Informação em Pauta

Fortaleza, CE v. 8 2023 ISSN 2525-3468
DOI: 10.36517/2525-3468.ip.v8i0.2023.92075.1-3

RESUMO DE TESE

SOUZA, Marcos de. **O comportamento de termos da Ciência da Informação por meio da modelagem de tópicos**. Orientação: Renato Rocha Souza. 2020. 405 f. Tese (Doutorado em Gestão e Organização do Conhecimento) – Programa de Pós-Graduação em Gestão & Organização do Conhecimento, Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2020. Disponível em: https://repositorio.ufmg.br/handle/1843/34292. Acesso em: 20 out. 2023.

O COMPORTAMENTO DE TERMOS DA CIÊNCIA DA INFORMAÇÃO POR MEIO DA MODELAGEM DE TÓPICOS

THE BEHAVIOR OF INFORMATION SCIENCE TERMS THROUGH TOPIC MODELING

RESUMO

O crescimento da pesquisa, ciência e tecnologia na perspectiva acadêmica tem contribuído para a produção de uma quantidade elevada de informações científicas produzidas em diversos formatos e tipos de documentos da comunicação científica. Levando em consideração a quantidade, variedade e complexidade de informações produzidas, tem sido cada vez mais necessário o uso de tecnologias e métodos para elaboração e produção de registros de informação, além da necessidade de produzir informações sobre informações. A Modelagem de Tópicos, constituída de métodos estatísticos/probabilísticos e recursos tecnológicos, utiliza modelos de algoritmos de aprendizagem que possibilita identificar padrões, organizar coleções, resumir conteúdos, extrair tópicos mais frequentes, identificar relações entre assuntos e mudanças realizadas ao longo do tempo em *corpora* de documentos. Partindo desse princípio, questiona-se: de que forma tem se apresentado, na segunda década do século XXI, os temas da produção científica brasileira na área da Ciência da Informação quando se comparado às áreas e disciplinas já estabelecidas na literatura por pesquisadores como núcleo da área? O objetivo geral buscou verificar a proximidade e o distanciamento entre os temas extraídos dos *corpora* de dados constituídos por documentos científicos com as áreas e disciplinas da Ciência da Informação estabelecidas na literatura. Dentre os objetivos específicos constam identificar, analisar e discutir o comportamento diacrônico dos termos extraídos dos *corpora* de dados, bem com suas respectivas relações, além de analisar e discutir os modelos de treinamento de extração de tópicos, selecionar os resultados significativos e validar junto à comunidade científica brasileira da Ciência da Informação. Justifica-se a importância desta pesquisa uma vez que a comparação entre estudos – mesmo que utilizando de metodologias e intervalos de tempo diferentes na composição de documentos – permite apresentar, por meio do mapeamento científico, novos resultados e prospectar diferentes cenários e perspectivas para a ciência estudada. Para a pesquisa empírica foram realizadas as etapas de coleta de dados e formação dos *corpora* de dados; preparação e pré-processamento referente à limpeza, manipulação, combinação e normalização dos dados; transformação dos dados referentes às operações matemáticas

e estatísticas aplicadas; modelagem e processamento, ao qual conecta os dados tratados aos modelos *Latent Semantic Indexing* e *Latent Dirichlet Allocation*; apresentação dos resultados por meio de sínteses textuais e gráficos interativos e estatísticos; validação dos resultados junto a pesquisadores da área estudada; e documentação gerada a partir dos resultados empíricos com o referencial teórico. Dentre os principais resultados constam: o comportamento parcialmente diferente entre o mapeamento científico das disciplinas do núcleo da Ciência da Informação encontrado na literatura com os resultados empíricos desta pesquisa; o comportamento diacrônico e surgimento de termos em pesquisas na área da Ciência da Informação, como *fake news*, *big data* e *machine learning*; a proximidade e o distanciamento entre disciplinas como Sistemas de Informação e Comunicação Científica Eletrônica; os melhores resultados na modelagem de tópicos realizada por meio do modelo *Latent Dirichlet Allocation*, levando em consideração o equilíbrio entre os pesos dos resultados e um maior número de bigramas e trigramas que contribuem para a uma melhor interpretação dos dados, realizada pelo indexador e validada pela comunidade científica.

Palavras-chave: modelagem de tópicos; alocação de Dirichlet latente; proximidade e distanciamento; comportamento diacrônico.

ABSTRACT

The growth of research, science and technology from an academic perspective has contributed to the production of a large amount of scientific information produced in various formats and types of scientific communication documents. Considering the amount, variety and complexity of information produced, it has been increasingly necessary to use technologies and methods for the elaboration and production of information records, in addition to the need to produce information about information. The Topic Modeling consisting of statistical / probabilistic methods and technological resources uses models of learning algorithms that make it possible to identify patterns, organize collections, summarize content, extract more frequent topics, identify relationships between issues and changes made over time in corpora of documents. Based on this principle, the question is: in what way has the themes of Brazilian scientific production in the area of Information Science been presented in the second decade of the XXI century when comparing the areas and disciplines already established in the literature by researchers as the core of the area? The general objective was to verify the proximity and the distance between the themes extracted from the data corps constituted by scientific documents and the areas and disciplines of Information Science established in the literature. Among the specific objectives were to identify, analyze and discuss the diachronic behavior of the terms extracted from the data corpora, as well as their respective relationships, and to analyze and discuss the training models for topic extraction, to select the significant results and to validate them with the Brazilian scientific community of Information Science. The importance of this research is justified since the comparison between studies, even if using different methodologies and time intervals in the composition of documents, allows presenting, through scientific mapping, new results and prospecting different scenarios and perspectives for the studied science. For the empirical research were carried out the steps data collection and formation of data corpora, preparation and pre-processing referring to cleaning, manipulation,

combination and normalization of data, transformation of the data referring to mathematical operations and applied statistics, modeling and processing to which connects the data treated with the Latent Semantic Indexing models, and Latent Dirichlet Allocation, presentation of the results through textual synthesis and interactive graphics and statistics, validation of the results with researchers in the studied area and documentation generated from the empirical results with the theoretical reference. Among the main results are the partially different behavior between the scientific mapping of the disciplines of the Information Science core found in the literature with the empirical results of this research; diachronic behavior and emergence of terms in research in the area of Information Science such as fake news, big data and machine learning; Proximity and distance between disciplines such as Information Systems and Electronic Scientific Communication; Better results in the modeling of topics using the Latent Dirichlet Allocation model taking into account the balance between the weights of the results and a greater number of bigrams and trigrams that contribute to a better interpretation of the data carried out by the indexer and validated by the scientific community.

Keywords: topic modelling; latent Dirichlet allocation; proximity and distance; diachronic behavior.

SOBRE O AUTOR

Marcos de Souza
Doutor em Gestão e Organização do Conhecimento pelo Programa de Pós-Graduação em Gestão e Organização do Conhecimento da Universidade Federal de Minas Gerais (UFMG).

E-mail: marcosdesouza82@gmail.com

ACESSO ABERTO

Financiamento: FAPEMIG

Recebido em: 2 set. 2023.
Aceito em: 20 out. 2023.
Publicado em: 20 out. 2023.

Como citar este resumo:

SOUZA, Marcos de. O comportamento de termos da Ciência da Informação por meio da modelagem de tópicos. **Informação em Pauta**, Fortaleza, v. 8, p. 1-3, 2023. DOI: 10.36517/2525-3468.ip.v8i0.2023.92075.1-3.