CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
CAMPUS TIMÓTEO

Mariana Araújo Tavares

SISTEMA DE APOIO A DECISÃO E RECUPERAÇÃO DE PONTOS COMERCIAIS: UM ESTUDO DE ALGORITMOS CLASSIFICADORES NO CONTEXTO DE INFORMAÇÃO GEOGRÁFICA VOLUNTÁRIA

Timóteo

2015

Mariana Araújo Tavares

SISTEMA DE APOIO A DECISÃO E RECUPERAÇÃO DE PONTOS COMERCIAIS: UM ESTUDO DE ALGORITMOS CLASSIFICADORES NO CONTEXTO DE INFORMAÇÃO GEOGRÁFICA VOLUNTÁRIA

Monografia apresentada à Coordenação de Engenharia de Computação do Campus Timóteo do Centro Federal de Educação Tecnológica de Minas Gerais para obtenção do grau de Bacharel em Engenharia de Computação.

Orientador: Leonardo Lacerda Alves

Timóteo

2015

Mariana Araújo Tavares

SISTEMA DE APOIO À DECISÃO E RECUPERAÇÃO DE PONTOS COMERCIAIS

Trabalho aprovado. Timóteo, 21 de agosto de 2017:

Leonardo Lacerda Alves
Orientador

Maurílio Alves Martins da Costa
Avaliador

Odilon Corrêa da Silva
Avaliador

Timóteo
2017

.

.

.

Resumo

Dados geográficos são usados para diversas áreas, porém produzir dados geográfico é caro. Uma alternativa é utilizar informações geográficas voluntárias (VGI) como base de dados.No entanto,por ser produzida por usuários não treinados,a informação geográfica voluntária pode apresentar imprecisão e duplo sentido além de outros fatores que podém dificultar sua classificação e é isso que motiva o estudo a ser apresentado. Foram estudados algoritmos classificadores com o objetivo de implementar pelo menos um para realizar a classificação de informações geográficas voluntárias. Neste trabalho será apresentado todo estudo feito para a realização do mesmo, bem como o processo de implementação do algoritmo baseado em Kmeans e os resultados ao utilizá - lo na classificação de VGIs.

Palavras-chave: kmeans, informação geográfica voluntária,Busca por similaridade.

Abstract

Geographic data is used for several areas, however producing geographic data is expensive. An alternative is to use voluntary geographic information (VGI) as a database. However, because it is produced by untrained users, voluntary geographic information can present imprecision and double meaning in addition to other factors that may hinder its classification and this is what motivates The study to be presented. We have studied classification algorithms with the objective of implementing at least one to perform the classification of voluntary geographic information. In this work will be presented all the study done to the same, as well as the process of implementation of the algorithm based on Kmeans and the results when using it in the classification of VGIs.

Keywords: Kmeans, voluntary geographic information, Search by similarity.

Lista de tabelas

Sumário

1 Introdução

Produzir informação geográfica é como representar computacionalmente o mundo real em termos de posicionamento com relação a um sistema de coordenadas, seus atributos não aparentes e das relações topológicas existentes, segundo (CâMARA, 1998) . Dados geográficos são uteis em diversas áreas como transportes, planejamento urbano, gestão, saúde pública, controle de desastres naturais, logística, marketing, entre outros.

No entanto, produzir esses dados é caro e consome tempo. Por isso é interessante permitir que a grande massa de pessoas produza esses dados através de informações voluntárias. Qualquer pessoa que possua conexão com a internet pode funcionar como um tipo de produtor de informação geográfica voluntária (VGI, de *Volunteered geographic information*).

Informações geográficas voluntárias apresentam como característica a imprecisão (HAKLAY; SINGLETON; PARKER, 2008), e por isso são difíceis de serem classificadas, o que sintetiza a motivação desta pesquisa.

1.1 Justificativa

Esta pesquisa justifica-se pela necessidade de implementar sistemas de recuperação de informações que deem suporte a sistemas de apoio a decisão no contexto de geomarketing, utilizando como base informações geográficas voluntárias (VGI).

Existem sistemas de apoio a decisão que utilizam informações geográficas recolhidas de bases governamentais que muitas vezes não atendem alguns requisitos necessários, ou utilizam informações recolhidas de grandes empresas, que geralmente são restritas a áreas metropolitanas que possuam um grande valor comercial, deixando de fora cidades de menor porte.

Nesse cenário é interessante dispor de alternativas. Por exemplo, a utilização de informações geográficas voluntárias, que já estão disponíveis na internet. O surgimento de aplicativos e redes sociais, que permitem que uma pessoa acrescente dados sobre o espaço em que vive, possibilitou a localização de informações importantes (BELL; COPE; CATT, 2007). Portanto, milhões de usuários não são mais vistos apenas como consumidores, mas como colaboradores e criadores de informação (BUDHATHOKI; BRUCE; NEDOVIC-BUDIC, 2008).

1.2 Problema

A seguinte questão constitui o problema dessa pesquisa: As classes de algoritmos usados em sistemas de apoio a decisão são eficientes para classificar informação geográfica voluntária?

As informações geográficas voluntárias podem ser imprecisas por serem geradas por usuários não treinados, e por isso podem apresentar erros ao serem classificadas. Também é preciso considerar diferenças entre informações que possuam a mesma sintaxe mas não a mesma semântica. Por exemplo, a palavra Consul, que pode representar um hipermercado na região do Vale do Aço e ao mesmo tempo pode representar apenas uma marca de produto.

1.3 Objetivos

Para responder ao problema proposto, o objetivo geral desse trabalho é verificar a eficiência de pelo menos uma classe de algoritmos classificadores, ao classificar um banco de dados de informações geográficas voluntárias recolhidas de redes sociais.

Mais especificamente objetivam-se:

- Definir métodos de refinamento e recolher informações voluntárias de uma rede social através de uma API;
- Implementar um algoritmo para classificar o banco de dados construído através do recolhimento das informações geográficas voluntárias;

1.4 Estrutura

Este trabalho está estruturado em seis capítulos, ordenados pelo momento em que foram concluídos dentro do ciclo de vida desta pesquisa:

- O capítulo 2 apresenta os procedimentos metodológicos através dos quais este trabalho se desenvolveu, o que inclui a formação das coleções de documentos para experimentação e avaliação, as etapas para o projeto do protótipo funcional, bem como as estratégias de validação.
- As bases teóricas são apresentadas em seguida, no capítulo 3, o que inclui marcos conceituais importantes para o contexto de algoritmos de busca por similaridade e para informações geográficas voluntárias.
- Os experimentos são descritos no capítulo 4, com detalhes sobre a coleta de dados e sobre o pré-processamento.
- Os resultados são analisados no capítulo 5.
- Finalmente, as considerações finais, limitações e trabalhos futuros são apresentados no capítulo 6.

2 Procedimentos metodológicos

Os procedimentos metodológicos são organizados em, Revisão de Literatura, Coleta de dados, Implementação do algoritmo e Avaliação dos Resultados.

2.1 Revisão da literatura

Os trabalhos coletados e estudados, apresentam estudos importantes sobre o sistema de apoio a decisão. Foram considerados trabalhos do ano de 1984 (Onde pode-se encontrar definições relacionadas ao desenvolvimento do projeto), a trabalhos publicados em 2015. Os trabalhos foram coletados no repositório ACM digital library e biblioteca digital da USP. Os termos usados para pesquisa foram, KNN , KNN reverso, R-tree, bicromático RKNN, e informação geográfica voluntária.

Portanto, a revisão de literatura apresentada no capítulo 3 resultou em informações sobre os trabalhos realizados sobre a informação geografica voluntária e também em informações sobre os algoritmos usados para classificar informações em um sistema de apoio a decisão, além disso apresenta conceitos importantes para compreender esses estudos.

2.2 Coleta de dados

Em seguida, foi reunida uma massa de dados que continha informações sobre endereços de estabelecimentos comerciais e clientes, e também características que ressaltavam a influência desses clientes aos estabelecimentos relacionados. Por exemplo, se um cliente constantemente faz uma postagem que está voltando do trabalho sempre em uma mesma ciclovia, esse cliente talvez fosse um cliente em potencial para uma padaria, ou seja, ele estaria voltando do trabalho e poderia fazer uma parada para comprar pães. As informações coletadas seriam informações concedidas voluntariamente por usuários que utilizam redes sociais e internet em geral.

2.3 Implementação dos algoritmos

Foi projetado e implementado um algoritmo classificador. O algoritmo foi avaliado por meio de testes de desempenho utilizando ampla variedade de conjuntos de dados. O algoritmo implentado foi baseado no algoritmo Kmeans, pois observou -se que o mesmo foi amplamente utilizado para analisar a similaridade entre palvras e documentos.

2.4 Avaliação dos resultados

Através dos resultados obtidos a partir do algoritmo de classificação, foram analisados os resultados para determinar se o algoritmo de classificação apresenta resultados satisfatórios para o objetivo deste trabalho.

3 Fundamentos históricos, teóricos e metodológicos

O principal objetivo desse capítulo é apresentar a revisão de literatura empregada neste trabalho, onde serão apresentados estudos já realizados anteriormente sobre temas relacionados a informação geográfica voluntária e algoritmos de consulta por similaridade. Também serão apresentados conceitos importantes para a compreensão dos mesmos.

3.1 Sistemas de informação geográfica voluntária

Os sistemas de informação geográfica voluntária (VGI) baseiam-se na criação de informações por meios voluntários. Goodchild (2007) usa para explicar esse conceito a expressão de que os cidadãos são como sensores , em um momento em que o uso intenso da internet e de dispositivos moveis são um fato, as pessoas podem compartilhar informação em qualquer lugar e a qualquer momento.

Sistemas VGI já adquiriram uma quantidade gigantesca de informação geo-referenciada. Um exemplo é que em maio de 2014 a Wikimapia continha mais de 23 milhões de objetos marcados por usuários registrados e hóspedes (RONZHIN, 2015). Mais de 4,5 milhões de artigos na Wikipédia tiveram geotags em agosto 2014(RONZHIN, 2015). Contudo, estes números são quase nada em comparação com o projeto OpenStreetMap, onde cerca de 1,7 milhões de usuários criaram mais de 2,5 bilhões de dados a partir de outubro 2014 (RONZHIN, 2015).

No entanto, apesar do fato dessa iniciativa gerar grande quantidade de informação, a qualidade e usabilidade do conteúdo é um assunto de debate. Uma observação a ser considerada é que a cobertura de dados não é completa ou consistente em todo o mundo. Áreas com maior densidade populacional recebem mais atenção dos usuários do que áreas com menor densidade (ZOOK et al., 2010). Em seus estudos, Leeuw et al. (2011) concluíram que as informações são mais precisamente classificadas por pessoas que possuem conhecimento do local do que por pessoas sem esse conhecimento.

Logo, quanto maior o número de pessoas vivendo em um determinado local, maior e, talvez, melhor, será o número de informações relacionadas a esse local (FLANAGIN; METZGER, 2008). Tuan (1975) afirma que isso é ocasionado pelo interesse natural que as pessoas têm em descrever o espaço conhecido por elas, o que pode justificar a observação feita por Zook et al. (2010). Castells (2003) faz importantes observações onde indica que há uma maior concentração de usuários em determinadas porções do espaço geográfico (DRAHOS, 1995). Além disso, em muitos casos, as pessoas coletam informações sem qualquer orientação ou instruções. Como resultado, a precisão desses dados é muitas vezes desconhecido, já que

não existem processos de garantia de qualidade para o recolhimento de dados (HAKLAY; SINGLETON; PARKER, 2008).

Outro fator é que pessoas diferentes podem categorizar o mesmo fenômeno de forma diferente. Dados geográficos voluntários frequentemente são captados de forma desestruturada, e em formatos diferentes e de confiabilidade duvidosa, o que torna a integração desses conjuntos de dados longe de ser trivial Ronzhin (2015) .

Para sanar a falta de precisão e consistência da informação geográfica voluntaria, Ronzhin (2015) propõe enriquecer semanticamente essas informações utilizando a núvem Linked Open Data (LOD) . Os dois primeiros passos do estudo incluem uma conversão dos dados para o Resource Description Framework (RDF), utilizando vocabulários e estabelecimento de ligações semânticas com entidades LOD relevantes. Para a realização de consulta foi utilizado a linguagem SPARQL, que é uma linguagem de consulta RDF, capaz de recuperar e manipular dados armazenados em Resource Description Framework. O modelo de consulta utilizado por Ronzhin (2015) seleciona todos os objetos localizados numa determinada área e em seguida, ordena os resultados com base na distância entre a geometria de um dado objeto.

O que pode ser notado é que tal consulta é um exemplo de uma consulta k-vizinhos mais próximo (kNN), porém SPARQL normalmente não suporta nativamente consultas KNN espaciais (PATROUMPAS; GIANNOPOULOS; ATHANASIOU, 2014). A abordagem cria uma carga computacional adicional, que poderia ser evitada se fosse utilizado um algoritmo kNN. Por exemplo, o software proprietário de vários produtos como o Oracle Spatial e gráficos têm uma pesquisa explícita kNN. (RONZHIN, 2015).

Diante disso, a próxima seção apresenta estudos sobre algoritmos como o KNN e outros pertencentes a mesma classe.

3.2 Consulta por Similaridade

A consulta por similaridade se aplica a vários problemas, entre eles o de apoio a decisão. A avaliação de similaridade é feita através de funções que medem a dissimilaridade, ou seja, a distância entre os objetos consultados. Essas funções ou algoritmos, recebem dois objetos de um mesmo domínio e retornam o valor da distância existente entre esses objetos.

Dentre os tipos de consulta por similaridade pode-se destacar a consulta por abrangência, a consulta aos k-vizinhos mais próximos e a consulta aos k-vizinhos mais próximos reversos (KORN; MUTHUKRISHNAN, 2000). Esses três tipos de consulta estão organizados nas próximas subsessões.

3.2.1 Consulta por abrangência

A consulta por abrangência Rq(Sq ,E) , recebe um valor Sq como parâmetro, que é o elemento central da consulta e também recebe um grau de dissimilaridade E. O algoritmo recupera todos os elementos da base de dados que diferem do elemento Sq por no máximo a dissimilaridade de E. (KORN; MUTHUKRISHNAN, 2000).

“Um exemplo de uma consulta por abrangência em uma base de dados de sequências genéticas é ‘Selecione as sequências de DNA que difiram desta sequência Sq dada por até 5 bases purínicas’ ,representada como Rq(Sq,5).” (OLIVEIRA, 2010).

3.2.2 Consulta aos k-vizinhos mais próximos

A consulta aos k-vizinhos mais próximos kNN (Sq, k) recebe como parâmetro o elemento Sq que representa o elemento central da busca e também o elemento k , que define a quantidade de elementos ao qual o elemento central Sq será comparado. Essa busca retorna os k elementos mais similares ao elemento Sq.(KORN; MUTHUKRISHNAN, 2000).

“Um exemplo de uma consulta aos k-vizinhos mais próximos em uma base de dados de sequências genéticas é ‘Selecione as 3 proteínas mais semelhantes a essa proteína Sq dada’,representada como knn(Sq,3).” (OLIVEIRA, 2010).

3.2.3 Consulta aos k-vizinhos mais próximos reversos

Segundo Korn e Muthukrishnan (2000) a consulta aos k-vizinhos mais próximos reversos RkNN (Sq, k) recebe como parâmetro o elemento Sq que representa o elemento central da busca e também o elemento k, que define a quantidade de elementos ao qual o elemento central Sq será comparado. Essa busca retorna os k elementos que possuam o elemento central Sq como um de seus k vizinhos mais próximos .

“Um exemplo de uma consulta aos k-vizinhos mais próximos reversos em uma base de dados de sequências genéticas é ‘Selecione as proteínas que tem esta proteína Sq ,dada como uma de suas 3 mais semelhantes.’, representada como Rknn(Sq,3).” (OLIVEIRA, 2010)

O RkNN, será abordado com mais detalhes na próxima sessão.

3.3 Conceitos sobre a Consulta Rknn

Korn e Muthukrishnan (2000) descrevem que dado um ponto q para consulta, os vizinhos mais próximos de q podem diferir substancialmente de todos os conjuntos de pontos para os quais q é um vizinho mais próximo. Esses pontos são chamados de vizinhos mais próximos reversos de q. Dado um elemento Sq a consulta recupera todos os elementos em volta de Sq, que possui Sq como seu vizinho mais influente.

3.3.1 Variantes

Korn e Muthukrishnan (2000) citam sobre algumas variantes que podem definir o modelo da consulta RkNN. Duas delas são o Modelo Bicromático e Modelo Monocromático. No modelo Bicromático existem pontos pertencente a categorias diferentes, por exemplo cliente e servidor os pontos podem, portanto, ser considerado como sendo de cores diferentes, vermelha ou azul. A consulta RNN agora consiste de um ponto em uma das categorias, a consulta por um ponto azul deve determinar os pontos vermelhos para o qual o ponto de consulta é o ponto mais próximo desse ponto azul. Formalmente, B denota o conjunto de pontos azuis e R

o conjunto de pontos vermelhos. Considerando um ponto azul como a consulta q. Já na versão monocromática todos pontos são da mesma categoria. Outras duas são o Modelo Estático e Dinâmico. O Modelo dinâmico é quando se deseja inserir ou excluir pontos do conjunto S e ainda apoiar a consulta RNN. Já o modelo estático não é modificado. (KORN; MUTHUKRISHNAN, 2000).

3.4 Algoritmos de Classificação

Korn e Muthukrishnan (2000) consideram uma versão básica do problema, usando o conjunto de dado de S que não é atualizado. A distância considerada entre dois pontos é a distância euclidiana a abordagem envolve duas etapas. Para cada ponto p pertencente ao conjunto S determina-se a distância euclidiana para o vizinho reverso de p em S denotada N(p) , gera-se um círculo (p , N(p)) onde p é o centro e N(p) é o raio. Na segunda etapa, para uma busca q, determina-se todos os círculo (p,N(p)) que contem q e retorna-se o seus centros p. Nessa etapa todos os vizinhos reversos de q são determinados. Se o ponto p é retornado na segunda etapa tem-se que q está no círculo (p,N(p)) , por isso a distância d(p,q) é menos que o raio N(p) , em outras palavras d(p,q) N(p) então o é q vizinho mais próximo reverso de p (equivalentemente, p é o vizinho reverso de q).

Foram considerados por Korn e Muthukrishnan (2000) espaços multidimensionais e sua proposta consiste em criar uma estrutura de indexação que armazena , para todo elemento o próprio elemento e a área coberta por sua região de influencia , a proposta foi de se usar o triangulo envolvente mínimo (MBR) de cada área coberta em uma r-tree, e criou uma arvore que foi chamada de Rnn – tree. Usando essa arvore, o vizinho reverso de um elemento de consulta sq pode ser obtido de maneira eficiente através de uma consulta sobre sq. Ela retorna todas as áreas contendo sq. Por ser apoiada na r-tree esta proposta trata apenas da situação em que k=1 em um espaço euclidiano (KORN; MUTHUKRISHNAN, 2000).

Além disso, Korn e Muthukrishnan (2000) desenvolveram algoritmos e estruturas de dados para buscas Rnn para conjuntos de dados estáticos e dinâmicos. Para o caso estático utilizaram a arvore RNN-tree para a realização de consultas RNN eficientes e consulta aos vizinhos mais próximos. Para o caso dinâmico eles utilizaram uma r-tree adicional para consultas NN pois a Rnn-tree é otimizada para consultas Rnn e não para consultas NN causando sobreposições das regiões em operações de deleção.

O método proposto foi baseado no caso monocromático, porém Korn e Muthukrishnan (2000) descrevem que pode ser adaptado para o caso Bicromático. A proposta feita por Korn e Muthukrishnan (2000) pré indexam todos os dados da base de dados segundo os vizinhos mais próximo naquele momento. Um problema observado é a necessidade da pré computação além de existir o problema de que quando um elemento do conjunto é removido ou um novo elemento é inserido os dados sobre os vizinhos mais próximos tem que ser recalculados e a estrutura de indexação pode precisar ser modificada pois se um novo elemento é inserido todos os elementos existentes que passam a tê-lo como elemento mais próximo tem que ser atualizados . Isso corresponde a atualizar os MBR de todos os elementos que originalmente cobrem o novo elemento. Se um elemento é removido toda estrutura deve ser refeita. Dessa

maneira essa proposta não é adequada para situações em que exista grande quantidade de inserções ou remoções.

Para sanar o primeiro problema observado, Yan e Lin (2001) propõem a integração do círculo NN em uma R-tree que indexa os próprios elementos de modo que a variante da R-tree resultante seja a RdNN-tree , ou seja , permite gravar informações sobre a distância do NN dos pontos em cada nó , o que evita uma manutenção extra na RNN-tree.

Em relação ao segundo problema, Satoni et all (2000) excluíram a necessidade da pré computação sugerindo que o espaço da busca fosse dividido em seis setores dessa forma a busca sessaria se o vizinho mais próximo do elemento da consulta não for o RNN do elemento da consulta. Porém a proposta feita não resolvia o problema do RNN Bicromático, então Satoni et all (2001) propuseram um algoritmo para tratar o caso. O algoritmo consiste em calcular os limites de uma região de influência dado um conjunto do tipo X armazenado em uma R-tree e um elemento de consulta Sq, depois utiliza a região de limite encontrada para recuperar os pontos do tipo B armazenados em outra R-tree, que estão dentro dessa região.

3.5 Agrupamento (Clustering)

Além dos algoritmos de classificação, existem os algoritmos de agrupamento que utilizam métricas de similaridade. O Agrupamento é um metodo que visa organizar grupos onde seus componentes são similares entre si. Na visão de K. Murty M. N. (1999) o agrupamento é a classificação não supervisionada de dados em grupos. K. Murty M. N. (1999) diz que o método de agrupamento possui amplo apelo e utilidade como uma das etapas na análise exploratória de dados. Uma de suas importantes aplicações é para a recuperação de informações.

Dias (2004) descreve o problema de clustering (agrupamento), onde dado um conjunto com *n* objetos *X = X1, X2, X3, ... Xn*, o problema de clusterização é a obtenção de um conjunto de *k* clusters *($k \leq n$)*, onde os objetos contidos em um cluster *Ci* possuam uma maior similaridade entre si do que com os *C = C1, C2,...,Ck*, objetos de qualquer um dos demais clusters do conjunto C. O algoritmo k-means é um algoritmo de Clustering. Os conceitos desse algoritmo serão apresentados na sub sessão seguinte.

3.5.1 K-means

Linden (2009) descreve que o algoritmo K-means busca minimizar a distância dos elementos a um conjunto de k centros dados por *X = X1, X2, X3, ... Xk* de forma iterativa. A distância entre um ponto pi e um conjunto de clusters, dada por *d(pi,)*, é definida como sendo a distância do ponto ao centro mais próximo dele. Esse algoritmo depende de um parâmetro k, que será definido pelo usuário k=número de clusters. Linden (2009) diz que a escolha do parâmetro k costuma ser um problema, tendo em vista que normalmente não se sabe quantos clusters existem a priori. O algoritmo do K-Means pode ser descrito de forma que:

1. Seja escolhido k distintos valores para centros dos grupos (possivelmente, de forma aleatória);

2. Associe cada ponto ao centro mais próximo;

3. Recalcule o centro de cada grupo;

4. Repita os passos 2-3 até nenhum elemento mudar de grupo.

3.6 Considerações Finais

Nesse capítulo foram discutidos conceitos sobre informações geográficas voluntárias, agrupamento, e classificação. Esses conceitos foram descritos com o objetivo de explicitar os métodos de consulta por similaridade mais comumente utilizados.

Nos estudos encontrados até o momento nenhum algoritmo de consulta por similaridade foi usado para classificar informações geográficas voluntárias (VGI).

Apenas o algoritmo Kmeans foi utilizado como objeto de estudo neste trabalho.

4 Descrição dos Experimentos

Neste capítulo é apresentado como foi realizado a coleta das informações geográficas voluntárias e o pré-processamento que tem como objetivo melhorar a recuperação dessas informações.

4.1 Coleta de Dados

Na etapa da coleta das Informações Geográfica Voluntárias foram coletadas postagens da rede social “Facebook” nos grupos de notícias da região do Vale do Aço: ’Portal Diário do Aço’, ’Shopping Vale do Aço’. Essas postagens relatam acontecimentos ocorridos nessa região. Através da coleta foi possível classificar as regiões conforme as características apresentadas e então definí-las como regiões em potencial para, por exemplo, a montagem de um estabelecimento comercial.

Antes que de fato fossem coletadas as postagens, foi necessária a realização de testes para compreensão de como seria feita a coleta. Utilizou -se uma ferramenta oferecida pelo próprio facebook chamada “Explorador da Graph API”, que se encontra disponível no Facebook. Essa ferramenta permite buscar dados que estão disponíveis no Facebook, como postagens, id, nome, dentre outros.

Para utilizar essa ferramenta foi necessário criar uma aplicação no Facebook for Developers, para que a mesma fornecesse uma chave. A chave é utilizada pela ferramenta “Explorador da Graph API”, para permitir as buscas nos perfis do Facebook.

Após testes utilizando a ferramenta, foi desenvolvido um código em PHP para que o mesmo retornasse as postagens, gravando–as em um banco de dados. A execução do arquivo ColetadeDados.php permite a busca das postagens em páginas do Facebook. Ele retorna um id, a página da qual foi retirado cada postagem e a própria postagem.

4.1.1 Descrição do programa ColetadeDados.php

O programa ColetadeDados.php foi desenvolvido na linguagem PHP. No código, é feita a abertura da conexão com o banco de dados e então feita a atribuição da chave que foi obtida, através da criação da aplicação no Facebook. Para fazer as requisições, a chave é validada, e inicia-se uma sessão.

Cada perfil do Facebook possui um id, ou seja, um código de identificação, e através da utilização desse código é definido de qual grupo do Facebook será realizada a coleta das postagens. Para capturar os dados, foi utilizado o método get(), onde são fornecidos o id do grupo e a lista dos campos desejados.

No código, foram utilizadas funções próprias da biblioteca Graph API.

Os dados retornados foram salvos na tabela tbl dadosipatinga. A tabela possui a seguinte estrutura:

ID – Que representa o identificador da postagem;

DS GRUPO – o nome do grupo da qual foi coletada a postagem;

FEED – Descrição da postagem.

4.2 Processamento dos Dados

De acordo com Barion (2008), o pré-processamento visa eliminar possíveis interferências na etapa de análise dos dados, como: stop list, sinonímia, polissemia e stem. Estes conceitos serão abordados na próxima subseção.

4.2.1 Conceitos

1. Stop List: Barion (2008) define que um sistema de mineração de dados associa uma stop list com um conjunto de documentos. Uma *stop list* é um conjunto de palavras definidas como *stop words* que são consideradas irrelevantes. Normalmente inclui artigos, preposições, conjunções e outras

   No exemplo:

   *"Ontem o **parque ipanema** estava **lotado** durante o **show**."*

   Para o seguinte estudo, as palavras relevantes seriam:

   - *Parque, Ipanema*, que representam uma localidade, ou seja, um referencial geográfico.
   - *lotado*, que representa aglomeração de indivíduos.
   - *show*, representando um evento ou acontecimento.

   As demais, seriam consideradas *stop words*.

2. Sinonímia: Pela definição de dicionário, sinonímia é uma particularidade das palavras que são sinônimas, ou seja, um conceito possui várias palavras.

   Por exemplo:

   Carro

   Que possui como sinônimos as palavras *automóvel, veículo*.

3. Polissemia: A polissemia é a definição dada para quando uma mesma palavra possui diferentes significados.

   Por exemplo:

   Exame

   A palavra acima pode representar mais de um sentido, pode ser um teste ou um procedimento médico.

4. Stem: Barion (2008) diz que o processo de Stemming é utilizado para remover todas as variações de palavra permanecendo somente a raiz. Na mineração de dados é preciso considerar palavras que compartilham um mesmo radical. É preciso identificar e analisar palavras que possuem pequenas variações sintáticas.

   Por exemplo:

   ***Droga**, **Droga**s, **Droga**do, **Droga**ria*.

   As palavras possuem o radical Droga em comum.

4.2.2 Contagem de palavras

As postagens foram agrupadas de acordo com a sua provável similaridade semântica ou seja, postagens pertencentes ao mesmo grupo serão parecidas semanticamente. Para que isso aconteça é necessário conhecer todas as palavras presentes na coleção de dados e com que frequência essas palavras aparecem.

Foi necessário, primeiramente, fazer uma contagem de palavras. Esse processo permitiu verificar quais eram as palavras mais frequentes na coleção de dados.

O programa ContaPalavras.php implementa a função descrita acima. Ele foi desenvolvido na linguagem PHP. Os dados da tabela tbl dadosipatinga foram agrupados em uma variável tratada para que não houvesse nenhum caracter especial. Dentro dessa variável foram contadas as ocorrências de cada palavra.

As palavras retornadas pela execução do programa ContaPalavras.php foram filtradas manualmente de forma que artigos definidos, artigos indefinidos, pronomes e palavras pouco relevantes não fizessem parte da coleção, visto que essas não são importantes para o objetivo desta pesquisa. Este grupo de palavras é a chamada Stop List.

As palavras foram gravadas em uma tabela do banco de dados tbl palavras todas. A tabela possui a seguinte estrutura:

ID – Que representa o identificador da palavra;

DS PALAVRA – descrição da palavra;

NR FREQUENCIA – frequência com a qual a palavra aparece na coleção.

4.2.3 Agrupamento de palavras

Em um sistema de agrupamento devem ser levados em consideração a sinonímia, a polissemia e o stem, e por isso foi feito o agrupamento manual, das palavras coletadas no processo de contagem de palavras.

Como citado anteriormente cada palavra possui uma frequência, ou seja, possui uma quantidade de vezes que aparece na coleção de dados. O grupo de palavras também possui uma frequência, que é formada pelo somatório das frequências das palavras pertencentes a esse grupo. Os grupos receberam o nome da palavra mais relevante pertencentes a ele.

4.2.4 Implementação TF-IDF

Antes da implementação do k means foi necessário implementar os valores tf–idf . O tf-idf é a frequência do termo inverso da frequência nos documentos. Esse cálculo tem o intuito de indicar a importância de uma palavra do documento em relação a uma coleção de documentos.

4.2.5 Algoritmo de Clustering KmeansM

O Algoritmo de Clusterização tem como objetivo reduzir a dimensionalidade dos dados coletados, ou seja quebrar a coleção de dados em grupos menores, onde esses dados agrupados apresentarão um certo grau de similaridade. Nesta subseção será apresentado como foi construído o algoritmo de clustering.

O algoritmo KmeansM.java foi implementado na linguagem java e é baseado no algoritmo de clustering K means ou K médias, sua ideia é fornecer uma classificação das informações de acordo com os dados. A Classificação do k-means é baseada em análise e comparações entre os valores numéricos dos dados.

No algoritmo desenvolvido, o usuário introduz o número de centroides desejado, e os grupos de palavras que serão verificados na coleção.

Antes do processo de clusterização pelo algoritmo kmeans, foram retiradas todas as amostras que não possuíam os grupos previamente escolhidos, visto que essas amostras não seriam relevantes no processo de clusterização e poderiam gerar interferências no mesmo.

Para calcular a distância entre os dados e os centroides foi usada a distância euclidiana: $\sqrt{((x1 - x2)^2 + (y1 - y2)^2)}$.

Para escolher os primeiros centroides foram escolhidas postagens com pesos diferentes, garantindo que não houvesse centroides iguais. Se o usuário escolher um número de centroides muito maior do qual existe de fato, esses centroides receberão peso zero.

5 Análise de Resultados

No capítulo anterior foram descritos os métodos desenvolvidos. Neste capítulo os métodos desenvolvidos foram aplicados para o pré-processamento e classificação das informações geográficas voluntárias.

5.1 Análise da Coleta de Dados

No processo de geração da base dados, foram coletadas um total de 11945 postagens. Essas postagens foram coletadas de grupos de notícias relacionadas a região do Vale do Aço. Os grupos utilizados para coleta são apresentados na tabela abaixo:

.

Tabela 1 – Quantidade de Postagens por Grupo

Grupos	Quantidade de Postagens
Portal Diário do Aço	8730
Shopping Vale do Aço	3215

Foram coletadas postagens importantes considerando o contexto do estudo. Muitas postagens possuem informações referente a bairros, estabelecimentos, cidades. E essas postagens serão consideradas informações geográficas voluntárias. Abaixo segue uma tabela exemplificando:

.

Tabela 2 – Postagens

Grupos	Postagens
Portal Diário do Aço	"Grupo arrombou loja no bairro Horto em Ipatinga..."
Shopping Vale do Aço	"O congestionamento começava no trevo do Ipatingão..."

Na postagem *"O congestionamento começava no trevo do Ipatingão..."*, podem ser retirados importantes informações para o estudo, como:

- Congestionamento - demonstrando que o local está cheio, com grande fluxo, provavelmente de automóveis.
- Ipatingão - Localização do congestionamento.

No entanto foi observado que existiam postagens que não possuiam nenhum referencial geográfico. Por exemplo:

"Deixaram para a última hora e o prazo termina nesta quarta-feira, dia 4."(Postagem retirada do Grupo: Portal Diário do Aço)

"Gastronomia: Arroz com brócolis http://t.co/hQhXhAtI"(Postagem retirada do Grupo: Shopping Vale do Aço)

Postagens como as descritas acima não são relevantes, pois não possuem dados necessários para serem consideradas informações geográficas. Para sanar essas e outras interferências, os dados coletados passaram por um pré-processamento que será descrito na seção seguinte.

5.2 Análise do Processamento dos Dados

Os dados coletados, mostrados na seção anterior, passaram por um processamento que será demonstrado nas subsessões seguintes.

5.2.1 Contagem de palavras

Todas as palavras pertencentes a coleção de dados foram contabilizadas somando um total de 148265 palavras. Entre elas encontram -se palavras que não são relevantes para o estudo apresentado, sendo essas as Stop Words. Foram consideradas como Stop Words, artigos, pronomes, nomes próprios, alguns verbos como ver, sentir, fazer, entre outros.

A tabela 3 apresenta apenas alguns exemplos de palavras que foram desconsideradas e a frequência com que elas aparecem na coleção de dados.

.

Tabela 3 – Contagem de Palavras - Stop Words

Palavras	Frequência
aquele	30
quando	187
vamos	85
pois	47
vara	19
sabem	15
vá	8
vejam	5
vao	5
umas	4
vanuza	3

Após a retirada das Stop words foi possível reduzir consideravelmente a quantidade de palavras a serem analisadas, somando um total de 16245 palavras. Após, foi feito o agrupamento das palavras restantes, que será demonstrado na próxima subseção.

5.2.2 Agrupamento de palavras

As palavras foram agrupadas respeitando os conceitos de sinonímia, polissemia e stem. Esse agrupamento foi feito justamente porque existem muitas palavras diferentes que possuem o mesmo sentido, ou um sentido próximo e por isso podem representam o mesmo conceito.

Por exemplo, nas frases *"Ontem o parque ipanema estava lotado durante o show."*, *"Exibição do Espetáculo Santinhas do Pau Oco no teatro do Shopping"*. Para o estudo apresentado não importa se a palavra é Show, Espetáculo ou Teatro. O que importa é que elas fazem parte de um mesmo grupo temático, no caso Entretenimento.

Abaixo, foram exemplificados três agrupamentos de palavras. Na tabela 6 que representa o grupo 10 - ASSALTO, é possível perceber os conceitos anteriormente citados.

As palavras: roubo e assalto pertencem a um mesmo grupo temático, ou seja, elas possuem sentidos próximos. Ainda nesse grupo é possível perceber as derivações dessas palavras: roubam, roubado, roubos, rouba, entre outras, e assaltam, assaltado, assaltaram, assaltou, entre outras.

.

Tabela 4 – GRUPO 76 - TIMOTEO

Palavras	Frequência
timoteo	2
timotinho	2
timirim	8
timoteo	2
timotinho	2
timóteo	509

.

Tabela 5 – GRUPO 57 - IPATINGA

Palavras	Frequência
ipatinguense	95
ipatinguenses	12
ipatingamg	4
ipatinga-mg	3
ipatinga-	3
ipatinga"	2
ipatinga	1645

.

Tabela 6 – GRUPO 10 - ASSALTO

Palavras	Frequência
assalto	112
assaltos	56
assaltantes	53
assaltante	31
assalta	16
assaltar	16
assaltou	10
assaltado	10
assaltaram	7
assaltadas	5
assaltada	4
assaltam	4
roubo	52
roubos	20
roubada	20
roubado	18
roubar	17
roubados	15
roubaram	11
roubou	8
roubalheira	5
roubadas	5
rouba	4
roubam	4

5.2.3 Implementação TF-IDF

Term Frequency (TF) representa a importância do termo, no caso da palavra, no documento, contabilizando quantas vezes ele apareceu no texto em questão.

Inversed Document Frequency (IDF) representa a importância global dessa palavra. Para seu cálculo leva-se em consideração toda a coleção de postagens. Ele é determinado pela frequência inversa da palavra.

Após a etapa de agrupamento das palavras foi feito o cálculo do TF-IDF. Esse cálculo é feito levando em consideração o grupo de palavras e as postagens coletadas.

Abaixo será exposta uma exemplificação de como foi realizado o calculo do TF-IDF, onde D1, D2, D3 representam três exemplos das postagens coletadas:

D1 = Assaltante que levou paulada na cabeça e foi preso em assalto frustrado no bairro Taúbas, Ipatinga, diz que não está arrependido.

D2 = Procuram-se dois assaltantes de postos de combustíveis em Timóteo.

D3 = Assaltaram, de novo, na rua Campinas, bairro Veneza.

P1,P2,P3 representam os grupos presentes nas postagens:

P1 = ASSALTO , P2 = TIMOTEO, P3 = IPATINGA

1. Calculando fi,j (freqüência bruta do termo i no documento j)

.

Tabela 7 – Calculo TFIDF - PASSO 1

i	P1	P2	P3
1 (D1)	2	0	1
2 (D2)	1	1	0
3 (D3)	1	0	0

2. Calculando max f i,j (freqüência do termo mais freqüente no documento j)

.

Tabela 8 – Calculo TFIDF - PASSO 2

i	max
1 (D1)	2
2 (D2)	1
3 (D3)	1

3. Calculando tfi,j = fi,j / max l f l,j (freqüência normalizada)

.

Tabela 9 – Calculo TFIDF - PASSO 3

i	P1	P2	P3
1 (D1)	1	0	0,5
2 (D2)	1	1	0
3 (D3)	1	0	0

4. Calculando Idfi = log2(N/ni) , a freqüência inversa dos termos i (inverse document frequency)

.

Tabela 10 – Calculo TFIDF - PASSO 4

i	P1	P2	P3
ni	2	1	1

N = 3 , total de documentos.

.

Tabela 11 – Calculo TFIDF - PASSO 4: log2(N/ni)

i	P1	P2	P3
idFi	0.58496250072	1.58496250072	1.58496250072

5. Calculando pesos finais wi,j = tfi,j . idfi

.

Tabela 12 – Calculo TFIDF - PASSO 5

i	P1	P2	P3
1 (D1)	0.58496250072	0	0.79248125036
2 (D2)	0.58496250072	1.58496250072	0
3 (D3)	0.58496250072	0	0

Após o cálculo do TF-IDF, foi aplicado o algoritmo de clusterização. Os resultados são demonstrados na subseção 5.2.4.

5.2.4 Análise do Algoritmo de Clustering - KmeansM

Para a etapa de agrupamento, foi utilizado o algoritmo kmeansM baseado no algoritmo Kmeans.

O algoritmo KmeansM permite agrupar postagens que sejam semanticamente parecidas, ou seja, são agrupadas de acordo com a significação das palavras. No entanto, se forem escolhidos grupos com palavras que representem localidades geográficas, essas postagens serão geograficamente semelhantes, permitindo que sejam retiradas das postagens informações sobre determinada localização.

No início do processo de clusterização o usuário pode inserir os grupos de palavras que desejar. Por exemplo, se o usuário inserir os grupos: 10 - Assalto, 57 - Ipatinga, 76 - Timóteo, serão buscadas em toda coleção postagens que possuam essas palavras e as demais postagens serão desconsideradas.

O usuário pode optar por não escolher os grupos de palavras, nesse caso serão levados em consideração todos os grupos de palavras.

Em seguida o usuário pode escolher a quantidade de centroides desejada.

Na primeira iteração o algoritmo seleciona os centroides aleatoriamente de acordo com a quantidade que o usuário escolheu. Nas demais iterações, os centroides são recalculados.

O algoritmo termina sua execução quando não há mais realocamentos das postagens nos clusters.

Nos anexos, são apresentados exemplos de clusters retirados dos processos de agrupamento.

Levando em consideração a postagem: "Assaltaram, de novo, na rua Campinas, bairro Veneza. Agora, o alvo foi uma construtora. Bandidos queriam dinheiro, mas fugiram apenas com um celular". Observou-se que quanto maior a quantidade de clusters melhor a qualidade do agrupamento.

No entanto, existe uma quantidade limite de Clusters possíveis, então não há diferença no agrupamento se for inserido um número muito grande de clusters que ultrapasse esse limite prático.

Utilizando todos os grupos de palavras e o banco de dados com as informações geográficas voluntárias, foi possível obter um máximo de 1194 clusters.

O algoritimo KmeansM foi executado em um computador com as seguintes especificações:

Fabricante Samsung Electronics, processador Intel(R) Core(TM) i5 - 5200U CPU 2.20GHz, memória RAM com 8 GB, e sistema operacional 64bits. O tempo gasto para processar a base com 11945 dados, com 10000 clusters foi de 32 minutos 39 segundos.

6 Conclusão

Este trabalho teve como objetivo verificar a adequação de algoritmos classificadores, ao classificar um banco de dados de informações geográficas voluntárias recolhidas de redes sociais.

Para isso, foi implementado um classificador Kmeans com o propósito de reunir postagens de usuários em temas compatíveis e em escopos geográficos compatíveis. Isso tem o potencial de favorecer atividades de recuperação de informação e de recomendação, essa última mais imediata.

O principal objeto resultante da execução deste trabalho foram os clusters de reportagens coletadas em redes sociais. Nos clusters, pode-se notar que reportagens com o tema similares, "assalto no Veneza"por exemplo, foram agrupados no mesmo cluster. Isso demonstrou que esse processo de filtragem utilizando informações geográficas voluntárias retiradas de redes sociais, juntamente com o classificador Kmeans, podem representar a intensidade que um certo evento ocorre em um local.

Através dessa constatação é possível dizer que o processo descrito no trabalho é uma opção eficiente para ser usada em diversos ramos, como exemplo, mostrar a incidência de uma doença como a Zica em um local, ou pontos com maior frequência de assaltos.

Porém, as principais contribuições deste trabalho, suas limitações e indicações de trabalhos futuros são independentes dos resultados obtidos na amostra usada. As contribuições são apresentadas na próxima seção.

6.1 Contribuições

O pré-processamento e a execução de um algoritmo classificador Kmeans possibilitaram um redução do volume de dados a ser processado, o que permite uma economia significativa de espaço de armazenamento e tempo de processamento. Esses dois recursos são escassos e muito exigidos em redes sociais e em bancos de dados geográficos tendo em vista o grande volume de mensagens e de localidades.

Os grupos de temas e de localidades observados mostram-se reduzidos. As notícias parecem girar em grupos muito limitados de temas, mesmo para localidades distantes. Por outro lado, na amostra usada, as principais referências geográficas adotadas foram a região metropolitana, o município, alguns pontos geográficos mais importantes (como estágio de futebol, rodovia, avenida principal, ou shopping), e alguns poucos bairros mais conhecidos. Isso torna especialmente mais simples o agrupamento tanto de temas quanto de localidades.

Adicionalmente, a presença de poucos grupos temáticos e de poucos grupos geográficos, facilita a identificação de postagens que constituiriam ruído na classificação, sem tema

ou sem localidade relevante.

A classificação se mostrou viável tanto pela identificação da simples ocorrência de palavras relacionadas quanto pela frequência com que as mesmas aparecem na descrição da reportagem. Apesar disso, o uso da frequência pode ser um indicador importante para aumentar a precisão da classificação em subcategorias mais específicas.

O custo de tempo de execução do Kmeans mostrou-se razoavelmente pequeno durante o processo de agrupamento, determinando sua viabilidade para este e para próximos trabalhos exploratórios com o mesmo objetivo.

6.2 Limitações

Embora o pré-processamento e o algoritmo classificador tenham sido úteis na economia de armazenamento e de tempo de processamento para a criação de grupos, este trabalho exploratório limitou-se a uma única região geográfica, muito pequena e simples, e em fontes de notícias pouco diversificadas.

Os grupos de temas e localidades consideraram as palavras semelhantes mas, por meio de algoritmos de classificação, notou-se que não foi possível relacionar o contexto no qual essas palavras foram descritas.

Títulos tais como "Condomínio fechado bem perto do Parque Estadual do Rio Doce"e "Vai fechar quase tudo no feriadão de natal e de ano novo. Veja o que abre e o que não abre esta semana"são bons exemplos. As duas notícias ficaram agrupadas no mesmo cluster, onde a palavra fechar e fechado parecem significar a mesma coisa, porém estão empregadas em contextos geográficos e semânticos diferentes, o que não pode ser tratado neste trabalho.

Quanto à adoção de algoritmo de classificação, a principal limitação deste trabalho é ter adotado apenas o algoritmo Kmeans. Outros algoritmos ou associação do Kmeans a outros algoritmos parecem promissores.

É possível melhorar os resultados de classificação, recuperação e recomendação utilizando o Kmeans associado com o Knn, ou seja, utilizar o método Knn dentro dos clusters produzidos pelo primeiro método. Como mencionado anteriormente, o Kmeans possibilitou uma pré-seleção agrupando as reportagens mais próximas. Em um cluster estão as reportagens que possuem palavras próximas, e que apenas por isso estão relacionada. O Knn buscaria os k vizinhos mais próximos da frase ou reportagem pesquisada, refinando a classificação.

Um outro método que poderia filtrar com maior eficácia é um kmeans recursivo que poderia gerar clusters de clusters. Ou seja, seria possível criar agrupamento de agrupamentos até que seus componentes sejam os mais parecidos possíveis. A importância dessa filtragem vai muito além de apenas auxiliar na escolha de possíveis locais para se abrir um estabelecimento, levando em conta a movimentação do lugar, trânsito e estabelecimentos próximos.

A inexistência de comparação com outros métodos também impede confirmar se o custo de tempo e de processamento em nossa classificação foi realmente a mais baixa possível. Apesar de nosso trabalho parecer ser o primeiro a ensaiar o uso de métodos de classi-

ficação clássicos para a categorização geográfica e temática, muitos estudos adicionais são necessários.

6.3 Trabalhos futuros

Este trabalho exploratório demonstrou a viabilidade do emprego de métodos de classificação clássicos na categorização de informação geográfica voluntária existente em redes sociais. Porém, como o trabalho teve escopos temático e geográfico muito limitados, novos estudos são necessários.

A presença de maior variedade de temas cria desafios adicionais para os classificadores e exige que outras estratégias para identificação de conceitos e temas sejam adotadas juntamente aos métodos classificadores.

Por outro lado, a presença de informação geográfica voluntária sobre diversas localidades diferentes produz ambiguidade, imprecisão e diferentes escalas geográficas que não parecem ser objeto de estudo dentro dos métodos clássicos de classificação.

Os métodos de classificação clássicos normalmente são usados após etapas de pré-processamento que parecem ser ainda mais importantes quando a coleção de dados foi produzida por humanos não especialistas, sem padrão de conformidade e em linguagem natural livre.

Trabalho futuro sobre as etapas de pré-processamento sobre informação geográfica voluntária é necessário para se avaliar com mais clareza a adequação dos métodos de classificação clássicos sobre esse tipo de informação.

Adicionalmente, a própria classificação por meio de diferentes métodos deve ser objeto de trabalhos futuros. Este trabalho avaliou apenas o Kmeans, em condições muito limitadas. Porém, outros métodos isolados ou em conjunto com o Kmeans potencialmente podem ajudar na recuperação e na recomendação de informação geográfica, especialmente no contexto de geomarketing, merecendo novos estudos.

A própria eficiência computacional dos métodos de classificação clássicos sobre informação geográfica precisa ser melhor investigada, criando referências de comparação para que trabalhos relacionados ao desempenho computacional possam acontecer.

Referências

BARION, D. L. E. C. N. Mineração de textos. *Revista de Ciencias Exatas e Tecnologia*, v. 3, p. 123–140, 2008. Citado nas páginas 19 e 20.

BELL, T.; COPE, A.; CATT, D. The third spatial revolution. In: *Workshop on Volunteered Geographic Information*. [S.l.: s.n.], 2007. Citado na página 8.

BUDHATHOKI, N.; BRUCE, B.; NEDOVIC-BUDIC, Z. Reconceptualizing the role of the user of spatial data infrastructure. *GeoJournal*, Springer Netherlands, v. 72, n. 3-4, p. 149–160, 2008. ISSN 0343-2521. Disponível em: <http://dx.doi.org/10.1007/s10708-008-9189-x>. Citado na página 8.

CASTELLS, M. *A Galáxia Internet: reflexões sobre a Internet, negócios e a sociedade*. [S.l.]: Zahar, 2003. Citado na página 12.

CâMARA, D. L. G. Sistemas de informaÇÃo geogrÁfica para aplicaÇÕes ambientais e cadastrais: Uma visÃo geral. *CONGRESSO BRASILEIRO DE ENGENHARIA AGRÍCOLA: CARTOGRAFIA, SENSORIAMENTO E GEOPROCESSAMENTO*, Poços de Caldas, v. 27, p. 59–88, 1998. Citado na página 8.

DIAS, L. S. O. C. R. Desenvolvimento e análise experimental de algoritmos evolutivos para o problema de clusterização automática em grafos orientados. Niteroi, 2004. Citado na página 16.

DRAHOS, P. Information feudalism in the information society. *The Information Society*, Taylor & Francis, v. 11, n. 3, p. 209–222, 1995. Citado na página 12.

FLANAGIN, A.; METZGER, M. The credibility of volunteered geographic information. *GeoJournal*, Springer Netherlands, v. 72, n. 3-4, p. 137–148, 2008. ISSN 0343-2521. Disponível em: <http://dx.doi.org/10.1007/s10708-008-9188-y>. Citado na página 12.

GOODCHILD, M. F. Citizens as sensors: the world of volunteered geography. *GeoJournal*, Springer, v. 69, n. 4, p. 211–221, 2007. Citado na página 12.

HAKLAY, M.; SINGLETON, A.; PARKER, C. Web mapping 2.0: The neogeography of the geoweb. *Geography Compass*, Wiley Online Library, v. 2, n. 6, p. 2011–2039, 2008. Citado nas páginas 8 e 13.

K. MURTY M. N., F. P. J. J. A. Data clustering: A review. *ACM Computing Surveys*, v. 31, p. 60, 1999. Citado na página 16.

KORN, F.; MUTHUKRISHNAN, S. Influence sets based on reverse nearest neighbor queries. In: TEXT RETRIEVAL CONFERENCE, 12., 2000, Dallas, Texas, USA. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. [S.l.]: ACM, 2000. Citado nas páginas 13, 14 e 15.

LEEUW, J. D. et al. An assessment of the accuracy of volunteered road map production in western kenya. *Remote Sensing*, Molecular Diversity Preservation International, v. 3, n. 2, p. 247–256, 2011. Citado na página 12.

LINDEN, R. Técnicas de agrupamento. *Revista de Sistemas de Informação da FSMA*, v. 4, p. 18–36, 2009. Citado na página 16.

OLIVEIRA, W. D. d. *Operação de busca exata aos K-vizinhos mais próximos reversos em espaços métricos*. 2010. Tese (Doutorado) — Universidade de São Paulo, 2010. Citado na página 14.

PATROUMPAS, K.; GIANNOPOULOS, G.; ATHANASIOU, S. Towards geospatial semantic data management: strengths, weaknesses, and challenges ahead. In: ACM. *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. [S.l.], 2014. p. 301–310. Citado na página 13.

RONZHIN, S. *Semantic enrichment of Volunteered Geographic Information using Linked Data: a use case scenario for disaster management*. 2015. Tese (Doutorado) — University of Twente, 2015. Citado nas páginas 12 e 13.

TUAN, Y.-F. Images and mental maps. *Annals of the Association of American Geographers*, Taylor & Francis, v. 65, n. 2, p. 205–212, 1975. Citado na página 12.

ZOOK, M. et al. Volunteered geographic information and crowdsourcing disaster relief: A case study of the haitian earthquake. *World Medical Health Policy*, Blackwell Publishing Ltd, v. 2, n. 2, p. 7–33, 2010. ISSN 1948-4682. Disponível em: <http://dx.doi.org/10.2202/1948-4682.1069>. Citado na página 12.

Anexos

ANEXO A –

O exemplo a seguir foi retirado do resultado do processo de agrupamento, onde utilizou -se todos os grupos de palavras e 10 mil centroides:

Cluster: 78

282- Assaltaram, de novo, na rua Campinas, bairro Veneza. Agora, o alvo foi uma construtora. Bandidos queriam dinheiro, mas fugiram apenas com um celular.

3035- Homem foi rendido por dupla, que depois de roubo de cordão no Veneza fugiu com um comparsa em um carro - http://goo.gl/JGP3jc

6659- Uma dupla de assaltantes é procurada, suspeita de atacar por duas vezes na rua Niterói, bairro Veneza II. O crime foi registrado por volta das 21h de domingo. Os proprietários da residência, Vicente de Paulo Silva, de 57 anos e Solange Ferreira da Consolação Silva, de 52 anos, relataram que estavam na varanda da residência, quando foram surpreendidos por dois homens, com idade aparente de 18 a 20 anos. Um deles estava de posse de uma arma de fogo. http://goo.gl/p0Rx98

ANEXO B –

O exemplo a seguir foi retirado do resultado do processo de agrupamento, onde utilizou -se todos os grupos de palavras e 100 centroides:

Cluster: 78

282- Assaltaram, de novo, na rua Campinas, bairro Veneza. Agora, o alvo foi uma construtora. Bandidos queriam dinheiro, mas fugiram apenas com um celular.

1126- Árvore com risco de queda preocupa moradores na rua Caxias do Sul, no bairro Veneza II http://goo.gl/K8670J

1350- Primeira batida do dia, no bairro Veneza: Vectra contra Astra - http://diariodoaco.com.br/notici 3/policia/vectra-bate-e-roda-sobre-ponte

1973- Carga excedente na altura deixou cinco postes quebrados e dezenas de casas sem eletricidade no bairro Veneza II http://diariodoaco.com.br/noticia/97645-3/policia/rede-eletrica-e-destruida-por-caminhao-no-veneza-ii

2636- Saldo da "bringuinha"entre Planalto e Veneza: Mais de 22 anos de cadeia para líderes de um dos bandos - http://goo.gl/rgBBln

3027- Apreendido no bairro Veneza, Kia Cerato com placas de Porto Alegre, mas que na verdade tinha sido furtado em Governador Valadares - Veja no Diário do Aço - http://goo.gl/XwxHOi

3035- Homem foi rendido por dupla, que depois de roubo de cordão no Veneza fugiu com um comparsa em um carro - http://goo.gl/JGP3jc

3124- Moradores de imóvel e representantes da Igreja do Evangelho Quadrangular no bairro Veneza protagonizam discussão acalorada http://goo.gl/6CEW9L

4689- Mundo cão. Reincidentes em crimes, menores de idade preparavam novos delitos quando foram abordados por policiais em um beco do bairro Veneza -http://goo.gl/2SXDtK

4947- Três jovens são alvejados a tiros no bairro Veneza. Um deles não resistiu a um ferimento na cabeça e morreu na manhã desta terça-feira. http://goo.gl/Hs4wGL

5273- Homem foi ao Veneza, em Ipatiga, tomar carro de devedor e acabou preso http://goo.gl/N1FZ7Y

5625- Denúncias anônimas via 190 levaram a Polícia Militar a apreender dois adolescentes suspeitos de tráfico de drogas no bairro Veneza. http://goo.gl/Qayvwx

5932- Assista às imagens da operação deflagrada hoje pela manhã para desocupar áreas públicas invadidas nos bairros Planalto II, Veneza II e Cidade Nova.

5948- Ordem judicial para desocupar áreas no Planalto, Veneza e Cidade Nova será cumprida na quinta-feira. Famílias participam de reunião e reclamam de demora em programas habitacionais. Entretanto, polícia não espera resistência - http://goo.gl/18qRG7

5971- Operação programada para esta semana vai acabar com a ocupação no Planalto, Veneza II e Cidade Nova - ?A orientação é a busca de uma solução, ou seja, uma desocupação voluntária, pacífica. Queremos que as pessoas entendam uma coisa. A única forma de permanecerem lá é reverter a ordem judicial. Se não houver uma mudança na decisão da Justiça, a área será devolvida aos municípios?, afirma comandante da PM. Tropa já está treinada e mobilizada. http://goo.gl/15ki70

5987- Tentativa de homicídio, hoje no bairro Veneza, já tem suspeito procurado pela polícia http://goo.gl/fc41U1

6659- Uma dupla de assaltantes é procurada, suspeita de atacar por duas vezes na rua Niterói, bairro Veneza II. O crime foi registrado por volta das 21h de domingo. Os proprietários da residência, Vicente de Paulo Silva, de 57 anos e Solange Ferreira da Consolação Silva, de 52 anos, relataram que estavam na varanda da residência, quando foram surpreendidos por dois homens, com idade aparente de 18 a 20 anos. Um deles estava de posse de uma arma de fogo. http://goo.gl/p0Rx98

6872- http://www.diariodoaco.com.br/noticias.aspx?cd=76656 - ACUSADO DE HOMICÍDIO NO VENEZA: 7108 Casa é destruída no Veneza.

Um incêndio na manhã desta quinta-feira (17) no bairro Veneza II quase terminou em tragédia. Cinco pessoas que estavam dentro do imóvel foram socorridas pelos vizinhos, que agiram antes da chegada do Corpo de Bombeiros.

http://www.diariodoaco.com.br/noticias.aspx?cd=75864

8224- Uma celebração na Igreja São Miguel, no bairro Veneza, lembrou na noite de ontem a morte do repórter Rodrigo Neto, ocorrida na madrugada de sexta-feira, 8. Vestidos com uma camiseta que lembra o episódio brutal, colegas de profissão e amigos marcaram presença na missa. Hoje completa uma semana que o profissional foi executado. As investigações sobre o caso prosseguem em sigilo e sem respostas até agora.

8287- Locomoção de peça de 160 toneladas, produzida pela Usiminas Mecânica, fechou a BR-381 das 8h às 10h30 entre o Horto e o Veneza I