PECHA KUCHA

Ciência de Dados aplicada ao Arca: desenvolvimento e disponibilização de ferramentas para recuperação da informação no Repositório Institucional da Fundação Oswaldo Cruz

Data Science applied to Arca: development and availability of tools for information retrieval in the Institutional Repository of Fundação Oswaldo Cruz

Ciencia de Datos aplicada al Arca: desarrollo y disponibilización de herramientas para recuperación de la información en el Repositorio Institucional de la Fundação Oswaldo Cruz

Marcel de Moraes Pedroso | marcel.pedroso@icict.fiocruz.br

Fundação Oswaldo Cruz (FIOCRUZ), Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT), Rio de Janeiro, RJ, Brasil

Jefferson da Costa Lima | jefferson.lima@icict.fiocruz.br

Fundação Oswaldo Cruz (FIOCRUZ), Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT), Rio de Janeiro, RJ, Brasil

Vinicius Belchior Assef Neto | vinicius.assef@icict.fiocruz.br

Fundação Oswaldo Cruz (FIOCRUZ), Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT), Rio de Janeiro, RJ, Brasil

Resumo

O repositório institucional Arca é o principal instrumento de realização do acesso aberto na Fundação Oswaldo Cruz, tendo como missão reunir, hospedar, preservar, disponibilizar e dar visibilidade à produção intelectual da Instituição. A diversidade temática e a complexidade institucional da Fundação fomentam um desafio metodológico relacionado a classificação e recuperação dos objetos digitais depositados e a governança dos metadados registrados pelas comunidades que integram o repositório. Em 2016 o mecanismo de busca do Arca contabilizou mais de 400 mil consultas. É necessário um sistema de Recuperação da Informação que atenda as especificidades de indexação do repositório e a crescente demanda por informação por parte dos usuários internos e externos a Fiocruz. Neste trabalho propomos a utilização de ferramentas de Ciência de Dados, especialmente técnicas de Mineração de Dados e Aprendizagem de Máquina com o objetivo de aprimorar a Recuperação da Informação, por meio da classificação automática de objetos digitais depositados no Arca e o desenvolvimento e disponibilização de sistema de RI baseado em métricas de qualidade relacionadas aos conceitos de precisão e revocação.

Palavras-chave: Ciência de Dados; Armazenamento e Recuperação da Informação; Mineração de Dados; Aprendizagem de Máquina; Repositórios Institucionais.

Abstract

The Arca institutional repository is the main instrument of open access at the Oswaldo Cruz Foundation, with the mission of gathering, hosting, preserving, making available and giving visibility to the institution's intellectual production. The thematic diversity and institutional complexity of the Foundation foster a methodological challenge related to the classification and retrieval of deposited digital objects and the governance of the metadata recorded by the communities that make up the repository. In 2016, the Arca search engine counted more than 400 thousand queries. An Information Retrieval system is needed that meets the specificities of indexing the repository and the growing demand for information from users internal and external to Fiocruz. In this work we propose the use of Data Science tools, especially Data Mining and Machine Learning techniques, with the objective of improving Information Retrieval by means of automatic classification of digital objects deposited in the Arca and the development and availability of the system of IR based on quality metrics related to precision and recall concepts.

Keywords: Data Science; Information Storage and Retrieval, Data Mining; Machine Learning; Institutional Repositories.

Resumen

El repositorio institucional Arca es el principal instrumento de realización del acceso abierto en la Fundación Oswaldo Cruz, teniendo como misión reunir, hospedar, preservar, poner a disposición y dar visibilidad a la producción intelectual de la Institución. La diversidad temática y la complejidad institucional de la Fundación fomentan un desafío metodológico relacionado con la clasificación y recuperación de los objetos digitales depositados y la gobernanza de los metadatos registrados por las comunidades que integran el repositorio. En 2016 el mecanismo de búsqueda del Arca contabilizó más de 400 mil consultas. Es necesario un sistema de Recuperación de la Información que atienda las especificidades de indexación del repositorio y la creciente demanda por información por parte de los usuarios internos y externos a Fiocruz. En este trabajo proponemos la utilización de herramientas de Ciencia de Datos, especialmente técnicas de Minería de Datos y Aprendizaje Automático con el objetivo de mejorar la Recuperación de la Información, a través de la clasificación automática de objetos digitales depositados en el Arca y el desarrollo y puesta a disposición del sistema de RI basado en métricas de calidad relacionadas con los conceptos de precisión y revocación.

Palabras clave: Ciencia de Datos; Almacenamiento y Recuperación de la Información; Minería de Datos; Aprendizaje Automático, Repositorios Institucionales.

Audiência

Pesquisadores, bibliotecários, docentes e discentes de instituições de ensino e pesquisa, bem como gestores de repositórios institucionais.

Proposta

O repositório institucional Arca[1] é o principal instrumento de realização do acesso aberto na Fundação Oswaldo Cruz[2], tendo como missão reunir, hospedar, preservar, disponibilizar e dar visibilidade à produção intelectual da Instituição. O Arca possui atualmente mais de 10 mil objetos digitais (artigos científicos e livros de seus pesquisadores, teses e dissertações de alunos dos seus cursos de pós-graduação, relatórios institucionais e recursos educacionais abertos).

A diversidade temática e a complexidade institucional da Fundação fomentam um desafio metodológico relacionado a classificação e recuperação dos objetos digitais depositados e a governança dos metadados registrados pelas comunidades (25 unidades técnico-científicas) que integram o repositório.

Em 2016 o mecanismo de busca do Arca registrou mais de 400 mil consultas. Nesse contexto, percebemos a necessidade de pensarmos em estratégias e sistemas de Recuperação da Informação[3] (RI) que

atendam as especificidades de indexação do repositório e a crescente demanda por informação por parte dos usuários internos e externos a Fiocruz.

Entendemos que o principal objetivo de um sistema de RI robusto é recuperar todos os documentos que são relevantes à nececisade de informação dos usuários e, ao mesmo tempo recuper o menor número possível de documentos irrelevantes. Esse equilíbrio não é uma tarefa metodologicamente trivial, nem tão pouco tecnologicamente simples.

No contexto de um sistema de RI complexo, a recuperação de dados consiste na identificação de quais documentos contêm as palavras-chave usadas pelo usuário para fazer uma consulta. Esta abordagem nem sempre é suficiente para satisfazer às necessidades do usuário, pois, em muitos casos, será interessante que o resultado também exiba os documentos que incluem sinônimos dos termos utilizados na consulta, além de apresentá-los em uma ordem que indique o quão relevantes eles são para aquela consulta.

Neste trabalho propomos o uso de ferramentas de Ciência de Dados[4], especialmente técnicas de Mineração de Dados e Aprendizagem de Máquina[5] com o objetivo de aprimorar a Recuperação da Informação, por meio da classificação automática de objetos digitais depositados no Arca e o desenvolvimento e disponibilização de sistema de RI baseado em métricas de qualidade relacionadas aos conceitos de precisão (documentos recuperados que são relevantes para a consulta) e revocação (documentos que são relevantes para a consulta e que são recuperados com êxito).

A Ciência de Dados é um conjunto de estratégias, ferramentas e técnicas que busca reunir equipes multidisciplinares formadas por pesquisadores com conhecimento substantivo do problema em análise, estatísticos, matemáticos e cientistas da computação. Trata-se de um campo de estudo bastante promissor e destaca-se pela capacidade de auxiliar a descoberta de informação útil a partir de grandes bases de dados, a tomada de decisão orientada por dados e a análise visual de grandes quantidades de informações[6].

Ela combina métodos tradicionais de análise com algoritmos sofisticados para processar grandes volumes de dados em formatos diversos (dados estrtuturados, semi-estruturados e não estruturados). O processo de análise, no âmbito da Ciência de Dados, envolve basicamente as fases de (i) coleta e ingestão: extração, transformação e carga; (ii) pré-processamento: seleção de registros, redução de dimensionalidade, normalização, criação de subconjuntos de dados; (iii) análise exploratória e mineração de dados: principalmente análises voltadas para classificação, associação, agrupamento, detecção de anomalias e predição; (iv) pós-processamento: interpretação de padrões, filtragem, visualização e acoplamento em sistemas de apoio a decisão e plataformas online para visualização.

Para o Arca, estão em desenvolvimento e disponibilização interfaces baseadas no ecossistema de softwares de código aberto disponíveis e implementados no âmbito da Plataforma de Ciência de Dados aplicada à Saúde do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde da Fundação Oswaldo Cruz visando (i) a indexação dos objetos digitais em base de dados não relacional (customização da ferramenta Elasticsearch e utilização do formato JSON - JavaScript Object Notation), (ii) proposição de classificação automática de objetos digitais depositados no Arca por meio de técnicas de Aprendizagem de Máquina (utilizando bibliotecas do R e Python), e (iii) sistema de Recuperação da Informação e Visualização de Dados (aplicativo Kibana) baseado em métricas de qualidade relacionadas a precisão e revocação.

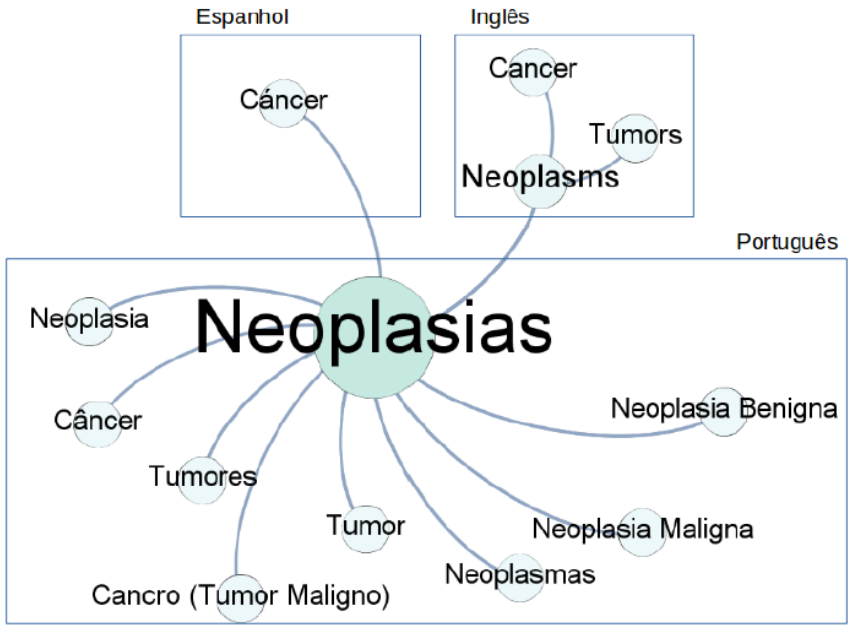
Como parte da pesquisa, desenvolvemos um estudo de caso[7] utilizando um subconjunto dos artigos, teses e dissertações depositados no Arca (4.707 objetos digitais extraídos em julho/2015). O objetivo do experimento foi a identificação automática de *n-grams* que pudessem melhor identificar os principais temas tratados em cada obra. *N-gram* é o termo genérico que define uma sequência contínua com n termos de um texto. Unigramas, bigramas e trigramas são os nomes dados aos *n-grams* com uma, duas e três palavras, respectivamente.

Em conjunto com os *n-grams* identificados, utilizamos o vocabulário Descritores em Ciências da Saúde[8] (DeCS). Ele é um vocabulário estruturado e trilíngue criado pela BIREME para servir como uma linguagem única na indexação de artigos, livros, relatórios técnicos, e outros tipos de materiais no campo da Saúde.

Depois de identificados os n-grams mais relevantes, verificamos se eles estão presentes no DeCS. Se estiverem, são associados ao documento como um descritor válido. De forma resumida, a seguir são listadas as etapas envolvidas no processo de extração de descritores. (i) Obtenção do Corpus; (ii) Pré-processamento dos documentos: Extração de dados dos arquivos PDF; Remoção de stopwords e de pontuação; Uso de stemming; Identificação de *n-grams* relevantes; Identificação do idioma principal do texto (inglês, português ou espanhol); (iii) Captura de dados do vocabulário Descritores em Ciências da Saúde (DeCS), e (iv) Cruzamento entre *n-grams* e o DeCS para a identificação de descritores para os documentos.

Como resultado inicial, podemos citar a melhora na revocação, com descritores em português sendo atribuídos aos títulos, mesmo para obras em inglês e espanhol, além da observância das remissivas definidas no DeCS. Na Figura 1, podemos observar uma parte dos termos que são agrupados sob o descritor Neoplasias. Com isso, reunimos obras correlacionadas e de idiomas diferentes sob um único termo.

Figura 1: *Remissivas em três idiomas para um descritor em português*



Fonte: Elaborada pelos autores com base em Lima (2016).

Das 4.707 obras analisadas, 63 tiveram o unigrama câncer identificado como relevante, mas uma busca pelo descritor Neoplasias retorna 103 obras. Um aumento superior a 60% no número de obras recuperadas. Este efeito é possível graças a combinação entre termos extraídos automaticamente e o uso de um vocabulário especializado.

A partir deste resultado inicial foi possível verificar o potencial para a aplicação de técnicas de Machine Learning na tarefa de classificação de textos. Em função do crescente volume de publicações, acreditamos que a pesquisa de métodos automáticos ou semiautomáticos para a classificação de textos se tornará uma importante ferramenta para o desenvolvimento e disponibilização de sistemas de RI baseados em métricas de qualidade relacionadas aos conceitos de precisão e revocação aplicadas ao Arca.

Referências

1. Fundação Oswaldo Cruz. Plano operativo: Arca repositório institucional: versão 1.1 setembro/2014. Rio de Janeiro, 2014. Consultado em abril de 2017.

2. Brasil. Fundação Oswaldo Cruz. Portaria 329/2014-PR. Instituir a Política de Acesso Aberto ao Conhecimento, visando garantir à sociedade o acesso gratuito, público e aberto ao conteúdo integral de toda obra intelectual produzida pela Fiocruz. Rio de Janeiro, março de 2014.

3. Baeza-Yates, R, Ribeiro-Neto, B. Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca. Bookman Editora, 2013.

4. Agarwal, R, Dhar, V. Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. Information Systems Research, Sep 2014, Vol.25(3), pp.443-448.

5. Castro, LN, Ferrari, DG. Introdução à Mineração de Dados. Conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

6. Instituto de Comunicação e Informação Científica e Tecnológica em Saúde. Plataforma de Ciência de Dados aplicada à Saúde. [homepage na internet]. Disponível em http://www.bigdata.icict.fiocruz.br/

7. Lima, JC. Análise lexicográfica da produção acadêmica da Fiocruz: uma proposta de metodologia. Dissertação - Rio de Janeiro: FGV/EMAP, setembro 2016.

8. Descritores em Ciências da Saúde da BIREME [homepage na internet]. Disponível em http://decs.bvs.br/