Estudo para integração entre a Plataforma Lattes a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) e o Banco de Teses e Dissertações da Capes

Study for integration between a Lattes Platform the Brazilian Digital Library of Theses and Dissertations (BDTD) and Bank of Thesis and Dissertations of Capes

Estudio para integración entre una Plataforma Lattes la Biblioteca Digital Brasileña de Tesis y Disertaciones (BDTD) y Banco de Tesis y Disertaciones de la Capes

Gabriel Lima Gomes | gabriel.lgo8@gmail.com
Instituto Brasileiro de Informação em Ciência e Tecnologia, Brasília/DF - Brasil

Washington Luís R. de Carvalho Segundo | wtonribeiro@gmail.com
Instituto Brasileiro de Informação em Ciência e Tecnologia, Brasília/DF - Brasil

Resumo

Este trabalho exibe uma estudo para integração entre a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), o Banco Teses e Dissertações da CAPES e a base nacional de currículos de pesquisadores (Plataforma Lattes). Ao todo foram analisados mais de 2 milhões de registros e foram adotados alguns procedimentos computacionais para coletar, normalizar e transformar os dados, além da aplicação de algoritmos de similaridade de strings para identificação de registros comuns entre as bases. Como resultado, observou-se que mais de 240 mil registros são estão na intersecção de BDTD e BTD CAPES, e que a Plataforma Lattes contém quase que a totalidade dos registros destas duas bases.

Palavras-chave: Algoritmos de Similaridade de strings; Bibliotecas Digitais de Teses e Dissertações; Repositórios Digitais; Base de teses e dissertações; Lattes;

Abstract

This work presents a study for integration between the Brazilian Digital Library of Theses and Dissertations (BDTD), CAPES's database of Theses and Dissertations and the national database of curriculum of researchers (Lattes platform). In all, more than 2 million records were analyzed and some computational procedures were adopted to collect, normalize and transform the data, as well as the application of string similarity algorithms to identify common registers between the databases. As a result, it has been observed that more than 240,000 records are at the intersection of BDTD and BTD CAPES, and that the Lattes Platform contains almost all of the records of these two databases.

Keywords: Strings Similarity Algorithms; Digital Libraries of Theses and Dissertations; Digital repositories; database of theses and dissertations; Lattes;

Resumen

Este trabajo muestra un estudio para la integración entre la Biblioteca Digital Brasileña de Tesis y Disertaciones (BDTD), el Banco Tesis y Disertaciones de la CAPES y la base nacional de currículos de investigadores (Plataforma Lattes). En total se analizaron más de 2 millones de registros y se adoptaron algunos procedimientos computacionales para recopilar, normalizar y transformar los datos, además de la aplicación de algoritmos de similitud de cadenas para identificación de registros comunes entre las bases. Como resultado, se observó que más de 240 mil registros están en la intersección de BDTD y BTD CAPES, y que la Plataforma Lattes contiene casi la totalidad de los registros de estas dos bases.

Palabras clave: Algoritmos de Similaridad de cadenas; Bibliotecas Digitales de Tesis y Disertaciones; Repositorios Digitales; Base de tesis y disertaciones; Lattes;

Introdução

A Biblioteca Digital Brasileira de Teses de Dissertações (BDTD) é uma rede que congrega repositórios brasileiros de acesso aberto que abrigam documentos do tipo tese ou dissertação, que foram produzidos no âmbito de programas de pós-graduação *strictu sensu* nacionais. Este consórcio nasceu em 2002, e vem sendo mantido por meio de uma iniciativa coordenada pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), que teve início com o apoio da agência nacional Financiadora de Estudos e Projetos (Finep)[1].

A Plataforma Lattes, que é a base nacional de currículos de pesquisadores do Brasil, surgiu em 1999 por meio de um projeto coordenado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), foi implementada por grupos universitários vinculados à Universidade Federal de Santa Catarina (UFSC) e a Universidade Federal de Pernambuco (UFPE), com contribuições também da empresa Multsoft.

O Banco de Teses e Dissertações da CAPES foi criado em 2002 pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) com intuito de abrigar os resumos de teses e dissertações dos programas de pós-graduação do país[2].

Nos últimos 5 anos intensificou-se a discussão política entre o IBICT, o CNPq e a CAPES, no sentido do estabelecimento de mecanismos comunicação, e interligação de dados, que façam com que a BDTD, a Plataforma Lattes e o Banco de Teses e Dissertações da CAPES interoperem, tendo a BDTD o papel de um *hub* central de informações sobre teses e dissertações brasileiras.

Objetivo

O presente trabalho tem por objeto a realização de estudo exploratório, de identificação de ferramentas e algoritmos para analisar a similaridade entre strings, que deverão subsidiar o intercâmbio e cruzamento de informações entre BDTD, Plataforma Lattes e Banco de Teses e Dissertações da CAPES.

Justificativa

É ponto pacífico que ao se abordar um assunto relativo a uma base nacional de currículos de pesquisadores, que venha a discussão o sistema da Plataforma Lattes. Fato que se justifica pelo grande volume de dados que esta base agrega, além da questão de que o cadastro de currículo na plataforma se tornou um pré-requisito para o ingresso a programas de pós-graduação e a obtenção de financiamento de bolsas e projetos nas principais agências de fomento nacionais e também nas estaduais.

Se tratando de bases de teses e dissertações, o Banco da CAPES é a iniciativa nacional que reúne o maior número de resumos e vem expandindo seu alcance para o acesso aos textos completos dos recursos que referencia. Neste quesito, a BDTD tem forte contribuição ao país pois agrega e viabiliza o acesso aberto a um volume representativo dos recursos digitais das teses e dissertações produzidas em âmbito nacional.

Faz-se necessário acrescentar que o uso BDTD tem sido extenso no Brasil e em países de língua portuguesa, pois de acordo com medições obtidas desde de outubro de 2016 por meio do *Google Analytics*, ocorrem em média 7 mil consultas diárias ao sistema de busca, sendo que 90% deste montante é realizado por indivíduos que estão em território nacional. Segunda, terceira e quarta posição são ocupadas respectivamente por Portugal, com 3,2%, Moçambique, com 2%, e Angola com 0,78% dos acessos.

Urge a necessidade de integração destas três grandes bases, e já estão disponíveis as ferramentas estatísticas necessárias, como a linguagem de programação R, e frameworks de manipulação de dados, em geral disponibilizados para a linguagem Python e R. Há ainda a mão protocolos de interoperabilidade como o *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH), e comunicação via *Application Programming Interfaces* (APIs) do tipo *Representational State Transfer* (REST) ou *Simple Object Access Protocol* (SOAP).

Metodologia

Para alcance do objetivo proposto, foi utilizada a linguagem de programação *R* para a coleta, tratamento e integração entre as bases[3,4]. O ambiente de desenvolvimento integrado (IDE) para o uso do R foi o *RStudio*. Esta IDE permite a execução de diferentes processos, como coleta de dados, pré-processamento, transformação, integração, aplicação de algoritmos de similaridade de *strings*, efetuação de cálculos estatísticos, e criação de diferentes gráficos, entres outras funcionalidades[5].

Os dados da BDTD foram baixados diretamente via URL (do *Apache-Sorl*) utilizando-se uma das funções do R. Estes registros foram baixados no formato CSV, no início de fevereiro de 2017, com um total de aproximadamente 470 mil registros. Já o do Banco da CAPES foi obtido via solicitação direta à agência, que forneceu no formato *Microsoft Access*, os registros com a abrangência de 1987 à 2016, totalizando mais de 900 mil entradas.

As coletas dos currículos Lattes ocorreram por meio de scripts concebidos na linguagem em *Python2*, e foram baixados via interface REST, da Plataforma Lattes. Foram coletadas informações pertinentes às teses e às dissertações, que ocorriam na seção de formação acadêmica, ou de orientações concluídas. Ao todo foram identificados quase 2 milhões e 200 mil registros de teses e dissertações.

Para a identificação de variantes do mesmo termo, e seguindo a mesma abordagem de Digiampietri *et al.*[6] e Mena-Chalco e Cesar Junior[7], utilizamos no presente estudo algoritmos de Levenshtein[8] e Jaro[9] de medição de distância entre *strings*. No ambiente do R, a biblioteca *Stringdist* foi utilizada para execução de funções já implementadas destes dois algoritmos.

Na etapa de limpeza e normalização dos dados foram aplicadas as seguintes regras: 1) os registros foram colocados todos em letras maiúsculas; 2) foram retirados espaços duplicados; 3) eliminaram-se caracteres especiais, caracteres numéricos e acentos, exceto vírgulas (,) e ponto-e-vírgulas (;) (estes por vezes estabeleciam separadores, e ordens de ocorrência no campo autor); 4) os nomes dos autores foram colocados em sua ordem direta. Por exemplo “Silva, João da” foi transformado para “João da Silva”.

No caso dos dados provindos a Plataforma Lattes identificou-se a necessidade da criação de um vocabulário controlado para o campo “Nome da Instituição”. A grande variação no preenchimento deste campo, mais de 6 mil tipos diferentes de entradas, tendo-se em mente o contexto brasileiro exibido na última avaliação da CAPES de 2016, o número instituições brasileiras que possuem programas de pós-graduação *strictu sensu* não deveria superar em muito o total de 500 ocorrências.

A performance do algoritmo de comparação seria drasticamente reduzida se não fosse possível agrupar as entradas de acordo com o campo da instituição de defesa. Para resolver este problema foi criada uma lista de instituições com identificadores, e sub-listas com as variações de preenchimento de cada instituição.

A lista inicial continha apenas 390 instituições, que foram aquelas filtradas das entradas dos Currículos Lattes que possuíam o preenchimento do identificador do Programa de Pós-Graduação da CAPES.

Foi criado então um arquivo no formato JSON, que foi enriquecido por meio da comparação do nome da instituição da lista e a variante encontrada em cada registro da lista do Lattes, utilizando-se o algoritmo de distância Jaro com valor de similaridade maior que 80%. Registros que não casavam nenhum nome principal ou variante foram identificados de instituição estrangeira, e portanto foram descartados da etapa seguinte de deduplicação interna e de cruzamento entre as bases.

Durante a deduplicação de cada base executou-se uma iteração com o algoritmo de Levenshtein com o valor de distância de similaridade menor que 3, e outra iteração com algoritmo Jaro com valor de distância de similaridade maior que 75%. Foram identificados aproximadamente 6.500 registros duplicados dentro da BDTD, 2.700 no Banco de Teses e Dissertações da CAPES e quase 400 mil registros na Plataforma Lattes. Neste último o mesmo registro pode aparecer na sessão de Formação e também em Orientações Concluídas.

A etapa seguinte foi a integração entre as bases da BDTD e Banco de Teses da CAPES, na qual se executou duas iterações, uma utilizando algoritmo de Levenshtein com valor de similaridade menor que 3, o que tardou aproximadamente 57h, e resultou em uma base de 1 milhão de registros. A execução com o algoritmo Jaro com similaridade mair que 85% durou 60h e resultou em uma base de mesmo volume. Esses algoritmos foram executados em uma máquina com o processador Intel i5, com 8GB de memória RAM.

A base BDTD / Banco CAPES foi então integrada à base da Plataforma Lattes utilizando-se algoritmo de Jaro com valor de similaridade maior que 75%. Esta execução durou 40h, e resultou em 1 milhão e 745 mil registros, após nova deduplicação. Esta execução foi realizada em uma máquina com 4 núcleos Intel Xeon e 12GB de memória RAM.

Resultados

Em resumo, a união dos registros de Teses e Dissertações da BDTD, do Banco da CAPES e da Plataforma Lattes, com aplicação de algoritmos de similaridade de *strings* para eliminação de réplicas, resultou em massa consolidada, em números exatos, de 1.745.138 registros. Entre estes, 412.487 são discriminados como do tipo Tese e 1.332.651 do tipo Dissertação. Identificou-se um total 498 instituições brasileiras com datas de defesa que vão de 1950 à 2017.

Verificou-se que o volume de registros comuns à BDTD e ao Banco da CAPES é de mais de 240 mil, o que representa aproximadamente 50% da BDTD e pouco mais de 25% do Banco da CAPES. Aproximadamente 100% dos registros de BDTD e CAPES ocorrem também na Plataforma Lattes, seja por declaração do próprio autor ou do orientador do trabalho.

Os principais desafios enfrentados ocorreram nas etapas de limpeza e normalização dos registros, e na escolha dos parâmetros a serem adotados na aplicação dos algoritmos de similaridade de *strings*. A variação de erros de preenchimento é extremamente grande em todas bases, e os valores parâmetros escolhidos para os algoritmos foram motivados pelos resultados obtidos em testes com amostras aleatórias, de no máximo alguns milhares de registros.

Conclusão e trabalhos futuros

Por se tratar de bases de grande volume, com dados não-estruturados, a infraestrutura computacional disponível influenciou o tempo de execução dos algoritmos. Nota-se que na última etapa, ao se utilizar uma máquina ligeiramente mais robusta, houve uma queda de mais de 30% do tempo consumido. Soluções

para esta questão podem estar não só relacionadas ao acréscimo do poder da infraestrutura, mas também ligadas à adoção de técnicas mais eficientes, como processamento paralelo distribuído.

Estabeleceu-se portanto um ponto de partida para a efetiva integração entres estas importantes bases nacionais, e como ação futura, pretende-se criar uma base de consulta centralizada que possa servir de validação dos dados declarados na Plataforma Lattes, que poderá também ser utilizada tanto pela comunidade científica de ciência da informação, ciência de dados e áreas correlatas, como pelo público não especializado.

Referências

1. BRASIL. IBICT. Biblioteca Digital Brasileira de Teses e Dissertações (BDTD). Disponível em: <http://bdtd.ibict.br>. Acessado em: 2017.
2. BRASIL. CAPES. Banco de Teses e Dissertações da CAPES. Disponível em: <http://bancodeteses.capes.gov.br/banco-teses/>. Acessado em: 2017.
3. IHAKA, Ross, R: Past and future history Computing Science and Statistics (1998): 392-396.
4. POULSON, Barton, R Succinctly. Syscfunsion, Technology Resource Portal. Acessado em: 2017.
5. GOMES, Gabriel Lima. Identificação de padrões de fraudes na obtenção da CNH, utilizando mineração de dados. Trabalho de Conclusão de Curso – IESB-DF. 2016.
6. DIGIAMPIETRI, Luciano A. et al. Minerando e caracterizando dados de currículos lattes. Em Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), 2012.
7. MENA-CHALCO, Jesús P.; Cesar Junior, Roberto M. Prospecção de dados acadêmicos de currículos Lattes através de Scriptlattes. Em Bibliometria e Cientometria: reflexões teóricas e interfaces. São Carlos (2013).
8. LEVENSHTEIN, Vladimir I. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, Vol. 10, N.8. 1966.
9. JARO, Matthew A. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association, Vol. 84, N. 406. Pag. 414-420. 1989.