Mitigando os Limites das Métricas Atuais de Avaliação de Estratégias de Modelagem de Tópicos

Antônio Pereira de Souza Júnior
antonio.pereira@aluno.ufsj.edu.br
Universidade Federal de São João Del Rei – UFSJ
Minas Gerais, Brasil

Felipe Augusto Resende Viegas
frviegas@dcc.ufmg.br
Universidade Federal de São João Del Rei – UFSJ
Minas Gerais, Brasil

Leonardo Chaves Dutra da Rocha
lcrocha@ufsj.edu.br
Universidade Federal de São João Del Rei – UFSJ
Minas Gerais, Brasil

ABSTRACT

Topic Modeling (TM) helps extract and organize information from large amounts of textual data by discovering semantic topics from documents. This master thesis delves into issues of *topic quality evaluation*, responsible for driving the advances in the TM field by assessing the overall quality of the topic generation process. Since traditional TM metrics capture the quality of topics by strictly evaluating the words that make up the topics, either syntactically (e.g., NPMI, TF-IDF Coherence) or semantically (e.g., WEP), we investigate whether we are approaching the limits of what the current evaluation metrics can assess regarding TM quality. For this, we perform a comprehensive experimental evaluation, considering three widely used datasets (ACM, 20News, and WOS) for which a natural organization of the collection's documents into semantic classes (topics) does exist. We contrast the quality of topics generated by four traditional and state-of-the-art TM techniques (i.e., LDA, NMF, CluWords, and BERTopic) with each collection's "natural topic structure". Our results show that, despite the importance of the current metrics, they could not capture some important idiosyncratic aspects of the TM task, in this case, the capability of the topics to induce a structural organization of the document space into distinct semantic groups. To mitigate such limitations, we propose incorporating metrics commonly used to evaluate clustering algorithms into the TM evaluation process, relying on some commonalities between TM and clustering tasks. Results highlight the effectiveness of clustering metrics in distinguishing the results of TM techniques compared to the datasets'*ground truth* (class organization). However, adopting additional evaluation metrics implies expanding the analysis space. Thus, as a third contribution, we propose consolidating the various metrics into a unified framework, using Game Theory for decision-making, specifically Multi-Attribute Utility Theory (MAUT). Our experimental results demonstrate that MAUT allows a more precise assessment of TM quality.

KEYWORDS

Modelagem de tópicos, Avaliação de modelagem de tópicos, Aprendizado de máquina, NLP, Mineração de dados

In: VI Concurso de Teses e Dissertações (CTD 2024). Anais Estendidos do XXX Simpósio Brasileiro de Sistemas Multimídia e Web (CTD'2024). Juiz de Fora/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2024.

ISSN 2596-1683

1 INTRODUÇÃO

1.1 Contexto e problema

Modelagem de Tópicos (MT) é uma das mais proeminentes e úteis abordagens não-supervisionada para extrair e organizar informações de grandes volumes de dados textuais. Apesar dos enormes avanços de novas estratégias, inclusive baseadas em Grandes Modelos de Linguagem (*Large Language Models* - LLM) de primeira geração (i.e., *transformers*) [2], o processo de avaliação da qualidade dos tópicos gerados e, consequentemente, da qualidade das técnicas em si, ainda é um desafio. Duas abordagens principais têm sido utilizadas na avaliação dessas estratégias, sintática e semântica, ambas restritas às palavras que sumarizam os tópicos.

A contribuição central da dissertação de mestrado do Antônio Pereira está em um estudo detalhado e aprofundado das limitações das métricas de avaliação atuais para MT. A partir da constatação das limitações, a dissertação propõe a adaptação de métricas de avaliação de tarefas de clusterização para o cenário de MT e uma metodologia de avaliação de MT mais abrangente, baseada em uma abordagem de avaliação multiperspectiva, oriunda da teoria de jogos, para avaliar a organização das palavras mais representativas em tópicos, bem como a capacidade das estratégias de MT em induzir uma organização estrutural dos documentos das coleções.

1.2 Aderência ao WebMedia

A aderência da presente dissertação de mestrado ao WebMedia começa pelo seu tema central, Modelagem de Tópicos, relacionado a diferentes tópicos de interesse: 1) Engenharia de Documentos, Modelos e Linguagens, 2) IA, Aprendizado de Máquina e Aprendizado Profundo e 3) Processamento de Linguagem Natural. Conforme detalhado nos subprodutos, a presente dissertação resultou na publicação de quatro artigos no WebMedia (dois em 2023, um em 2022 e um em 2021), sendo que um dos trabalhos publicados no Webmedia 2023 foi eleito o melhor artigo completo do evento. Tratam-se de fatos contundentes de que a dissertação trata de um problema de claro interesse para as áreas de Sistemas Multimídia e Web.

2 METODOLOGIA

2.1 Definição do Problema

Definimos o problema a ser tratado investigando se estamos nos aproximando dos limites das métricas atuais em relação à qualidade de avaliação dos tópicos no contexto de MT. Realizamos um experimento abrangente, considerando três coleções de dados amplamente utilizadas na literatura, para as quais o tópico (classe) de

cada documento é conhecido (i.e., ACM, 20News e WOS). Comparamos a qualidade dos tópicos gerados por quatro das principais estratégias de MT (i.e., LDA, NMF, CluWords e BERTopic) com a estrutura de tópicos prévia de cada coleção. Nossos resultados mostram que, apesar da importância das métricas de avaliação atuais, estas não conseguiram captar alguns aspectos idiossincráticos importantes da MT, indicando a necessidade de propor novas métricas que considerem, por exemplo, a estrutura e organização dos documentos que compõem os tópicos.

2.2 Objetivo

Mitigar as limitações das métricas de avaliação de MT atuais propondo/adaptando e avaliando novas métricas.

2.3 Solução

Propomos adaptar métricas comumente utilizadas para avaliar algoritmos de clusterização [3], uma vez que existem semelhanças significativas entre as estratégias de MT e de clusterização, como sua natureza não-supervisionada e o objetivo de agrupar elementos semelhantes: Silhouette Score, Calinski-Harabasz e BetaCV. No entanto, isto implica expandir o espaço de análise por meio da inclusão de um novo conjunto de métricas. Assim, propomos também a consolidação das várias métricas, que consideram tanto a qualidade das palavras que compõem os tópicos (tradicionais) como a estrutura organizacional dos documentos, num resultado unificado multiperspectiva, utilizando metodologia Multiattribute Utility Theory (MAUT) [1] (MAUT). Trata-se de uma metodologia comum em Teoria de Jogos para tomada de decisão, que visa auxiliar na escolha de uma alternativa entre várias opções, considerando múltiplos atributos e preferências conflitantes. A MAUT se fundamenta em três conceitos principais: atributos, funções de utilidade e pesos. Na proposta utilizada neste trabalho, os atributos correspondem aos valores de todas as métricas utilizadas (i.e., NPMI, TF-IDF Coherence, WEP, Silhouette Score, Calinski-Harabasz e Beta CV), a função de utilidade é a min-max [4] e os pesos, que representam a importância relativa de cada atributo, foram iguais para todas as métricas. Com essa adaptação, os tópicos gerados por cada uma das técnicas de MT podem ser avaliados sob diferente perspectivas.

2.4 Avaliação

A avaliação experimental proposta considerou as mesmas três coleções de dados (i.e., ACM, 20News e WOS), primeiramente comparando os tópicos gerados pelas mesmas estratégias de MT (i.e., LDA, NMF, CluWords e BERTopic), com a estrutura prévia de classes dos documentos de cada coleção, o *ground truth*, considerando a três métricas adaptadas do contexto de clustering (i.e., Silhouette Score, Calinski-Harabasz e BetaCV). Analisando os resultados obtidos, conforme esperado, a estrutura do *ground truth* demonstrou o melhor desempenho nessas métricas, com valores superiores aos encontrados nas estratégias de MT. Ao contrário das métricas tradicionais de MT que apresentaram resultados contraditórios, quando os resultados do *ground truth* eram comparados com os tópicos gerados pelas técnicas de MT, as métricas de clusterização, sob esse aspecto, foram bem consistentes, evidenciando sua eficácia na distinção dos resultados dos algoritmos de MT e do *ground truth*. Por outro lado, para ambos os conjuntos de métricas, tradicionais de MT e de clusterização, não houve um consenso que apontasse qual a melhor das estratégias de MT para todas as coleções. A partir dessa constatação, todos os algoritmos de MT foram novamente avaliados na mesma configuração experimental, dessa vez considerando a MAUT. Especificamente, nossos resultados permitiram observar os avanços semânticos gerados pelo uso de *word embeddings* em algumas estratégias de MT, além da solidez e consistência da construção de tópicos por meio de estratégias de fatorização de matrizes.

3 AVANÇO NO ESTADO-DA-ARTE

O principal avanço da dissertação está na identificação das limitações das métricas atuais de avaliação de estratégias de MT que impactam diretamente na evolução dessas estratégias. Apresentamos uma proposta inicial de novas métricas e também de uma metodologia de avaliação multiperspectiva por meio da MAUT que abrem caminho para que futuras pesquisas desenvolvam e refinam ainda mais o processo de avaliação de estratégias de MT, garantindo uma comparação justa, permitindo melhorar a geração e organização de tópicos em grandes conjuntos de dados textuais.

4 CONCLUSÃO

Neste trabalho, voltamos nossa atenção para um importante desafio no contexto da MT, que é a avaliação de tópicos gerados pelas diversas estratégias. As métricas tradicionais capturam a qualidade dos tópicos avaliando estritamente as palavras que construíram os tópicos, sintaticamente (ou seja, NPMI, TF-IDF *Coherence*) ou semanticamente (ou seja, WEP). Este trabalho demonstrou empiricamente, através de experimentos extensivos, que o conjunto atual de métricas negligência um importante aspecto que geralmente é esperado por estratégias de MT, a capacidade de organizar semanticamente o espaço de documentos em grupos significativos.

Foi proposto então, a utilização de métricas empregadas na avaliação de estratégias de clusterização, que quantificam a qualidade da estrutura e organização dos documentos que compõem os tópicos. Foi adotado também uma abordagem que tem o potencial de integrar ambos os conjuntos de métricas, MAUT, resultando em uma avaliação unificada e mais abrangente. Essa estratégia permitiu avaliar melhor as várias estratégias de MT presentes na literatura.

Como trabalho futuro, pretendemos empregar a metodologia de avaliação desenvolvida para contrastar os algoritmos de MT utilizados aqui com a geração de tópicos por meio de modelos *Large Language Model* (LLM). Nosso objetivo é estender as análises que tratam dos tópicos induzidos pelas estratégias de MT. Considerando outras coleções e sistemas de classificação, bem como estender também nossa avaliação para estratégias de tópicos hierárquicas, considerando avaliação de Modelagem de Tópicos Hierárquica (MTH).

AGRADECIMENTOS

Este trabalho foi financiado por CNPq, CAPES, Fapemig e AWS

REFERÊNCIAS

[1] Rodrigo Carvalho, Nícollas Silva, Luiz Chaves, Adriano C. M. Pereira, and Leonardo Rocha. 2019. Geographic-categorical diversification in POI recommendations. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web, WebMedia 2019, Rio de Janeiro, Brazil, October 29 - November 01, 2019*. ACM, 349–356.

[2] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[3] Gokhan Kul, Duc Thanh Anh Luong, Ting Xie, Varun Chandola, Oliver Kennedy, and Shambhu Upadhyaya. 2018. Similarity metrics for SQL query clustering. *IEEE Transactions on Knowledge and Data Engineering* 30, 12 (2018), 2408–2420.

[4] C Saranya and G Manikandan. 2013. A study on normalization techniques for privacy preserving data mining. *International Journal of Engineering and Technology (IJET)* 5, 3 (2013), 2701–2704.