



COOCORRÊNCIA DE PALAVRAS-CHAVE EM DADOS ABERTOS DA CAPES: teses e dissertações em Ciência da Informação

**KEYWORD CO-OCCURRENCE NETWORKS: theses and dissertations in
Information Science**

Francis Bento Marques 
Universidade Federal de Minas Gerais

Yuri Bento Marques 
Universidade Federal de Minas Gerais

Benildes Coura Moreira dos Santos Maculan 
Universidade Federal de Minas Gerais

RESUMO

O grande volume de dados produzidos, armazenados e disponibilizados para acesso tornou os computadores imprescindíveis para transformá-los em informação processável pelo homem. Com a mineração de textos, as palavras extraídas podem ser utilizadas no apontamento de relações entre elementos textuais internos ou externos a eles. Neste estudo, apresenta-se uma pesquisa em andamento que busca a descoberta de padrões de coocorrência de palavras-chave nas dissertações e teses do domínio da Ciência da Informação brasileira, utilizando técnicas de inteligência artificial aplicadas aos dados abertos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior(Capes). A metodologia se caracteriza como de natureza aplicada, com objetivos exploratórios de descritivos, e com procedimentos de análise quanti-qualitativos. O método utilizado é quantitativo, na forma de um estudo métrico, com a análise baseada nos princípios da teoria dos grafos. Espera-se que os resultados evidenciem a possibilidade de parcerias entre pesquisadores, tendências de pesquisa, temas pouco explorados, entre outros elementos.

Palavras-Chave: Padrões de coocorrência, Dados Capes, Inteligência Artificial.

ABSTRACT

The large volume of data produced, stored and made available for access made computers essential to transform them into processable information by man. With text mining, the extracted words can be used in pointing out the relationships between textual elements that are internal or external to them. This study presents an ongoing research that seeks to discover patterns of co-occurrence of keywords in dissertations and theses in the domain of Brazilian Information Science, using artificial intelligence techniques applied to open data from the Coordination for the Improvement of Personnel of Higher Level (Capes). The methodology is characterized as applied in nature, with exploratory and descriptive objectives, and with quantitative and qualitative analysis procedures. The method used is quantitative, in the form of a metric study, with the analysis based on the principles of graph theory. It is expected that the results show the possibility of partnerships between researchers, research trends, underexplored themes, among other elements.

Keywords: Co-occurrence standards, Data Capes, Artificial Intelligence.

1. INTRODUÇÃO

O avanço tecnológico vem permitindo gerar, coletar e armazenar uma grande quantidade de dados. No início da década de noventa do século XX havia previsão de que a quantidade de informações iria dobrar a cada 20 anos (FAYYAD, 2001). O grande volume de dados tornou os computadores imprescindíveis para transformá-los em informação processável pelo homem. Dentro dessa realidade, as técnicas para analisar dados vêm crescendo, mas ainda a extração de conhecimento em bases de dados é um desafio (TAN; STEINBACH; KUMAR, 2009).

A mineração de texto é um método de descoberta de conhecimento dentro da mineração de dados (WANG *et al.*, 2013). Ela é uma solução na compreensão das informações básicas providas por um ou mais textos, por meio de algoritmos estruturados. Com essa técnica, as palavras extraídas podem ser utilizadas no apontamento de relações entre elementos textuais internos ou externos aos textos, e esses resultados são empregados em estudos de descoberta de conhecimento, dentre outras aplicações (LEE; YI; PARK, 2008).

Duriau, Reger e Pfarrer (2007) apresentam um método frequentemente utilizado para a análise de conteúdo. Este método adota uma técnica de descrição sistemática, objetiva e quantitativa do conteúdo de um texto. Ele é aplicado para conhecer diversas áreas de conhecimento, e, neste estudo, selecionamos a Ciência da Informação (CI). Segundo Zins (2007), a CI está em constante mudança, o que leva seus pesquisadores a revisitar conceitos e analisar suas bases construtivas. Araújo (2014) aponta que a CI conta com seis subáreas: fluxos da informação científica, representação e recuperação da informação, estudos de usuários, gestão do conhecimento, economia política da informação e estudos métricos da informação.

Dentro dos estudos métricos, o objetivo desta investigação é a descoberta de padrões de coocorrência de palavras-chave nas dissertações e teses da CI brasileira, utilizando técnicas de inteligência artificial (IA) aplicadas aos dados abertos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes). Com isso, considera-se que será possível sugerir parcerias entre pesquisadores, tendências de pesquisa, temas pouco explorados, entre outros.

Depois desta breve introdução, este artigo segue a seguinte estrutura: a seção 2 faz uma introdução aos conceitos de bibliometria e IA; a seção 3 apresenta a metodologia, indicando a caracterização da pesquisa e dos dados, ferramentas e as técnicas empregadas nas análises; a seção 4 traz os resultados parciais, e, por fim, a seção 5 expõe as considerações finais.

2. FUNDAMENTAÇÃO TEÓRICO-METODOLÓGICA

2.1. Bibliometria

As análises bibliométricas são usadas para rastrear a produção acadêmica, geralmente definida como artigos de periódicos ou patentes, e, depois, comparar e classificar o estado do conhecimento com gráficos e mapas. Ao longo dos anos, isso vem se tornando uma referência amplamente aceita (GODIN, 2006).

Le Coadic (2004) afirma que o objetivo da bibliometria é mensurar as ações de gestão dos documentos. Para Araújo (2006, p. 12), a bibliometria é a “técnica quantitativa e estatística de medição dos índices de produção e disseminação do conhecimento científico”, que também utiliza matemática, para expor características da literatura de distintos meios de comunicação.

Um elemento textual importante, que pode fornecer uma interpretação concisa do conteúdo de trabalhos acadêmicos, são as palavras-chave. É usual ver a técnica de nuvens de palavras-chave para evidenciar conceitos mais relevantes. Porém, as nuvens de *tags* exibem apenas a frequência de palavras-chave, e não demonstram as relações que ocorrem entre elas (FELDMAN; DAGAN; HIRSH, 1998). Na maioria das abordagens usando esse elemento textual, o principal atributo é a relevância da palavra dentro do texto (ZHU *et al.*, 2015).

2.2. Inteligência Artificial aplicada na bibliometria

A Inteligência Artificial (IA) é um ramo da Ciência da Computação, usado com muitos sentidos diferentes. Esta tecnologia está motivada pela construção de técnicas para auxiliar na análise, compreensão ou até na visualização de grandes quantidades de dados, estes oriundos de aplicações científicas ou empresariais. A IA também é empregada na descoberta de conhecimento, a partir da identificação de padrões, associações, mudanças e anomalias provindas de uma base de dados ou repositórios. A tecnologia pode ser usada por empresas na tomada de decisões ou por pesquisadores, para a descoberta de uma vacina, por exemplo (TSAI, 2011).

A IA “faz parte da vida das pessoas por meio do oferecimento de ferramentas que facilitam tarefas diárias”, sendo muito utilizada “para interpretar os dados [...], sendo capaz de guardar, cruzar e analisar informações em uma quantidade maior que os próprios seres humanos, bem como em menor tempo” (SANTOS; CAMILO; MELLO, 2018, p. 53). Para cumprir essas tarefas, a “máquina tem que ser capaz de reconhecer a linguagem natural

falada pelos humanos (...) sendo capaz de realizar complexas análises morfológicas, sintáticas, semânticas e contextuais sobre a informação que recebe” (GARCÍA, 2012, p. 53). Segundo o autor, “o processamento da linguagem natural, o PNL (NLP, Natural Language Processing) é um ramo da Inteligência Artificial que se ocupa das capacidades de comunicação dos computadores com os humanos, utilizando a sua própria linguagem” (2012, p. 53).

3. METODOLOGIA

Nesta seção apresentam-se a caracterização da pesquisa, as ferramentas utilizadas, os dados trabalhados na pesquisa, o modelo de análise e os procedimentos empregados para alcançar os objetivos.

Seguindo as orientações de Gil (2008), esta pesquisa se caracteriza como de natureza aplicada, com objetivos exploratórios de descritivos, e com procedimentos de análise quanti-qualitativos.

Foram utilizadas duas ferramentas: o Gephi e o algoritmo Force Atlas 2.

1. O Gephi é um *software* utilizado em estudos e análise de dados de rede, que busca as relações entre grupos de pessoas, instituições, eventos, entre outros, e para o entendimento de temáticas tais como transmissão de doenças, difusão de ideias, inovação e mudanças de conceitos ao longo do tempo (CHERVEN, 2015). É uma ferramenta interativa para análise e visualização de gráficos, que permite estudos mais detalhados, sem a necessidade de experiência com programação (KHOKHAR, 2015).
2. O algoritmo Force Atlas 2, um melhoramento do Force Atlas, faz um equilíbrio entre a qualidade do layout e a velocidade de cálculos computacionais, resultando em considerável melhora do algoritmo no que se refere à substituição da simulação direta de soma, considerando apenas a força de repulsão entre cada nó. O algoritmo foi escolhido, pois permite agrupar os nós interligados e afastar os que não estão interligados.

Os insumos textuais trabalhados se referem aos dados abertos da Capes, que se alinham à Lei de Acesso à Informação (LAI), que instituiu a Política de Dados Abertos nacionais, em vigor desde maio de 2012. O inciso III do art. 2º do Decreto 8.777/2016, define como dados abertos aqueles estruturados, acessíveis sob licença aberta e livre, a humanos e processáveis por máquina, desde que creditando-se a autoria ou a fonte dos dados (BRASIL, 2016). Esses dados devem satisfazer a alguns atributos jurídicos, assegurando o uso, reuso e a

distribuição livre, com a adoção de uma licença específica, devendo obedecer às boas práticas para que sejam disponibilizados na web (LÓSCIO; BURLE; CALEGARI, 2016). A Capes é uma agência ligada ao Ministério da Educação e Cultura, criada com o objetivo de garantir a formação especializada, atendendo aos empreendimentos públicos e privados, tendo a função de avaliar, certificar e financiar a pós-graduação no Brasil. Sendo uma instituição pública, segue a LAI e disponibiliza dados abertos sobre a pós-graduação, razão pela qual foi selecionada como insumo de coleta de dados. De acordo com o portal (CAPES, 2019), estão disponíveis quatro categorias de dados: Acessos ao Portal de Periódicos, Bolsas e Auxílios, Avaliação da Pós-Graduação Stricto Sensu e Servidores, Contratos, Dotação e Execução Orçamentária, totalizando 47 conjuntos de dados.

O modelo de análise quantitativo é o da coocorrência das palavras-chave catalogadas no conjunto de dados da Capes. Na análise qualitativa são utilizados os princípios da teoria dos grafos, na qual “um grafo é composto por um conjunto de objectos ligados entre si”, sendo esses objetos “denominados de vértices, enquanto que as ligações, caracterizando uma relação entre os objectos, são denominadas de arestas” (GUERREIRO, 2012, p. 5). As medidas de análise dependerão do tipo de análise pretendida, e, segundo o autor, a análise feita em nível do nó evidencia o papel de um elemento específico na rede global (importância; centralidade; prestígio), e a análise feita a nível da própria rede permite perceber a estrutura mais ampla que gerou a rede.

Os procedimentos metodológicos foram divididos em quatro etapas: 1) exploração e coleta de dados sobre a pós-graduação brasileira; 2) análise e seleção da amostra; 3) padronização dos dados e agrupamentos; 4) análises das conexões.

4. RESULTADOS PARCIAIS

Etapa 1: exploração e coleta de dados sobre a pós-graduação brasileira: coleta dos dados ocorreu entre maio e julho de 2021. Estavam disponíveis 47 conjuntos de dados abertos no Portal. Destes, 3 conjuntos referentes ao “Catálogo de Teses e Dissertações - Brasil”, cobrindo o período de 1987 a 2019.

Etapa 2: análise e seleção da amostra: os dados foram analisados e foi escolhido o conjunto de dados do Catálogo de Teses e Dissertações - Brasil, com período de 2017 a 2019 (mais recente). Após prévia avaliação, selecionaram-se os dados da Grande Área de Conhecimento da “Ciência da Informação”, obtendo-se 1.116 trabalhos (teses e dissertações), representando 0,42% do total contido nos arquivos. Foram mantidos os seguintes dados: Ano Base de

Coleta dos Dados, Nome da Produção (título), Descrição do Abstract, Descrição do Keyword, Nome do Subtipo da Produção, Nome do Orientador, Descrição do Resumo e Nome da Instituição de Ensino.

Etapa 3: padronização dos dados e agrupamentos: a desambiguação dos dados foi realizada fazendo-se a normalização quanto a acentos e caracteres especiais. Utilizaram-se técnicas de PNL (Stemming; Lemmatization) para agrupar as palavras-chave. Por exemplo: as palavras “BIBLIOTECAS” e “BIBLIOTECA” foram agrupadas como “BIBLIOTECA”; “INSTITUTO FEDERAL DE EDUCAÇÃO” e “INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA” foram agrupadas como “INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA”.

Etapa 4: análises das conexões: gerou-se uma matriz de coocorrência, onde os nós são as palavras-chave dos documentos. O grau de um nó representa a quantidade de vezes que uma mesma palavra-chave aparece. Quanto mais vezes uma palavra-chave se repetir, maior será sua representação visual no grafo. As arestas são a coocorrência das palavras-chave em um determinado trabalho acadêmico. A espessura da aresta indica a força do relacionamento, isto é, a quantidade de vezes que duas palavras foram citadas em um mesmo trabalho.

5. CONSIDERAÇÕES FINAIS

Esta pesquisa se enquadra dentro dos estudos métricos, que busca analisar padrões de coocorrência de palavras-chave nas dissertações e teses da área da CI, e, com os resultados, espera-se possibilitar parcerias entre pesquisadores, elencar tendências de pesquisa e evidenciar temáticas pouco exploradas.

REFERÊNCIAS

ARAÚJO, C. A. Bibliometria: evolução histórica e questões atuais. *Em questão*, Porto Alegre, v. 12, n. 1, p. 11-32, jan./jun. 2006.

ARAÚJO, C. A. Fundamentos da Ciência da Informação: correntes teóricas e o conceito de informação. *Perspectivas em Gestão & Conhecimento*, João Pessoa, v. 4, n. 1, p. 57-79, jan./jun. 2014.

BRASIL. Decreto nº. 8.777, de 11 de maio de 2016. Institui a Política de Dados Abertos do Poder Executivo federal. *Diário Oficial [da] República Federativa do Brasil*, Poder Executivo, Brasília, DF, 12 maio 2016. Seção 1, p. 21.

CHERVEN, K. *Mastering Gephi network visualization: produce advanced network graphs in gephi and gain valuable insights into your network datasets*. Birmingham: Packt Publishing, 2015.

- COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR (CAPES). *Dados Abertos Capes*. Brasília, 2019. Disponível em: <https://dadosabertos.capes.gov.br>. Acesso em: 24 jun. 2021.
- DURIAU, V. J.; REGER, R. K.; PFARRER, M. D. A content analysis of the content analysis literature in organization studies: research themes, data sources, and methodological refinements. *Organizational Research Methods*, [S.l.], v. 10, n. 1, p. 5-34, jan. 2007.
- FAYYAD, U. Knowledge discovery in databases: an overview. In: DŽEROSKI, S.; LAVRAČ, N. (Ed.). *Relational Data Mining*. Berlin: Heidelberg, 2001. p. 28-47.
- FELDMAN, R.; DAGAN, I.; HIRSH, H. Mining text using keyword distributions. *Journal Of Intelligent Information Systems*, [S.l.], v. 10, n. 3, p. 281-300, mar. 1998.
- GARCÍA, A. *Inteligência Artificial: fundamentos, prática y aplicaciones*. Madrid: RC Libros, 2012.
- GIL, A. C. *Métodos e técnicas de pesquisa social*. 6. ed. São Paulo: Atlas, 2008.
- GODIN, B. On the origins of bibliometrics. *Scientometrics*, [S.l.], v. 68, n. 1, p. 109-133, jul. 2006.
- GUERREIRO, A. J. C. *Análise de redes sociais: aplicação a uma rede de clientes*. 2012. 73f. Dissertação (Mestrado em Análise de Dados e Sistemas de Apoio à Decisão) - Faculdade de Economia, Universidade do Porto, Porto, 2012.
- KHOKHAR, D. *Gephi Cookbook: over 90 hands-on recipes to master the art of network analysis and visualization with gephi*. Birmingham: Packt Publishing, 2015.
- LE COADIC, Y. F. *A Ciência da Informação*. 2. ed. Brasília: Briquet de Lemos, 2004.
- LEE, H.; YI, G.-S.; PARK, J. C. E3Miner: a text mining tool for ubiquitin-protein ligases. *Nucleic Acids Research*, Bethesda, v. 36, n. 1, p. 416-422, maio 2008.
- LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. Boas Práticas para Dados na WEB: desafios e benefícios. *Revista Principia : Divulgação Científica e Tecnológica do IFPB, João Pessoa*, v. 1, n. 32, p. 9-26, dez. 2016.
- SANTOS, B. R. P.; CAMILO, E. S.; MELLO, M. R. G. Big data e Inteligência Artificial: aspectos éticos e legais mediante teoria crítica. *Complexitas: Revista de Filosofia Temática*, Belém, v. 3, n. 1, p. 50-60, jan./jun. 2018.
- TAN, P.; STEINBACH, M; KUMAR, V. *Introdução ao datamining: mineração de dados*. Rio de Janeiro: Ciência Moderna, 2009.
- TSAI, H.-H. Research trends analysis by comparing data mining and customer relationship management through bibliometric methodology. *Scientometrics*, [S. l.], v. 87, n. 3, p. 425-450, mar. 2011.
- WANG, L. *et al.* G-Hadoop: mapreduce across distributed data centers for data-intensive computing. *Future Generation Computer Systems*, [S.l.], v. 29, n. 3, p. 739-750, mar. 2013.
- ZHU, L. *et al.* Keywords co-occurrence mapping knowledge domain research based on the theory of Big Data in oil and gas industry. *Scientometrics*, [S.l.], v. 105, n. 1, p. 249-260, ago. 2015.
- ZINS, C. Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society For Information Science And Technology*, [S.l.], v. 58, n. 4, p. 479-493, abr. 2007.

