INSTITUTO FEDERAL Goiás
14° Seminário de Iniciação Científica e Tecnológica

Sistemas de Recomendação em Instituições de Ensino e Pesquisa

João Victor Cangerana Rocha (IC), Daniel Xavier Sousa (PQ), Alexandre Bellezi Jose (PQ)

PIBITI - Câmpus Anápolis
daniel.sousa@ifg.edu.br, alexandre.jose@ifg.edu.br

***Palavras Chave**: Sistemas de Recomendação; Recuperação de Informação; Representação de Dados*

Introdução

Os algoritmos de Sistemas de Recomendação têm se destacado no cenário científico e comercial devido sua capacidade de filtrar documentos relevantes considerando um contexto de grande volume de dados. Junto a isso, entendemos como uma necessidade (principalmente em instituições multicampi) recomendar trabalhos científicos aos pesquisadores de uma intuição de forma a maximizar a cooperação e conhecimento de seus pares, que em muitos casos desconhecem o fato de colegas estudarem o mesmo tópico de pesquisa. Assim, este projeto apresenta uma ampliação da ferramenta já conhecida como IFG-PRODUZ (ifgproduz.ifg.edu.br), aplicando a funcionalidade de recomendação de produções científicas aos pesquisadores do IFG. No intuito de termos uma ferramenta de melhor performance, apresentamos avaliações de diversas estratégias de representação de dados dos currículos Lattes. Considerando TF-IDF e variações de Word2Vec, nossos resultados mostram que o modelo TI-IDF apresentou a melhor precisão na recomendação de produções científicas.

Metodologia

Os modelos de recomendação se diferenciam principalmente por: filtragem colaborativa, baseada em conteúdo e uma mistura de ambos. Devida à falta de avaliação das produções bibliográficas entre os pesquisadores, aplicamos a filtragem baseada em conteúdo. A intuição é que um pesquisador A deve ter interesse nos artigos do pesquisador B, se houver alta similaridade entre os artigos de A e artigos de B. Assim, nosso trabalho avalia duas formas bastantes aplicáveis na área de Recuperação de Informação (RI) para computar a similaridade entre os artigos. TF-IDF (MANNING, 2008), que considera os termos de um documento como uma relação independente entre outros termos. E ainda Word2Vec (MIKOLOV, 2013), que constrói uma representação denominada de **embedding**, que quantifica a informação do termo junto a outros termos, aprendendo o contexto das palavras. No intuito de ampliar nossas análises, executamos as estratégias Word2Vec considerando somente a base de dados do IFG-PRODUZ, executando uma construção própria dos **embeddings**, e ainda fazendo uso de dados já treinados a partir da Wikipedia.

Resultados e Discussão

A Tabela 1 apresenta a precisão dos nossos resultados, considerando métricas já conhecidas da área de RI. No caso, o modelo do TF-IDF apresentou melhor precisão na recomendação. Acreditamos que esses resultados se justificam, pois embora os currículos descrevem diversas áreas do conhecimento, os termos em cada área apresentam contextos bem definidos, ou seja, poucos casos em que duas palavras, de mesma sintaxe, se relacionem de forma diferentes com as outras palavras. Notamos que o modelo de Word2Vec treinado durante o projeto (com currículos Lattes) obteve um resultado similar ao modelo pré-treinado com a base do Wikipedia (base maior e geral). Acreditamos que o treinamento com palavras específicas dos currículos tenha ajudado a obter um modelo representativo.

Tabela 1: Resultados obtidos pelos modelos.

	MAP	MRR	NDCG
TD-IDF	0.807	0.844	0.740
Word2vec Treinado	0.803	0.816	0.735
Word2vec Pré-Treinado	0.800	0.809	0.731

Conclusões

Este trabalho apresenta uma análise de representação de dados no contexto de Sistemas de Recomendação para filtrar produções científicas aos pesquisadores de uma instituição. Nosso trabalho mostra que a representação do TF-IDF se mantém como melhor precisão neste contexto.

Agradecimentos

Agradecemos ao IFG por todo suporte e bolsa fornecida.

MIKOLOV, Tomas et al. Distributed representations of words and phrases and their compositionality. In: **Advances in neural information processing systems**. 2013.

MANNING, Christopher D.; et a.l. Scoring, term weighting and the vector space model. **Introduction to Information retrieval**, 2008.