UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA

Ad Nunes Ribeiro

LIGAÇÃO DE DADOS ESPAÇO-TEMPORAIS E TEXTUAIS DE MÍDIAS SOCIAIS COM LOCAIS DE INTERESSE VISITADOS

Florianópolis

2017

Ad Nunes Ribeiro

LIGAÇÃO DE DADOS ESPAÇO-TEMPORAIS E TEXTUAIS DE MÍDIAS SOCIAIS COM LOCAIS DE INTERESSE VISITADOS

Tese submetida ao Programa de Graduação em Ciência da Computação para a obtenção do Grau de Bacharael em Ciência da Computação.
Orientador: Prof. Dr. Renato Fileto
Coorientador: Me. Italo Lopes Oliveira

Florianópolis

2017

Catalogação na fonte elaborada pela biblioteca da
Universidade Federal de Santa Catarina

A ficha catalográfica é confeccionada pela Biblioteca Central.

Tamanho: 7cm x 12 cm

Fonte: Times New Roman 9,5

Maiores informações em:
http://www.bu.ufsc.br/design/Catalogacao.html

Todo pensamento é uma causa, e toda condição é um efeito

John Joseph Murphy

RESUMO

Dados de mídias sociais são gerados constantemente e representam uma relevante fonte de informação atualizada nos dias atuais. Todavia, mídias sociais incluem dados semi-estruturados (e.g. coordenadas geográficas opcionais, indicação de local) e não estruturados (e.g. textos de postagens) que precisam ter sua semântica correta explicitada. Uma das formas de fazer isso é através de anotações semânticas que ligam porções relevantes de dados (e.g. um tweet inteiro, menções a entidades nomeadas) a descrições com semântica bem definida em base de dados (e.g. locais de interesse do OpenStreetMap) ou de conhecimento (e.g. DBpedia, LinkedGeoData). Este trabalho se propõe a estudar, selecionar, desenvolver e testar algoritmos para ligar dados de mídias sociais que possuam coordenadas geográficas e texto (e.g. tweets georreferenciados) a lugares de interesse visitados, descritos em bases de dados ou de conhecimento com semântica bem definida. Esta tarefa pode não ser tão simples em certas situações, porque coordenadas costumam ter precisão limitada e pode haver uma grande quantidade de locais de interesse próximos à coordenada da postagem e até mesmo locais de interesse com coordenadas praticamente sobrepostas (e.g. em um edifício com vários andares). Desta forma, foi implementado um algoritmo para ligação de dados sobre movimento que considera dois raios distintos: um menor em que basta proximidade para ligar com confiança e outro maior em que se requer também menção ao local visitado no texto para aumentar a confiabilidade da ligação. Em ambos os casos, ambiguidades são resolvidas escolhendo o local de interesse com rótulo lexicalmente mais similar ao local mencionado no texto da postagem.

Palavras-chave: Postagens em mídias sociais. Dados ligados abertos. Dados geoespaciais. Ligação espacial. Enriquecimento semântico.

ABSTRACT

Social media data are constantly generated and represents a relevant source of up-to-date information nowadays. However, they include semi-structured data (e.g. optional geographical coordinates, indication of place) and unstructured data (e.g. post texts) that need to have their correct semantics made explicit. One way to do this is through semantic annotations that links relevant portions of data (e.g. an entire tweet, mentions to named entities) to well-defined descriptions in databases (e.g. OpenStreetMap sites of interest) or knowledge bases(e.g. DBpedia, LinkedGeoData). This work aims to study, select, develop and test algorithms for linking data from social media that have geographic coordinates and text (e.g. georeferenced tweets) to visited places of interest, described in databases or knowledge bases with well defined semantics. This task may not be easy in certain situations because coordinates are usually limited in accuracy and there may be a large number of places of interest close to the coordinates of a post and even places of interest with pratically superimposed coordinates (e.g. in a building with several floors). Thus, an algorithm was implemented to link movement data that considers two distinct radios: a smaller one in which proximity is enough to connect with confidence, another larger one in which it is also required a mention to a visited place in the text to increase the reliability of the linking. In both cases, ambiguities are resolved by choosing the place of interest with a label that is most similar to the location mentioned in the posting text.

Keywords: Posts. Social Media. Linked Open Data. Geographic Data. Spatial Linking. Semantic Enrichment.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

LISTA DE SÍMBOLOS

LISTA DE CÓDIGOS

LISTA DE ALGORITMOS

SUMÁRIO

1 INTRODUÇÃO

Todos os dias, a humanidade cria em média 2.5 quintilhões de bytes de dados e aproximadamente 90% dos dados no mundo de hoje foram criados apenas nos últimos dois anos (IBM, 2012). O surgimento da Internet e a sua popularização através do globo fez com que uma enormidade de dados fossem criados e disponibilizados online, vindos de lugares diversos e a todo instante.

Chamamos de dados sobre movimento qualquer coleção de posições espaço-temporais de objetos móveis. Tais dados podem ser capturados por sensores (e.g. GPS , GSM) e / ou sistemas de informação (e.g. mídias sociais) (FILETO et al., 2015b). Postagens georreferenciadas de mídias sociais são consideradas dados sobre movimento, pois suas coordenadas referem-se supostamente á posições de onde foram enviadas. Além disso, as posições podem possuir relação com o conteúdo textual da postagem. Todavia, coordenadas geográficas associadas a postagens, quando disponíveis, costumam ter baixa precisão, seja devido a problemas nos dispositivos de coleta (e.g. oclusão do sensor GPS), seja por erros às vezes intencionalmente inseridos pelo usuário e/ou servidor por questão de privacidade.

Dados sobre movimento podem ser associados a locais de interesse (PoIs, do inglês Place/Point of Interest), os quais podem ser tomados, por exemplo, de bases de dados (e.g. OpenStreetMap (OSM)[1], bases de PoIs da Google[2] ou do FourSquare[3]) ou bases de conhecimento (e.g. LinkedGeoData (LGD)[4], GeoNames[5]). Muitas das bases que contêm PoIs são de acesso público e gratuito[6] (e.g. OSM), além de serem baseadas em padrões de sistemas abertos (e.g. LGD). Todavia, os dados geoespaciais (e.g. coordenadas, geometria) associados aos PoIs de tais fontes também costumam apresentar problemas de precisão espacial. Assim, muitas vezes, não é trivial ligar dados sobre movimento aos locais efetivamente visitados pelos objetos móveis que geraram tais dados, especialmente em regiões com alta densidade de PoIs, que às vezes podem ser pequenos (e.g. banca de jornal, caixa eletrônico, ponto de ônibus), muito próximos um dos outros ou até geograficamente so-

[1] http://www.openstreetmap.org/
[2] https://developers.google.com/places/
[3] https://foursquare.com/
[4] http://linkedgeodata.org/
[5] http://www.geonames.org/
[6] Dos exemplos apresentados, somente as bases da Google e do Foursquare não estão disponíveis integralmente para uso livre

brepostos (e.g. apartamentos, salas ou lojas em um edifício com vários andares). Desta forma, ligar corretamente dados sobre movimento a PoIs visitados pode se revelar um desafio.

A obtenção de dados sobre movimento também pode apresentar desafios, principalmente quando se trata postagens de mídias sociais, onde existe uma política de privacidade e muitas informações não estão disponíveis ao público. O trabalho descrito nesta monografia procura utilizar uma quantidade limitada de informações. Esta característica é o principal diferencial para os outros trabalhos correlatos que utilizam informações sobre o usuário da postagem, muitas vezes confidenciais.

A ligação entre dados espaço-temporais e PoIs descritos em bases de dados ou de conhecimento espacial faz parte do processo de enriquecimento semântico de dados, e a utilização de métodos de ligação beneficia uma grande quantidade de áreas e aplicações da atualidade (e.g. controle de tráfego, segurança pública e marketing). Um bom enriquecimento semântico oferece a aplicações dados com semântica bem definida em maior quantidade e com maior precisão.

1.1 OBJETIVOS

1.1.1 Geral

O objetivo geral deste trabalho é desenvolver, implementar e testar métodos que visam a associação de dados sobre movimento que estão atrelados a dados textuais, tais como postagens georreferenciadas em mídias sociais, a PoIs que tenham sido visitados ao longo do tempo.

1.1.2 Específicos

Os objetivos específicos desse trabalho são:

1. Melhor compreensão do estado da arte em relação a ligação de dados espaço-temporais e textuais a PoIs através de uma revisão bibliográfica com ênfase em mídias sociais;
2. Pesquisar, avaliar e adaptar métodos da literatura, e desenvolver novos métodos para associar dados sobre movimento que estão atrelados a textos a PoIs visitados;
3. Auxiliar na coleta, organização e preparação de dados para experimentos, incluindo tweets e coleções de PoIs;

4. Validar os métodos selecionados através de experimentos que meçam seu tempo de execução e a qualidade dos resultados gerados, utilizando medidas como precisão e cobertura;

5. Organizar o acesso aos dados enriquecidos semanticamente para alimentar trabalhos futuros, tais como a construção de *data warehouses* para análise de tais dados.

6. Redigir artigos e monografia descrevendo o trabalho e resultados obtidos.

1.2 METODOLOGIA

O trabalho foi desenvolvido no Laboratório para Integração de Sistemas e Aplicações (LISA) da UFSC. As subseções a seguir descrevem os materiais, métodos, ferramentas, etapas e o cronograma para o desenvolvimento deste trabalho.

1.2.1 Ferramentas e Meios

Em busca de facilitar o estudo e desenvolvimento deste trabalho, foram utilizadas algumas ferramentas prontas e gratuitas, assim como meios já conhecidos na literatura. Tais ferramentas e meios estão descritos a seguir:

- O LISA conta com um extenso banco de dados contendo tweets enviados do território nacional. Tal banco será usado como fonte de dados sobre movimento.

- A base de dados OSM será utilizada como fonte de PoIs a serem usados no enriquecimento dos dados sobre movimento.

- Métodos e algoritmos para ligar dados sobre movimento a PoIs serão desenvolvidos em linguagem Python, através da IDE (*Integrated Development Environment*) Eclipse[7] .

- Tanto dados sobre movimento quanto os PoIs usados no enriquecimento semântico serão armazenados em um banco de dados PostgreSQL[8] com a extensão PostGIS[9] para gerenciamento

[7] http://www.eclipse.org/
[8] https://www.postgresql.org/
[9] http://PostGIS.net/

de dados espaciais. Métodos de indexação e execução eficiente de junções espaciais disponíveis no PostGIS constituem a base inicial dos métodos de enriquecimento semântico a serem desenvolvidos.

- Bibliotecas de similaridade léxica, fonética, semântica e contextual serão usadas para efetuar desambiguação quando houver vários PoIs candidatos nas proximidades de uma postagem.

1.2.2 Etapas

1. **Fundamentação teórica:** o primeiro passo consiste no estudo de uma série de conceitos básicos necessários para a realização deste trabalho.
2. **Revisão da literatura:** revisão sistemática da literatura, buscando outros trabalhos similares ou com tema em junção espacial de dados sobre movimento com PoIs.
3. **Implementação:**
   - Codificação de métodos para ligação de *tweets* com PoIs extraídos do OSM.
   - Validação das fases dos métodos propostos, assim como análise quanto à eficiência e qualidade dos resultados obtidos.
4. **Análise dos resultados:** Apresentação e análise crítica dos resultados obtidos.

1.3 ORGANIZAÇÃO DO TRABALHO

Este trabalho está dividido em 6 capítulos. O capitulo 2 traz a fundamentação teórica em dados sobre movimento, dados de mídias sociais e locais de interesse. O capítulo 3 traz a revisão bibliográfica no assunto e os trabalhos correlatos. O 4º capítulo descreve em detalhes a solução proposta neste trabalho. O capítulo 5 apresenta detalhes de implementação e relata os experimentos efetuados para validar a proposta. Finalmente, o capítulo 6 faz a reflexão sobre os resultados obtidos e traz propostas para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Para um melhor entendimento do nossos estudos e experimentos se faz necessária a apresentação de alguns temas como dados de mídias sociais e dados geográficos, abordando principalmente o campo da semântica de Locais de Interesse (PoI).

2.1 DADOS GEOGRÁFICOS

Um dado geoespacial representa a forma, a localização e/ou atributos descritivos de algum objeto no planeta Terra, seja ele natural ou construído pelo homem (e.g. prédios, oceanos, shoppings). Os dados espaciais normalmente são armazenados de acordo com algum sistema de coordenadas geográficas e podem ser manipulados com operadores topológicos (e.g. disjunto, toca, intercepta, corte), de medição espacial (e.g. área, perímetro) e de posição relativa (e.g. ao norte de, ao sul de).

Tais dados normalmente são acessados, analisados e manipulados por sistemas de informação geográfica (Geographic Information Systems - GIS). Um GIS integra hardware, software, dados geográficos e pessoas permitindo ver, entender, consultar e interpretar dados para revelar relações, padrões e tendências.

Dados geoespaciais podem ser divididos quanto à sua natureza, em 3 classes: tabular, vetores e matrizes. O formato tabular corresponde ao grupo de dados armazenados em tabelas, possuindo campos alfa-numéricos.

O segundo formato é o vetorial, onde cada dado possui suas características representadas por pelo menos 1 par de coordenadas. Este formato permite uma visão esteticamente mais agradável através de geometrias, porém dados em constante atualização normalmente não são armazenados em vetores, visto que algoritmos para manipulação de vetores são complexos.

A figura 1 ilustra as geometrias básicas de dados vetoriais bidimensionais: pontos (representados por latitude e longitude), linhas e polígonos (ambos representados por sequências de pontos geográficos).

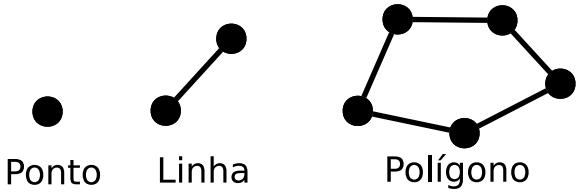

Figura 1 – Exemplo de geometrias representadas por vetores

O terceiro formato corresponde ao matricial, também chamado de *raster*. Nele, dados estão armazenados em linhas e colunas que representam uma superfície do mundo, na qual cada célula corresponde a um valor do atributo representado.

Utilizado para dados de satélites e sensores remotos, o modelo matricial é mais fácil e rápido de ser realizado, pois a representação matricial é análoga a imagens raster, e células podem ser tratadas como pixel a serem visualizados. Porém este modelo possui desvantagens quanto a precisão dos dados, visto que ela depende do tamanho da superfície representada, do número de células usadas para representá-la e do número de bits utilizados para representar cada célula. Como a superfície é dividida em células, quanto maior a célula menor a resolução da representação e o custo da velocidade de processamento e armazenamento dos dados crescem de acordo com a resolução da representação. A figura 2 ilustra uma superfície de mundo armazenada no modelo matricial.

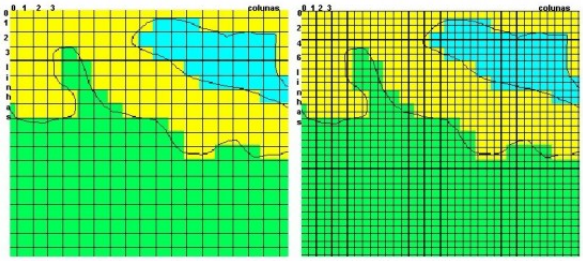

Figura 2 – Exemplo de formato matricial

Neste trabalho, PoIs são representados por pontos geográficos associados a atributos alfa-numéricos que referem-se a propriedades relevantes que auxiliam na especificação do que é cada PoI. O crescimento da publicação de dados geoespaciais na Web (por consequência, o aumento de publicação de PoIs) fez com que fosse criado o consórcio OGC[1] (Open Geospatial Consortium) para a regulamentação e padronização de conteúdos e serviços geo-espaciais. O OGC hoje congrega mais de 527 empresas, agências governamentais e universidades ao redor do mundo. Além disso, atualmente, com o enriquecimento dos dados geoespaciais com outros dados da web semântica, iniciativas que transformam PoIs em dados ligados abertos (Linked Open Data - LOD), como o LinkedGeoData, vem crescendo cada vez mais.

2.1.1 Locais De Interesse

Um PoI é um dado geográfico que algum individuo possa achar útil ou interessante (GAO; CAO; FAN, 2016). Dados geoespaciais publicados seguindo padrões LOD estão sendo armazenados em diversas bases (e.g. LinkedGeoData, base de dados da Google). Dados geoespaciais publicados na Web, tais como dados ligados e em formatos RDF, XML JSON são semi-estruturados segundo *tags* (etiquetas) que classificam ou descrevem informações relevantes de cada PoI. Um PoI pode ter varias *tags*, e cada uma é representada na forma de tupla $<key,value>$. Essas *tags* podem representar propriedades ou relações, tais como nome (e.g. propriedade rdf:label ou dc:title), geometrias em algum sistema de coordenadas geográficas, classes de PoIs e relações de generalização-especialização entre tais classes (hierarquia de subsumption).

Os PoIs utilizados neste trabalho foram retirados do OSM, uma base de dados geográficos colaborativa e livre. O dados do OSM podem ser intercambiados na linguagem de marcação XML (eXtensible Markup Language) e são organizados e mantidos através de páginas wiki[2](CODESCU et al., 2014), que listam as *tags* mais populares e especificam a sua função de uso. Tais fontes procuram facilitar a edição e manipulação dos dados mantidos no OSM, de forma a expandir e melhorar o próprio banco. O código descrito abaixo mostra um PoI representado no OSM.

[1] http://www.opengeospatial.org/ogc

[2] http://wiki.openstreetmap.org/wiki/Tags e http://wiki.openstreetmap.org/wiki/Features

```
<?xml version="1.0" encoding="UTF-8"?>
<osm version="0.6" generator="CGImap␣0.6.0
(1918␣thorn-03.openstreetmap.org)" copyright=
"OpenStreetMap␣and␣contributors" attribution=
"http://www.openstreetmap.org/copyright"
license="http://opendatacommons.org/licenses/
odbl/1-0/">
 <way id="129453071" visible="true" version="3"
 changeset="11489151" timestamp=
"2012-05-03T14:14:31Z" user="Koyaani"
uid="506085">
  <nd ref="1428242371"/>
  <nd ref="1428242394"/>
  <nd ref="1739255084"/>
  <nd ref="1739255081"/>
  <nd ref="1428242341"/>
  <nd ref="1428242351"/>
  <nd ref="1428242371"/>
  <tag k="amenity" v="bank"/>
  <tag k="building" v="yes"/>
  <tag k="name" v="Ita "/>
  <tag k="opening_hours" v="Mo-Fr␣10:00-16:00"/>
  <tag k="wheelchair" v="yes"/>
 </way>
</osm>
```

Listing 2.1 – XML de um PoI no OSM

2.1.2 Dados Geográficos no PostGIS

Para armazenar e manipular tais dados, o banco de dados escolhido foi o PostgreSQL[3] juntamente com sua extensão PostGIS[4] para a manipulação de dados geoespaciais. Tal banco foi escolhido devido sua licença permitir o uso gratuito do mesmo, além de ser um dos Sistema de Gerenciamento de Banco de Dados (SGBD) mais utilizados no mundo[5].

PostGIS é uma extensão livre (*open source*) que permite a representação de dados geométricos (e.g. pontos, polígonos, segmentos de pontos, coleções geométricas) e inclui operações espaciais com resultados escalares (e.g área, distância, perímetro) ou operações espaciais com resultados geométricos (e.g. buffer, união ou intersecção de geometrias), operadores topológicos e outros operadores com resultados booleanos (e.g. simetria, touch, intersect, overlaps, inside), entre outros recursos relevantes, tais como indexação espacial de objetos em

[3] https://www.postgresql.org/
[4] http://postgis.net/
[5] https://db-engines.com/en/ranking

espaços multi-dimensionais. A extensão foi registrada em 2006 pelo OGC e obedece ao padrão de acesso simples (*simple features access*).

O PostGIS possui duas classes de dados espaciais: "geometry"e "geography". A principal diferencia entre os duas é a base de referência de cada um, enquanto na primeira classe o sistema de referência é o espacial, ou seja um plano, na segunda classe o sistema é uma esfera. A terra por ser uma esfera faz com que as operações com a segunda classe de dados mais precisas, porém mais complexas visto que qualquer cálculo precisa levar em consideração a curvatura da terra.

A classe de dados "geometry"é o formato padrão para dados geométricos e possui uma vasta quantidade de funções. Como dito anteriormente, sua precisão pode apresentar problemas, por não levar em consideração a curvatura da terra e dentre seus principais tipos de objeto estão:

- Point
- Linestring
- Multipoint
- Multilinestring
- Multipolygon
- Polygon
- Geometrycollection

A figura 3 apresenta a hierarquia de tipos que compõe a classe "Geometry".

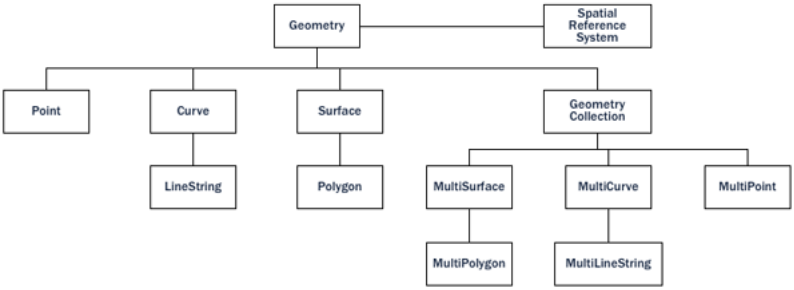

Figura 3 – Hierarquia da classe "Geometry"

Junção espacial é uma operação disponível em alguns sistemas de dados geográficos para efetuar a junção entre duas ou mais tabelas através das coordenadas geográficas presentes em algum dos atributos das tabelas. Esta ação é uma das ações existente em banco de dados geográficos que permite que os dados sejam analisados sob uma ótica geoespacial.

Para auxiliar na especificação e na realização de junções, o PostGIS conta com alguns operadores topológicos (alguns com resultados booleanos), operador buffer e indexação. Dentre os principais operadores topológicos estão:

- ST_Disjoint(obj1, obj2): Analisa se existe algum ponto em comum entre os dois objetos, retornando "True"caso não haja.
- ST_Intersects(obj1, obj2): Analisa se existe alguma intersecção entre os dois objetos, retornando "True"caso haja.
- ST_Within(obj1, obj2): Analisa se o objeto 1 está completamente dentro do objeto 2, retornando "True"caso esteja.
- ST_DWithin(obj1, obj2, dist): Analisa se o objeto 1 está em uma distância máxima "dist"do objeto 2, retornando "True"caso esteja.
- ST_Crosses(obj1, obj2): Verifica se dois objetos se cruzam e retorna verdadeiro caso ocorra

- ST_Buffer(obj1,dist): Retorna uma geometria que cobre todos os pontos à uma distância "dist"do objeto

O PostGIS também traz suporte à indexação, que aumenta a eficiência de trabalhos com grandes volumes de informação. A indexação de geometrias no PostGIS é feita por uma estrutura genérica de índice (GIST), que se baseia nas caixas limites do objeto (*Bounding Box*). Algumas operações espaciais utilizam esses retângulos diretamente, a figura 4 exemplifica o uso para enquadrar geometrias.

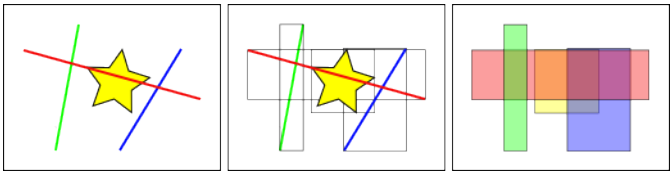

Figura 4 – Bounding Boxes

2.2 DADOS DE MÍDIAS SOCIAIS

Mídias sociais são canais on-line que possibilitam a interação e o relacionamento entre usuários (AGICHTEIN; CASTILLO; DONATO, 2008), gerando grandes quantidades de dados. O surgimento de tais redes pode ser correlacionado com a evolução da web e de seus usuários: inicialmente, os usuários que eram apenas consumidores de material criado por uma pequena porcentagem de outros usuários passaram a produzir material próprio e compartilhar-los em redes sociais.

Os dados a serem enriquecidos semanticamente neste trabalho foram retirados do Twitter[6], devido à grande fonte de usuários e pelo conteúdo das postagens ser definido como de uso público. O Twitter é umas das maiores rede sociais da atualidade, com dados de uso como 313 milhões de usuários ativos mensalmente, 1 bilhão de usuários únicos no mês e mais de 40 línguas suportadas [7].

Um *tweet* (uma postagem de um usuário no Twitter) essencialmente é composto pelo texto da postagem (text), o usuário que realizou a postagem (user) e podem ou não conter uma localização (*position(p)*) atrelada ao dado.

[6] https://twitter.com/
[7] https://about.twitter.com/company

O texto da postagem possui informações valiosas para entendermos o contexto da postagem. É possível extrair do texto as palavras de maior importância semântica, como menções a lugares, nome de pessoas, nome de obras literárias, etc. Tais palavras são denominadas palavras relevantes, e podem ser utilizadas no processo de enriquecimento semântico para atribuir semântica bem definida ao *tweet*, atribuindo metadados a essas postagens. Na figura 5 temos a palavra revelante do texto circulada em azul e a localização circulada e apontada em vermelho.

Figura 5 – Exemplo de Tweet

Apesar da quantidade de palavras relevantes presentes em um texto de mídia social, a extração desta nem sempre pode ser feita de maneira simples. Cada texto possui suas próprias características, gerando uma heterogeneidade e a falta de um padrão fixo. Entre tais características, há algumas que apresentam desafios para a compreensão do texto, citadas abaixo:

- **Tamanho da postagem**: Postagens de mídias sociais em geral possuem um limite de tamanho pequeno, fazendo com que usuários passassem a usar abreviações como forma de colocar mais informações no texto. A desambiguação dessas abreviações nem sempre pode ser feita de forma trivial, visto que muitas delas são criadas pelo próprio usuário e/ou a mesma abreviação pode ter vários significados a depender do usuário, o que dificulta a catalogação das mesmas.

O twitter historicamente possuía um limite de 140 caracteres por postagem, o que incluía além das letras do texto, outros tipos de conteúdo(eg., fotos, GIFs(*Graphics Interchange Format*), vídeos, enquetes). Porém a partir de 2016, tais tipos de conteúdo saíram da contagem de caracteres. A figura 6 mostra a contagem de caracteres feita.

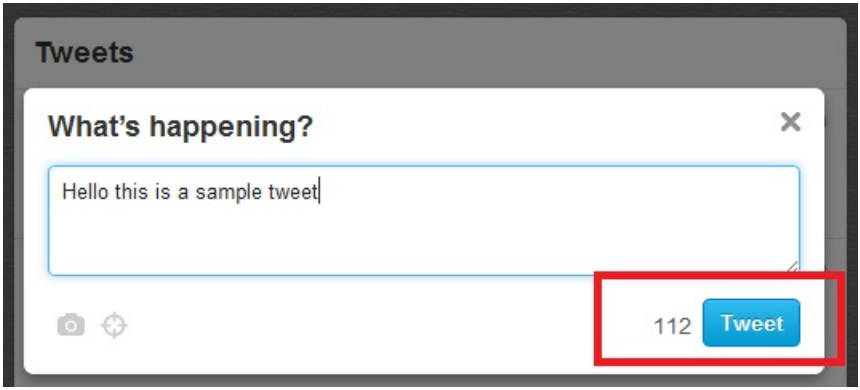

Figura 6 – Contagem de caracteres do Twitter

- **Linguagem coloquial**: A linguagem comumente utilizada nas redes sociais é a não formal, o que implica em uma grande quantidade gírias no texto, erros de digitação e as faltas propositais de pontuação e acentuação. O problema dessas gírias é que elas estão em constante criação e desuso, conforme o tempo passa e a sociedade se transforma.

  A figura 7 exemplifica gírias como a palavra "me", que significa "meu", e erros de digitação com as palavras: "i"(verbo "ir"), "corta"(verbo "cortar") e o substantivo "cabelo"que está no singular, e deveria estar no plural. Alem disso, podemos ver que a preposição "que"está abreviada como "q".

Figura 7 – Gírias e erros de digitação

- **Palavras sem importância**: Grande parte de uma postagem não possui importância alguma para a extração das palavras relevantes. Além disso, postagens muitas vezes apresentam imitações de ruídos e/ou palavras sem sentido. A figura 8 mostra um *tweet* feito através da conta pessoal do atual presidente dos Estados Unidos da América (EUA), onde a palavra "covfefe"não possui nenhum sentido ou significado.

Figura 8 – Exemplo de ruídos e palavras sem sentido

- **Entidades locais e desconhecidas**: Muito das referências usadas pelos usuários partiram através do conhecimento local, e por serem conhecidos apenas pelas pessoas da região ou são lugares recentemente abertos, tais entidades não existem na maioria das bases de conhecimentos usadas no processo.

3 TRABALHOS CORRELATOS

Este capítulo busca resumir quatro trabalhos realizados na área que foram considerados mais relacionados e relevantes para o foco deste TCC, mostrando relações entre eles e o trabalho descrito neste documento. Todos os trabalhos estão relacionados de alguma forma a PoIs, localização geográfica ou Mídias Sociais.

O principal trabalho correlato é o de (FILETO et al., 2015b), ele serve de base para este TCC e por esse motivo é o trabalho que possui maior semelhança com este. O trabalho apresenta um conjunto de definições sobre dados de movimento e usa estas definições para formar a ontologia descrita no artigo. As principais definições propostas neste trabalho são listadas abaixo.

1. Sequência de Posições de um objeto em movimento (Moving Object's position sequence (MOPS)).

2. Segmento de Movimento (Movement Segment (MS)) de um MOPS

3. Hierarquia sobre os MS (Moviment Segment Hierarchy (MSH)) de um determinado MOPS

4. Faceta Semântica

5. Anotação de Movimento

Além dessas definições, o artigo aborda o processo de enriquecimento semântico de dados de movimento, explicitando suas etapas e apresentando um algoritmo para a ligação dos dados. O algoritmo leva em consideração a distância entre os dados e a similaridade textual, porém difere do algoritmo apresentado neste TCC pois considera apenas um único raio.

O trabalho de (WU et al., 2015), assim como este TCC e o trabalho de (FILETO et al., 2015b), aborda o campo de anotações semânticas para dados de movimento de mídias sociais buscando definir a localização do usuário usando o conteúdo textual da postagem, porém não leva em consideração a distância entre a posição da postagem e o local em questão. Além do texto da postagem, o sistema descrito no artigo utiliza-se dos históricos de postagens do usuário e de métodos de contagem das palavras relevantes que ocorrem em determinada localização. Dentre esses metodos estão:

- Baseado na densidade: Considera a proporção de menções à uma mesma palavra em torno de uma área
- Baseado na frequência: Considera a quantidade de vezes em que uma palavra foi mencionada em um mesmo local

Em (FURLETTI et al., 2013) e (KIM; PEREIRA; ZHAO, 2014) os autores visam determinar as atividades realizadas por um objeto móvel nos PoIs visitados pelo mesmo. Apesar da identificação da atividade realizada por um objeto móvel não ser o foco deste trabalho, ambos trabalhos propõem métodos para identificar os PoIs visitados pelo objeto móvel, possibilitando a inclusão dos mesmos como trabalhos correlatos. Em (FURLETTI et al., 2013), os autores consideram uma distância que o objeto móvel está disposto a andar para chegar no PoI e métodos probabilísticos baseados no modelo gravitacional de Newton. Por fim, (KIM; PEREIRA; ZHAO, 2014) utilizam de dados estatísticos para determinar as posições que um objeto móvel parou para realizar alguma atividade. Apesar de tal abordagem não retornar os possíveis lugares visitados pelo objeto móvel, ao determinar as posições onde possivelmente ocorreu alguma atividade, é possível utilizar a mesma para otimizar nossa proposta e outras propostas para melhor a precisão e, consequentemente, reduzir o número de falsos-positivos.

Embora todos esses métodos associem postagens aos PoIs visitados (de onde as supostas postagens foram enviadas) há variações de precisão e cobertura nos resultados dos diferentes métodos . Aqueles que utilizam mais informações tendem a gerar melhores resultados. Por exemplo, métodos que utilizam perfis de usuário e de atividades tipicamente realizadas em PoIs, como os indicados nas duas últimas linhas da tabela 1. Todavia, esses resultados são conseguidos em certos casos e a informação adicional utilizada nem sempre está disponível. O método proposto neste TCC tenta obter melhores resultados sem tais informações adicionais, sendo o único a utilizar raios distintos de proximidade da postagem com os PoIs candidatos, além de informação textual, no processo de desambiguação. As Tabelas 2 e 3 continuam com o comparativo entre este trabalho e os trabalhos correlatos.

Tabela 1 – Saída e entrada de cada algoritmo/sistema proposto

Trabalho	**Entrada**	**Saída**
TCC	PoIs, Post, Raio menor, Raior maior	(Post,PoI$_{visit}$)
(FILETO et al., 2015b)	PoIs, Post, Raio	(Post,PoI$_{visit}$)
(WU et al., 2015)	PoIs, Usuário, Post	(Post,PoI$_{visit}$)
(FURLETTI et al., 2013)	PoIs, Usuário, Post	(Post,PoI$_{visit}$)
(KIM; PEREIRA; ZHAO, 2014)	Atividades, Usuário, PoIs, Post	(Post,PoI$_{visit}$)

Tabela 2 – Utilização de PoIs, Texto da Postagem e Distância entre PoI-Usuário em métodos para ligação a PoIs

Trabalho	**PoIs**	**Texto**	**Dist.**	**Múlt. Raios**
TCC	Sim	Sim	Sim	Sim
(FILETO et al., 2015b)	Sim	Sim	Sim	Não
(WU et al., 2015)	Sim	Sim	Não	Não
(FURLETTI et al., 2013)	Sim	Não	Sim	Não
(KIM; PEREIRA; ZHAO, 2014)	Sim	Não	Sim	Não

Tabela 3 – Utilização de métodos relacionados à atividades

Trabalho	**Prob.**	**Densidade**	**Frequência**
TCC	Não	Não	Não
(FILETO et al., 2015b)	Não	Não	Não
(WU et al., 2015)	Não	Sim	Sim
(FURLETTI et al., 2013)	Sim	Não	Não
(KIM; PEREIRA; ZHAO, 2014)	Não	Não	Sim

4 PROPOSTA

Para atingir os objetivos especificados na Seção 1.1, este trabalho propõe um algoritmo baseado em junção espacial e similaridade léxica para associar dados sobre movimentos atrelados a dados textuais com PoIs visitados pelo objeto móvel.

Neste trabalho, somente a utilização de junção espacial não é o suficiente para determinar o PoI visitado por um usuário/objeto móvel. Isso ocorre porque que os PoIs em um determinado espaço estão, por muitas vezes, sobrepostos (e.g. lojas distintas em diferentes andares de um shopping) e/ou em grande quantidade (e.g. centros comerciais, prédios). Nestas situações a precisão das coordenadas geográficas das postagens e dos PoIs por vezes é insuficiente para fazer as ligações corretas. Assim, é necessário utilizar o texto presente nas postagens para desambiguar o PoI visitado pelo usuário.

Os textos presentes em postagens podem fazer menções diretas (e.g. nome do local) e/ou indiretas (e.g. palavras relevantes relacionadas semanticamente ao local) ao PoI visitado pelo usuário. Para menções diretas, medidas de similaridade léxica (e.g. Jaro-Winkler, Damerau-Levenshtein) podem ser usadas para comparar a menção com os nomes de superfície dos PoIs retornados pela junção espacial. Para as menções indiretas, ferramentas e técnicas de anotação semântica (e.g. Babelfy, DBpedia-Spotlight) podem ser usadas para enriquecer semanticamente o texto com anotações que apontam para recursos semanticamente bem-definidos, e as semânticas presentes em tais recursos podem ser comparada com as semânticas presentes nos PoIs.

O principal foco deste trabalho é a desambiguação dos PoIs nos casos de alta densidade dos mesmos. A desambiguação é feita através da comparação das distância geográfica entre o par $< postagem, PoI >$ e/ou através das menções diretas presentes no texto da postagem. O uso de menções indiretas foi deixado para trabalhos futuros.

4.1 ALGORITMO PROPOSTO

O algoritmo proposto visa ligar uma postagem a PoIs visitados,considerando primeiro aqueles mais próximos, dentro de um raio passado por parâmetro em graus decimais (*Decimal Degrees*) e realiza a desambiguação selecionando a ligação cujo nome da entidade citada no texto possua a maior similaridade léxica com o nome de superfície

do PoI.

Os dados ligados são ordenados de acordo com o ID da postagem, para em seguida ser feita a utilização do algoritmo de desambiguação. Quando não há menção a PoI no texto e apenas 1 PoI candidato dentro de um raio menor a ligação com este PoI é considerada correta.

O primeiro passo do algoritmo é realizar uma junção espacial entre uma tabela contendo as postagen e uma tabela contendo PoIs da mesma região geográfica. A consulta mostrada no *Listing* 4.1 realiza a junção (JOIN) duas tabelas SQL de acordo com a distância das coordenadas geográficas, selecionando as seguintes colunas:

- post.id: ID da postagem (String)
- post.text: Texto da postagem (String)
- nodes.id: ID do POI (String)
- name: Nome do POI (String)
- dist: Distância entre as coordenadas da postagem e do PoI (Double, Decimal degree)

```
SELECT tweet.id, tweet.text, nodes.id, nodes.tags -> name AS name,
 distance(nodes.geom.t.geom) AS dist
FROM public.tweet INNER JOIN public.nodes
ON ST_DWithin(nodes.geom,tweet.geom,largestRadius)
WHERE nodes.tags ? name
ORDER BY tweet.id, dist ASC;
```

Listing 4.1 – Código SQL da junção espacial pela distância

Após a junção espacial, o algoritmo separa as ligações de acordo com o ID da postagem, a distância entre a postagem e o PoI candidato. O primeiro grupo consiste de pares onde a distância entre a postagem e o PoI é inferior a um raio menor $raio_1$, enquanto o segundo grupo consiste de pares cuja distância entre a postagem e o PoI supera $raio_1$ mas é inferior a um raio maior ($raio_2 > raio_1$). Ambos os raios são parâmetros do algoritmo proposto. Quando há apenas um PoI candidato a ligação a uma postagem no primeiro grupo, tal ligação é considerada. Caso contrário a desambiguação dos candidatos é feita segundo a ordem de distância, mas considerando apenas candidatos que tenham uma similaridade léxica de menção no texto da postagem com o nome do PoI superior ao limiar (threshold) *similaridade*.

O algoritmo é dividido em duas etapas: Agrupamento por ID e por distância (distinção em 2 raios) e a fase de Comparação de similaridade.

4.1.1 Agrupamento por ID e por distância

A primeira etapa resulta em uma tabela onde cada tupla retornada é uma ligação feita entre a postagem e um PoI e está ordenada de acordo com o ID da postagem. O algoritmo percorre a tabela organizando cada ligação de acordo com a distância entre a entidade e o PoI, separando em dois grupos:

- Primeiro Raio: Área circular com raio menor;
- Segundo Raio: Área circular com um raio maior do que o Primeiro Raio.

A partir disso, o algoritmo entra em processo de decisão da melhor ligação feita, seguindo alguns critérios:

1. Caso haja somente um único PoI no raio, o algoritmo considera essa ligação como sendo a melhor e não realiza a comparação de similaridade;
2. Caso haja algum único PoI no primeiro raio, o algoritmo automaticamente ignora as ligações feitas no segundo raio;
3. Caso haja mais de um PoI segundo raio em questão (menor ou maior), algoritimo verifica a similaridade léxica entre a menção ao lugar presente na postagem e os nomes de superfície dos PoIs.

Agrupamento por ID e por distância

Input : $T = (\{T_0, T_1, ..., T_n\})$ // Postagens de mídias sociais.
$P = \{P_0, P_1, ..., P_p\}$ // PoIs.
$raio_1$ // Raio menor em decimal degrees.
$raio_2$ // Raio maior em decimal degrees.

Output: SA // Tupla PoI visitado e Post (Post, PoI$_v$*isitado*).

begin
SA $\leftarrow \emptyset$
SJ $\leftarrow \pi_{t \leftarrow T.*, p \leftarrow P.*, geoDist}(T \bowtie_{(geoDist \leftarrow dist(T.geom, P.geom)) < \text{raio}_2} \text{P}))$ //*PoIsePosts*
foreach $t \in T$ **do**
  $G_1 \leftarrow \emptyset$;
  // grupo raio 1 vazio
  $G_2 \leftarrow \emptyset$;
  // grupo raio 2 vazio
  foreach *(t,p,geoDist)* $\in SJ$ **do**
    if $geoDist < raio_1$ **then**
      $G_1 \leftarrow G_1 \cup (t, p, geoDist)$;
    else
      $G_2 \leftarrow G_2 \cup (t, p, geoDist)$;
    end
  end
  if *len*$(G_1) > 0$ **then**
    if *len*$(G_1) == 1$ **then**
      $SA \leftarrow SA \cup G_1$;
    else
      $SA \leftarrow SA \cup comparacaoSimilaridade(G_1)$;
    end
  else
    if *len*$(G_2) == 1$ **then**
      $SA \leftarrow SA \cup G_2$;
    else
      $SA \leftarrow SA \cup comparacaoSimilaridade(G_1)$;
    end
  end
end
return SA;

Algorithm 1: O Algoritmo proposto

4.1.2 Comparação de similaridade textual

Com múltiplos PoIs dentro dos raios, se faz necessário a comparação de similaridade léxica entre os nomes de superfície. Como em um primeiro momento temos o texto completo da postagem, o algoritmo realiza o tratamento deste texto separando apenas o que ele julga como o nome do local, que é chamado de menção à entidade.

O grande problema da realização desta etapa são as diferentes formas encontradas no texto dessas postagens, que por muitas vezes não seguem padrão algum, podendo possuir palavras não encontradas no dicionário formal da língua (e.g. gírias, abreviações, palavras estrangeiras) e/ou palavras com múltiplos significados.

No escopo deste trabalho, somente postagens feitas pelo aplicativo Foursquare no Twitter foram utilizadas, visto que tais postagens obedecem à um padrão fixo, facilitando o tratamento do texto e a extração das menções a entidade. Este padrão será abordado de maneira mais precisa no Capítulo 5.

Após a extração da menção à um local do texto da postagem, a similaridade léxica é calculada linha por linha, e é dada através da distancia Jaro-Winkler. Essa distância correlaciona o tamanho das strings, o numero de correlações entre caracteres e o numero de transposições entre o nome da entidade retirada do texto da postagem e o nome do PoI presente no valor da tag do PoI com chave "nome". A escolha desta medida de distância léxica foi feita devido à experiência prévia obtida pelo laboratório onde esta monografia foi desenvolvida.

Comparação de similaridade Textual

Input : G = $\{(t, p_0, geoDist_0),(t, p_1, geoDist_1),...,(t, p_n, geoDist_n)\}$ // Set Tupla PoI visitado e Post (Post, $\text{PoI}_v isitado$).

Output: g // Tupla PoI visitado e Post (Post, $\text{PoI}_v isitado$).

begin

g $\leftarrow \emptyset$

$entidade \leftarrow palavraRelevante(t.text)$;

// palavra relevante do texto da postagem

$proximidade$ = -1/;

foreach $(t, p, geoDist) \in G$ **do**

  $proximidade \leftarrow$ $distance.get\_jaroWinkler\_distance(entidade, p.nome)$;

  if $similaridade < proximidade$ **then**

    $similaridade = proximidade$;

    $g \leftarrow (t, p, geoDist)$;

  end

end

return *g;*

Algorithm 2: Comparação de similaridade Textual

5 EXPERIMENTOS

Os experimentos foram feitos utilizando um protótipo do algoritmo (algoritmo 1) para ligação de Tweets georreferenciados a POIs, que recebe como parâmetros: o raio de distância máxima entre os dados de entrada (raio maior) e um raio menor. Tal protótipo foi implementado em linguagem Pyton, sobre a estensão PostGIS versão 2.3 do SGBD PostgreSQL versão 9.5, a qual provê primitivas para representação e manipulação eficiente de extensões geográficas, incluindo implementação eficiente de junção espacial baseada em índices espaciais do GIST. Tais experimento foram executados sobre um computador com processador i3, disco de 500 gigabytes e 2 gigabytes de RAM. Estes experimentos usaram como entrada a totalidade de PoIs e postagens as bases de dados descritas na seção 5.1.

Os experimentos foram repetidos 10 vezes para efeito de validação e o algoritmo foi analisado de 3 maneiras distintas. A primeira maneira foi através da realização de uma sequência de experimentos onde o raio menor é variado, começando inicialmente com raio de tamanho 0,0003dd (cerca de 33,396 metros) e decrementando esse raio em 0,00005dd. Estes experimentos mantiveram o raio maior fixo em 0,001dd (cerca de 111,2m) e usaram como entrada a totalidade de PoIs e postagens dos bancos, que serão descritos na seção 5.1. No experimento com raio menor inexistente (raio menor = 0), o algoritmo deste trabalho equivale ao algoritmo descrito no artigo baquara 2 e os resultados foram usados como base de comparação.

A segunda forma de análise foi quanto à quantidade de postagens presentes no banco de dados. Nestes 8 experimentos, foram utilizados a totalidade dos PoIs presentes no banco, e possui como raio menor o valor 0,0003dd. Este experimento procurou mostrar o crescimento do tempo de execução do algoritmo quando utilizados em bases maiores.

Por último, foi realizado o experimento para medir o grau de acerto das ligações feitas pelo algoritmo. Este experimento contou com a totalidade dos PoIs e postagens do banco, o parâmetro do raio menor foi de 0,0003dd e uma amostra deste experimento, com aproximadamente 2100 ligações feitas, foi analisada à regra ouro.

5.1 BASES DE DADOS UTILIZADAS

5.1.1 Postagens

Como é de se imaginar pela quantidade de usuários do twitter dita na seção 2.2, a quantidade de tweets no mundo é enorme e grande parte deles não possuem coordenadas geográficas ou obedecem à um padrão específico de texto. Para reduzirmos esta quantidade e garantir um padrão para as postagens, apenas tweets feitos através do aplicativo Foursquare foram utilizados no estudo.

Foursquare é um serviço de rede social que tem como objetivo compartilhar e receber informação sobre locais de interesse. Tweets feitos através deste aplicativo possuem um forma fixa: Uma mesma frase padrão (sujeito,verbo,preposição), twitter do local (optativo), nome do local e elementos textuais (optativo).

I am at @lisaufsc Laboratório LISA.
<Frase Padrão><Twitter do local opcional><Menção ao local><Elementos Textuais optativos>

O padrão adotado pelo Foursquare faz com que o tratamento do texto e extração da entidade nomeada seja feito de uma maneira mais fácil, já que o nome da entidade está sempre após a preposição "at"(em). O resto do tratamento realizado é a retirada de elementos do texto que podem vir após o nome da entidade, elementos que são identificados por começarem com caracteres específicos:

- Endereço ("(")
- Website ("http")
- Nome de acompanhantes ("w/")

Inicialmente foram utilizados dois bancos de dados para realização dos experimentos: o banco de postagens de mídias sociais do LISA e um banco local de PoIs retirados do OSM. O banco do LISA contempla não apenas postagens do foursquare, fazendo com que seja necessário o tratamento dessas postagens em tempo de execução, o que aumenta consideravelmente este tempo. Para a resolução deste problema foi feita a extração de tweets unicamente feitos através do foursquare, e a inserção destes no banco local junto com os POIs. A figura 9 ilustra a parte usada do esquema de classe do banco de postagens do LISA.

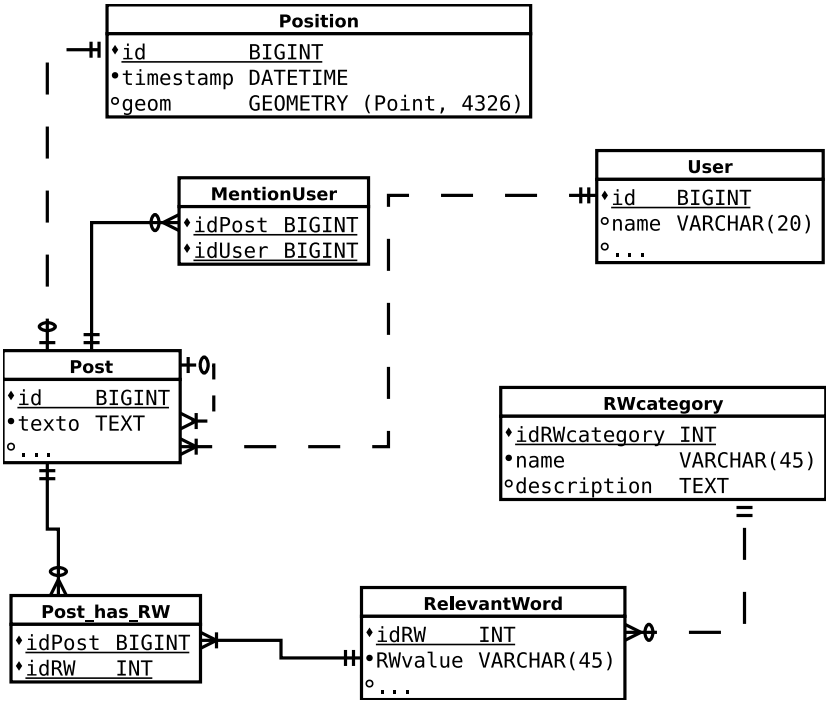

Figura 9 – Esquema do banco de dados do LISA

Nos experimentos foram utilizados 500 mil tweets oriundos do Foursquare, coletados em uma área (Minimum Bounding Retangle(MBR)) em torno do território brasileiro entre 2010 e 2015. Todos são compostos por: um id(id), uma localização geográfica($geom$), um timestamp(t) e um texto($text$).

5.1.2 PoIs

Foram utilizados 55.604.765 PoIs coletados em um MBR em torno do território brasileiro, todos retirados do OSM na data de 23 de março de 2017 às 21 horas e 43 minutos. Os PoIs são compostos essencialmente por:

- id(*id*)
- timestamp(*t*)
- localização geográfica(*geom*)
- Conjunto de tags(*tags*)

Além dos PoIs, o banco também contém o registro de:

- Sequência de PoIs(*way*) feitas por um usuário
- Usuários(*user*)

Os PoIs estão organizados no banco de dados do OSM em 4 tabelas SQL listadas abaixo e mostradas na figura 10:

- nodes: Dados relativos ao PoI.
- ways: Dados de movimento (Sequência de PoIs visitados).
- way_nodes: Relação entre um PoI e uma sequência de PoIs visitados.
- users: Dados relativos ao usuário criador da sequência e do PoI.

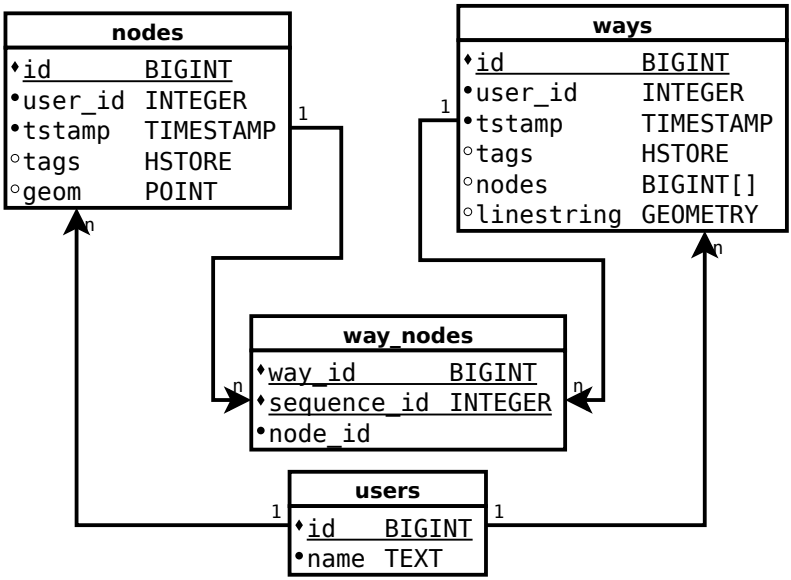

Figura 10 – Esquema do banco de dados do OSM

5.2 RESULTADOS

Foram realizados experimentos com 500 mil tweets retirados do foursquare, com 34% sendo ligados à um POI. Foi-se constatado uma significante diminuição no tempo de execução entre as primeiras e a última implementação do algoritmo. No experimento feito com a totalidade dos bancos (PoIs e Postagens), o tempo de query corresponde a 50% do total, enquanto os outros 50% corresponde ao tempo do algoritmo em si. O estudo mostra que quanto maior a quantidade de Tweets, menor a relevância do tempo da query para o tempo total.

5.2.1 Variação do raio menor

Esta sequência de experimentos foi realizado com toda a base de dados de PoIs e Postagens descritos na subseção 5.1, realizando sucessivas ligações dessas bases com alterações no parâmetro do raio

menor do algoritmo.

Inicialmente foi considerado a precisão usual dos GPS's atuais, cerca de 32 metros. O PostGIS possui como medida o grau decimal, fazendo com que seja necessário o arredondamento desse valor para 0,0003dd (cerca de 33,396 metros). A sequência de experimentos usa este valor como raio menor inicial, e diminui 0,00005dd a cada experimento feito até que o raio chegue a 0. Com o raio menor igual a 0, o algoritmo se torna igual ao algoritmo descrito no artigo (FILETO et al., 2015b).

Estes experimentos buscam encontrar informações sobre:

- Quantidade de tweets ligados ao total e em cada raio
- Quantidade de tweets ligados somente por proximidade espacial
- Quantidade de PoIs encontrados ao total e em cada raio
- Número máximo de PoIS em um único raio
- Distância entre PoI e tweet
- Tempo de execução total, e em cada fase do algoritmo

A informação coletada mostrou uma proporção direta entre o tamanho do raio e a quantidade de ligações entre postagens e PoIs únicos dentro do raio, ou seja, ligações onde não foi necessário o processo de desambiguação. Essa ligações contribuem para um tempo de execução menor, visto que os cálculos de distância textual não são feitos. Em comparação com o algoritmo proposto pelo artigo (FILETO et al., 2015a), o algoritmo proposto possui 36% de ligações com PoIs únicos a mais. A tabela 4 e a figura 11 mostram a quantidade de ligações feitas em cada raio, e as ligações com PoIs únicos.

	Quantidade de ligações efetuadas		
Raio menor (m)	**No Raio menor**	**No Raio maior**	**Único POI**
33,4	60739 (35,69%)	109455 (64,31%)	96960 (56,97%)
27,8	49029 (28,81%)	121165 (71,19%)	94038 (55,25%)
22,2	37975 (22,31%)	132219 (77,69%)	89594 (52,64%)
16,7	26140 (15,36%)	144054 (84,64%)	84281 (49,52%)
11,1	14074 (8,27%)	156120 (91,73%)	78072 (45,87%)
Baquara 2 (0)	0	170194	71167 (41,82%)

Tabela 4 – Ligações feitas nos experimentos

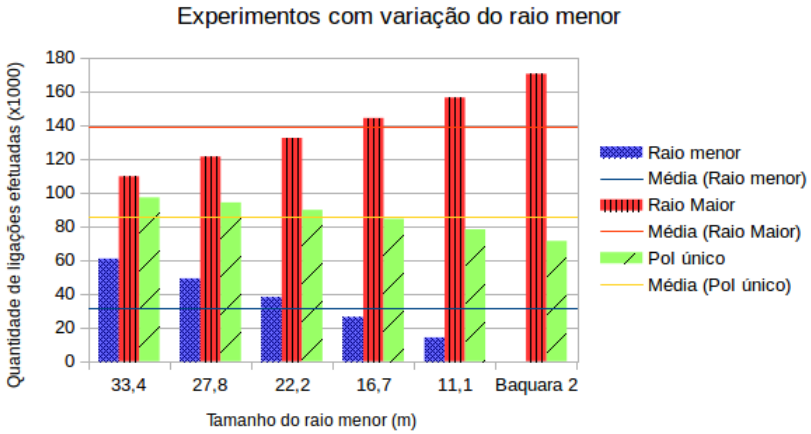

Figura 11 – Quantidade de ligações efetuadas com raio maior == 111,2m

A proporção entre o tamanho do raio e a média de distâncias das ligações se mostra inversamente proporcional. A tabela 5 e a figura 12 mostram os números adquiridos.

	Quantidade de PoIs encontrados		
Raio menor (m)	**No Raio menor**	**No Raio maior**	**Dist. Média (m)**
33,4	107106 (14,66%)	623625 (85,34%)	56,638
27,8	77907 (10,66%)	652824 (89,34%)	57,9844
22,2	54701 (7,49%)	676030 (92,51%)	59,32244
16,7	34378 (4,70%)	696353 (95,30%)	60,7769
11,1	16068 (2,20%)	714663 (97,80%)	62,583
Baquara 2 (0)	0	730731	64,1114095075

Tabela 5 – PoIs encontrados nos experimentos

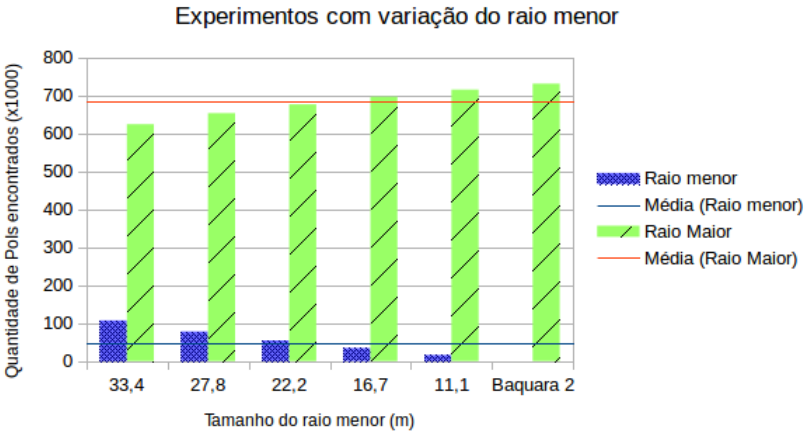

Figura 12 – Quantidade de ligações efetuadas com raio maior == 111,2m

Como mostrado na tabela 4, o tamanho do raio menor possui influência na quantidade de ligações com PoIs únicos dentro do raio. Por serem ligações feitas com um menor tempo de execução em relação às ligações comuns, o tempo de execução total se torna inversamente proporcional ao tamanho do raio menor. Dividindo este tempo em tempo de junção espacial e tempo de algoritmo, os números mostram uma maior importância do tempo de query e essa importância é proporcional ao tamanho do raio menor. A tabela 6, juntamente com a figura 13 mostram o tempo gasto com cada etapa em cada experimento.

	Tempo médio		
Raio menor (m)	**Junção Espacial (s)**	**Desambiguação s()**	**Total (s)**
33,4	96,55 (59,46%)	65,82 (40,54%)	162,37
27,8	116,76 (62,23%)	70,85 (37,77%)	187,61
22,2	116,72 (59,38%)	79,83 (40,62%)	196,55
16,7	125,3 (60,84%)	80,7 (39,16%)	206
11,1	126,49 (58,93%)	88,15 (41,07%)	214,64
Baquara 2	105,94 (54,45%)	88,63 (45,55%)	194,57

Tabela 6 – Tempo médio de execução dos experimentos

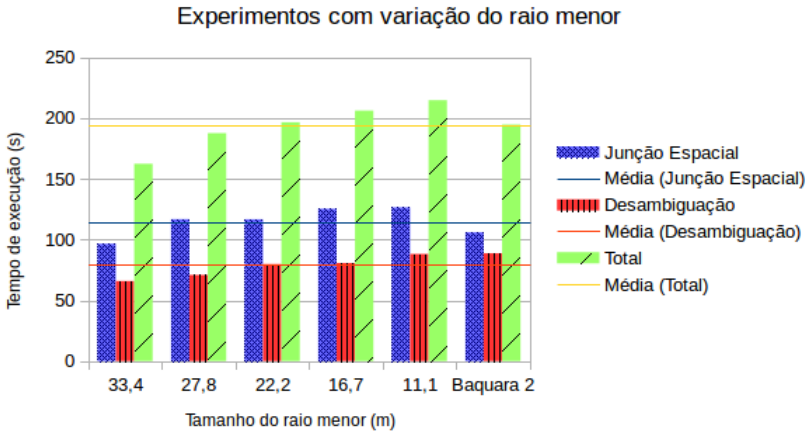

Figura 13 – Tempo de execução da proposta vs Baquara 2

No gráfico acima é possível perceber um menor tempo de junção espacial no algoritmo Baquara 2, isto se deve ao fato de que ele possui apena um único raio, ou seja, o algoritmo verifica apenas uma vez a localização do PoI. Esse tempo economizado passa a ser menor que o tempo economizado através de ligações com um único PoI quando o raio maior passa a ser de 33,4m.

5.2.2 Variação na quantidade de postagens

Esta bateria de experimentos foi composta por 8 experimentos, todos com o parâmetro passado para o raio menor no valor de 0,0003dd (33,4 m) e para o raio maior o valor de 0,001dd (111,2m). Os experimentos utilizaram parcialmente os dados de postagens e totalmente os dados de PoIs descritos na subseção 5.1. A tabela 7 mostra a entrada (Postagens) e a saída (Ligações) de cada experimento.

Experimento	Postagens (x1000)	Ligações (x1000)
1	80	27226
2	140	47730
3	200	68163
4	260	93658
5	320	107832
6	380	129648
7	440	150158
8	500	170194

Tabela 7 – Experimentos com variação na quantidade de postagens

Estes experimentos possuem como objetivo capturar o tempo de execução total e de cada etapa do algoritmo em cada situação de entrada do algoritmo. Cada experimento na tabela 8 foi executado 10 vezes para a retirada da média de cada tempo.

Tabela 8 – Tempo médio de execução

	Tempo médio		
Exeperimento	Junção Espacial (s)	Desambiguação (s)	Total (s)
1	58,65 (84,75%)	10,55 (15,25%)	69,2
2	70,20 (79,22%)	18,41 (20,78%)	88,61
3	79,17 (73,68%)	28,27 (26,32%)	107,44
4	87,63 (71,01%)	35,77 (28,99%)	123,40
5	95,99 (69,27%)	42,58 (30,73%)	138,57
6	96,51 (65,59%)	50,64 (34,41%)	147,15
7	96,37 (62,63%)	57,49 (37,37%)	153,86
8	96,54 (59,46%)	65,82 (40,54%)	162,36

Como é possível ver na tabela 8, o tempo de junção espacial representa uma importância maior em relação ao tempo total em experimentos com quantidade menor de postagens. Experimentos com uma grande quantidade de postagens, a fase do algoritmo propriamente dito se torna um grande fator no tempo total, enquanto os tempos de junção espacial de cada experimento tendem a permanecer próximos a partir de 320 mil postagens. A figura 14 ilustra esse comportamento.

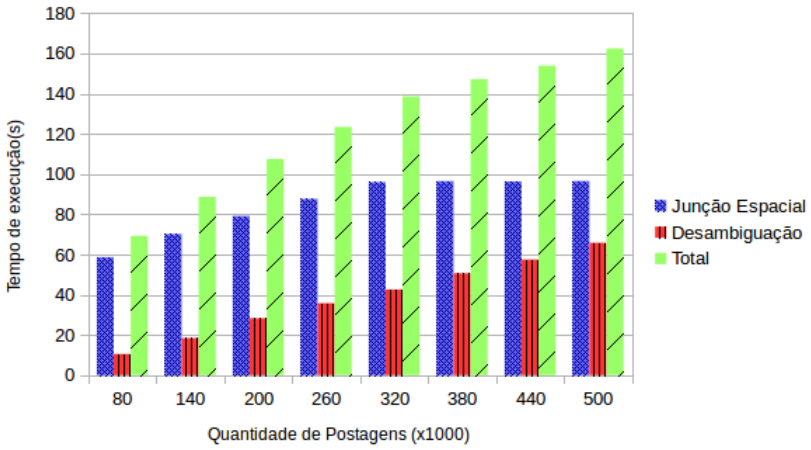

Figura 14 – Escalabilidade da proposta

Como visto na subseção 5.2.1, o algoritmo proposto por este trabalho se mostrou mais eficiente que o algoritmo proposto em (FILETO et al., 2015a) quanto ao tempo de execução, com uma economia através do uso de proximidade espacial. Observando as tabelas 6 e 8, é possível projetar que a porcentagem de ganho, que já era de 30% em favor do algoritmo deste trabalho, se torne maior com quantidades maiores de dados.

5.2.3 Amostra com regra ouro

A partir do experimento feito com a totalidade de PoIs e postagens, com o raio menor de 0,0003 foi retirado uma amostra de 2110 ligações. Nesta amostra foi comparado manualmente as ligações buscando decidir se o nome do PoI e a entidade encontrada no texto da postagem fazem referência ao mesmo lugar. A tabela 9, e as figuras 15 e 16 mostram estatísticas dessa amostra.

Tabela 9 – Dados da amostra

Dados	**Quantidade**
Total de postagens ligados	2110
Ligações aceitas no raio menor	592 (28,06%)
Ligações aceitas no raio maior	1518 (71,94%)
Postagens ligadas somente por proximidade espacial	1135 (53,79%)
PoIs encontrados ao total no raio menor	998
PoIs encontrados ao total no raio maior	7152
Número máximo de PoIS em um único raio	67
Distância média entre PoI e postagem ligada	62,24m
Tempo de junção espacial	125,492
Tempo do algoritmo	0,947s (0,75%)
Tempo de execução total	126,439s (99,25%)

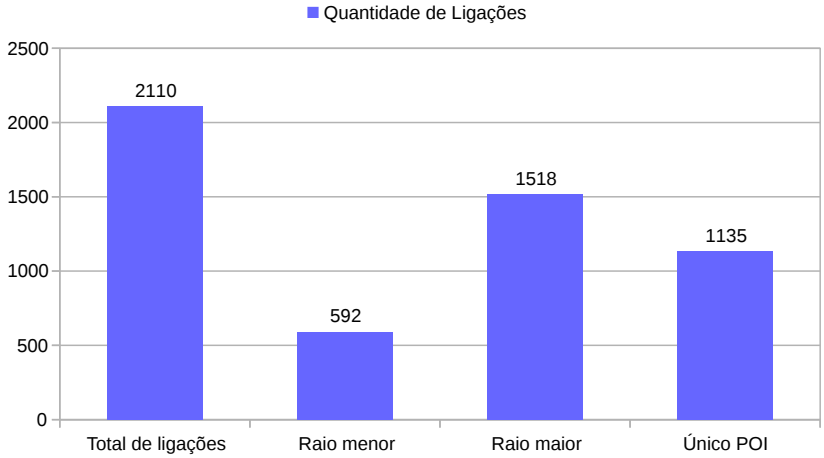

Figura 15 – Caracterização das ligações efetuadas

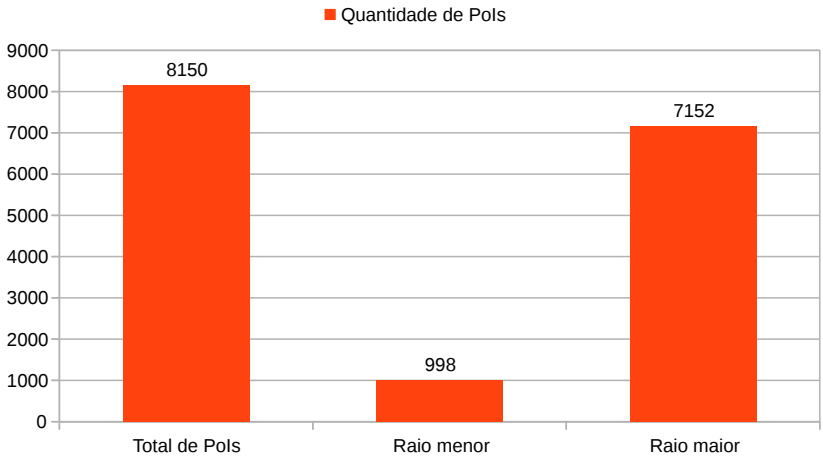

Figura 16 – Caracterização dos PoIs ligados

O material analisado foi classificado em 4 categorias distintas:

- Acertos: corresponde às ligações onde o conteúdo textual da postagem única faz referência ao nome do PoI.
- Erros: corresponde às ligações onde o conteúdo textual da postagem única não faz referência ao nome do PoI e foi utilizado a comparação de similaridade textual.
- Único PoI: O conteúdo textual da postagem única não referencia o nome do PoI, porém não foi utilizado a comparação de similaridade. Os dados dessa categoria não são considerados erros pois é um comportamento previsto pelo algoritmo.
- Conteúdo textual repetido: O conteúdo textual da postagem é repetido e desconsiderado.

A tabela 10 e a figura 17 representam essas classes quanto ao número de ligações.

Tabela 10 – Classes da amostra

Dados	Quantidade
Total de ligações	2110
Ligações corretas	208
Ligações erradas	451
Ligações com único PoI	242
Ligações com conteúdo textual repetido	1209

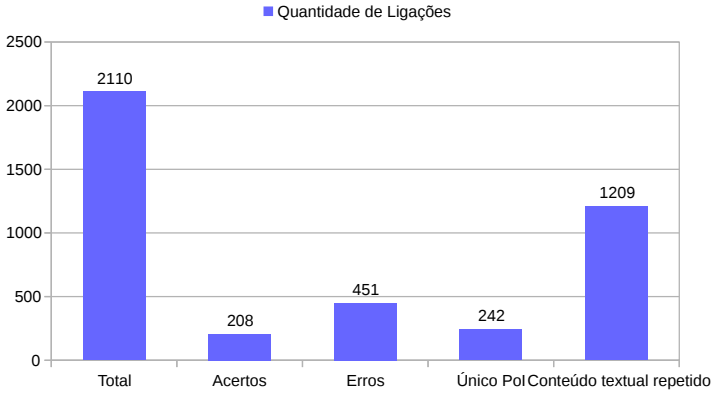

Figura 17 – Acurácia das ligações efetuadas (raio menor = 33,4m,raio maior 111,2m)

Para obter as estatísticas desconsideramos a quarta categoria, visto que os textos são repetidos e analisamos cada categoria distintamente. Esta categoria corresponde à cerca de 57% da amostra, uma verificação mais profunda nestas postagens mostrou que eram postagens diferentes e o texto repetido é devido ao padrão do FourSquare que não possibilita ao usuário editar a menção ao local.

O algoritmo obteve como resultado correto cerca de 25% das ligações não repetidas, e essa quantidade equivale a aproximadamente 10% da amostra total. Analisando os acertos e comparando os dois raios, percebemos que os números entre os dois raios são similares, e que em um pouco menos da metade dos acertos não foi necessário a comparação de similaridade. A tabela 11 e a figura 18 mostram as

estatísticas dos acertos, dividindo eles de acordo com o raio em que o PoI se encontra, e a quantidade de vizinhos do mesmo.

Tabela 11 – Acertos

Tipo da Ligação	Quantidade
Total	208
No raio menor	51
No raio menor e único PoI	54
No raio maior	61
No raio maior e único PoI	42

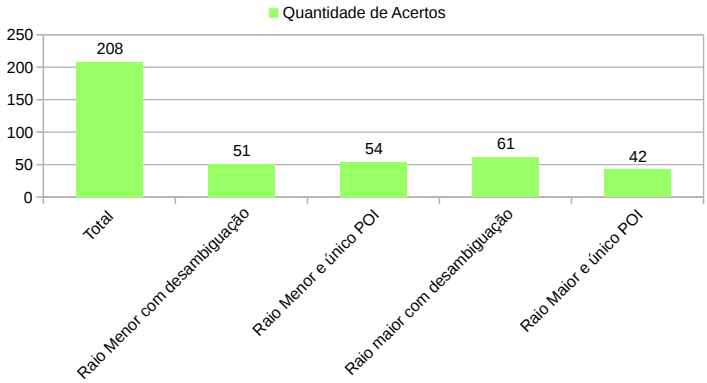

Figura 18 – Quantidades de acertos (raio menor = 33,4m, raio maior = 111,2m

O estudo mostrou que a quantidade de erros equivale a 50% das potagens não repetidas (25% da amostra total), e uma análise maior sobre esses dados mostra o fator desses erros acontecerem. Os erros podem ser classificados como:

- Causados por erro de comparação de similaridade textual.
- Causados devido ao PoI correto estar em outro raio.
- Causados pela falta do PoI correto.

O primeiro tipo de erro ocorre devido ao erro no cálculo da distância Jaro entre a menção do texto e o nome do PoI. Este erro de cálculo acontece devido às características de um texto de mídias sociais, descritos anteriormente na seção 2.2. O segundo tipo é causado pela falha da precisão das coordenadas da postagem, onde a localização desta possui PoIs não visitados à uma distância dentro do raio menor e o PoI visitado se encontra além deste raio.

O último tipo de erro ocorre pela falha nos dados de entrada, e não estão relacionados ao algoritmo em si. Apesar da grande quantidade de PoIs utilizado neste trabalho, a base de dados não possui todos os PoIs existentes neste planeta, além disso, muitas destas postagens possuem menções a entidades locais e portanto desconhecidas para o banco de dados. A tabela 12 e o gráfico da figura 19 mostram os número obtidos, dividindo também as ligações de acordo com o raio em que o PoI se encontra e a quantidade de vizinhos do mesmo.

Tabela 12 – Erros

Tipo da Ligação	**Quantidade**
Total	451
No raio menor	78
No raio menor e único PoI	75
No raio maior	260
Devido ao nome	13
Devido ao raio	25

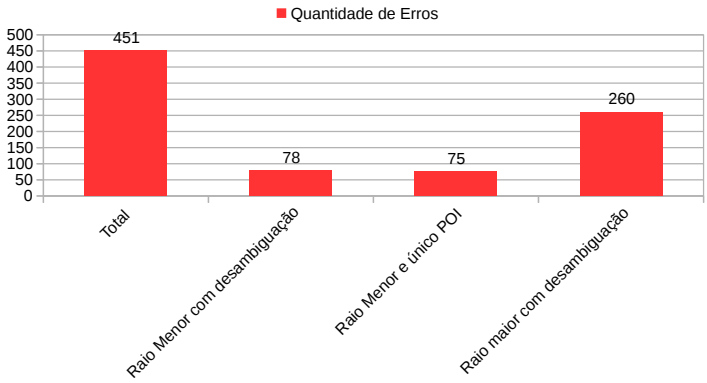

Figura 19 – Ligações erradas (raio menor = 33,4m, raio maior = 111,2m)

Os erros causados pelo algoritmo são compostos pelas duas primeiras classes, e representam cerca de 8,5% dos erros (5,54% devido ao raio 5,54 e 2,88% devido ao nome). Quando olhamos esse número na quantidade total de postagens não repetidas, o valor passa a ser de pouco mais de 4%(2,77% devido ao raio 5,54 e 1,44% devido ao nome), mostrando a eficácea do algoritmo. Esses números são mostrados pelas figuras 20 e 21.

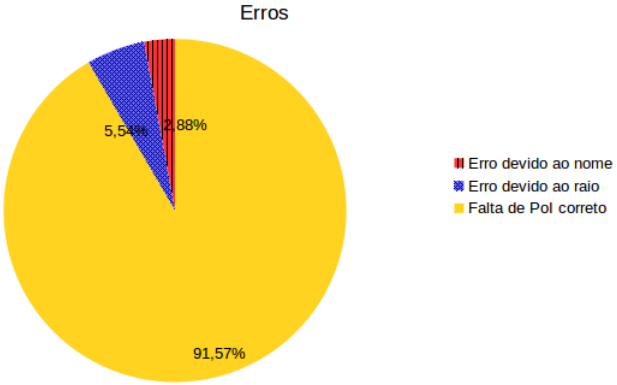

Figura 20 – Porcentagem de ligações erradas para cada causa (raio menor = 33,4m, raio maior = 111,2m)

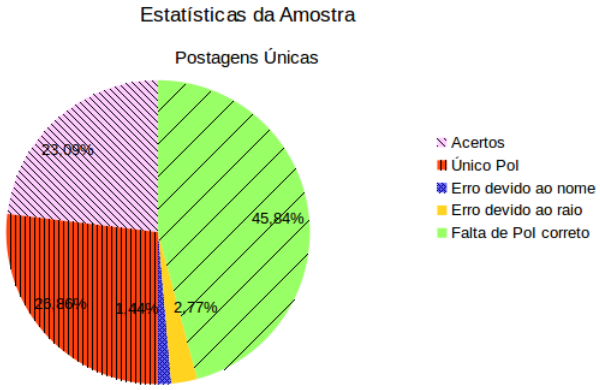

Figura 21 – Porcentagem de Ligações erradas para cada causa em relação ao total de postagens únicas (raio menor = 33,4m, raio maior = 111,2m)

6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho propôs um algoritmo capaz de realizar a ligação de dados espaciais com postagens de mídias sociais geo-referenciadas e realizou a implementação de um protótipo do algoritmo. Tal protótipo foi analisado quanto a variação dos dados de entrada (quantidade de postagens), variação do parâmetro passado (raio menor) e quanto ao seu grau de acerto (amostra).

Os experimentos mostraram que o algoritmo possui uma baixa taxa de erro (4% das ligações não repetidas) e um tempo de execução viável em muitas aplicações práticas. O algoritmo se mostrou mais rápido que o apresentado no artigo do baquara 2, com a tendência de maiores ganhos quando usado em bases maiores.

Para trabalhos futuros a mudança de bases de dados se torna bastante importante. Experimentos mostraram que a base de dados de postagens que seguiam o padrão de texto do Foursquare possuíam uma grande quantidade de postagens com conteúdo textual repetido. Entretanto, as mesmas possuíam IDs diferentes, o que leva a crer que as referências contidas nas postagens não foram escritas pelos usuários, e sim pelo próprio aplicativo Foursquare.

Além desta sugestão, segue uma lista para trabalhos futuros:

- Utilização de outras bases de postagens e PoIs (linkedGeoData, base de PoIs do Google).
- Utilização de outras técnicas de comparação de similaridade, extração de características do texto (e.g., entidades nomeadas, palavras relevantes).
- Validação do algoritmo expandido com vários tipos de dados e padrões de texto.
- Utilização do algoritmo com outras ferramentas (e.g. outros bancos de dados).

DIREITOS AUTORIAIS

O autor é o único responsável pelo conteúdo do material impresso incluído no trabalho.

REFERĘNCIAS

ABNT. Nbr 9241-11: Requisitos ergonômicos para trabalho de escritórios com computadores. In: *Parte 11 – Orientações sobre Usabilidade*. Rio de Janeiro, Brazil: [s.n.], 2002.

AGICHTEIN, E.; CASTILLO, C.; DONATO, D. Finding high-quality content in social media. 2008.

CAMBOIM, P. *ARQUITETURA PARA INTEGRAÇÃO DE DADOS INTERLIGADOS ABERTOS À INDE-BR*. 2013. <http://bdtd.ibict.br/vufind/Record/UFPR_b305290996977f7d37f11-07442f53199/Details>. Acessado em 23/11/2016.

CODESCU, M. et al. Ontology-based route planning for openstreetmap. 2014. <https://www.researchgate.net/publication/265203560_OSMonto_-An_Ontology_of_OpenStreetMap_Tags>. Acessado em 16/05/2017.

FILETO, R. et al. Semantic enrichment and analysis of movement data: Probably it is just starting! *Sigspatial*, 2015.

FILETO, R. et al. The baquara 2 knowledge-based framework for semantic enrichment and analysis of movement data. *Data & Knowledge Engineering*, v. 98, p. 104–122, 2015.

FURLETTI, B. et al. Inferring human activities from gps tracks. *UrbComp '13 Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, 2013.

GAO, J.; CAO, B.; FAN, H. Point of interest data storage using ontology. *3rd International Conference on Systems and Informatics*, p. 1122–1126, 2016.

HABIB, M. B.; KEULEN, M. van. Information extraction for social media. 2014.

IBM. *IBM What is big data? - Bringing big data to the enterprise*. 2012. <http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html>. Acessado em 22/11/2016.

JORENTE, V.; OLIVEIRA, B. e. Information design and information science: A possible approach. *XVI ENANCIB Informação, Memória e Patrimônio: do documento as redes*, 2015.

KIM, Y.; PEREIRA, F. C.; ZHAO, F. Activity recognition for a smartphone based travel survey based on cross-user history data. *Pattern Recognition (ICPR), 22nd International Conference on*, 2014.

LóSCIO, B. *Dados, Integração de Dados e Dados Interligados*. 2012. <http://dados.gov.br/wp/wpcontent/uploads/2012/10/workshop_BSB_02.pdf>. Acessado em 3/12/2016.

MEIJ, E.; WEERKAMP, W.; RIJKE, M. de. Adding semantics to microblog posts. 2012.

MORSTATTER, F.; GAO, H.; LIU, H. Discovering location information in social media. 2015.

SANG, T. K.; F, E.; MEULDER, F. D. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, v. 4. Association for Computational Linguistics, p. 142–147, 2003.

UFSC. *Coordenação de Projetos*. Departamento de Informática e Estatística - INE. <http://projetos.inf.ufsc.br>. Acessado em 20/11/2016.

UFSC. *O Departamento*. Departamento de Informática e Estatística - INE. <http://ine.ufsc.br/o-departamento/>. Acessado em 20/11/2016.

UFSC. *Regimento interno para elaboração de Trabalhos de Conclusão de Curso (TCC) de Ciências da Computação*. Departamento de Informática e Estatística - INE. <*https* : *//projetos.inf.ufsc.br/arquivos/RI* − *TCC* − *CCO*$_v$*23maro.doc*>. Acessado em 20/11/2016.

UFSC. *Sistema – TCC – UFSC*. Departamento de Informática e Estatística - INE. <http://tcc.inf.ufsc.br>. Acessado em 20/11/2016.

WU, F. et al. Semantic annotation of mobility data using social media. *WWW '15 Proceedings of the 24th International Conference on World Wide Web*, p. 1253–1263, 2015.