na rede indica a existência de grupos que publicam isoladamente.

3. scriptLattes

O scriptLattes é um sistema de código aberto desenvolvido por Mena-Chalco e Cesar-Jr (2009) que extrai dados de currículos registrados na Plataforma Lattes e a partir desses dados disponibiliza uma série de relatórios e gráficos com os totais de cada tipo de produção, incluindo um gráfico de colaboração referente aos currículos coletados.

O processo feito pelo scriptLattes para extração dos dados é apresentado na Figura 1. Primeiramente é necessário fornecer a lista dos identificadores dos currículos Lattes. A partir desta lista o sistema faz o *download* do currículo em formato HTML de cada identificador constante na lista. Após o *download* dos currículos são extraídos os dados necessários aos relatórios que o sistema disponibiliza. Como última etapa do processo são gerados os relatórios de saída. É importante salientar outras importantes funções do scriptLattes executadas entre esses processos, como o filtro das produções por ano e a eliminação de redundância entre produções.

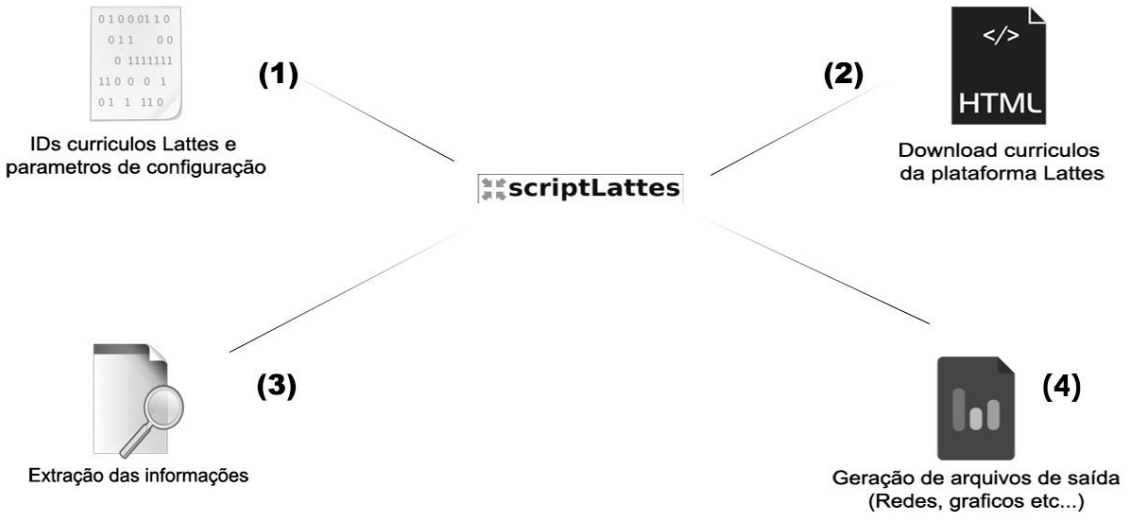

Figura 1. Processo de extração dos currículos

3.1 Extensões desenvolvidas no scriptLattes

O scriptLattes na sua versão atual possui várias funções de filtro que são parametrizadas em seu arquivo de configuração, tais como o filtro são por data (por pesquisador ou global), filtro que evita a extração de algum tipo de produção ou filtros que não evitam a inclusão de algum tipo de informação nos relatórios de saída.

Neste trabalho foi implementado um novo filtro[3] que busca em tempo de execução os registros do currículo a partir do ano de entrada do pesquisador em determinada instituição. O sistema já oferece um filtro por data para cada pesquisador, porém ele é estático e necessita do conhecimento prévio do ano de entrada do pesquisador na universidade.

Na Figura 2 pode-se visualizar como funciona o filtro do ano de entrada na universidade. Primeiramente é informado no arquivo de configuração do scriptLattes o nome da universidade a ser filtrada. É feito o *download* do currículo expandido no formato HTML. Na extração de dados, os dados de atuação profissional, caso o pesquisador seja um professor, ou de instituição de ensino, caso seja um aluno, são extraídos do HTML. Com a informação do ano de entrada e saída do pesquisador é possível filtrar apenas as publicações contidas no período em que o mesmo possui vínculo com a instituição filtrada.

Figura 2. Processo de filtragem das produções pelo ano de entrada na instituição

Outra funcionalidade implementada foi a extração de palavras chaves contidas em cada produção bibliográfica dos currículos com o objetivo de gerar uma nuvem de termos (*tag cloud*). Na Figura 3 pode-se ver o processo de geração da nuvem de termos. Para buscar essas palavras chaves é necessário efetuar o *download* dos currículos

[3] As funcionalidades implementadas podem ser acessadas no repositório: https://bitbucket.org/DouglasGiordano/script-lattes-unipampa

expandidos de cada pesquisador. Quando o download termina as informações do HTML são extraídas, cada produção bibliográfica tem suas palavras chaves armazenadas em tempo de execução em um atributo tipo lista. Após a extração de dados ser feita, cada uma das listas de produções é percorrida e é gerado um arquivo de saída contendo as palavras chaves encontradas nas produções dos pesquisadores.

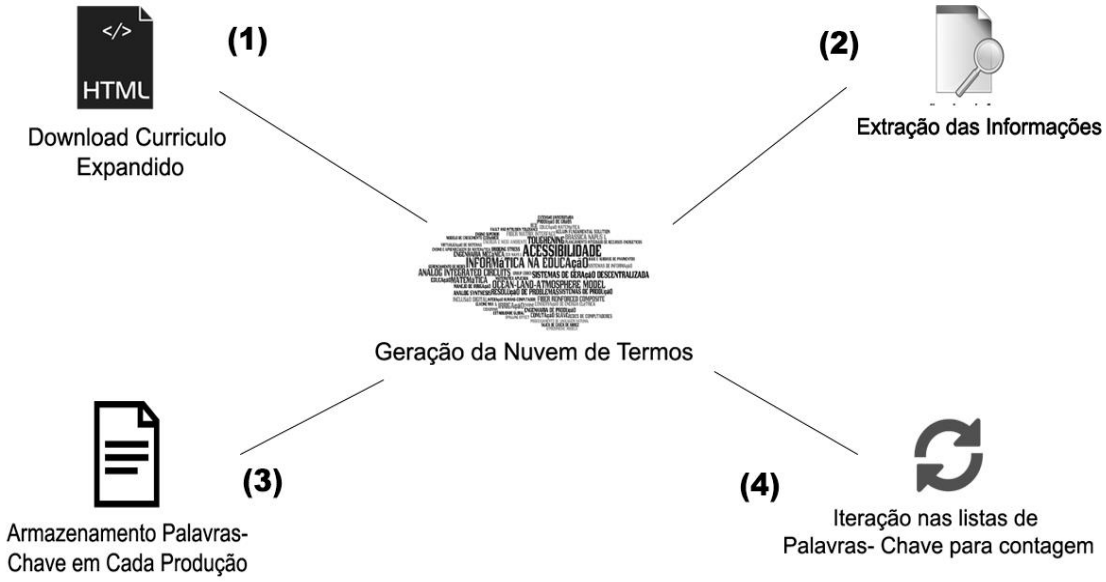

Figura 3. Processo de extração de palavras chave e geração da nuvem de termos

Para a geração gráfica da nuvem de termos foi utilizado a API(*Application Programming Interface*) PyTagCloud na linguagem Python. A API foi modificada com a alteração da forma de contagem das palavras e também para limitar o número de palavras exibidas na nuvem.

4. Gephi

Gephi é um software de código aberto utilizado para analisar redes. Ele permite acesso fácil e amplo aos dados de rede e permite a importação, visualização, filtragem, navegação e agrupamentos dos dados (*clustering*). Com a sua utilização analistas de dados conseguem criar hipóteses, descobrem padrões intuitivamente, isolam estrutura singulares, etc. Além disso, permite o trabalho com conjuntos de dados complexos e através de seus vários algoritmos de *layout* produz valorosos resultados visuais. (BASTIAN; HEYMANN; JACOMY, 2009).

Com ele é possível calcular diversas métricas de ARS, tais como centralidade, densidade, diâmetro, número de componente, detecção de comunidades (modularidade), etc.

5. Utilização das ferramentas em um estudo de caso

A utilização das ferramentas de código aberto scriptLattes e Gephi para análise da colaboração científica de um grupo de pesquisadores foi demonstrada através de um estudo de caso em um campus de uma universidade.

Com o scriptLattes foi possível criar o arquivo de rede contendo a lista de todos os pesquisadores e as relações de coautoria entre os mesmos, assim como criar a nuvem de termos (*tag cloud*) resultante das palavras chaves do conjunto de produções dos pesquisadores. A partir do arquivo de rede, a ferramenta Gephi possibilitou o cálculo de algumas métricas de rede e sua visualização. Na Figura 4 pode-se ver a relação entre as duas ferramentas, onde o scriptLattes gera um arquivo de rede com extensão GDF e a ferramenta de análise exploratória de dados Gephi importa esse arquivo e cria a rede.

Durante a utilização das ferramentas alguns problemas foram encontrados, como currículos desatualizados, problemas de filtro no *download* dos currículos. Entretanto esses problemas foram contornados com o aviso aos pesquisadores sobre possíveis currículos desatualizados e com correções no scriptLattes. Depois de resolvido esses problemas a rede foi criada com um maior grau de segurança e foi possível analisar as relações de colaboração entre os pesquisadores da universidade, cujos dados são apresentados na seção Resultados.

Figura 4. Relação entre as ferramentas scriptLattes e Gephi

5.1 Procedimentos Metodológicos

Os procedimentos utilizados no estudo de caso estão divididos em duas etapas: Coleta, extração e correlação de dados e Análise de rede.

5.1.1 Coleta, extração e correlação de dados

Nesta primeira etapa inicialmente se coletou os identificadores de currículo Lattes de todos os professores e alunos ativos na universidade, totalizando 81 professores e 793 alunos. Para a realização da coleta, extração e a correlação dos dados das produções em coautoria dos currículos foi utilizado o scriptLattes. A ferramenta gerou o arquivo de rede (.gdf) contendo os nodos e as relações de coautoria da rede.

5.1.2 Análise de rede

O arquivo de rede gerado anteriormente foi importado pelo Gephi e a partir dessa importação foi gerada a rede de colaboração científica (rede de coautoria) dos professores e alunos que é visualizada na Figura 5. No Gephi foram calculadas algumas métricas de ARS apresentadas na seção a seguir.

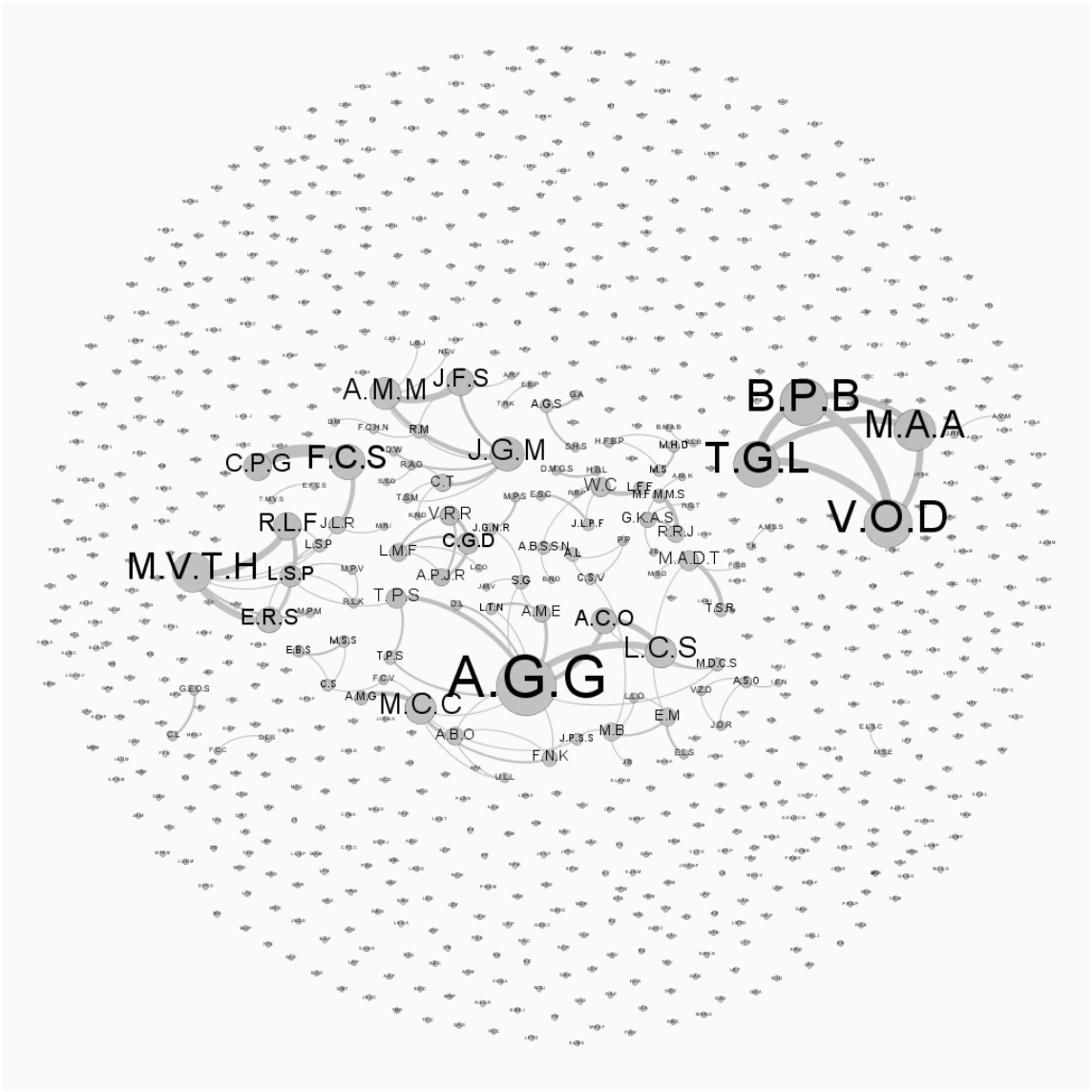

Figura 5. Rede de coautoria dos pesquisadores da Universidade

6. Resultados

Na rede de coautoria foram identificados 874 nodos (81 professores e 793) alunos e 128 relações. A densidade geral encontrada foi 0 onde o valor máximo é 1.0 e o grau médio de colaboração foi 0,2.

Os autores foram classificados segundo a medida de centralidade de grau que leva em consideração o peso das relações (C.G. c/ Peso) e que determina o grau de colaboração entre os atores da rede. A tabela 1 apresenta os cinco autores com maior grau de colaboração.

A análise de rede revelou a existência de 779 componentes isolados, onde o maior componente (componente gigante) possui 54 autores e representa 6,1% da rede. O componente gigante da rede é apresentado na Figura 5 com destaque para os autores com maior centralidade de grau com peso.

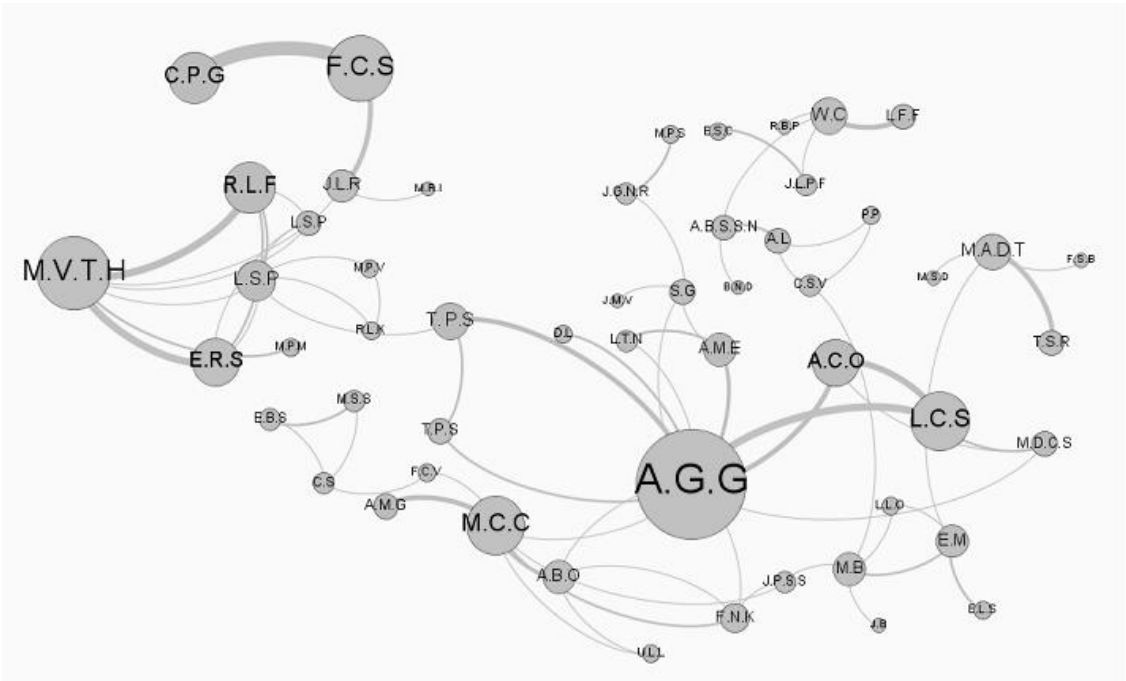

Figura 6. Componente gigante da rede de coautoria dos pesquisadores da Universidade

Tabela 1. Ranking dos pesquisadores da universidade

Ranking	Autor	Nº Produções	C.G	C.G. c/Peso	Aluno/Professor
1	A. G. G.	44	12	27	Professor
2	B.P.B	15	3	21	Aluno
3	V.O.D	14	3	21	Professor

4	T.G.L	9	3	20	Aluno
5	M.A.A	7	3	18	Aluno

O número total de produções bibliográficas dos pesquisadores a partir do ano de ingresso na instituição é 1501. A partir dessas produções foram extraídas as palavras chaves e contabilizada a frequência das mesmas. Com isso foi possível gerar uma nuvem de termos (*tag cloud*) com os 50 termos de maior relevância. Através da Figura 7 é possível visualizar as palavras "Informática na Educação" (19) e "Acessibilidade" (18) como os termos com maior frequência.

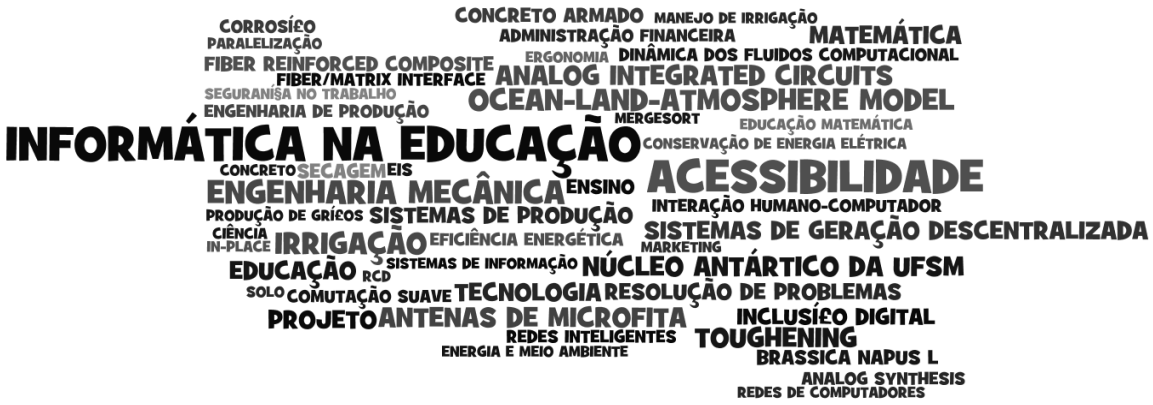

Figura 7. Nuvem de termos das produções

7. Discussão e Conclusão

Este trabalho apresentou a utilização conjunta de duas ferramentas - scriptLattes e Gephi - para análise de rede de colaboração científica, bem como algumas extensões implementadas na ferramenta scriptLattes. Para demonstrar o uso das mesmas foi realizado a análise da rede de colaboração científica de professores e alunos de um campus de uma universidade.

No que tange as ferramentas, este trabalho, assim como outros já citados, confirma que elas são complementares e adequadas para a realização de estudos de colaboração. Uma das extensões implementadas filtra as produções pelo ano de entrada do pesquisador em uma determinada instituição, permitindo assim a definição de um escopo de análise mais refinado. A implementação da nuvem de termos permite uma análise além da topografia da rede, ou seja, a análise da semântica ou do conteúdo das produções dos pesquisadores.

No que tange aos resultados da análise da rede em si, observou-se um grande número de componentes ou nodos isolados (779) que comparado com o total de nodos da rede (874) indica que um percentual considerável de pesquisadores não produziu conjuntamente com outros pesquisadores do campus.

Destaca-se que os resultados podem apresentar um percentual de erro devido a inserção incorreta de dados no currículo Lattes e também refletem o estado inicial da pesquisa.

Como trabalhos futuros, pretende-se continuar a extensão do scriptLattes com a implementação de funcionalidades relacionadas ao grafo de colaboração que a ferramenta já disponibiliza, tais como o perfil de colaboração de determinado nodo do grafo, que pode ser obtido pela relação das palavras chaves das produções envolvidas nas colaborações do nodo, e tornando-as disponíveis para a comunidade que o utiliza. Em relação ao trabalho de análise de rede pretende-se ampliar o escopo de análise para os outros 10 campi da referida universidade buscando identificar as relações de colaboração entre os mesmos.

Referências

BASTIAN, M.; HEYMANN, S.; JACOMY, M. Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media, 2009.

FERRAZ, R. R. N.; QUONIAM, L.; ALVARES, L. M. A. R. Avaliação de redes multidisciplinares com a ferramenta scriptlattes: os casos da nanotecnologia, da dengue e de um programa de pós-graduação Stricto Sensu em Administração. Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, v. 19, n.40, p.67-98, mai./ago., 2014.

FERRAZ, R. R. N.; QUONIAM, L. A utilização da ferramenta computacional Scriptlattes para avaliação das competências em pesquisa no Brasil. Prisma.com, v., n. 41, 2013.

FREEMAN, L. Centrality in Social Networks: Conceptual Clarification. Social Networks, 1(3), 215–239, 1979.

MOORE, C.; NEWMAN, M. E. J. Epidemics and percolation in small-world networks. Phys. Rev. E 61, 5678–5682, 2000.

MENA-CHALCO, J. P.; CESAR-JR., R. M. scriptLattes: An open-source knowledge extraction system from the Lattes platform. Journal of the Brazilian Computer Society, v. 15, n. 4, p. 31-39, 2009.

MENA-CHALCO, J. P.; DIGIAMPIETRI, L. A.; LOPES, F. M.; CESAR, R. M. Brazilian bibliometric coauthorship networks. Journal of the Association for Information Science and Technology, v.65, n.7, p.1424–1445, 2014.

NEWMAN, M. E. Coauthorship networks and patterns of scientific collaboration. Proceedings of the National Academy of Sciences of the United States of America, v. 101, n. Suppl 1:5200-5205, april 2004.

SCOTT, J. Social Network Analysis. A Handbook. 2nd edition. SAGE Publications: London, 2000.

WASSERMAN, S.; FAUST, K. Social Network Analysis: methods and applications. Cambridge University Press. Structural analysis in social the social sciences series, v. 8, (1994) 1999.