Detectando linhas de pesquisa emergentes com análise de redes via algoritmo de agrupamento hierárquico

Roney Fraga Souza, Ivette Luna e Rosangela Ballini
Instituto de Economia
Universidade Estadual de Campinas
Campinas, São Paulo, Brasil
Emails: roneyfraga@gmail.com, ivette@eco.unicamp.br, ballini@eco.unicamp.br

Abstract*—Este trabalho tem como objetivo a aplicação de algoritmos de clusterização hierárquico para detectar linhas de pesquisa emergentes, que buscam mensurar eficiência utilizando o método** *Data Envelopment Analysis - DEA. O processo básico para analisar as fontes de pesquisa emergentes em DEA, utiliza algoritmos de agrupamentos. Este trabalho procura analisar o algoritmo de agrupamento topológico, o qual é baseado na ideia de modularidade, aplicado a uma base de dados de citação. Uma abordagem alternativa a este método de agrupamento é a proposta de utilização do algoritmo aglomerativo hierárquico o qual é baseado na fusão de agrupamentos com base em medidas de proximidade. As etapas para a aplicação dos algoritmos de clusterização são: coleta de dados de citação na plataforma de dados bibliográficos** *Web of Science***; construção de redes de citação para cada ano; exclusão dos artigos que não citam outros artigos da rede, mantendo apenas o componente principal da rede; divisão da rede em grupos via método de agrupamento; determinação do peso de cada artigo em seu grupo e nos demais grupos da rede; e, por fim, analisar o conteúdo de cada grupo via análise de** *Natural Language Processing***.**

I. Introdução

Uma das habilidades básicas exercida pelos homens é o agrupamento de objetos semelhantes para produzir uma classificação [1]. Diante do grande volume de informações disponíveis na atualidade, a cada dia é mais difícil identificar padrões comuns entre as informações. No mundo científico o mesmo acontece, dificultando a identificação das diferentes abordagens estudadas em uma linha de pesquisa, ou mesmo, a identificação de diferentes vertentes em uma metodologia. Recentemente em algumas áreas do conhecimento, tais como, semicondutores de nitreto e medicina regenerativa, análises de linhas de pesquisa emergentes foram realizadas [2], [3]. Nestes estudos, redes de citação foram particionadas em grupos utilizando o método de agrupamento topológico [4], visando o acompanhamento da evolução dos temas de pesquisa em cada grupo, sendo as redes de citação visualizadas com termos centrais para cada grupo.

Este trabalho tem como objetivo analisar as fontes de pesquisa emergentes em *Data Envelopment Analysis* - DEA, utilizando análise de redes e algoritmos de agrupamentos hierárquicos. DEA é uma técnica para mensurar eficiência que surgiu em 1978 [5]. A análise de eficiência via DEA pode ser aplicada a empresas comerciais, governos e agências sem fins lucrativos, como escolas, hospitais, unidades militares, forças policiais, sistemas de justiça criminal, bem como países e regiões. Por definição essas unidades são chamadas de *Decision Making Unit* - DMU. A avaliação de eficiência resulta em uma pontuação de desempenho que varia entre zero e um, expondo o grau de eficiência obtido pela DMU. Para chegar a essas pontuações, DEA também identifica as fontes e as quantidades de ineficiência em cada insumo e produto para cada DMU [6]. No estudo de Emrouznejad, Parker e Tavares, 30 anos de literatura em DEA foram analisados [7], evidenciando um crescimento médio de 226 artigos por ano entre 1995 e 2003, e de 360 artigos entre 2004 e 2006. No total os autores contabilizaram mais de 4.000 publicações de artigos ou capítulos de livros. O elevado número de textos publicados a cada ano evidência a expansão do interesse acadêmico no tema. Contudo, dificulta o entendimento do estado da arte da análise de eficiência com DEA. Cabe destacar que os autores apenas quantificaram os textos publicados que utilizaram DEA.

O presente trabalho propõe, inicialmente, o estudo e a aplicação do método de agrupamento topológico, proposto por Newman [4], em uma rede de citação de fontes de pesquisa em DEA. Esta técnica de agrupamento é baseada em um método de otimização de uma função modular que depende do número de conexões e de vértices, visando a fusão de grupos com maior similaridade. Como uma proposta alternativa a esta técnica de agrupamento, surge o interesse da aplicação do método de agrupamento hierárquico proposto por Yager [8], o que é um método de agrupamento abseado em medidas de proximidade para a formação de grupos e a fusão destes.

Após esta breve introdução, a próxima seção apresenta as etapas necessárias para a construção da base de dados necessária para a aplicação dos métodos de clusterização. A seguir, na seção III, é descrito, brevemente, os métodos de agrupamentos topológicos e o proposto por Yager. Na seção IV, são apresentados os resultados da aplicação do método topológico a base de dados e propostas futuras de trabalho.

II. Construção da Base de Dados

O presente trabalho busca aplicar uma técnica de agrupamento hierárquico sobre uma rede de citação que permita identificar as fontes de pesquisa emergentes em DEA. Para tal objetivo é necessário a cosntrução da base de dados, seguindo os seguintes passos:

1. Coleta de dados de citação das fontes *Science Citation Index* e *Sciences Citation Index* compilados pelo *Institute for Scientific Information* - ISI, que mantém bancos de dados de citações que cobrem milhares de revistas acadêmicas e oferece serviços de banco de dados bibliográficos.

Para realizar esta etapa, utilizou-se o *Web of Science*, que é uma interface para o usuário final baseada na *web* de bases de dados de citação do ISI. A coleta foi realizada com base nos seguintes critérios de busca: *Topic=("sfa" AND "efficiency*") OR Topic= ("data envelopment analysis*") OR Topic=("dea" AND "efficiency*") OR Topic=("farrell efficiency*") OR Topic= ("Productive Efficiency*") AND Year Published= (1950-2013)*. Como resultado da busca foram obtidos 7.114 textos.

2. Construção de redes de citação para cada ano. Nesta etapa foram considerados como vértices, ou nós, os artigos e as citações as arestas ou conexões entre os nós, formando assim uma rede de citações.

3. Eliminação dos artigos que não citam os demais textos da rede, mantendo apenas o componente principal da rede. Esta etapa tem como objetivo excluir textos que não fazem parte da rede. A coleta dos textos que foram excluídos da rede é devido à amplitude dos termos de busca.

Após finalizada a etapa da construção da rede de citações, o próximo passo é a aplicação dos métodos de agrupamentos, os quais são descritos brevemente na próxima seção.

III. Métodos de Agrupamentos Hierárquicos

Neste trabalho dois métodos de agrupamentos são de interesse: o primeiro, frequentemente usado na literatura, é o método de agrupamento topológico, proposto por Newman [4]; o segunda método é o algoritmo de aglomerativo hierárquico (HAG) proposto por Yager [8]. Os dois métodos são descritos brevemente a seguir.

O primeiro método de agrupamento a ser analisado é o proposto por Newman, sendo uma técnica baseada em medidas topológicas para a formação dos grupos. Este método evidência aglomerações coesas, com densidade elevada de ligações em cada grupo, tendo sido empregado no trabalho [2] para detecção de pesquisas emergentes em redes de citação. Este algoritmo de agrupamento é hierárquico, sendo baseado na ideia de modularidade, a qual mede quando a divisão é adequada, no sentido de que há muitas arestas em um grupo e poucos entre os grupos.

Neste trabalho, aplicou-se o o método proposto por A. Clauset, M. E. J. Newman e C. Moore [9] para detecção dos grupos. Este método é um aprimoramento do algoritmo proposto por M. E. J. Newman [10]. Vale destacar que o algoritmo utiliza uma técnica de otimização, *fast greedy* a qual a função objetivo é a função modular que depende do número de vértices e de arestas. O tempo de processamento em uma rede com n vértices e m arestas é igual a $O(m\ d\ log\ n)$, sendo d a profundidade do dendrograma. Como em aplicações reais, as redes são esparsas e hierárquicas, com $m \sim n$ e $d \sim \log n$ este algoritmo de agrupamento hierárquico o empo de execução é essencialmente linear, $O(n\ log^2\ n)$. Neste artigo apenas o método de clusterização propsoto por Newman foi aplicado a rede de citações

O segunda método de clusterização que se pretende aplicar é o método de agrupamento hierárquico proposto por [8], como um método alternativo ao algoritmo de Newman. Este algoritmo consiste em uma técnica de agrupamento que pode ser subdividida em métodos de aglomeração que realizam uma série de fusões sucessivas dos n indivíduos em grupos, particionando os n indivíduos sucessivamente em subgrupos mais similares por meio de medidas de proximidade [1]. Ao dispor a rede de citação em grupos, pode-se detectar uma frente de pesquisa que consiste em um grupo de artigos.

Este algoritmo pode ser resumido em três passos: i) Seleção de um conjunto de grupos iniciais; ii) Determinação dos pares de grupos mais similares segundo uma medida de proximidade; iii) repetição do passo ii) até que um critério de parada seja satisfeito. A forma como é processada cada uma dessas etapas influenciarão os resultados do processo de agrupamento. A exploração de cada uma dessas três etapas serão objeto de pesquisa futura.

Após a aplicação do método de clusterização, as próximas etapas são: visualização da rede de citação usando o algoritmo *large graph layout* - LGL [11], e determinação do peso de cada artigo no grupo e seu coeficiente de participação. Por fim, ocorre a extração dos termos característicos de cada grupo por filtragem linguística, usando os resumos dos trabalhos e aplicando *Natural Language Processing* - NLP [12] [13].

O passo de determinação do peso de cada artigo no grupo e seu coeficiente de participação corresponde ao grau de importância de cada artigo, representado na rede pelos nós. A determinação de sua relevância dentro do grupo [14], pode ser representada por:

$$z_i = \frac{K_i - \bar{K}_{s_i}}{\sigma_{K_{s_i}}} \tag{1}$$

em que K_i é o número de conexões de um nó i relacionado a outros nós no seu grupo s_i, $\bar{K}_{s_i}$ é a média de K sobre todos os nós em s_i, e $\sigma_{K_{s_i}}$ é o desvio padrão de K em s_i. Dessa forma, o grau de relevância z_i de um nó no grupo é alto se o número de ligações do nó no grupo é alto.

Para analisar a participação entre grupos, que evidencia o quão distribuídas estão as conexões de um nó i entre os diferentes grupos, deve-se analisar:

$$P_i = 1 - \sum_{s=1}^{N_M} \left(\frac{K_{is}}{K_i} \right)^2 \tag{2}$$

sendo K_{is} o número de ligações de um nó i para outros nós no grupo s, e K_i o grau total do nó i, ou seja, o número de ligações conectados ao nó i. O coeficiente de participação P_i próximo de um, indica que suas conexões estão uniformemente distribuídas entre todos os grupos, e próximo de zero indica que todas as conexões estão dentro de seu próprio grupo [14], [3].

Para analisar o conteúdo de cada grupo, deve-se utilizar *Natural Language Processing* - NPL, que permite detectar o tema de pesquisa de cada grupo. Os termos são extraídos do resumo dos artigos utilizando filtros linguísticos [12], [3]. Os filtros extraem os termos demarcados por elementos que, inseridos na oração, constituem uma unidade significativa, mantendo entre si relações de dependência e de ordem.

Por fim, cabe destacar que os maiores desafios da realização deste estudo estão na aplicação de técnicas de diferentes linhas do conhecimento e na implementação dos métodos de agrupamentos assim como NLP.

IV. Resultados

A técnica de agrupamento baseada em medidas topológicas, foi aplicada para detecção de grupos em uma rede de citação construída a partir da *Web of Science*. Nota-se que ao aplicar a técnica de agrupamento 26 grupos foram detectados, cabendo destacar que a maioria dos grupos apresentam poucos trabalhos científicos, conforme Tabela I. A Figura 1 apresenta a visualização da estrutura de rede obtida. Pode-se observar que alguns grupos apresentam um grande número de citações, representadas pelos nós, e outros grupos são menores.

TABLE I. Número de grupos e quantidade de trabalhos científicos (vértices) por grupo.

Grupo	Vértices	Grupo	Vértices	Grupo	Vértices
1	1705	18	3	19	5
2	336	10	27	20	6
3	1686	11	19	21	2
4	1235	12	15	22	7
5	72	13	8	23	4
6	539	14	6	24	3
7	55	15	5	25	2
8	49	16	6	26	3
9	49	17	4		

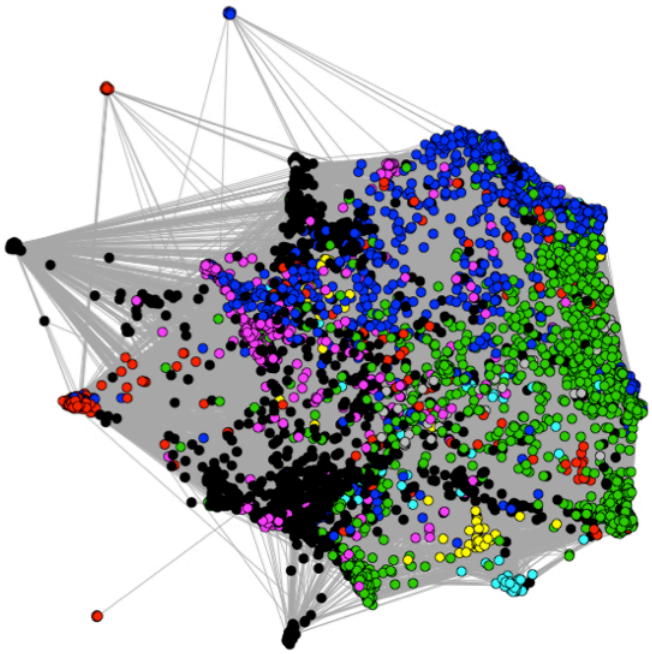

Fig. 1. Rede com os grupos destacados por cores.

Como trabalhos futuros propõe-se o estudo e aplicação do algoritmo de clusterização hierárquico proposto por Yager para a detecção de de rede de citação. Deve-se destacar que as etapas para o desenvolvimento deste algoritmo são flexíveis, e os critérios para a diisão dos grupos são determiandos via medidas de proximidade.

Agradecimentos

Os autores agradecem ao CNPq e à CAPES pelo auxílio financeiro.

References

[1] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*, 5th ed., ser. Wiley Series in Probability and Statistics, W. A.Shewhart and S. S. Wilks, Eds., jan 2011. [Online]. Available: http://onlinelibrary.wiley.com/book/10.1002/9780470977811

[2] N. Shibata, Y. Kajikawa, Y. Takeda, and K. Matsushima, "Detecting emerging research fronts based on topological measures in citation networks of scientific publications," *Technovation*, vol. 28, no. 11, pp. 758–775, 2008.

[3] N. Shibata, Y. Kajikawa, Y. Takeda, I. Sakata, and K. Matsushima, "Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications," *Technological Forecasting and Social Change*, vol. 78, no. 2, pp. 274–282, 2011.

[4] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, p. 066133, 2004.

[5] A. Charnes, W. W. Cooper, and E. Rhodes, "Measuring the efficiency of decision making units," *European Journal of Operational Research*, vol. 2, no. 6, pp. 429–444, 1978.

[6] W. W. Cooper, L. M. Seiford, and K. Tone, *Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software*. Springer Science+ Business Media, 2007.

[7] A. Emrouznejad, B. R. Parker, and G. Tavares, "Evaluation of research in efficiency and productivity: A survey and analysis of the first 30 years of scholarly literature in dea," *Socio-Economic Planning Ssciences*, vol. 42, no. 3, pp. 151–157, 2008.

[8] R. R. Yager, "Intelligent control of the hierarchical agglomerative clustering process," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 30, no. 6, pp. 835–845, 2000.

[9] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, no. 6, p. 066111, 2004.

[10] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 026113, 2004.

[11] A. T. Adai, S. V. Date, S. Wieland, and E. M. Marcotte, "Lgl: creating a map of protein function with an algorithm for visualizing very large biological networks," *Journal of Molecular Biology*, vol. 340, no. 1, pp. 179–190, 2004.

[12] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms:. the c-value/nc-value method," *International Journal on Digital Libraries*, vol. 3, no. 2, pp. 115–130, 2000.

[13] H. Mima, K. Frantzi, and S. Ananiadou, "The c-value/example-based approach to the automatic recognition of multi-word terms for cross-language terminology," in *Proc. Int. Joint Workshop on Cross-Language Issues in AI. Held at 5th Pacific Rim Int. Conf. on Artificial IntelligenceI (PRICAI'98)*, 1998, pp. 10–21.

[14] R. Guimera and L. A. N. Amaral, "Functional cartography of complex metabolic networks," *Nature*, vol. 433, no. 7028, pp. 895–900, 2005.