Buskaki Empresas - Ferramenta para busca de dados abertos de empresas curitibanas

Everton S. B. Junior[1], Wilian Cavassin[1], Nádia P. Kozievitch[1], Matheus Biscaya Gutierrez[1]

[1]Universidade Tecnológica Federal do Paraná (UTFPR)
Avenida Sete de Setembro, 3165
Departamento Acadêmico de Informática – DAINF – Curitiba – Brasil

{evertonjunior.1997,wiliancavassin,matheusgutierrez}@alunos.utfpr.edu.br,

nadiap@utfpr.edu.br

Abstract. *The search of Legal Person data is a relevant process to many entities. There are numerous reasons to the search, like consult trusted data sources for potential business partners, studies related to urban development, comply with legal obligations and in many cases just collect data about specific enterprise. In this context, this article presents a tool to search for information from Curitiba companies that provides another option than CNPJ, based on Geographic Information Systems (GIS) and smart cities. The project uses Open Data and applies textual comparison algorithms to increase the scope of the search results.*

Resumo. *A busca de dados de Pessoa Jurídica é um processo relevante para diversas entidades, seja para consultar uma fonte confiável de dados sobre potenciais parceiros de negócios, para estudos relacionados a desenvolvimento urbano ou para simplesmente coletar informações sobre determinada empresa. Neste contexto, este artigo apresenta uma ferramenta para busca de informações de empresas curitibanas que disponibilize outras formas de consulta além do CNPJ, baseada em Sistemas de Informação Geográfica (SIG) e cidades inteligentes. A ferramenta utiliza dados abertos e aplica algoritmos de comparação textual para aumentar a abrangência dos resultados da busca.*

1. Introdução

O Nota Paraná[1] é um programa do governo do Paraná de combate à sonegação fiscal, que consiste em uma estratégia de devolução de parte do imposto pago por empresas na venda de produtos ou serviços. O consumidor pode cadastrar seu CPF na nota fiscal para receber esse retorno financeiro ou doar a nota para entidades sociais depositando-a em urnas localizadas nos estabelecimentos. Posteriormente elas são recolhidas pelas entidades, que recebem a contribuição monetária para cada nota que cadastram no sistema Nota Paraná.

A Associação de Amigas da Mama [2] (AAMA) (uma ONG curitibana destinada a apoiar e ajudar mulheres diagnosticadas com câncer de mama) é um exemplo de ONG que

[1]http://www.notaparana.pr.gov.br/
[2]https://www.amigasdamama.org.br/

pode ser beneficiada pelo programa. E para realizar esta etapa, a ONG utiliza planilhas (Figura 1) preenchidas manualmente para realizar a busca de CNPJs para a entrada no sistema Nota Paraná.

	ANA LEIA (ABIGAIL) ASSISTIDA = 6 Estabelecimentos				
1	DO VALE FARMACIA	R. Sargento Luiz G. M. Ribas	630		Uberaba
2	SUPERMERCADOS GÓES	R. Sargento Luiz G. M. Ribas	646		
3	MERCADO RIO CRISTALINO	R. Amaury Mauad Guerios	134		Uberaba
4	MH Supermercado	R. Maurício Thá	165		
5	MINIPREÇO - JUMBO COM. DE UTILIDADES	Uberaba - 3013-1101			
6	Auto Posto Marfim Ltda	Av. Comendador Franco	7434		Uberaba
	ANI - Filha da ABIGAIL = 2 Estabelecimentos				
1	HIPERFARMA	R. Pernambuco	316		
2	SUPERMERCADOS GOES	R. Pernambuco	316		
	CASSIA - AMIGA DA ABIGAIL = 13 Estabelecimentos				
1	ATENAS MATERIAIS DE CONSTRUÇÃO LTDA.	Av. Prefeito Erasto Gaertner	1611		Bacacheri
2	Auto Posto Base Aerea	Av. Prefeito Erasto Gaertner	1600		Bacacheri
3	Restaurante e Eventos Curitiba Gourmet	Av. Prefeito Erasto Gaertner	1573		Bacacheri

Figura 1. Planilha AAMA, 1º semana de outubro 2019

Um dos motivos que dificultam a pesquisa por tais informações é que toda empresa possui dois nomes: Razão Social e Nome Fantasia. A Razão social é um nome único e exclusivo da pessoa jurídica, oficializado na Junta Comercial [3], geralmente relacionada ao tipo da empresa. O Nome Fantasia é o nome popular de uma empresa, que pode, ou não, ser igual à sua razão social, sua principal função é relacionada à divulgação dela, visando o maior aproveitamento da sua marca e das estratégia de marketing e vendas. Assim, determinada empresa pode se apresentar de formas diferentes, de acordo com o alvo da comunicação, como pode ser visto na Figura 2.

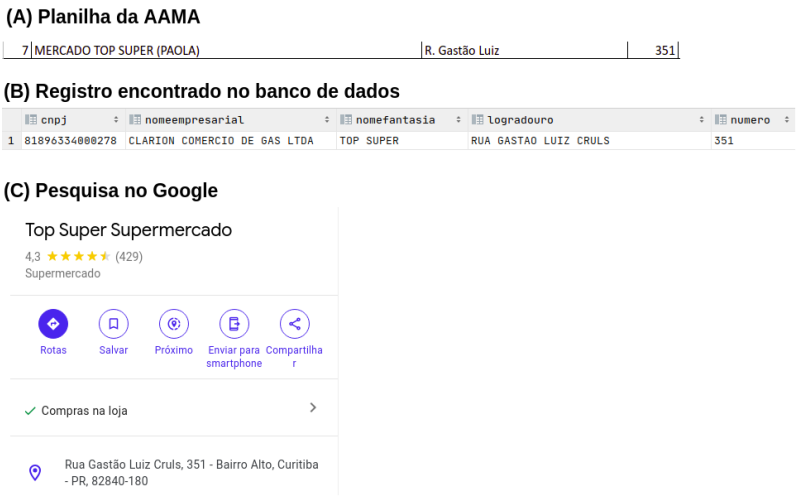

Figura 2. Comparação de apresentações da empresa Top Super

Este artigo apresenta uma ferramenta para busca de informações de empresas curitibanas, disponibilizando vários tipos de busca, além da visualização do resultado, login e acesso às buscas históricas. A ferramenta baseia-se em sistema de informação geográfica (SIG) e cidades inteligentes. O projeto utiliza dados abertos fornecidos pela Receita Federal[4] e aplica o Algoritmo de Levenshtein para aumentar a abrangência dos resultados da busca.

[3] http://www.juntacomercial.pr.gov.br/
[4] https://receita.economia.gov.br/

2. Trabalhos Relacionados

Existem diversas ferramentas para consulta de dados de pessoas jurídicas, desde as mais simples e gratuitas às ferramentas pagas, que provêm uma consulta com uma maior quantidade de informações empresariais, inclusive seus sócios.

As ferramentas gratuitas podem ser separadas em dois grupos: sites e aplicativos. Dentre os sites, destacam-se o Portal REDESIM [5] e o site da Junta Comercial do Estado de São Paulo [6] (JUCESP), dentre os aplicativos pode-se citar: o CNPJ[7], disponibilizado pela Receita Federal, e o Consulta CNPJ [8]. As ferramentas pagas para realização desta consulta, como a API do Serpro e a consulta do Serasa Experian, cobram por cada consulta realizada.

Dentre as ferramentas gratuitas citadas anteriormente, apenas duas possibilitam que a consulta seja feita de outra forma além do CNPJ: o Portal REDESIM e o site da JUCESP permitem consultas pelo nome da empresa. Para a utilização da ferramenta do Portal é necessário um certo nível de conhecimento técnico: é necessário criar um cadastro, lidar com captchas, saber diferenciar Razão Social de Nome Fantasia e o único filtro de localização é por unidade da federação. A base de dados da JUCESP, apesar de possuir registros de empresas de todo país, conta com maior acervo das empresas do estado de São Paulo, não contemplando a maioria das empresas de outros estados. Nenhuma dessas ferramentas aplica um algoritmo robusto para realizar a pesquisa, sendo capazes apenas de retornar os registros com uma correspondência do texto inserido.

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, Hernández and Stolfo 1995, Rahm and Do 2000], definida 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.

O processo de *Entity matching* é particularmente desafiador para entidades que são muito heterogêneas e de qualidade de dados limitada em relação à integridade e consistência de seus atributos. A Tabela 1 ilustra alguns das diferentes maneiras de se referir a um mesmo endereço na base de CNPJs da Receita Federal. Podemos assumir que esses dados cadastrais são inseridos manualmente por funcionários e podem conter vários problemas de qualidade, tais como erros de digitação, abreviações e diferentes classificações de Tipo do Logradouro.

Para Garg e Singla [Singla and Garg 2012] O problema da comparação de strings consiste em comparar duas strings, uma é o texto T [1...n], a string principal fornecida, e a outra é o padrão P [1...m], para ser combinada com a principal, dado m $\leq$ n. A correspondência de strings é usada de forma variada em aplicações, como por exemplo num esquema de banco de dados ou num sistemas de rede. Existem duas técnicas principais de

[5] https://consultacnpj.redesim.gov.br/

[6] https://www.jucesponline.sp.gov.br/pesquisa.aspx?IDProduto=7

[7] https://play.google.com/store/apps/details?id=br.gov.fazenda.receita.pessoajuridica

[8] https://play.google.com/store/apps/details?id=com.consultacnpj

Tabela 1. Diferentes referencias para um mesmo endereço na base de CNPJ da Receita Federal

Tipo do Logradouro	Logradouro
ALAMEDA	DOUTOR MURICY
RUA	DR MURICI
RUA	DR MURICY
ALAMEDA	DR MURICY
ALAMEDA	DR. MURICY

correspondência de strings, uma é a correspondência exata e a outra é a correspondência aproximada.

Dada a natureza do problema relacionado a este projeto, foram implementados e analisados os algoritmos de comparação textual aproximados. Para efetuar os filtros de busca propostos, apesar da eficiência do operador LIKE, do operador de igualdade e demais operadores lógicos em consultas SQL, eles são limitados em bases de dados onde ocorreram erros de digitação ou em buscas fonéticas. Mesmo que os dados estejam consistentes e as informações tenham sido cadastradas corretamente, a falha humana pode ocorrer no momento em que o usuário digita a informação que deseja buscar [Ruberto and Antoniazzi 2017]. No momento em que um grande volume de dados é armazenado, a Lógica Fuzzy auxilia no reconhecimento de padrões para que estes dados se tornem informações úteis aos usuários [Chen et al. 2016].

Em teoria da informação, existem diversas métricas para avaliar semelhanças em strings, optou-se por duas métricas populares, ambas levam o nome de seus autores, para avaliar strings em semelhança de forma, a distância de Hamming[Hamming 1950] e a distância de Levenshtein[Levenshtein 1966]. A distância de Hamming conta o número mínimo de substituições necessárias para editar uma string s até transformá-la em uma string t[Russell and Norvig 2016]. Levenshtein considera inserções e deleções também, deste modo a distância de Levenshtein produzirá distâncias menores ou iguais à distância de Hamming. Salientamos aqui que abreviações como ”mal”e ”marechal”são beneficiadas por essa propriedade.

O termo *Soundex* cobre variações de um algoritmo desenvolvido e patenteado em 1918 por Russell e Odel em 1918[Russell and Odell 1918]. A essência desse algoritmos e suas derivações ao longo dos anos é a converção de uma palavra em um código, ao qual consiste na primeira letra da palavra, seguida por três (ou mais em algumas derivações) dígitos. Esses dígitos são atribuídos de acordo com um agrupamento pré-determinado de consoantes, onde os grupos consonantais compartilham características fonéticas. Este é o conceito-chave por trás do Soundex: uma constante relação entre letras e sons, que objetiva que palavras com sons semelhantes sejam atribuídas ao mesmo código. Sendo o ”Soundex”uma técnica com origem na língua inglesa, necessitaram-se adaptações, mudando a tabela de códigos, baseando-se na adaptação proposta por Ruberto e Antoniazzi [Ruberto and Antoniazzi 2017].

O algoritmo Metaphone foi criado em 1990 por Lawrence Philips [Philips 1990] como uma alternativa para resolver deficiências relacionadas ao Soundex. Mais tarde, 10 anos depois, o autor lançou outra versão, chamada Double Metaphone [Philips 2000]. O termo ”Double”é explicado pela capacidade do algoritmo poder retornar dois códigos

para uma string, um primário e um secundário, sendo considerado mais complexo que seu antecessor.

3. A Ferramenta

A ferramenta tem por objetivo realizar vários tipos de busca, além da visualização do resultado, login e acesso às buscas históricas.

A Figura 3 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.

Do lado do servidor da aplicação foi utilizado o web server Puma versão 4.3.7, com as seguintes tecnologias: Cascading Style Sheets (CSS) e HyperText Markup Language (HTML), JavaScript, Ruby[9] versão 2.2.3, Framework da linguagem de programação Ruby, Ruby on Rails[10] 5.0.0.1, PostgreSQL[11] versão 9.6.13 x64, PostGIS[12] versão 2.3.1, Ruby on Rails GMaps4Rails[13] , Ruby on Rails Geocoder[14], Ruby on Rails ActiveRecord PostGIS Adapter[15], Ruby on Rails Net SSH Gateway[16].

A base de dados utilizada pela aplicação fez uso de um servidor AMD EPYC 7401 24Core 48-threads x64 2.8GHz, e a aplicação foi hospedada em uma plataforma nuvem Heroku[17], em um container de categoria *free*, com 512MB de RAM dedicada.

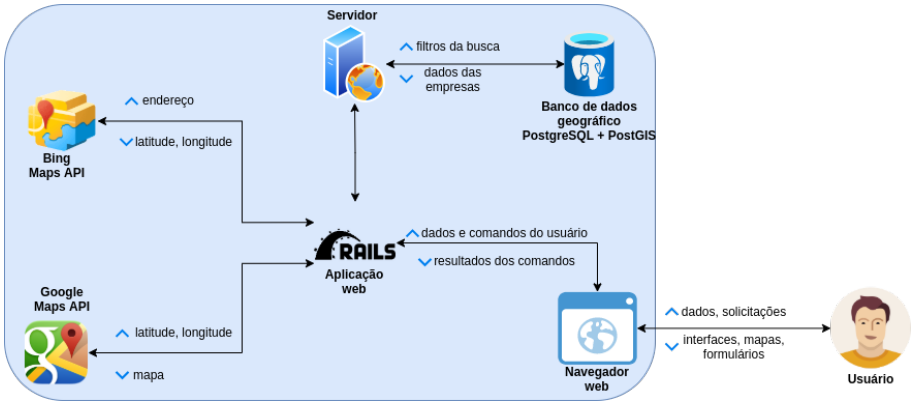

Figura 3. Arquitetura da ferramenta.

Os dados utilizados foram obtidos através do site da Receita Federal[18], e detalhes podem ser obtidos em [Bichibichi et al. 2018]. A Figura 4 apresenta o esquema de dados utilizados pela ferramenta.

[9] https://www.ruby-lang.org/pt/
[10] https://rubyonrails.org/
[11] https://www.postgresql.org/
[12] https://postgis.net/
[13] https://rubygems.org/gems/gmaps4rails/versions/2.1.2?locale=pt-BR
[14] https://rubygems.org/gems/geocoder
[15] https://rubygems.org/gems/activerecord-postgis-adapter
[16] https://rubygems.org/gems/net-ssh-gateway
[17] https://www.heroku.com/
[18] https://receita.economia.gov.br/orientacao/tributaria/cadastros/cadastro-nacional-de-pessoas-juridicas-cnpj/dados-publicos-cnpj

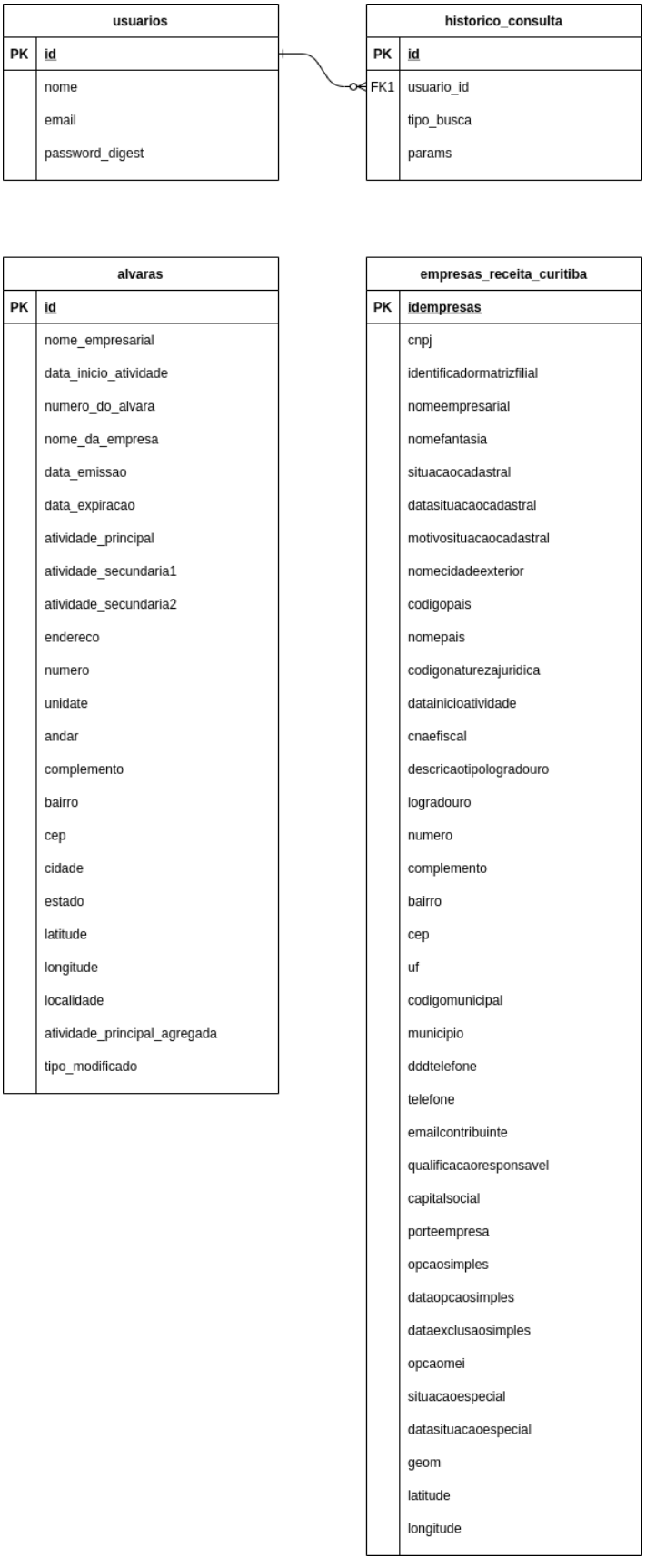

Figura 4. Esquema do banco de dados.

Inicialmente é necessário um cadastro e uma conta de acesso para o usuário aceder ao sistema. Ao informar os parâmetros da busca e submeter o formulário, o servidor realiza as operações de busca de empresa, descritas no Algoritmo 1.

Algoritmo 1 Busca por nome da empresa

Input: nome da empresa a ser pesquisada
Output: lista de empresas compatíveis

1: $salvar\_historico\_buscas(nome)$
2: $nome \leftarrow remover\_acentos(nome)$
3: $nome \leftarrow para\_letra\_maiuscula(nome)$
4: $lista\_empresas \leftarrow executa\_query\_busca(nome)$
5: **for** $empresa$ em $lista\_empresas$ **do**
6: **if** $empresa.latitude$ é nulo OU $empresa.longitude$ é nulo **then**
7: $endereco\_empresa \leftarrow formatar\_endereco\_para\_api\_endereços(empresa)$
8: $resultados\_geocoder \leftarrow consulta\_api\_geocodificacao(endereco\_empresa)$
9: $endereco\_georreferenciado \leftarrow encontrar\_por\_cep(resultados\_geocoder, empresa.cep)$
10: $empresa.latitude \leftarrow endereco\_georreferenciado.latitude$
11: $empresa.longitude \leftarrow endereco\_georreferenciado.longitude$
12: $atualizar\_empresa(empresa)$
return $lista\_empresas$

Com intuito de aumentar a abrangência da busca, possibilitando contornar possíveis erros de digitação ou abreviações, foi decidido utilizar estratégias de comparação textual aproximada. Para isso, comparamos a performance de três algoritmos com dados da tabela de empresas: Distância de Levenshtein[Levenshtein 1966], Soundex[Russell and Odell 1918] adaptado ao português brasileiro, e Metaphone-pt_BR[Jordão and Rosa 2012]. Detalhes podem ser verificados em [Junior 2020]. O Algoritmo de Levenshtein foi selecionado para o posterior uso. Vale salientar que a adição desse algoritmo não causou grande impacto de performance de consultas.

O processo da escolha geocodificação foi feita com base em análises de custo, acurácia, licenciamento e tecnologias compatíveis, optando-se pelo Geocoder[19]. A escolha do serviço ideal não é um processo simples. Muitos dos serviços gratuitos demonstraram uma precisão ruim durante os testes, o Nominatim[20] demonstrou erros de até 2Km em certos casos, na Figura 5 podemos observar a diferença do resultado da geocodificação de três serviços, Bing, Nominatim e GoogleMaps, para o mesmo endereço: Avenida Sete de Setembro 3561, Centro, Curitiba - PR. Por ser uma aplicação web sem restrição de uso, utilizar um serviço *pay-as-you-go*, como a GoogleMaps Geocoding API[21], que pode gerar altos custos inesperados à aplicação ao exceder a quota de uso gratuita.

Segundo Davis e Fonseca [Davis and Fonseca 2007], para que a geocodificação tenha maior assertividade antes de realizar a consulta em um serviço de geocodificação, o endereço precisa ser formatado afim de facilitar a correspondência das informações no

[19] `https://github.com/alexreisner/geocoder`
[20] `https://wiki.openstreetmap.org/wiki/Nominatim`
[21] `https://developers.google.com/maps/documentation/geocoding/overview`

banco de dados do serviço. Cada API retorna os resultados em um formato diferente, então para cada teste era necessário adequar o *parsing* da resposta da API.

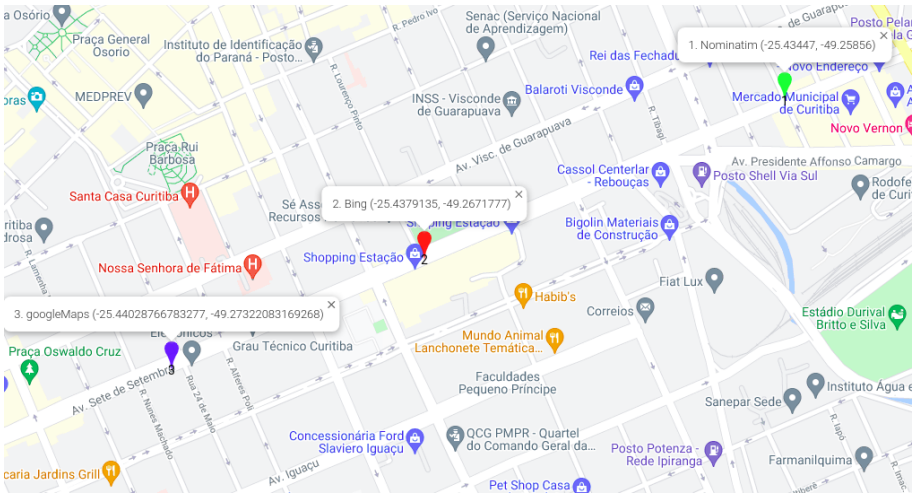

Figura 5. Comparação de serviços de geocoding.

Dois grandes desafios foram verificados no decorrer do projeto: (i) aumentar a abrangência da busca sem afetar muito sua assertividade (ou seja, era necessário que o algoritmo de busca fosse capaz de retornar resultados aproximados dos dados inseridos, a ponto de contornar possíveis erros de digitação). Para isso foram analisados 3 algoritmos de comparação textual e o algoritmo escolhido foi o da Distância de Levenshtein [Levenshtein 1966]; (ii) diminuir o uso de quota dos serviços de geocodificação através do aproveitamento de dados já existentes em uma outra entidade na base de dados.

A ferramenta permite tanto a busca por nome (Figura 6), quanto a busca por cnpj. Além disso, lista as empresas por bairro e por rua. Quando a empresa é encontrada, seus detalhes podem ser visualizados também (Figura 7).

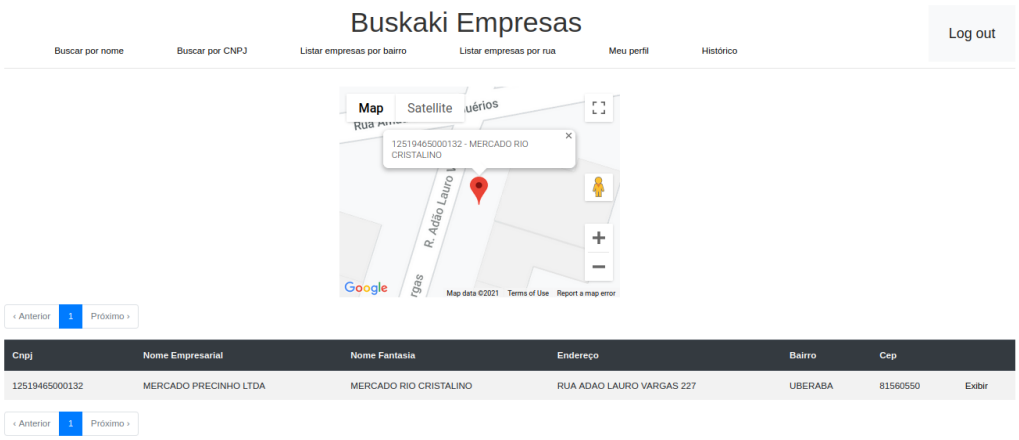

Figura 6. Interface do protótipo Buskaki.

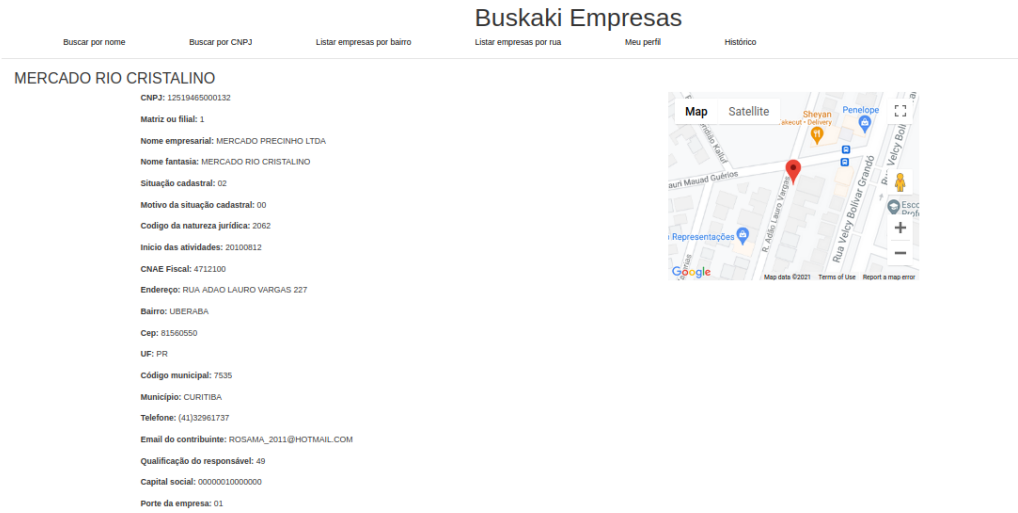

Figura 7. Interface do protótipo - Detalhes da empresa encontrada.

O aplicativo está disponível online[22] e um vídeo de explicação pode ser acessado através da plataforma youtube [23].

4. Conclusão

A ferramenta Buskaki Empresas utiliza conceitos conceitos de geoprocessamento e cidades inteligentes para consultar os dados abertos de empresas curitibanas. Ela aplica o Algoritmo de Levenshtein com dados abertos da Receita Federal e informações fornecidas pela prefeitura do município. Apesar de existirem aplicações com propostas semelhantes, em geral, estas estão sob responsabilidade de empresas privadas.

Para auxiliar a limitação do escopo e a escolha das tecnologias que seriam utilizadas, foi realizada uma revisão bibliográfica sobre os principais conceitos que sustentam o projeto, verificando o funcionamento de aplicações que os usam. A análise e modelagem resultaram no projeto de uma aplicação web que é uma ferramenta que permite a consulta e visualização das informações de empresas curitibanas, com o auxílio de um mapa dinâmico para facilitar o reconhecimento do registro desejado.

Como trabalhos futuros, podemos citar a implementação de uma nova funcionalidade responsável por atualizar automaticamente a base de dados através de um *crawler*, a atualização automática da fonte de dados (Receita Federal), testes unitários e junto ao usuário final, sugestões de empresas baseadas no histórico de busca do usuário, atualização de informações de empresas com o Google e análise de requisitos.

5. Agradecimentos

Os autores gostariam de agradecer a Universidade Tecnológica Federal do Paraná (DIREC01/2020), o Instituto Brasileiro de Geografia e Estatistica (IBGE), o Instituto de Pesquisa e Planejamento Urbano de Curitiba (IPPUC) e a Prefeitura de Curitiba.

[22] `https://buskaki-empresas.herokuapp.com`
[23] `https://youtu.be/mWmuQ49mnoI`

Referências

Bichibichi, Y. S., Kozievitch, N. P., and Carvalho, R. A. (2018). Análise de evolução de emissão de alvarás próximos a dois shoppings em curitiba. In *Anais da XIV Escola Regional de Banco de Dados*. SBC.

Chen, S.-M., Cheng, S.-H., and Lan, T.-C. (2016). A novel similarity measure between intuitionistic fuzzy sets based on the centroid points of transformed fuzzy numbers with applications to pattern recognition. *Information Sciences*, 343:15–40.

Cohen, W. W., Kautz, H., and McAllester, D. (2000). Hardening soft information sources. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 255–259.

Davis, C. A. and Fonseca, F. T. (2007). Assessing the certainty of locations produced by an address geocoding system. *Geoinformatica*, 11(1):103–129.

Hamming, R. W. (1950). Error detecting and error correction codes. *The Bell System Technical Journal*, XXIX(2):147–160.

Hernández, M. A. and Stolfo, S. J. (1995). The merge/purge problem for large databases. *ACM Sigmod Record*, 24(2):127–138.

Jordão, C. C. and Rosa, J. L. G. (2012). Metaphone-pt_br: The phonetic importance on search and correction of textual information. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 297–305. Springer.

Junior, E. S. B. (2020). Buskaki Empresas - Ferramenta para busca de dados abertos de empresas curitibanas. Monografia (Engenharia da Computação), UTFPR.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

Philips, L. (1990). Hanging on the metaphone. *Computer Language*, 7(12):39–43.

Philips, L. (2000). The double metaphone search algorithm. *C/C++ users journal*, 18(6):38–43.

Rahm, E. and Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13.

Ruberto, D. L. V. G. and Antoniazzi, R. L. (2017). Análise e comparação de algoritmos de similaridade e distância entre strings adaptados ao português brasileiro. In *Anais da XIII Escola Regional de Banco de Dados*. SBC.

Russell, R. and Odell, M. (1918). Soundex patent 01 261 167.

Russell, S. and Norvig, P. (2016). *Artificial Intelligence: A Modern Approach*. Pearson.

Singla, N. and Garg, D. (2012). String matching algorithms and their applicability in various applications. *International journal of soft computing and engineering*, 1(6):218–222.