Fall 8-7-2023

Dados de Alvarás - Uma Abordagem de Integração para Busca Textual

Bruno Guillen

Gabriel V. de Santana

Nádia P. Kozievitch

Dados de Alvarás - Uma Abordagem de Integração para Busca Textual

Artigo em Desenvolvimento

Bruno Guillen
Federal University of Technology - Paraná
brunoguillen@alunos.utfpr.edu.br

Gabriel V. de Santana
Federal University of Technology - Paraná
gabrielsantana@alunos.utfpr.edu.br

Nádia P. Kozievitch
Federal University of Technology - Paraná
nadiap@utfpr.edu.br

Abstract

Search and integration of business license data is a key process for many entities, whether to consult a reliable source of data on potential business partners or for studies related to urban development. Despite being available as open data, currently there is no established data standard and there is no free platform to consult them. In this direction, this work in progress uses open data from business licenses provided by the Municipality of Curitiba and data from Federal Revenue of Brazil to carry out the search for Legal Entity data. Through the new approach, using entity matching with Sørensen-Dice and Jaccard Similarity algorithms, the objective is to improve the textual search and present the result through a web interface.

Keywords

Text Comparison Algorithms, Open Data, Sørensen-Dice Algorithm, Jaccard Similarity.

Resumo

A busca e integração de dados de alvarás é um processo primordial para diversas entidades, seja para consultar uma fonte confiável de dados sobre potenciais parceiros de negócio ou para estudos relacionados a desenvolvimento urbano. Apesar de estarem disponíveis como dados abertos, atualmente não há um padrão estabelecido dos dados e não há uma plataforma gratuita para consultá-los. Nesta direção, este trabalho em andamento utiliza dados abertos de alvarás fornecidos pelo Município de Curitiba e dados da Receita Federal do Brasil. para realizar a busca de dados de Pessoa Jurídica. Através de uma nova abordagem, utilizando *entity matching* com os algoritmos de Sørensen-Dice e Similaridade de Jaccard, o objetivo é aprimorar a busca textual e apresentar o resultado através de uma interface web.

Palavras-chave

Algoritmos de comparação textual, Dados Abertos, Algoritmo Sørensen-Dice, Similaridade de Jaccard.

Introdução

Segundo Lemos (Lemos 2013), nos anos 90, o debate entre as novas tecnologias de informação e comunicação (TIC) e o espaço urbano estava prestes a cunhar o termo “cidades digitais”. O objetivo era implantar nesse espaço urbano uma infraestrutura digital eficiente, com intuito de fomentar processos inovadores nas estruturas de governo, nas empresas e no comércio. Nesta direção, há varios desafios que podem ser mencionados, como acesso (API, arquivos, design), computacional (escalabilidade, integração,

interoperabilidade), dado (qualidade, duplicação, semântica), temporalidade, formato, legislação, licenciamento e segurança.

Curitiba, capital do estado do Paraná, faz parte do grupo de cidades nomeado C40[1], que tem como objetivo a melhora da qualidade de vida e proteção ambiental. Curitiba, que é uma cidade planejada, é o centro econômico do estado do Paraná e apresenta um dos maiores PIB[2] do país. A partir da década de 1980, o poder público vem realizando investimentos no planejamento e execução de modernizações nos sistemas de infraestrutura da cidade, com especial atenção às questões de mobilidade e dados abertos (como os dados de alvarás). Em paralelo, ONGs (como a Associação de Amigas da Mama[3]) arrecadam dinheiro através de doações de notas fiscais de empresas disponibilizadas pela Nota Paraná[4], e carecem de sistemas que integrem este tipo de dado.

Nesta direção, este trabalho em andamento utiliza dados abertos de alvarás fornecidos pela Receita Federal do Brasil e do Município de Curitiba para auxiliar o processo de cadastro para auxílio a entidades sociais do Nota Paraná como a ONG Amigas da Mama. Ele aprimora o sistema apresentado em (Junior 2021), desenvolvendo uma aplicação web utilizando uma nova abordagem, através de *entity matching* com os algoritmos de Sørensen-Dice e Similaridade de Jaccard.

Trabalhos Relacionados

Conceitos recentes como Cidades Inteligentes, Computação Urbana e Sistemas de Informação Geográfica estão sendo discutidos em diversos fóruns internacionais, utilizando temas como sustentabilidade e uso eficiente das infraestruturas da cidade. Alguns exemplos de fontes de dados relacionados incluem (i) dados estatísticos oficiais; (ii) sensor sem fio redes; (iii) a infraestrutura da cidade; e (iv) as informações geográficas voluntárias compartilhadas na Internet. Nesta direção, a técnica de *Entity Matching* (também conhecido como identificação de duplicatas ou resolução de entidades) é uma tarefa crucial para integração e limpeza de dados (Cohen et al. 2000). Definido como a tarefa de identificar entidades (objetos, instâncias de dados) que se referem à mesma entidade no mundo real, as entidades a serem resolvidas podem residir em fontes de dados distribuídas, tipicamente heterogêneas, ou em uma única fonte de dados, por exemplo, em um banco de dados ou no armazenamento de um mecanismo de busca. Junior (Junior 2021), utilizou o conceito de *entity matching* para a integração de dados. Ele cita que um mesmo endereço pode aparecer de diversas maneiras nos dados da Receita Federal do Brasil, dificultando sua integração (como Alameda Doutor Muricy, Rua Doutor Muricy, Rua Dr Murici, Alameda Dr Murici). No mesmo artigo, dados são utilizados em uma interface web, com algoritmos fonéticos e de Levenshtein. Nesta direção, como os algoritmos fonéticos foram desenvolvidos para identificar palavras que soam de forma similar, não foram tão assertivos para lidar com falta de padronização na classificação dos endereços (ex. Rua ←→ Alameda) e com abreviações (ex. Marechal ←→ Mal.), enquanto o algoritmo de Levenshtein, foi capaz de assimilar a correspondência entre registros com poucos erros de digitação e algumas abreviações. Uma outra abordagem nesta direção é distância de Hamming (Chan et al. 2020), que mede a distância dos elementos na própria *string*, sendo funcional apenas com *strings* do mesmo tamanho. O algoritmo, entretanto, torna o seu uso limitado, apesar de possuir sua eficiência para esses casos. Uma outra abordagem a ser citada é a distância de Damerau-Levenshtein (Zhao and Sahni 2017): uma comparação de caractere (uma extensão do algoritimo de levenshtein) que além das transações normais considera transposições entre os caracteres adjacentes. Já a similaridade de Jaccard (Jalal et al. 2022) calcula a diferença dos diversos subsets existentes nas strings que serão utilizadas sendo vários n-gramas de informações a serem calculados. Podemos ainda citar o Coeficiente de Sørensen-Dice (Li et al. 2020) (métrica de proporção através dos n-gramas que são presentes nas strings) que serão comparadas ou a combinação Sørensen-Dice Jaccard (compara termos com tamanhos diferentes e mitiga problemas com *typo* em sua *string)*. Nesse sentido, devido ao limite de tempo, este trabalho analisa somente dois algoritmos: Sørensen-Dice e Similaridade de Jaccard para aprimorar o *matching* e o tratamento textual existente em (Junior 2021). Esses algoritmos são conhecidos

[1] https://www.c40.org/
[2] https://cidades.ibge.gov.br/brasil/pr/curitiba/pesquisa/38/47001?tipo=ranking
[3] https://www.amigasdamama.org.br/
[4] http://www.notaparana.pr.gov.br

por sua eficácia em comparar e relacionar textos semelhantes, o que pode ser útil na sincronização de dados em cidades inteligentes.

O protótipo

O protótipo tem como objetivo analisar a utilização dos algoritmos Sørensen-Dice e Similaridade de Jaccard para aprimorar o *matching* e o tratamento textual existente no Buskaki (Junior 2021). A Figura 1 apresenta a arquitetura da ferramenta. Utilizando um dispositivo móvel ou computador, o usuário realiza uma interação com a página web, que se conecta ao servidor e retorna os dados solicitados, vindos do banco de dados. O protótipo utiliza as seguintes tecnologias: PostgreSQL[5],Angular[6], Python, Pandas, SQLalchemy[7], psycopg2, Google API (para a visualização espacial dos dados) e Selenium (para o crawler que atualiza a base de dados e corrige os erros de endereço).

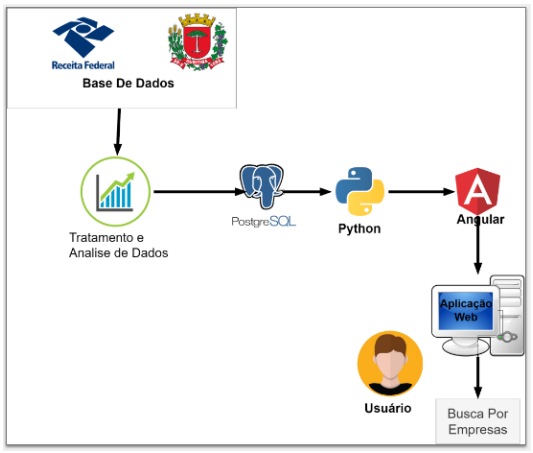

Figura 1. Arquitetura do Protótipo

Os dados utilizados foram obtidos através do site da Receita Federal[8], e alvarás[9] da cidade de Curitiba (detalhes podem ser obtidos em (Junior 2021) e (Bichibichi et al. 2018)). A tabela empresas_receita_curitiba possui 585.624 registros de empresas curitibana, até o ano de 2019. Os dados incluiam atributos como nome empresarial, nome fantasia, logradouro, número, bairro, cep, data inicio de atividade, e cnpj. A tabela alvara possui 515.876 registros até o ano de 2019, e através do logradouro, os dados são georreferenciados (note que o cnpj não está presente).

Inicialmente é realizado um cadastro e uma conta de acesso para o usuário aceder ao sistema. Consideraremos como usuário um membro de uma ONG, que busca dados de empresas através de uma nota fiscal. Ao informar os parâmetros da busca e submeter o formulário, o servidor realiza as operações de busca de empresa. Com intuito de aumentar a abrangência da busca, possibilitando contornar possíveis erros de digitação ou abreviações, são utilizadas estratégias de comparação textual aproximada. Para isso, será comparada a performance de dois algoritmos com dados da tabela de empresas: Algoritmo de Sørensen-Dice (Li et al. 2020) e Similaridade de Jaccard (Jalal et al. 2022). O Algoritmo de Sørensen-Dice (Figura 2) foi desenvolvido em PL/pgSQL, e faz a comparação textual de maneira estatística que mede a similaridade de duas Strings através da comparação de seus bigramas. Já o Algorítmo de Similaridade de Jaccard (Figura 3) foi desenvolvido em PL/pgSQL, é uma medida estatística

[5] https://www.postgresql.org
[6] https://angular.io
[7] https://www.sqlalchemy.org
[8] https://receita.economia.gov.br/orientacao/tributaria/cadastros/cadastro-nacional-de-pessoas-juridicas-cnpj/dados-publicos-cnpj
[9] https://www.curitiba.pr.gov.br/dadosabertos/busca/

utilizada com intuito de mensurar a similaridade e a diversidade de conjuntos. Neste trabalho é utilizado para comparar as strings sejam endereços ou os possíveis nomes das empresas, e possui como como entrada duas string que deseja se comparar e e como retorno o índice de similaridade de Jaccard. As próximas etapas (além da comparação dos algoritmos), incluem as novas funcionalidades de buscas (como a busca geral, cnpj, endereço, bairro ou nome, como ilustrado na Figura 4), a visualização e um tratamento de dados como utilização de um *crawler* para que se corrija endereços incorretos nos base de dados. Para a avaliação preliminar do protótipo, será submetido aos usuários de teste um questionário[10] com uma pergunta sobre a busca geral, busca avançada e por cnpj.

ALGORITMO 1: Algoritmo de Sørensen-Dice

Input: $S1$ e $S2$, dois conjuntos de elementos
Output: *coef SorensenDice*: Coeficiente de Sørensen-Dice entre $S1$ e S
1: *intersecao* ← $|s1 \cap s2|$
2: *uniao* ← $|s1| + |s2| -$ *intersecao*
3: *coef SorensenDice* ← $(2 \cdot intersecao)/uniao$
4: retorna *coef SorensenDice*

Figura 2. Algoritmo de Sørensen-Dice.

ALGORITMO 2: Algoritmo de Similaridade de Jaccard

Input: $S1$ e $S2$, dois conjuntos de elementos
Output: *simJ* , a similaridade de Jaccard entre $S1$ e $S2$
1: *contIntersecao* ← 0
2: *contU niao* ← *tamanho*($S1$) + *tamanho*($S2$)
3: para *elemento* em $S1$ faça
4: se *elemento* está em $S2$ então
5: *contIntersecao* ← *contIntersecao* + 1
6: *contU niao* ← *contU niao* − 1
7: finaliza se
8: finaliza para
9: *simJ* ← *contIntersecao*/*contU niao*
10: retorna *simJ*

Figura 3. Algoritmo de Similaridade de Jaccard.

Dentre os desafios a serem enfrentados pelo protótipo podemos listar: 1) a capacidade de retornar resultados aproximados dos dados inseridos, a ponto de contornar possíveis erros de digitação ou padronização (para isso são necessários vários testes de mesa para a definição de parâmetros); 2) a diminuição da quota dos serviços de geocodificação, já que a maioria dos serviços são pagos ou tem limitação de uso; 3) a falta de padronização quanto ao uso de abreviações nos dados de endereços

[10] https://docs.google.com/forms/d/e/1FAIpQLSdDEdhqLefr7TtrsRYVbTjMRtdKKZr_9X0ZHQHXm22aUm4Bqw/viewform

fornecidos pela Receita Federal e pela prefeitura de Curitiba, dificultando o *matching* de registros iguais presentes em ambas as tabelas (como Alameda Dom Pedro II, Rua Dom Pedro II, Alameda D Pedro II, Rua D Pedro II).

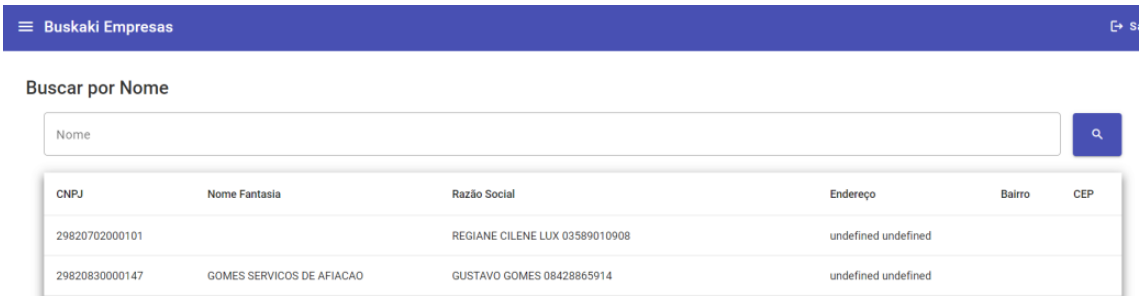

Figura 4. Protótipo da tela de busca por nome.

Conclusão

Este trabalho em andamento utiliza dados abertos de alvarás fornecidos pela Receita Federal do Brasil e do Município de Curitiba para auxiliar o processo de cadastro para auxílio a entidades sociais do Nota Paraná como a ONG Amigas da Mama. Ele aprimora o sistema apresentado em (Junior 2021), desenvolvendo uma aplicação web utilizando uma nova abordagem, através da abordagem de *Entity Matching,* com os algoritmos de Sørensen-Dice e Similaridade de Jaccard. Apesar de existirem aplicações com propostas semelhantes, em geral, estas estão sob responsabilidade de empresas privadas. A aplicação web terá buscas diferenciadas (bairro, nome, cpf, avançada), visualização e login de usuário. Como trabalhos futuros, podemos citar testes com usuários, e testes com mais tipos diferenciados de dados.

Agradecimentos

Os autores gostariam de agradecer o IPPUC, URBS, Prefeitura de Curitiba e o projeto *Smart City Concepts in Curitiba* (2019-04893) patrocinado pela VINNOVA (*Sweden's innovation agency*).

Referências

Bichibichi, Y et al. 2018, “Análise de evolução de emissão de alvarás próximos a dois shoppings em Curitiba,” in ´ Escola Regional de Banco de Dados (ERBD), 14., 2018, Rio Grande. Anais. Porto Alegre: Sociedade Brasileira de Computação, 2018 . ISSN 2595-413X.

Chan, T. M., Golan S., Kociumaka T., Kopelowitz T., Porat E., 2020, “Approximating text-to-pattern hamming distances. In: Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing. New York, NY, USA: Association for Computing Machinery, 2020. (STOC 2020), p. 643–656. ISBN 9781450369794.

Cohen W. W., Kautz H., McAllester D.,2000, “Hardening soft information sources,” in Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 255–259.

Jalal, A.; Jasim, A.; Mahawish, A. A. A, 2022, “A web content mining application for detecting relevant pages using jaccard similarity”. IJECE, v. 12, p. 6461–6471, 12.

Lemos, A. 2013. “Cidades inteligentes,” GV EXECUTIVO, vol. 12, no. 2, pp. 46–49.

Li, X., Wang, C., Zhang, X., Sun, W., 2020, “Generic sao similarity measure via extended sørensen-dice index”. IEEE Access, v. 8, p. 66538–66552, 2020

Junior, E. S. B. 2021. “Buskaki empresas - ferramenta para busca de dados abertos de empresas curitibanas”, Monografia (Engenharia da Computação), UTFPR, 2021.

Zhao C., Sahni,S., 2017, "Efficient computation of the Damerau-Levenshtein distance between biological sequences," 2017 IEEE 7th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), Orlando, FL, USA, pp. 1-1, doi: 10.1109/ICCABS.2017.8114295.