UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de São José do Rio Preto

Diogo Lemos Guimarães

Framework para prospecção de dados espaciais baseado em semântica apoiado por ontologias

São José do Rio Preto

2015

Diogo Lemos Guimarães

Framework para prospecção de dados espaciais baseado em semântica apoiado por ontologias

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, Área de Concentração – Computação Aplicada, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Campus de São José do Rio Preto.

Orientador: Prof. Dr. Carlos Roberto Valêncio

São José do Rio Preto

2015

Guimarães, Diogo Lemos.
Framework para prospecção de dados espaciais baseado em semântica apoiado por ontologias / Diogo Lemos Guimarães. -- São José do Rio Preto, 2015
84 f. : il., gráfs., tabs.

Orientador: Carlos Roberto Valêncio
Dissertação (mestrado) – Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências, Letras e Ciências Exatas

1. Computação. 2. Mineração de dados (Computação) 3. Sistemas de dados espaciais. 4. Ontologias (Recuperação da informação)
5. Semântica - Processamento de dados. 6. Banco de dados. I. Valêncio, Carlos Roberto. II. Universidade Estadual Paulista "Júlio de Mesquita Filho". Instituto de Biociências, Letras e Ciências Exatas. III. Título.
CDU – 518.721

Ficha catalográfica elaborada pela Biblioteca do IBILCE
UNESP - Câmpus de São José do Rio Preto

Diogo Lemos Guimarães

Framework para prospecção de dados espaciais baseado em semântica apoiado por ontologias

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, Área de Concentração – Computação Aplicada, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Banca Examinadora

Prof. Dr. Carlos Roberto Valêncio
UNESP – São José do Rio Preto
Orientador

Profa. Dra. Rogéria Cristiane Gratão de Souza
UNESP – São José do Rio Preto

Profa. Dra. Marilde T. Prado Santos
UFSCar – São Carlos

São José do Rio Preto
14 de janeiro de 2015

Sobald wir lernen uns selbst zu vertrauen, fangen wir an zu leben

Johann Wolfgang von Goethe

RESUMO

Com a popularização de dispositivos que permitem a obtenção de dados espaciais, como pontos geográficos, velocidade e direção, aumentou-se consideravelmente o interesse pela captura, armazenamento e interpretação desses dados. Com isto, cada vez mais, um número maior de aplicações demostram interesse nesse tipo de dado e necessitam de bases de dados próprias para armazenar as informações. Estas bases são conhecidas como bases de dados espaciais. Com o objetivo de obter informações relevantes destes dados, algoritmos de prospecção de dados espaciais foram desenvolvidos e vem avançando com o intuito de, por exemplo, melhorar a qualidade dos resultados obtidos. Todavia, os algoritmos atuais desconsideram que os pontos geográficos estão em determinadas regiões, que por si só, fornecem informações semânticas relevantes. Com o objetivo de aprimorar os resultados de algoritmos, o uso de ontologias permite adicionar semântica e expressar o conhecimento sobre um domínio específico. O trabalho desenvolvido apresenta uma abordagem que permite, por meio do uso de ontologia, estender algoritmos espaciais para utilizarem um novo atributo durante o processo de criação de agrupamentos, o coeficiente semântico do ponto. Através do *framework* desenvolvido é possível adaptar algoritmos para utilizarem essa abordagem possibilitando gerar resultados mais relevantes.

Palavras-chaves: Mineração de Dados Espaciais, Ontologia, Semântica, Banco de Dados

Abstract

With the popularity of devices it became possible to easily obtain spatial data, such as geographic points, speed and direction, and because of that it also increased considerably the interest in obtaining, storing and analyzing such data. Therefore, a larger number of applications demonstrate interest in this type of data requiring it's own databases for storing this kind of information. These bases are known as spatial databases. In order to obtain relevant information from these data, spatial data algorithms were developed and has been advancing in order to, for example, improve the quality of the results. However, current algorithms disregard that geographical points are in certain regions, which in itself, provide relevant semantic information. In order to improve the results of algorithms, the use of ontologies allows to express semantics and knowledge about a particular domain. This work presents an approach that allows, through the use of ontology, extend spatial algorithms to use a new attribute in the process of creating groups, the semantic coefficient point. Through the developed framework it is possible to adapted algorithms to use this approach enabling generate more relevant results.

Keywords: Spatial data mining, ontology, semantics, data base

Sumário

Lista de Tabelas

Lista de Figuras

Siglas e Abreviações

CEREST	Centro de Referência de Saúde do Trabalhador
CLARANS	*Clustering Large Applications based on Randomized Search*
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise*
EPS	*Episolon*
GIS	*Geographic Information System*
GBD	Grupo de Banco de Dados
GPS	*Global Positioning System*
IRI	*International Resource Identifier*
KDD	*Knowledge Discovery in Database*
OWL	*Web Ontology Language*
POI	*Points of Interest*
SDM	*Spatial Data Mining*
SGBDs	Sistemas Gerenciadores de Banco de Dados
SPARQL	*SPARQL Protocol and RDF Query Language*
SIVAT	Sistema de Informação e Vigilância de Acidentes do Trabalho
SQL	*Structured Query Language*
RDF	*Resource Description FrameWork*
RDFS	*RDF Schema*
URI	*Uniform Resource Identifier*
URL	*Uniform Resource Locator*
URN	*Uniform Resource Name*
W3C	*World Wide Web Consortium*

1. Introdução

1.1. CONSIDERAÇÕES INICIAIS

O uso da computação torna-se cada vez mais indispensável e frequente, sendo utilizado nas mais diversas áreas para armazenar dados e posteriormente extrair informações. O cenário atual é considerado como rico em dados, porém pobre em informações, deixando claro que o maior desafio é a obtenção e extração de informações não triviais e desconhecidas em conjuntos de dados (HAN; KAMBER, 2011).

Com o intuito de solucionar este desafio e possibilitar a extração e prospecção de dados interessantes de uma base de dados, foi estipulado o processo de KDD, *Knowledge Discovery in Database*, onde uma das etapas é a aplicação do processo de *Data Mining*, ou prospecção de dados (OSEI-BRYSON, 2012). Existem diversas técnicas e algoritmos disponíveis na literatura que utilizam diferentes abordagens, porém com um único objetivo em comum, permitir a extração de conhecimento inicialmente implícito e desconhecido.

Com a evolução da computação e dos aparelhos eletrônicos disponíveis, tais como GPS e aparelhos móveis, a obtenção de dados espaciais tornou-se mais frequente (MENNIS; GUO, 2009). Informações espaciais são de interesse por possibilitar uma nova gama de análise e informações previamente não abordadas nos demais algoritmos, permitindo análises mais ricas e melhor entendimento de uma determinada situação (MENNIS; GUO, 2009).

1.2. MOTIVAÇÃO E ESCOPO

A prospecção de dados espacial ou *spatial data mining* - SDM é uma área de pesquisa relacionada à prospecção de dados, que se refere à extração de conhecimento implícito, a relações espaciais e a outros padrões que não estão explícitos em uma base de dados espacial (WANG et al., 2009). Devido a grande quantidade de dados geográficos existentes e a diversidade dos mesmos é impossível realizar uma análise manual nessas bases de dados (YASMIN, 2012). Com o aumento cada vez mais frequente dessas bases se torna importante a

definição de novas abordagens no processo de extração de conhecimento em um âmbito espacial.

Existem diversas técnicas para a prospecção de dados espaciais, sendo uma destas técnicas a de agrupamento de dados (JIN; MIAO, 2010). Tais técnicas consideram normalmente apenas o cálculo de distância entre os pontos para formação dos *clusters*. Algoritmos mais recentes foram adaptados para considerar também, além dos atributos *espaciais*, os atributos não espaciais, por meio do cálculo de similaridade, durante a etapa de formação dos *clusters* (BAE et al., 2009; HAN; KAMBER; TUNG, 2001), e com isso, contribuir com resultados mais interessantes.

Apesar disto, entende-se que os resultados obtidos não são ótimos, principalmente quando diversos pontos são analisados pelo algoritmo devido à grande quantidade de agrupamentos gerados. Este fato, causa uma dificuldade na etapa de análise dos resultados, visto que são gerados uma grande quantia de agrupamentos, quando na verdade, apenas alguns são realmente relevantes para o usuário final, além da necessidade de um especialista na área para analisar e filtrar os resultados obtidos.

O uso de ontologias por sua vez, está se tornando cada vez mais frequentes em diversas aplicações relacionadas a ciência da computação (CUGLER; YAGUINUMA; SANTOS, 2010). Apesar de sua principal aplicação ser relacionada a *Web* Semântica, por meio da estruturação dos dados disponíveis na *Web* (HENDLER, 2009; LIANG et al., 2010), esse uso não é restrito e é aplicado nas mais diversas áreas da computação, como integração de dados, extração de semântica de textos e *Big Data* (LIN et al., 2013)(NIMMAGADDA; DREHER, 2013). Portanto, devido a sua grande diversidade é possível aplicar ontologias para enriquecer o processo de prospecção de dados proporcionando assim, resultados mais interessantes para o usuário final.

1.3. OBJETIVOS E METODOLOGIAS

Neste trabalho propõe-se uma abordagem para estender algoritmos de prospecção de dados multirrelacionais em bases de dados georreferenciadas, por meio da criação de um *framework* adaptável, que utiliza conhecimentos definidos em ontologias para garantir uma qualidade maior dos resultados gerados.

Diferente das demais abordagens, são consideradas as informações semânticas da região referente a localização do ponto, neste caso verifica-se se um determinado ponto é

relevante ou não de acordo com a semântica implícita em sua posição geográfica. Com isso, pretende-se otimizar os resultados retornados por meio de *clusters* mais refinados e relevantes.

Para o trabalho desenvolvido tem-se as seguintes hipóteses:

- Algoritmos atuais de prospecção de dados geram agrupamentos, às vezes, irrelevantes para o usuário;
- Com o uso de ontologias propõe-se adicionar informação semântica ao ponto de acordo com sua região, enriquecendo o processo de prospecção;
- Considerando as informações implícitas na região referente ao ponto é possível gerar agrupamentos mais relevantes, reduzindo a quantidade de agrupamentos irrelevantes.

A metodologia de desenvolvimento envolveu diversas etapas, com seu início por meio de estudo prático e aplicação de ferramentas de prospecção de dados espaciais em bases de dados reais. Em seguida, fez-se o levantamento bibliográfico de técnicas de extração de conhecimento e suas principais desvantagens. Após o estudo, deu-se o desenvolvimento do *framework* de modo a utilizar ontologias para aprimorar o resultado obtido da mineração de dados espaciais, visto que tal carência ficou evidente durante o estudo prático.

Como metodologia para a validação da estratégia proposta neste trabalho, a mesma é aplicada durante a etapa de extração de conhecimento do algoritmo *MR-Clustering* (ICHIBA, 2013) *e VDBSCAN+* (MEDEIROS, 2014), algoritmos multirrelacionais de prospecção de dados espacial.

Portanto, norteado pelas hipóteses descritas, esta proposta visa obter resultados mais apurados e realistas no final do processo de prospecção dos dados, espera-se que os resultados sejam relevantes e facilite o trabalho do usuário final responsável por analisar os resultados gerados pelos algoritmos.

1.4. ORGANIZAÇÃO DO TRABALHO

O trabalho está organizado em cinco capítulos descritos a seguir:

- **Introdução** – Contempla uma introdução ao tema abordado, assim como a motivação para a realização do trabalho e os objetivos e metodologias para a realização do mesmo.
- **Fundamentação Teórica**– Neste capítulo são apresentados conceitos básicos e

informações do estado da arte na área de *data mining* espacial e ontologias, assim como trabalhos correlatos em ambas às áreas;

- **Aplicação de ontologias em prospecção de dados espaciais** – Neste capítulo é apresentada a abordagem utilizada para a criação do *framework*, no qual será detalhada a arquitetura proposta assim como as principais etapas para o desenvolvimento do mesmo.
- **Testes e Resultados** – São apresentadas as propostas de testes que serão realizados com o intuito de validar o algoritmo desenvolvido;
- **Conclusão** – Conclusão do trabalho e trabalhos futuros.

2. Fundamentação Teórica

2.1. CONSIDERAÇÕES INICIAIS

Neste capítulo são apresentados dois conceitos primordiais para o entendimento do trabalho proposto: o primeiro conceito trata-se de prospecção de dados espaciais, para isso serão apresentados os conceitos relacionados à prospecção de dados, contemplando informações do processo KDD, dados geográficos e as diversas abordagens dos algoritmos espaciais, assim como demais conceitos fundamentais para compreensão dos mesmos.

O segundo conceito, também importante, é o de ontologia, onde serão introduzidos métodos e ferramentas para armazenamento em triplas, linguagens utilizadas para buscar e atualizar ontologias, estrutura e categorias de ontologias, as diversas possibilidades de aplicação entre outros.

Por fim, é apresentado um levantamento com alguns trabalhos correlatos, contendo algoritmos de prospecção de dados espaciais e aplicações de ontologias nas mais diversas áreas.

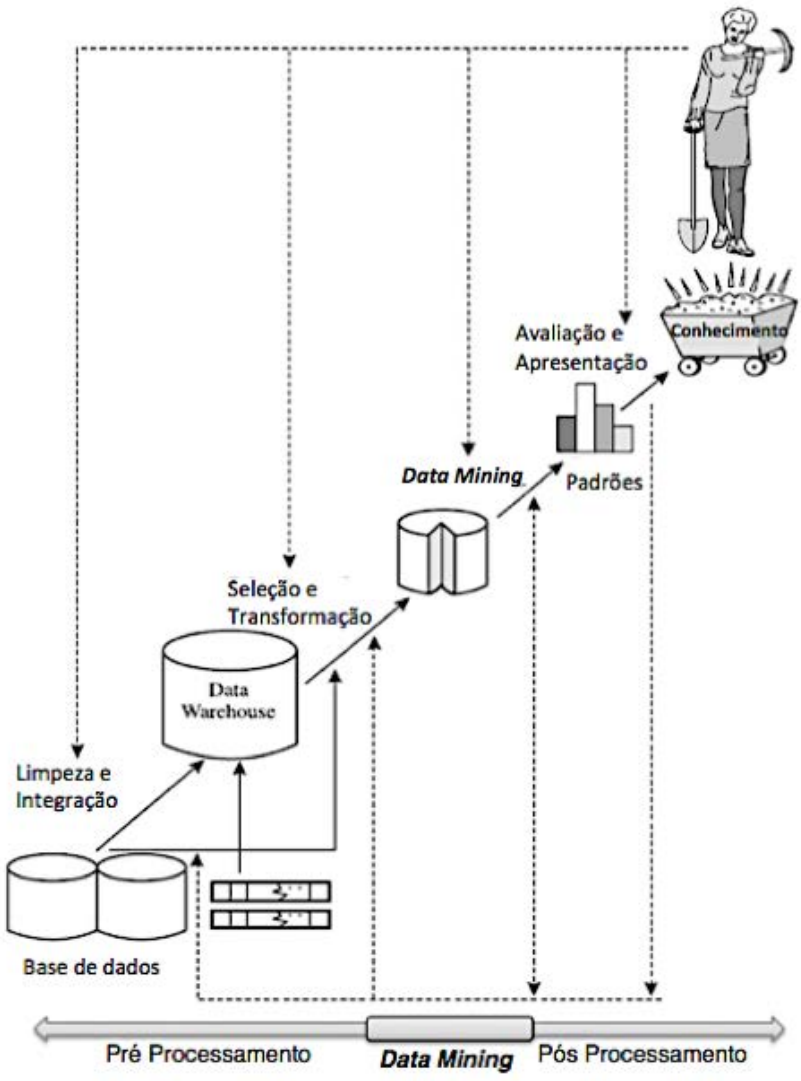
2.2. PROSPECÇÃO DE DADOS

O processo de descoberta de conhecimento, ou do inglês *knowledge discovery in databases (KDD),* permite a extração de conhecimento de diversos domínios e áreas, incluindo registros médicos, financeiro, empresariais, comerciais, entre outros, com o objetivo de descobrir relações e padrões globais presentes em grandes bases de dados.

O KDD pode ser definido como um processo não trivial com o objetivo de identificar padrões válidos, novos, potencialmente úteis e, finalmente, compreensíveis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Este processo, exibido na Figura 1, é organizado em três grandes etapas que possuem passos específicos e bem definidos, que devem ser executados corretamente para que os dados extraídos sejam confiáveis e interessantes, a saber:

- Pré-processamento dos dados;
- *Data Mining*;
- Pós-processamento dos resultados.

Figura 1: Processo de Descoberta de Conhecimento



Fonte: Adapatada e traduzida de (HAN; KAMBER, 2011)

Na etapa de pré-processamento ocorrem as atividades de preparação da base para a prospecção dos dados, portanto, efetua-se atividades relacionadas à limpeza, integração e seleção de dados. Essa etapa é primordial para obter um bom resultado no final do processo de extração de conhecimento, visto que é responsável por consolidar os dados que serão analisados e remover as inconsistências e informações duplicadas das bases de dados.

Uma vez que os dados estão padronizados e não exista mais inconsistência na base de dados, o processo de prospecção de dados, também conhecido como mineração de dados, é

então iniciado. Existem diversas abordagens para a prospecção de dados, como algoritmos de agrupamento, classificação, predição e regras de associação (HAN; KAMBER, 2011). Cada algoritmo irá analisar a base de dados com suas particularidades e então retornar o resultado de forma bruta, exibindo os padrões encontrados para o usuário.

Por fim, na etapa de pós-processamento é realizada a análise dos dados brutos retornados da etapa de *data mining* e os mesmos são apresentados para o usuário de uma forma que permita interpretar os dados da maneira desejada, permitindo que consiga analisar e efetuar tomadas de decisões a partir da informação obtida.

2.2.1. DATA MINING MULTIRRELACIONAL

Há várias estratégias que podem ser adotadas para a aplicação da prospecção de dados. A maioria dos algoritmos busca por dados considerados tradicionais e que estão armazenados em apenas uma tabela (DOU et al., 2008), porém, as bases de dados atuais estão organizadas em estruturas multirrelacionais, ou seja, com múltiplas tabelas (DOMINGOS, 2003).

Portanto, para aplicar a prospecção de dados convencional é necessário efetuar o processo de junção ou agregação das tabelas existentes, porém ambos os processos podem ocasionar em perda de semântica dos dados disponíveis na base de dados ou até mesmo ocasionar na perda de informação interessante (KNOBBE; BLOCKEEL, 1999; VALÊNCIO et al., 2011).

Para evitar problemas deste tipo, algoritmos multirrelacionais vêm sendo desenvolvidos com o objetivo de realizar a busca pelos dados diretamente na base de dados, sem a necessidade de efetuar a etapa de junção dos dados em apenas uma tabela.

Normalmente, algoritmos de prospecção de dados multirrelacional são desenvolvidos tendo como base algoritmos de prospecção de dados tradicional, e pelo mesmo motivo as mesmas técnicas utilizadas podem ser estendidas para a prospecção de dados espacial.

Um dos pontos negativos em relação à aplicação de técnicas multirrelacionais é o aumento no custo computacional para execução do algoritmo. Este fato ocorre devido à necessidade da criação de estruturas de dados específicas para armazenar e buscar pelas informações espalhadas nas diversas tabelas da base de dados, para depois extrair o conhecimento desejado (VALÊNCIO et al., 2011). Porém, o custo computacional adicional pode ser justificado pela qualidade e integridade dos resultados obtidos.

2.3. ANÁLISE DE DADOS ESPACIAIS

Além de aplicações tradicionais, o processo de descoberta de conhecimento também pode ser utilizado em bases de dados espaciais, onde considera-se, durante a análise dos dados, tanto os dados convencionais quanto os dados geográficos presentes nas relações entre os objetos.

2.3.1. DADOS ESPACIAIS

O avanço de novas tecnologias proporcionou uma mudança no contexto de extração das bases de dados, devido ao surgimento dos dados espaciais (YUESHUN, 2009). Devido à diversidade de representação e interpretação dos dados espaciais, estes apresentam grandes desafios durante o processo de extração de conhecimento, pois o dado passa a conter informações referentes à localização, distância, região, direção e demais características específicas referentes aos dados espaciais (JI et al., 2010).

Estes desafios estão normalmente ligados a complexidade única que dados espaciais apresentam em relação aos dados relacionais convencionais, apresentando normalmente três limitações (MENNIS; GUO, 2009):

- A maioria dos métodos atuais apresentam uma perspectiva limitada ou em um modelo relacional específico. Se a perspectiva escolhida ou modelo utilizado não for o apropriado para o fenômeno analisado, a análise irá no máximo mostrar que não existem relações interessantes, invés de sugerir uma alternativa mais apropriada;
- Muito dos métodos atuais não conseguem processar uma grande quantidade de dados. Visto que não são implementados, considerando as características únicas dos dados espaciais. Outro agravante é o tamanho dos dados que precisam ser manipulados, visto que em algumas aplicações é necessário realizar análise de imagens de satélites que possuem alta-definição;
- Devido a novos tipos de dados, como trajetória de objetos em movimento, e novas necessidades por parte das aplicações, novas abordagens precisam ser consideradas para extrair informações relevantes.

2.3.2. PROSPECÇÃO DE DADOS ESPACIAIS

Prospecção de dados espaciais pode ser definida como uma utilização extensiva de métodos estatísticos, tecnologia de reconhecimento de padrões, métodos de inteligência artificial, conhecimento de máquina etc. Com o objetivo de extrair conhecimento compreensível, interessante, oculto e inicialmente desconhecido de bases de dados espacial, como dados de gerenciamento, negócios ou de sensores remotos (HONGFEI; XIAOYAN, 2011).

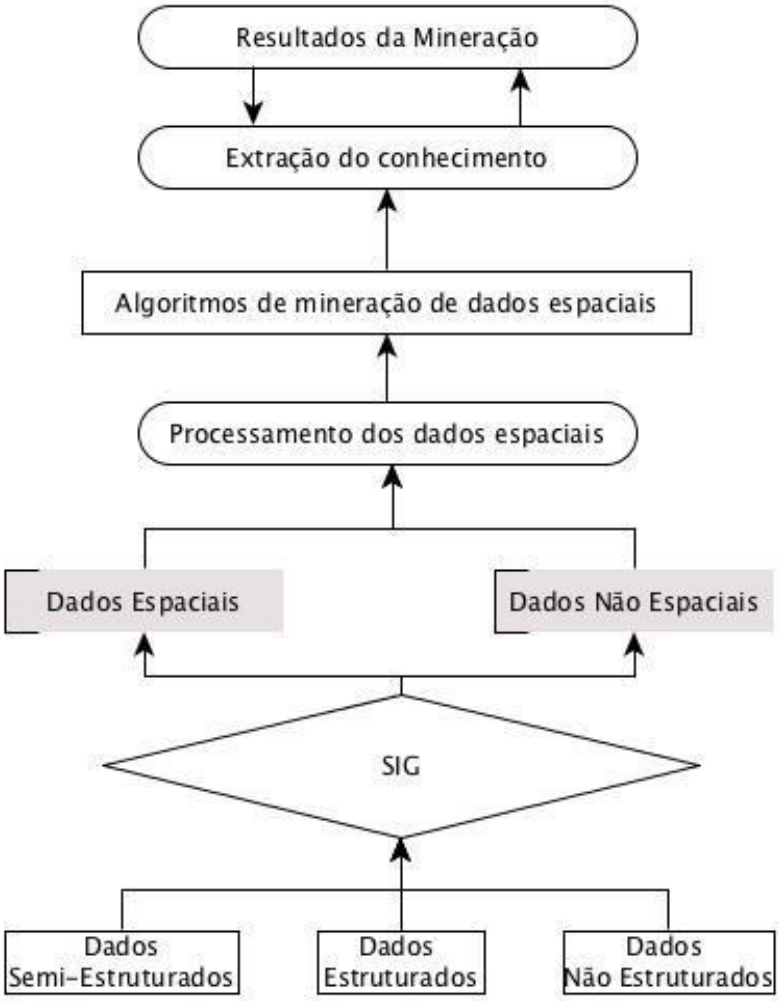
A extração de dados interessantes com dados espaciais é considerada mais complexa por apresentar, além das características específicas de dados espaciais já apresentadas anteriormente, características como grande quantidade de dados, não linearidade, ambiguidade e possibilidade de trabalhar com escalas diferentes. Devido a este motivo, as técnicas atuais apresentam algumas limitações, que estão listadas a seguir (JIN; MIAO, 2010):

- A maioria dos algoritmos de prospecção de dados espaciais são apenas uma conversão e adequação dos algoritmos tradicionais de prospecção de dados. Portanto, não consideram as peculiaridades e características dos dados espaciais. Por apresentar tipos de dados diferentes dos tradicionais, onde normalmente a prospecção é realizada em dados como *strings*, inteiros, datas etc, os algoritmos tradicionais são normalmente incapazes de analisar corretamente os dados espaciais;
- A eficiência dos algoritmos espaciais não é alta. A descoberta de padrões para este tipo de algoritmo não é refinada. Frequentemente, a quantidade de dados é muito alta o que obriga a remoção de dados que não tem ligação com o objetivo da prospecção para de maneira eficiente reduzir a dimensão do problema;
- Não existe uma linguagem SQL padronizada para efetuar consultas em bases de dados espaciais. Normalmente, são utilizadas extensões dessas linguagens que permitem consultar e adicionar dados espaciais em bases relacionais;
- O usuário possui dificuldades em controlar o processo de prospecção de dados espacial, o que torna difícil a utilização do conhecimento de um especialista durante o processo;
- O conhecimento adquirido é limitado pois, as aplicações focam na solução de apenas um problema específico;
- Não possui uma integração confiável com outros sistemas, dessa maneira, a utilização das técnicas atuais se torna muito restrita, pois, normalmente, para a aplicação de tal técnica é necessário a conexão com uma base de dados específica.

Caso deseje aplicar a descoberta de conhecimento em um campo mais amplo, haverá a necessidade de integração com fontes de dados distintas.

A aplicação da prospecção de dados espacial normalmente é realizada em conjunto com os atributos descritivos - não-espaciais - permitindo assim que estas informações sejam consideradas durante todo o processo de prospecção (GAZOLA; FURTADO, 2007). Na Figura 2 está representado um fluxograma de como é realizada a extração de conhecimento em um âmbito espacial com o auxilio de sistemas de informação geográficas – SIG, ou do inglês *geographic information system*, GIS .

Figura 2: Modelo para extração de conhecimento espacial com auxílio GIS



Fonte: Traduzida de (HONGFEI; XIAOYAN, 2011)

O uso do SIG é importante devido à possibilidade de utilização de suas poderosas funções espaciais, que ajudam a encontrar as relações dos objetos entre diferentes camadas, o que permite verificar se existe intersecção entre formas, calcular distâncias entre pontos, áreas

de regiões, entre outras.

Além disso, a utilização do SIG permite melhor controle durante a etapa de seleção dos dados e do pós-processamento. Isto ocorre, pois o usuário consegue visualizar os objetos espaciais projetados em mapas, o que facilita a manipulação, análise e a compreensão das informações encontradas.

2.4. MÉTODOS DE EXTRAÇÃO DE CONHECIMENTO ESPACIAL

Nesta seção são apresentados alguns dos métodos de extração de conhecimento em base de dados espacial, apesar de várias categorias de métodos existirem na literatura, este trabalho apresenta os métodos utilizados com maior frequência (JIN; MIAO, 2010).

2.4.1. AGRUPAMENTO ESPACIAL

O método de agrupamento espacial, também conhecido como *clustering*, tem como objetivo agrupar objetos de acordo com suas similaridades. A estratégia faz com que os dados de cada agrupamento formado tenham atributos semelhantes entre si e os mais diferentes possíveis dos demais agrupamentos (HAN; KAMBER, 2011; VAARANDI, 2003).

Algoritmos de agrupamento são usados normalmente para análise de dados, quando os dados estão organizados em grupos. Os métodos de agrupamento podem ser classificados em dois grupos (HAN; KAMBER; TUNG, 2001). O primeiro são os métodos de particionamento, que são utilizados com bastante frequência, que tem como principio particionar um conjunto de objetos de um determinado universo, em *k clusters*. Os algoritmos mais conhecidos são o *k-means* (LLOYD, 1982), o *EM* e o *k-memoid* (KAUFMAN; ROUSSEEUW, 1990) com algoritmos mais recentes como o CHSMST (HE; ZHAO; SHI, 2011) e VDBSCAN (LIU; ZHOU; WU, 2007).

A outra categoria é formada por métodos hierárquicos, que organizam os dados em uma hierarquia formando um *dendrogram*, que é uma árvore que divide a base de dados recursivamente em conjuntos menores, que pode ser formado de acordo com duas abordagens, "*bottom-up*" e "*top-down*". Alguns dos principais métodos com a abordagem hierárquica são *AGNES* e *DIANA* (KAUFMAN; ROUSSEEUW, 1990), que foi uma das primeiras abordagens, seguidos pelos métodos *CURE*, *CHAMELEON* e *WARDs* (MENNIS;

GUO, 2009).

2.4.2. REGRAS DE ASSOCIAÇÃO ESPACIAL

As regras de associação foram inicialmente originadas com o intuito de descobrir regularidades em grandes bases de dados. As regras de associação tradicionais são da forma X → Y, Se X então Y, e para verificar se a regra encontrada é útil e forte são definidos valores limitantes conhecidos como suporte e confiança (AGRAWAL; IMIELIŃSKI; SWAMI, 1993). De modo similar ao que ocorre em bases de dados relacionais, as regras de associação podem ser utilizadas em bases de dados espaciais.

Regras de associação espacial são formadas por predicados não espacial e espacial, sendo que alguns dos predicados espaciais utilizados são, por exemplo: disjuntos, dentro, *perto_de, longe_de, insersecção_com, sobreposição_com*, etc. (MENNIS; GUO, 2009). Portanto, uma regra de associação espacial pode ter os seguintes formatos: S → N ou N → S. Sendo S o conjunto de predicados espaciais e N representa o conjunto de predicados não espaciais (MILLLER; HAN, 2009; SANTOS; AMARAL, 2004).

2.4.3. CLASSIFICAÇÃO ESPACIAL

Por último, o método de classificação espacial tem como princípio agrupar os dados em classes ou categorias de acordo com suas propriedades. A diferença básica em relação ao método de agrupamento espacial é que no método de classificação é necessário um conjunto de dados de teste para validar e avaliar o desempenho do modelo de treinamento. Normalmente, para esse modelo são utilizadas árvores de decisão, redes neurais artificiais, funções de descriminação linear etc. (JIEHAI; WEI, 2010).

A classificação espacial diferencia-se da tradicional por não considerar apenas os atributos do objeto analisado, mas também os atributos dos vizinhos e suas relações espaciais, permitindo entender o relacionamento entre os objetos e prever o comportamento de outros (SANTOS; AMARAL, 2004).

2.5. ONTOLOGIAS

Grande parte dos algoritmos de prospecção de dados espaciais contemplam uma etapa de análise manual, onde um especialista no universo que a prospecção está sendo aplicada, irá inferir o seu conhecimento as regras encontradas, ajudando o algoritmo a obter melhores resultados futuros e na montagem final das regras.

Esta etapa é importante, pois algoritmos tradicionais não conseguem classificar ou interpretar conhecimentos que são triviais para o especialista e, portanto, não interessantes para o resultado final.

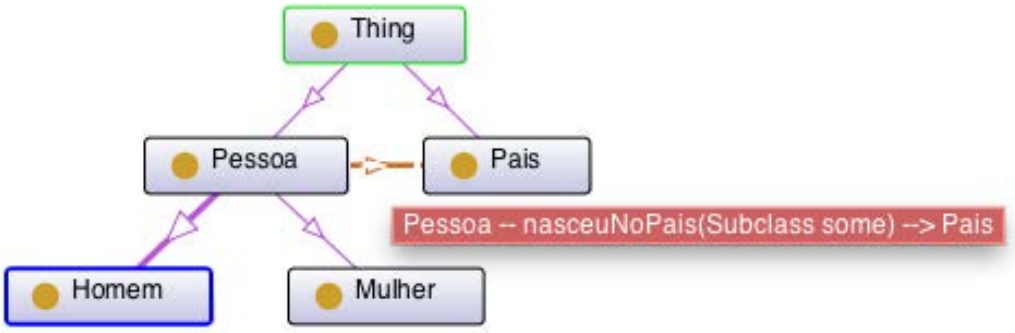
Com o intuito de contornar esta etapa, na qual depende de intervenção humana, algoritmos começaram a utilizar técnicas de classificação, que permitem interpretar e analisar a semântica das regras encontradas, porém não é possível fazer muitas inferências devido à limitação das classificações, visto que classificação é uma forma simples de ontologia (BERMAN, 2013).

Ontologias são representações teóricas sobre um conjunto de objetos, propriedades de objetos e relações entre objetos que são possíveis em um domínio de conhecimento específico (CHANDRASEKARAN; JOSEPHSON; BENJAMINS, 1999). Portanto, é possível definir um vocabulário para a ontologia e adicionar semântica entre as relações entre os objetos definidos na mesma, e com isso, conseguir resultados mais interessantes do que utilizando métodos de classificação convencionais, onde a principal limitação é que cada classe só pode possuir uma classe pai. No caso de ontologias, são construídas de modo a permitir que um objeto esteja conectado a mais de uma classe diretamente (BERMAN, 2013).

Outra definição de ontologia pode ser descrita como uma conceitualização de termos de um vocabulário normalmente especializado em um domínio ou assunto e não o vocabulário em si. Portanto, uma ontologia deve representar de forma semântica a ideia do domínio a ser tratado, sendo irrelevante se o conteúdo está em inglês ou português, sua representação deve ser a mesma (CHANDRASEKARAN; JOSEPHSON; BENJAMINS, 1999).

O uso de ontologias tornou-se bastante popular por proporcionar um entendimento em comum e compartilhado de um conhecimento e poder ser compartilhado entre pessoas e aplicações para os mais diversos usos (LIAO; CHEN; HSU, 2009). Na Figura 3 é ilustrada a representação de uma ontologia criada no programa *Protégé*, ferramenta de código aberta desenvolvida pela universidade de Stanford.

Figura 3: Exemplo de Ontologia criada utilizando o Protégé



Fonte: Do autor

2.5.1. WEB SEMÂNTICA

Um dos principais motivos para a popularização do uso de ontologias veio devido à necessidade de melhorar a distribuição e navegação das informações na *Web*, dessa maneira, diversos *sites* ou aplicações *on-line* podem compartilhar um conhecimento específico de maneira padronizada. Com essa abordagem é possível garantir uma *Web* mais inteligente e integrada com níveis de informações mais detalhados e sem incoerência entre informações (PARREIRAS, 2012). Uma definição formal de web semântica pode ser dada como um conjunto de padrões e boas práticas para compartilhar dados e a semântica entre esses dados pela *Web* para uso de aplicações (DUCHARME, 2013).

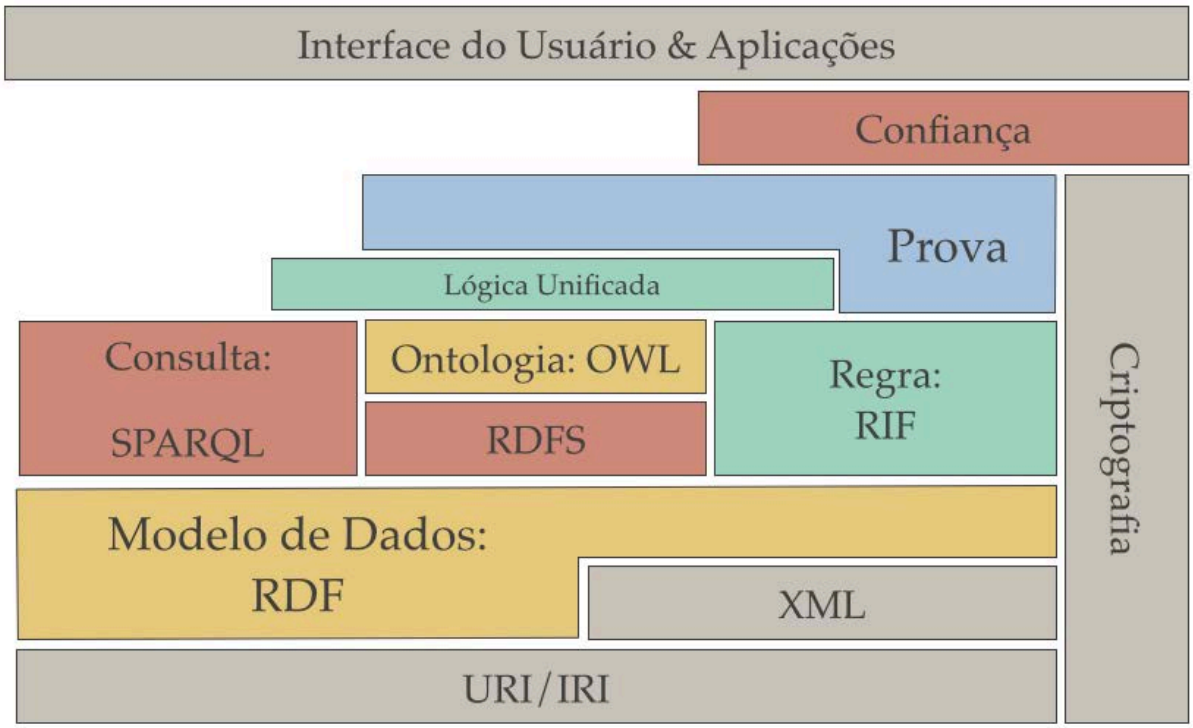
Para deixar mais claro as vantagens da *Web* Semântica, imagine que um *site* de um evento divulgue uma lista de hotéis para os participantes deste evento. Para o usuário, a informação disponibilizada é um conjunto de caracteres, no qual consegue-se identificar e interpretar como nomes de hotéis e seus respectivos endereços. Entretanto, o *site* ou aplicativo do evento que está divulgando poderia utilizar-se da *Web* Semântica para interpretar e entender que aquele conjunto de caracteres representam hotéis e suas posições geográficas.

Com isso, seria possível exibir informações adicionais compartilhadas por meio de ontologias, como por exemplo, disponibilidade de vagas, média de preço da hospedagem, distância do hotel em relação ao local do evento e outras informações extras contidas na ontologia. Dessa maneira, os dados disponibilizados não são usados apenas para visualização, mas para melhorar a qualidade da informação e explorar uma variedade de novos serviços inteligentes (YONG-GUI; ZHEN, 2010).

Na Figura 4, pode ser observada a arquitetura da *Web* Semântica e suas diversas camadas. Apesar da grande popularização no uso ontologias ser relacionado a Web Semântica, diversas outras aplicações também utilizam ontologias para diversas finalidades, como integração de dados, prospecção de dados, interpretação de textos, etc.

As camadas bases apresentadas na Figura 4 são importantes para a maioria das aplicações e são destacadas na seção seguinte.

Figura 4: Camadas da Web Semântica

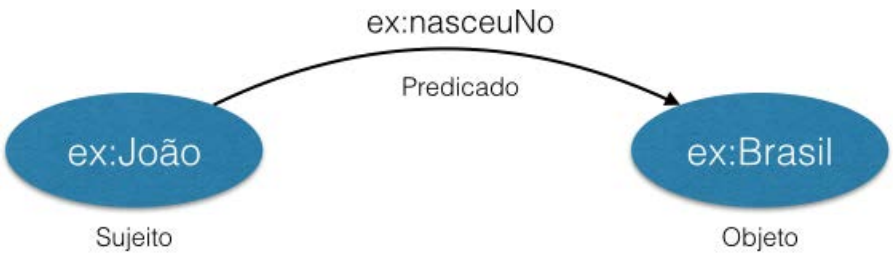


Fonte: Adaptada e traduzida de (PARREIRAS, 2012)

2.5.2. CAMADAS INFERIORES

RDF (*Resource Description FrameWork*) é a camada base e serve como suporte para as demais tecnologias (BHATIA; JAIN, 2011). *RDF* representa um modelo de armazenamento de dados onde as informações são expressas em forma de triplas. A tripla é formada por sujeito, predicado e objeto, na Figura 5 está representada a estrutura de uma tripla.

Figura 5: Ilustração de uma tripla RDF



Fonte: Do autor

Cada item da tripla é representado por uma URI, *Uniform Resource Identifier*, que tem como objetivo identificar unicamente cada recurso de uma ontologia específica. O primeiro tipo de *URI* foram as URLs (*Uniform Resource Locator*), que permitem que usuários acessem uma página *Web* unicamente por meio de um identificador ou endereço único, apesar de inicialmente não ser conhecida por esse termo.

Os termos foram evoluindo com o tempo e atualmente, considera-se o termo URI ou IRI para representar um recurso de uma ontologia enquanto o termo URL para o endereço de uma página (DUCHARME, 2013).

Como dito antes, o *RDF* é um modelo de dados onde as informações estão organizadas em triplas e são representadas seguindo uma das sintaxes disponíveis como, N-*triples*, *RDF/XML*, *Turtle* (BECKETT et al., 2014) e outros. Por ser relativamente nova, início de 2014, a sintaxe *Turtle* foi oficialmente recomendada pela *W3C*[1]. Na Figura 6 está representada a tripla exibida na Figura 5 na sintaxe *Turtle*.

Figura 6: Exemplo de tripla definida em Turtle

```
@prefix : <http://qbd.ibilce.unesp.br/onto/Exemplo#>

:i1001 rdf:type :Pais;
                :nome "Brasil" .
:i1000 rdf:type :Pessoa;
       :nome "José" ;
                :nasceuNo :i1001 .
```

Fonte: Do autor

[1] W3C - The World Wide Web Consortium Disponível em <http://www.w3c.org/2001/sw>. Acessado em 10/12/2013

Outra característica do *RDF* é o suporte a *tags* para um determinado valor literal. Sendo bastante utilizado, por exemplo, para definir que diferentes valores do tipo *strings* representam o mesmo recurso em idiomas diferentes. Um exemplo do uso de *tags* está disponível na Figura 7.

Figura 7: Identificação de *labels* em RDF

```
:i1001 rdf:type :Pais ;
                :nome "Brasil"@pt-br ;
                :nome "Brazil"@en-US.
```

Fonte: Do autor

Por ser a primeira camada, o *RDF* é, em termos de semântica, limitado, visto que o principal objetivo da camada é garantir a estrutura básica para o compartilhamento das informações. Os conceitos de semântica vão sendo adicionados nas próximas camadas como *RDFS* e *OWL* (ALLEMANG; HENDLER, 2011).

A partir do *RDFS* (RDF Schema) são introduzidos os conceitos de classes e superclasse, a possibilidade de definir uma hierarquia entre as propriedades e também estipular para uma propriedade quais classes podem relacionar-se com elas, por meio do uso de domínios e alcance (*domain* e *range*, do inglês).

Pode-se então concluir que uma ontologia é, portanto, formada de uma sequência de axiomas e fatos e é composta, normalmente, pelos seguintes elementos (COELHO, 2012; HORRIDGE et al., 2007):

- Entidade ou Indivíduos – Representam objetos de um domínio de interesse, ou seja, são responsáveis por identificar um indivíduo na ontologia. Por exemplo, João, Maria, Pedro, são casos de indivíduos para uma classe de Pessoas, enquanto Brasil, EUA e Chile são indivíduos de País;
- Propriedades – Que podem ser divididas em dois grupos, propriedades de objetos e propriedades de dados. A primeira, são relações binárias em indivíduos, ou seja, elas são responsáveis por vincular dois indivíduos de classes. Seguindo o mesmo exemplo, a propriedade *nasceuNo* pode vincular os indivíduos José e Brasil. Por sua vez, a propriedade de dado liga um indivíduo a um dado literal, por exemplo, a propriedade *nome*, que irá conectar a entidade Brasil a uma *string* com o valor “Brasil”;
- Classes – Representam um conjunto de indivíduos. Elas podem estar organizadas

de forma a criar uma taxonomia, permitindo subclasses e superclasses.

Apesar do conceito de tripla já existir na camada do *RDF* apenas por meio do RDFS é possível identificar e adicionar a semântica necessária, por exemplo, para informar que :Pais e :Pessoa são classes, basta adicionar as seguintes triplas, como exemplificado na Figura 8.

Figura 8: Exemplo de adição do conceito de classe disponível no RDFS

```
@prefix  rdfs: <http://www.w3.org/2000/01/rdf-schema#>

:Pais rdf:type rdfs:Class.
:Pessoa rdf:type rdfs:Class.
```

Fonte: Do autor

Como é possível perceber, a estrutura do *RDF* é altamente escalável e expansível, visto que foi necessário apenas adicionar outras triplas contendo as informações desejadas fazendo com que adicionar novas informações ou conhecimento seja um processo simples, onde toda informação é adicionada por meio da adição de uma nova tripla.

Por sua vez com a *OWL* (*Web Ontology Language*), adiciona-se mais um nível de semântica às triplas, sendo o padrão estipulado pela *W3C* em constante desenvolvimento, com a última versão liberada em 2012 com várias melhorias. A *OWL 2* permite definir características para as propriedades de objeto e criar restrições de valores mais elaboradas. Existem 4 propriedades possíveis para as propriedades de objeto e a partir da *OWL 2* foram adicionadas 3 inversas dessas propriedades, sendo no total 7 propriedades de objeto:

- **Funcional -** Se uma propriedade é funcional, significa que um indivíduo *x* só pode estar conectado a no máximo um indivíduo *y*. Um exemplo onde essa função pode ser aplicada seria em uma propriedade *:possuiPai;*
- **Funcional inversa -** Já na funcional inversa, um determinado indivíduo *y* só pode estar relacionado com um indivíduo *x.* Um exemplo de propriedade seria *:éPaiDe;*
- **Simétrica -** Se uma propriedade for simétrica representa que se um indivíduo *x* está conectado a um *y*, então *y* também está conectado a *x*. Um exemplo de função *:éAmigoDe*;
- **Assimétrica -** No caso de uma propriedade assimétrica, a ontologia não permite que se *x* está relacionado com *y* que *y* tenha o mesmo relacionamento com *x*;

- **Reflexiva -** Propriedades reflexivas consideram que o indivíduo deve estar conectado com ele próprio;
- **Irreflexiva -** No caso de propriedades irreflexivas o indivíduo não pode estar relacionado com ele próprio. Um exemplo seria a propriedade de objeto *:éCasadoCom*;
- **Transitiva -** No caso de uma propriedade transitiva, um indivíduo *x* conectado a um *y* que por sua vez se conecta a outro indivíduo *z*, implica que *x* está conectado a *z*. Como exemplo tem-se a propriedade *rdfs:subClassOf.*

2.5.3. CONSULTAR INFORMAÇÕES EM ONTOLOGIAS

O método de armazenamento e disposição dos dados de uma ontologia difere quando comparado com uma base de dados relacionais, visto que toda informação é disposta por meio de conjuntos de triplas. Para exemplificar melhor as diferenças, na Tabela 1 é exibida uma tabela de um modelo relacional da entidade “Pessoa” e na Tabela 2 as mesmas informações, porém no modelo *RDF*.

Tabela 1: Entidade "Pessoa" - Modelo Relacional

Pessoa				
ID	Nome	Sexo	Cidade	Escolaridade
1	Maria	F	SJRP	Superior Completo
2	João	M	Belo Horizonte	Superior Incompleto

Tabela 2: Representação da Entidade "Pessoa" em triplas RDF

Sujeito	Predicado	Objeto
:Pessoa1	:nome	“Maria”
:Pessoa1	:sexo	“F”
:Pessoa1	:cidade	“SJRP”
:Pessoa1	:escolaridade	“Superior Completo”
:Pessoa2	:nome	“João”
:Pessoa2	:sexo	“M”
:Pessoa2	:cidade	“Belo Horizonte”
:Pessoa2	:escolaridade	“Superior Incompleto”

As ontologias definidas utilizando *OWL* nada mais são do que conjuntos de triplas, por esse motivo é necessário utilizar algum *software* que armazene as triplas e consiga interpretar as classes e propriedades da *OWL* para que se consiga então, fazer inferências sobre os dados (DUCHARME, 2013).

Para isso, é necessário utilizar *softwares* específicos conhecidos como *TripleStore* ou *RDF Store* que são responsáveis por armazenar e interpretar a semântica implícita, alguns dos mais utilizados são: *Virtuoso*, *Sesame*, *Apache Jena* e *Parliamet* (LE et al., 2012).

O *SPARQL* foi confeccionado para consultar *RDF*, e por isso a consulta é descrita em padrões de triplas, que são similares as do *RDF*, porém podem conter variáveis. Um exemplo simples de consulta pode ser analisado na Figura 9.

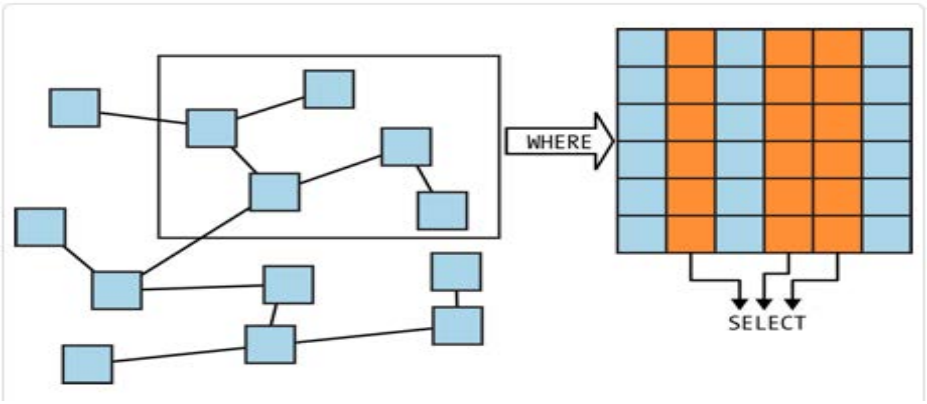
Figura 9: Exemplo de consulta SPARQL

```
SELECT ?paisNasceu
WHERE
{:Joao :nasceuNo ?paisNasceu}
```

Fonte: Do autor

Neste caso, a consulta está retornando, caso exista, o objeto da tripla onde o sujeito é :Joao e o predicado :nasceuNo. Na Figura 10 está ilustrado como é o processo de consulta em uma ontologia. No caso a cláusula *WHERE* é responsável por extrair as informações do conjunto de dados enquanto a *SELECT* seleciona quais informações serão exibidas.

Figura 10: Exemplificação de consulta SPARQL



Fonte: Retirada de (DUCHARME, 2013)

Por esse motivo, mesmo que o usuário não conheça a estrutura da ontologia é possível realizar consultas e descobrir novas informações. Por exemplo, supondo que além da propriedade de objeto *:nasceuNo*, a ontologia também possua a *:moraEm*, logo, mesmo que o usuário desconheça essa informação, é possível obter utilizando a seguinte consulta exibida na Figura 11.

Figura 11: Consulta SPARQL para obter qualquer propriedade relacionada a País

```
SELECT ?propriedade ?pais
WHERE
{:Joao ?propriedade ?pais.
?pais rdfs:Class :Pais}
```

Fonte: Do autor

Neste caso, a consulta irá retornar todas as propriedades de objeto e os valores (instâncias) que *:Joao* tem com qualquer indivíduo da classe *:Pais*. Além disso, existem diversas outras funcionalidades da linguagem *SPARQL* para otimizar e aumentar os parâmetros de busca. Assim como *OWL*, o *SPARQL* continua em grande desenvolvimento, recentemente as especificações da *W3C* foram atualizadas para a versão 1.1 permitindo inserir, atualizar e remover triplas, invés de apenas realizar consultas.

2.5.4. CLASSIFICAÇÃO DE ONTOLOGIAS

Com a aplicação de ontologias tornando-se cada vez mais frequente, tornou-se necessário classificá-las de modo a organizar os diferentes tipos existentes. Dessa maneira, é possível classificá-las quanto a sua função, grau de formalismo, aplicação, estrutura e finalmente quanto ao conteúdo. Apesar de possuir diversos tipos de classificação, normalmente as mais utilizadas são as de função e conteúdo (ALMEIDA; BAX, 2003). A primeira refere-se ao tamanho e tipo de estrutura da conceptualização, que podem ser separada em três níveis (VAN HEIJST; SCHREIBER; WIELINGA, 1997):

- **Ontologia terminológica** representa termos formais representados em um domínio específico. Por exemplo, uma ontologia com as descrições dos termos utilizados na área de meio ambiente é uma ontologia terminológica;
- **Ontologia informacional** especifica a estrutura dos registros do banco de dados. Neste nível, preocupa-se com o armazenamento das informações no banco de dados, sem se preocupar com a representação semântica dos dados;
- **Ontologia de modelo de conhecimento** representa a conceptualização do conhecimento. Se comparada com as ontologias informacionais, as ontologias de conhecimento apresentam uma estrutura interna muito mais rica e complexa. São também utilizadas para um determinado uso particular do conhecimento que descrevem.

A classificação por conteúdo refere-se ao conteúdo descrito na ontologia, sendo organizada nos seguintes níveis (BRAVO, 2010; MIZOGUCHI; VANWELKENHUYSEN; IKEDA, 1995):

- **Ontologias de domínio** são aquelas que representam conceitos mais gerais que não representam um único domínio, mas conhecimentos gerais como espaço e tempo;
- **Ontologias de tarefas** representam os conceitos de um domínio específico a partir da generalização das ontologias de alto nível. Estão relacionados a conceitos como “medicina”, “comércio” etc.;
- **Ontologias comuns ou gerais** contêm todas as definições que são necessárias para modelar o conhecimento de uma aplicação particular. Normalmente, são a mistura dos conceitos das ontologias de alto nível e as de domínio.

2.5.5. DESAFIOS

A utilização de ontologias traz diversas vantagens já ressaltadas, como permitir a representação de um domínio e o compartilhamento do mesmo. Porém, existem diversos desafios e problemas que precisam ser enfrentados durante a confecção e utilização de ontologias, alguns destes desafios estão listados (CONTRERAS; CORCHO; GÓMEZ-PÉREZ, 2009; HEPP, 2008; SHVAIKO; EUZENAT, 2008):

- Interação – é essencial que humanos consigam entender as especificações de uma ontologia, tanto quanto durante o processo de criação – *design* – quanto para a criação de consultas e anotações;
- Escolha da ontologia – devido à complexidade de alguns domínios é difícil fazer a escolha de uma ontologia que agregue todo o conhecimento;
- Criação e evolução das ontologias – Ontologias não devem ser estática, é necessário evoluir e estender os conceitos da mesma, do mesmo modo que esses conceitos evoluem com o tempo em seu domínio. Manter a ontologia atualizada com conceitos atuais é considerado um desafio;
- Mapeamento e interoperabilidade – Outro desafio refere-se a dificuldade de efetuar o mapeamento de termos em diferentes ontologias, permitindo que as mesmas consigam se comunicar de forma transparente. Existem vários trabalhos que propõe soluções referente ao mapeamento e alinhamento de ontologias;
- Bibliotecas de ontologias – Um dos principais objetivos da ontologia é permitir o compartilhamento de ontologias, porém sem um padrão definido, ontologias de um mesmo conhecimento podem ter o contexto diferente;
- Desempenho e Escalabilidade – Enquanto sistemas gerenciadores de bases de dados (SGDBs) possuem um alto grau de maturidade e fornecem alta performance e escalabilidade, repositórios de ontologias ainda não conseguem fornecer o mesmo desempenho;
- Metodologia de desenvolvimento – Apesar de existirem trabalhos e pesquisas bem consolidadas que apresentam possíveis metodologias, tais como as apresentadas por Uschold (USCHOLD; GRUNINGER, 2004), *Methondology* (FERNÁNDEZ; GÓMEZ-PÉREZ; JURISTO, 1997), *On-to-Knowledge* (STAAB; STUDER, 2001) ainda não existe um padrão bem definido e uma dificuldade em executar essas metodologias corretamente.

2.6. APLICAÇÃO DE SEMÂNTICA NO *DATA MINING*

Ontologias representam os conceitos e relações entre os objetos de domínios específicos (ELSAYED et al., 2007). Existe uma utilização vasta de técnicas para adicionar significado aos dados, principalmente devido à necessidade da criação da Web Semântica, projeto da *W3C* para adicionar semântica nas páginas *web*, permitindo uma padronização, facilitando assim a obtenção de informação na *WEB* (VENANCIO; FILETO; MEDEIROS, 2003).

A utilização de ontologias é bastante diversificada e está presente em diversas outras áreas. Em relação a aplicação em bases de dados, a semântica pode ser aplicada no processo de integração, limpeza e prospecção de dados. Dessa forma é possível atribuir significado e conhecimento às relações e aos objetos. Considera-se que no contexto de prospecção de dados, o conhecimento é descoberto e durante a construção de ontologia o conhecimento é adquirido (HWANG, 2004).

A utilização de ontologias durante o processo de *data mining* permite que o conhecimento seja reutilizado e automatizado, sem a necessidade de uma análise dos dados manual e que é refeita no fim de cada execução (KUO et al., 2007). Além disso, também traz vantagens em relação à qualidade dos dados obtidos, visto que poderá ignorar resultados não interessantes e ressaltar os mais interessantes de acordo com seu significado semântico.

Algoritmos de prospecção de dados não são genéricos ao ponto de serem utilizados em qualquer domínio de aplicação, portanto é possível utilizar ontologias para estender algoritmos de prospecção de dados para qualquer ambiente utilizando ontologias de domínio específicas. Por exemplo, considerando-se acidentes de trabalho, um acidente em uma área residencial, tem relações e características distintas de acidentes em áreas comercias ou industriais e o algoritmo poderá ressaltar os agrupamentos que realmente merecem destaques.

A abordagem utilizada varia de acordo com a intenção dos algoritmos propostos, podendo utilizar a ontologia no pós-processamento dos dados, após a etapa de prospecção de dados (HWANG, 2004), ou ainda durante a própria execução do algoritmo, com o intuito de direcionar melhor os resultados encontrados. Cada abordagem tem suas vantagens e fica como papel da ontologia basicamente guiar o algoritmo de prospecção de maneira que seja interessante de acordo com o domínio que está sendo analisado e proporcionar o contexto para o qual os dados extraídos devam ser analisados (HWANG, 2004).

2.7. TRABALHOS CORRELATOS

Como ressaltado nos capítulos anteriores o uso de ontologias é bastante diversificado, sendo utilizado nas mais diversas áreas da computação, desde integração de dados, limpeza, *Big Data* e prospecção de dados.

Atualmente existe uma tendência em utilizar ontologias para expressar o contexto de um domínio e alguns trabalhos estão utilizando tal tecnologia com o intuito de adicionar semântica no processo de prospecção de dados (ONGENAE et al., 2013).

Peng Yue (YUE et al., 2013) desenvolveu um trabalho para descoberta de recursos geo-espaciais complexos em imagens. Para isso, é utilizado ontologias que definem a semântica desses recursos complexos e uma ontologia para definir as características espaciais. Na ontologia de recursos complexos é possível definir, por exemplo, que uma escola é formada por prédios, por uma região com grama, quadras, entre outros.

O algoritmo de mineração de imagem irá extrair então as informações básicas da imagem, como identificando prédios, estradas, grama e rios. Com essa informação a ontologia é consultada retornando a informação do recurso complexo.

O trabalho desenvolvido por Arnaud (VANDECASTEELE; NAPOLI, 2012), considera a quantidade de desastres naturais marítimos e o aumento de piratas, para apresentar uma ferramenta que realiza detecção de comportamentos anormais marítimos com o intuito de auxiliar na solução e interpretação destes problemas. Os sistemas atuais necessitam analisar o estado atual do navio, para isso necessitam analisar uma grande quantia de dados, a complexidade existente entre as diferentes situações e os sistemas atuais não consideraram as características espaciais existentes. O trabalho propõe a utilização de ontologia em conjunto com sistemas de informação georreferenciada para solucionar estes problemas.

Panče (PANOV; DZEROSKI; SOLDATOVA, 2008) apresenta uma abordagem interessante quanto ao uso de ontologias, desenvolvendo uma meta-ontologia denominada *OntoDM*, onde se cria uma ontologia para representar o processo de prospecção de dados, com o objetivo de solucionar problemas mais genéricos e não apenas um problemas específico. Uma outra vantagem do uso da abordagem é a padronização nos resultados obtidos. A *OntoDM* possui entidades básicas do processo de *data mining* como: tipos de dados, tarefas de prospecção, generalização e algoritmos. Após estes itens definidos pode-se pensar em entidades mais complexas como: cenários de prospecção de dados.

O trabalho de Max Braun (BRAUN; SCHERP; STAAB, 2010) propõe a criação de pontos semânticos onde são utilizadas ontologias para coletar, definir e armazenar as

informações de pontos de interesses, *POIs*, em sistemas de mapas. Os pontos de interesse são determinados por meio de uma ontologia de vocabulários, onde estão definidos termos como as categorias dos pontos e as suas relações existentes. Dessa maneira o trabalho fornece uma otimização em relação a busca dos pontos, retornando aqueles de maior interesse para o usuário. Além disso, como os termos são definidos por meio de ontologias é possível que novas informações sejam adicionadas e alteradas de forma colaborativas e estruturada por vários usuários.

Por sua vez, o trabalho (BRAVO, 2010) consegue de maneira automática, analisar sintaticamente textos disponíveis na *Web* e estruturar os elementos gramaticais do texto criando ontologias para os termos. Para isso, o algoritmo utiliza um analisador sintático (PALAVRAS) e uma ontologia base da língua portuguesa.

Por último, os autores (VENANCIO; FILETO; MEDEIROS, 2003), propuseram um trabalho que permite aplicar ontologias de objetos geográficos para facilitar navegações em *GIS*. Os autores alegam que as ferramentas atuais, por não apresentarem suporte à semântica acabam perdendo informação e referência do contexto das informações ao executar *zoom in* no mapa. Neste caso, ao aplicar uma ontologia em um sistema de navegação de mapas, ao realizar uma busca por um texto não causaria ambiguidades, pois seria considerado o contexto em questão que o usuário encontra-se, por exemplo, entre a cidade São Paulo e o estado São Paulo.

Com os trabalhos apresentados conclui-se que o uso de ontologias é bastante comum nas mais diversas áreas da computação, e fornecem aos algoritmos novas abordagens que resultam em um resultado com maior qualidade. Porém, as abordagens atuais tratam o problema da semântica pontualmente, por meio de uma alteração em um algoritmo base.

O trabalho desenvolvido, *framework* OntoSDM, difere dos demais trabalhos apresentados por apresentar um modelo genérico de uso de ontologias para aplicação de algoritmos de prospecção de dados por agrupamentos de dados. Além disso, permite que a alteração e refinamento dos resultados ocorra durante a etapa de geração dos agrupamentos e não posteriormente como parte da etapa de pós-processamento.

2.8. CONSIDERAÇÃO FINAL

Neste capítulo foram abordados os principais conceitos relacionados à prospecção de dados espaciais e apresentadas as principais formas de extração de conhecimento -

agrupamento, classificação e regra de associação - apresentando para cada uma exemplos de algoritmos disponíveis na literatura. Outro conceito introduzido é o de ontologias, representações semânticas de um domínio, que cada vez mais tornam-se mais populares e está presente em diversas áreas da computação.

No final do capítulo foram apresentados trabalhos correlatos, existentes na literatura que representam o estado da arte relacionado a ontologias e prospecção de dados. Com os trabalhos correlatos apresentados neste capítulo é possível comprovar a diversidade de aplicações que utilizam ontologias para melhorar resultados de algoritmos de prospecção, permitir integração entre fontes distintas e outras vantagens.

3. OntoSDM: Framework para uso de ontologias

3.1. CONSIDERAÇÕES INICIAIS

Os algoritmos de prospecção de dados espaciais existentes na literatura, durante a etapa de formação dos *clusters*, consideram apenas as distâncias entre os pontos e a similaridade das características. A fim de melhorar os resultados obtidos foi elaborado um *framework* que, por meio do uso de ontologias, é capaz de extrair a semântica da região de um determinado ponto por meio das definições contidas na ontologia atribuir um novo atributo para o ponto, seu coeficiente semântico.

Neste capítulo é detalhado o processo de desenvolvimento da ferramenta, assim como as tecnologias utilizadas e a abordagem utilizada durante a aplicação no algoritmo de prospecção de dados espaciais multirrelacional *MR-Clustering* (ICHIBA, 2013) e CHSMST+.

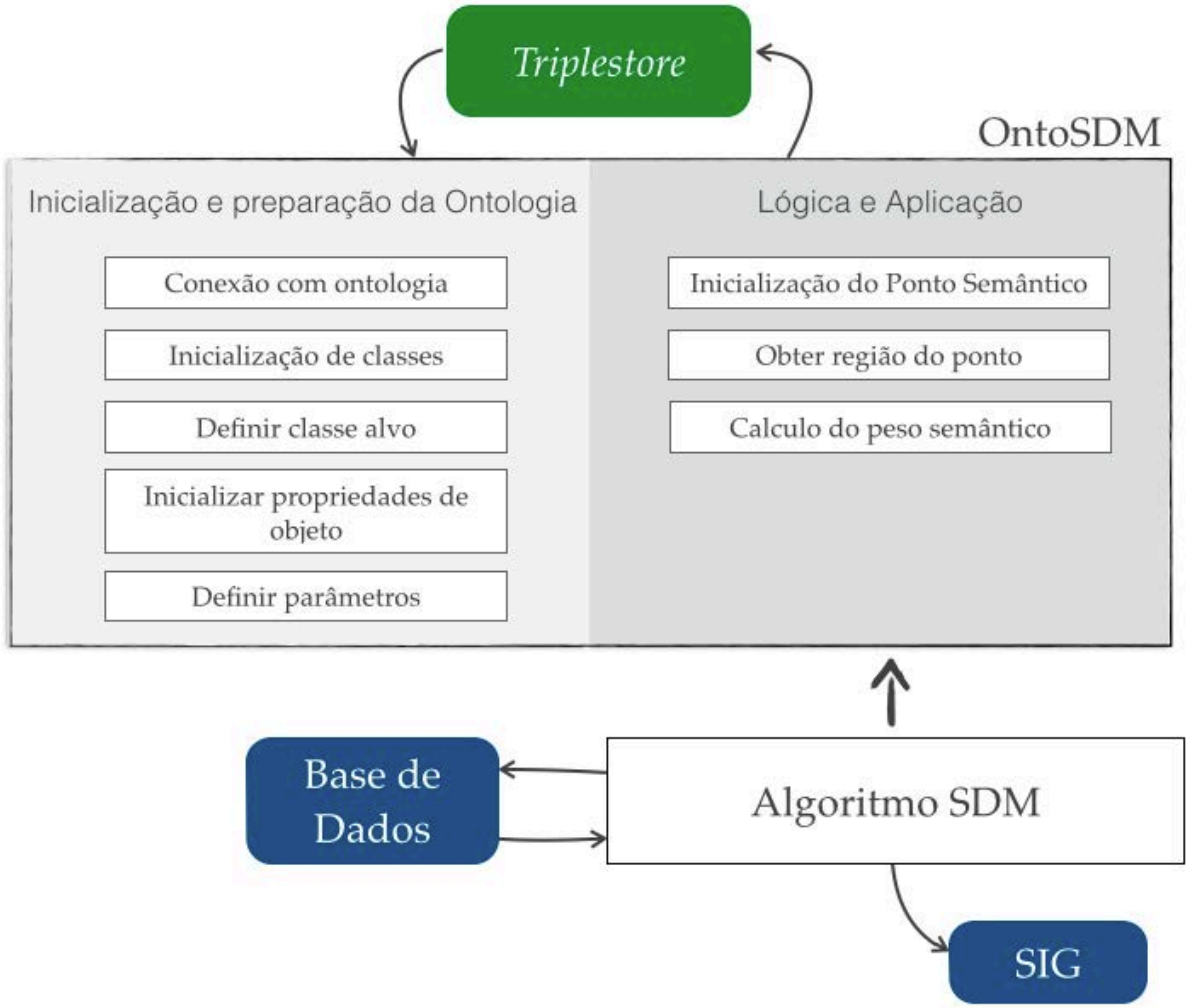
O objetivo é contribuir para que os resultados do algoritmo executado retornem *clusters* mais interessantes, com isso facilita-se a análise e a interpretação dos resultados.

3.2. ARQUITETURA PROPOSTA

Desde o início, a intenção é propor uma arquitetura que possa tornar adaptável a maioria dos algoritmos atuais de *data mining* espacial, sendo necessário efetuar poucas alterações no algoritmo original para que o mesmo consiga adaptar-se e utilizar a ontologia desejada por meio do *framework* desenvolvido.

Além da portabilidade para qualquer algoritmo, também é importante que a solução consiga adaptar-se a qualquer ontologia de domínio desejada, desde que a mesma possua algumas informações específicas para que o *framework* consiga comunicar-se corretamente. Na Figura 12 está representada de forma simplificada a arquitetura desenvolvida.

Figura 12: Arquitetura da abordagem OntoSDM



Fonte: Do autor

Como ilustrado na Figura 12, a abordagem OntoSDM se comunica com o algoritmo SDM base, de modo a estender suas funcionalidades. O *framework* é organizado em duas etapas, a primeira é responsável pela inicialização e preparação da ontologia e a segunda é responsável pela Lógica e Aplicação. Como detalhado anteriormente, a *TripleStore* é responsável por armazenar triplas *RDF* e possibilitar que a mesma possa ser interpretada como uma ontologia e consultada utilizando a linguagem *SPARQL*.

Portanto, a primeira etapa é responsável por efetuar a conexão com a *TripleStore*, tratar possíveis erros de conexão e obter as classes e propriedades de objetos existentes na ontologia. Devido ao fato do *framework* ser bastante genérico, após a inicialização das classes é necessário o usuário definir todos os parâmetros de configurações, como por exemplo, realizar o mapeamento da ontologia com a base relacional, para isso deve-se informar quais classes representam as tabelas da base de dados, indicar a classe com a informação

geográfica, definir quais propriedades de atributos que devem ser consideradas e seus respectivos pesos que devem ser levados em consideração pelo algoritmo.

A segunda etapa, Aplicação e Lógica, ocorre durante a execução do algoritmo SDM base. Sempre que um ponto é analisado, sua região é pesquisada na ontologia e suas informações semânticas são retornadas, portanto é possível efetuar o cálculo do peso semântico, que é utilizado posteriormente pelo algoritmo de prospecção de dados espaciais. Uma vez que o algoritmo possui a informação do peso semântico, pode-se adotar diferentes estratégias referente a como essa informação será utilizada.

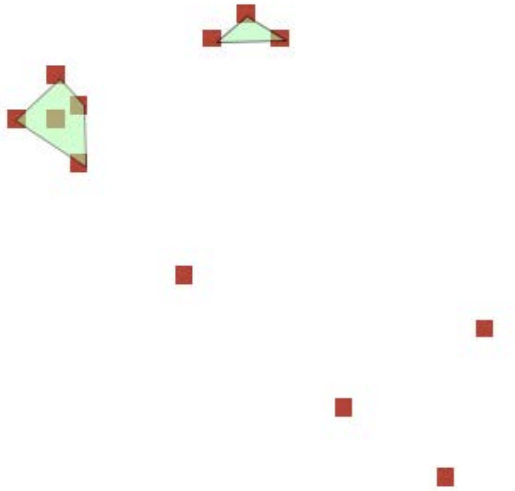
Na próxima seção é apresentada com mais detalhes cada uma das etapas da abordagem descrita ateriormente.

3.3. FRAMEWORK - ONTOSDM

O principal diferencial e vantagem do trabalho apresentado em relação aos demais trabalhos existentes na literatura é o fato da utilização de ontologias para adicionar informação semântica no momento da formação dos *clusters*.

Como já foi descrito, os algoritmos convencionais não fazem uso do suporte semântico e uma das possibilidades para se alcançar um refinamento dos resultados, é com o uso de ontologias para expressar a semântico do domínio, sendo considerado um caminho viável e interessante. Para entender melhor a importância desse tipo de informação, considere um algoritmo qualquer de prospecção de dados espaciais da categoria de agrupamentos. Apenas para ilustração do exemplo, após a etapa de prospecção, os agrupamentos gerados podem ser observados na Figura 13, onde cada ponto representa uma ocorrência e os polígonos verdes representam os agrupamentos gerados.

Figura 13: Exemplo de agrupamento

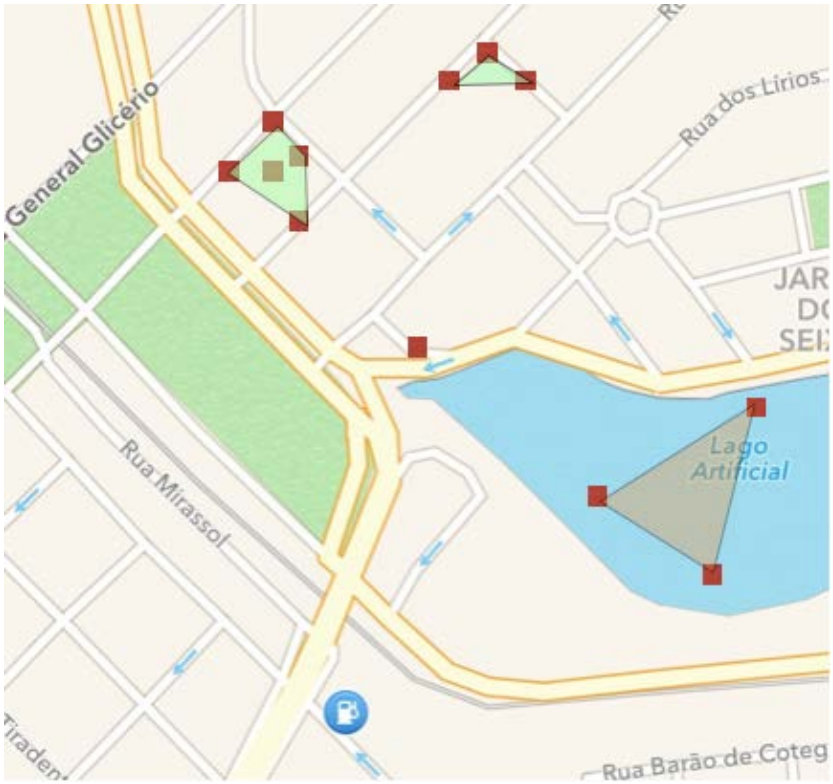


Fonte: Do autor

Os algoritmos atuais são capazes de formar agrupamentos baseados em densidade, ou seja, considerando a distância que cada ponto encontra-se de outro, ou até mesmo da similaridade dos atributos não espaciais, por meio da comparação de quão similares são os atributos não espaciais entre cada um dos pontos. Todavia, deve-se considerar também que os pontos geográficos estão localizados em uma determinada região ou área específica e essa região é relevante para a formação ou não do agrupamento e, portanto, deve ser também considerada. Normalmente, é necessário utilizar um especialista para definir se um determinado agrupamento é realmente interessante ou não, porém essa é uma tarefa manual e bastante custosa.

Ao analisar a Figura 14, na qual os pontos são exibidos sobrepostos de sua posição no mapa, pode-se obter uma interpretação totalmente diferente dos possíveis agrupamentos. Supondo que, por exemplo, os pontos que estão em um lago possam apresentar uma relevância superior e, então devam ser considerados em um agrupamento, ou ainda, pontos que estão em uma região onde um determinado padrão é frequente possam ser desconsiderados. Devido a esse fato, é importante que algoritmos que realizam a prospecção de dados espaciais considerarem também as informações semânticas implícitas em uma determinada área.

Figura 14: Exemplo de agrupamento considerando suas regiões

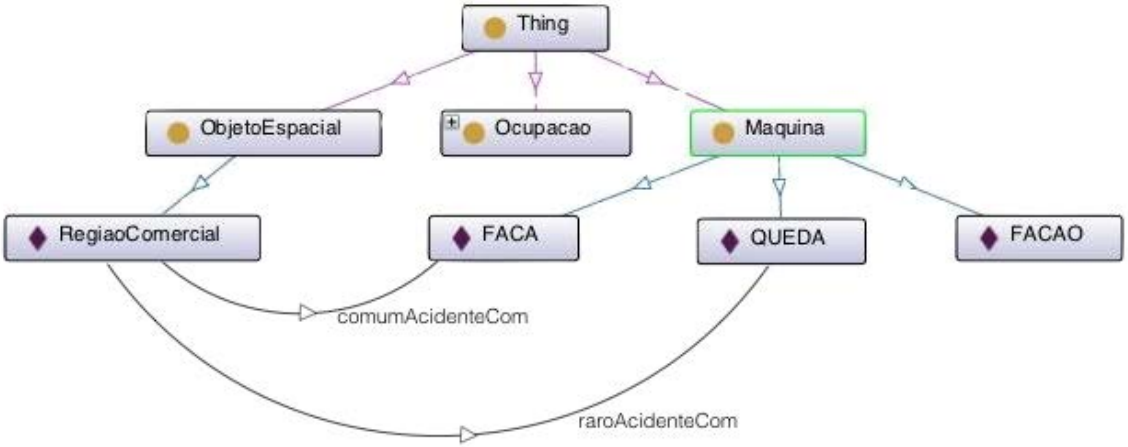


Fonte: Do autor

Com isso, a primeira etapa da abordagem é a definição das diferentes regiões que devem ser consideradas. Essa etapa é definida inteiramente na ontologia de domínio que será utilizada para o suporte ao algoritmo, pois assim adiciona-se a possibilidade de adaptação do algoritmo a qualquer cenário que se deseja aplicar a análise de dados.

Para criar ou adaptar uma ontologia compatível com a ferramenta, é necessário adicionar uma propriedade de dados do tipo *string* que irá armazenar as informações dos múltiplos polígonos que formam a região na classe referente a região da ontologia. Uma vez identificada a classe responsável por armazenar as regiões também é necessário relacionar as instâncias das diferentes regiões com propriedades de objetos, sejam novas ou já existentes na ontologia original. As propriedades de objeto serão utilizadas, posteriormente, para o cálculo do valor semântico de cada ponto. Um modelo de como a ontologia pode ser organizada está disposto na Figura 15.

Figura 15: Representação de um modelo de Ontologia



Fonte: Do autor

Uma vez definida a ontologia deve-se disponibilizar a mesma e ajustar os parâmetros de conexão da ferramenta desenvolvida. No trabalho, foi utilizado a solução *OpenRDF Sesame*[2] para armazenar a ontologia e permitir o acesso e consultas na ontologia criada por meio de bibliotecas *Java*.

Após a definição e configuração da ontologia é possível iniciar a etapa de inicialização e preparação com a utilização do *framework*. Para isso, o usuário deve definir os parâmetros de conexão com a ontologia. Assim que possuir uma conexão com a *triplestore*, a ontologia será consultada para retornar todas as classes e subclasses existentes e então listá-los para o usuário.

O usuário deve então informar qual das classes é a responsável por armazenar a informação referente a região. A ontologia será consultada novamente retornando todas as propriedades de objeto para a classe alvo. Por fim, o usuário deverá informar para cada propriedade de objeto que deseja ser considerado um peso semântico equivalente. Além disso, o usuário deve informar qual será o limite do coeficiente semântico, para que o algoritmo consiga definir se o ponto é interessante ou não, através deste coeficiente de corte.

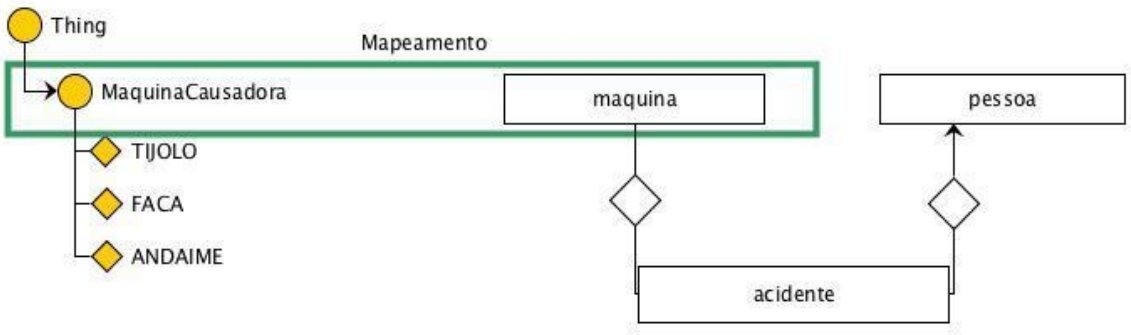
O valor referente aos pesos semânticos variam de -1 a 1, sendo -1 representando propriedades menos interessantes e 1 as mais relevantes. Após definir os pesos para cada propriedade apenas os pontos que possuírem o coeficiente maior do que o limite definido são considerados relevantes.

Caso o usuário deseje, é possível mapear quais classes da ontologia referem-se a quais tabelas da base de dados. Essa etapa é opcional, caso o usuário não realize o mapeamento os

[2] Open RDF – Disponível em < http://www.openrdf.org>. Acessado em 28/04/2014

itens serão pesquisados em toda a ontologia. Essa etapa é importante quando a ontologia é muito extensa, podendo evitar ambiguidade e melhorar o tempo de consulta na ontologia. Na Figura 16 é possível visualizar um modelo do mapeamento entre uma classe da ontologia "*MaquinaCausadora*" com a tabela "*maquina*" de um modelo relacional.

Figura 16: Mapeamento de Classes da Ontologia com Tabelas no Modelo Relacional

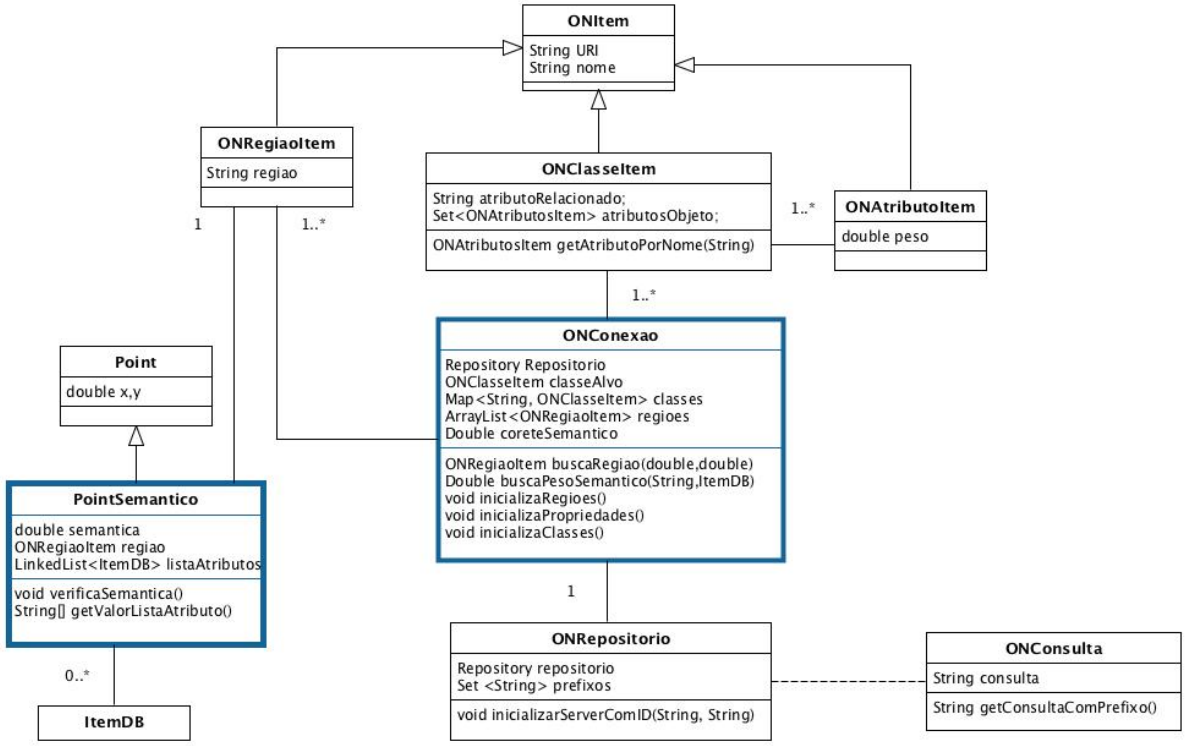


Fonte: Do autor

A etapa seguinte, referente a Lógica e Aplicação, ocorre durante a execução do algoritmo base, onde ao invés do ponto convencional, contendo apenas as informações referentes a sua coordenada, é inicializado o ponto semântico, em que é adicionado um novo atributo referente ao seu peso semântico.

Na Figura 17 é exibido o modelo de classe do *framework* OntoSDM. As duas principais classes estão destacadas, sendo a primeira, ONConexao a responsável por efetuar a conexão com um repositório e realizar as diversas consultas necessárias durante o processo. A outra classe é a *PointSemantico*, sendo a responsável por possuir a informação semântica do ponto e realizar o cálculo do mesmo.

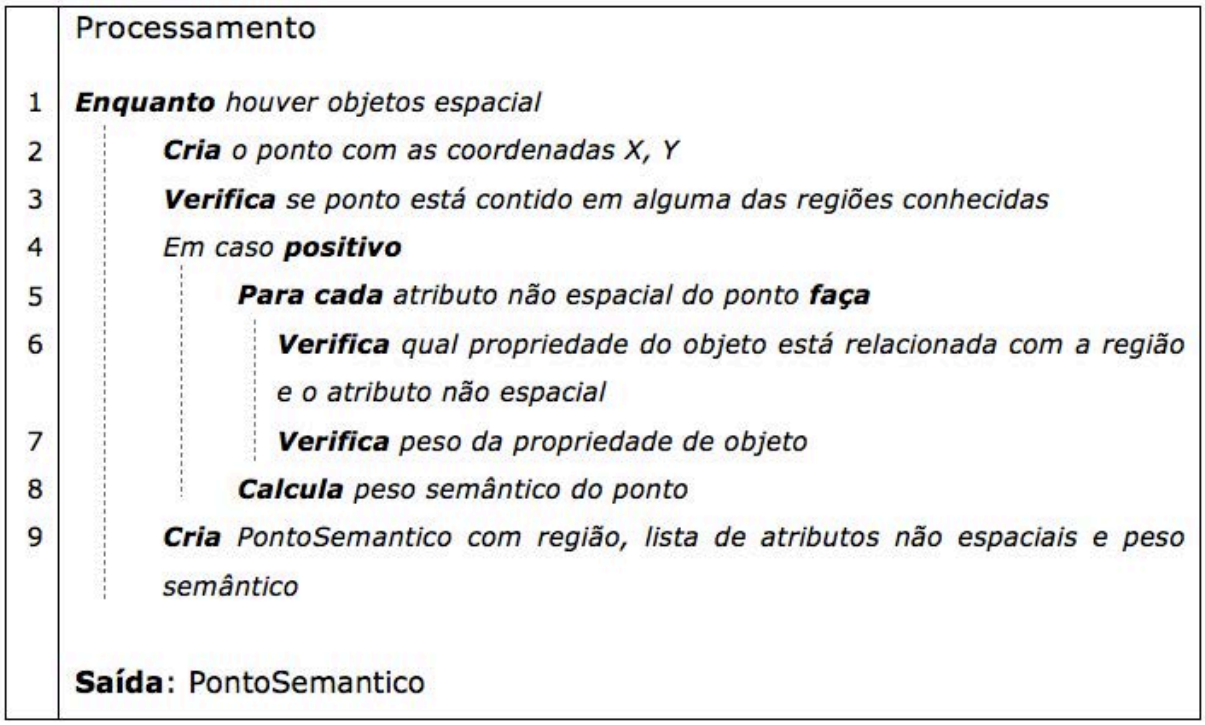
Figura 17: Modelo de Classe – Framework OntoSDM



Fonte: Do autor

Como é possível notar na Figura 17, a classe *PointSemantico* deve ser uma subclasse da classe *Point* e, portanto, herda todos os atributos e métodos já implementados, não alterando o funcionamento do algoritmo original. Deste modo, quando o algoritmo inicializar os pontos consultados na base de dados, irá inicializar o *PointSemantico*, com isso obtêm-se a região na qual o ponto encontra-se e o seu valor semântico para aquela região. O algoritmo completo para a inicialização dos pontos semânticos pode ser conferido na Figura 18.

Figura 18: Algoritmo com etapa de criação do ponto semântico



Fonte: Do autor

Esta é uma das principais etapas do algoritmo, tendo em vista que é a etapa responsável por criar os pontos semânticos.

Inicialmente, o algoritmo irá receber o objeto espacial, ponto, definido na base de dados que é analisado e cria-se o ponto com as coordenadas *x* e *y*. Posteriormente, descrito na linha **3,** a coordenada do ponto é comparada com cada uma das regiões conhecidas. Para isso, foi utilizado a função *ST_Within* do *PostGIS*, que permite verificar se um dado espacial está contido em outro. Como todas as regiões são conhecidas no momento das configurações dos parâmetros, aproveita-se para transferir essa informação contida na ontologia para classes, evita-se assim que a ontologia seja consultada sem necessidades. Com isso, é possível percorrer a lista de regiões e verificar se o ponto está contido em alguma das regiões conhecidas.

Caso o ponto esteja em alguma região conhecida, cada atributo não espacial do mesmo é analisado. Primeiramente, verifica-se na ontologia se a mesma possui alguma informação referente ao atributo em relação a região que o ponto está localizado. A lista de propriedades de objeto é retornada e então é verificado se os pesos para as respectivas propriedades foram definidos pelo usuário. Esta etapa refere-se as linhas **6-8** do algoritmo exibido na Figura 18. Essa etapa só é possível devido à forma de armazenamento das informações na ontologia ser

no formato de triplas, dessa maneira não há a necessidade de possuir um conhecimento prévio do que se está buscando exatamente. Devido a este fato, é possível que a consulta *SPARQL*, exibida na Figura 19, consiga encontrar todas as propriedades de objeto que estejam vinculadas com a região e o item desejado.

Figura 19: Consulta SPARQL responsável por retornar propriedades de objetos desejadas

```
SELECT DISTINCT ?propriedadeObjeto
WHERE {nomeRegiao  ?propriedadeObjeto  item.getNome().
       ?propriedadeObjeto rdf:type owl:ObjectProperty}
```

Fonte: Do autor

Assim que a lista de propriedades de objeto é obtida, é possível realizar o cálculo do valor semântico de acordo com os pesos definidos pelo usuário na etapa de configuração dos parâmetros. O algoritmo irá então verificar o peso de cada propriedade de objeto e calcular o valor semântico para aquele ponto.

O peso semântico varia de -1 a 1, onde o valor -1 representa baixa relevância semântica e o 1 representa alta relevância. As fórmulas para o cálculo do valor semântico estão representadas em (1), (2) e (3). A fórmula (1) indica o valor de **A(p),** onde para um determinado ponto **p** que possui atributos não espaciais $\mathbf{o_i}$ tal que **i = 0...n,** é realizada a somatória dos atributos não espaciais que possuem relevância negativa, o^-**.**

O valor de B(p) representado na fórmula (2) é obtido por meio da somatória dos atributos não espaciais que possuem relevância positiva, $\mathbf{o^+}$, para um determinado ponto **p**. Finalmente, é realizado o cálculo do valor semântico, **S(p)**, para o ponto **p,** representado na fórmula (3).

$$A(p) = \sum_{i=i}^{n} o_i^- \quad (1)$$

$$B(p) = \sum_{i=i}^{n} o_i^+ \quad (2)$$

$$S(p) = \frac{(A(p)+B(p))}{\max\{|A(p)|,B(p),1\}} \quad (3)$$

Aplicando essa abordagem, introduz-se um novo conceito para o ponto geográfico, que pode ser considerado em algoritmos de prospecção espacial, seu valor semântico em relação a

região que o mesmo está localizado. Por fim, o algoritmo já possui todas as informações necessárias, como região que se encontra o ponto, lista de atributos não espaciais e, principalmente, a relevância semântica do ponto para uma determinada região.

Portanto, deve-se agora considerar qual será a abordagem que o algoritmo base irá utilizar em relação à informação do ponto semântico e como irá se comportar, pode-se realizar diferentes estratégias como, desconsiderar o ponto imediatamente, caso o mesmo não seja interessante, ou ainda, utilizar o valor semântico para adicionar novos *clusters*, utilizar o valor semântico junto com outros parâmetros para tomar a decisão de adicioná-lo ou não a um determinado agrupamento entre outras abordagens possíveis.

3.4. ALGORITMOS SDM BASES

Para validar a abordagem utilizada, foi decidido adaptar o algoritmo *MRClustering* (ICHIBA, 2013) e o CHSMST+ (MEDEIROS, 2014) para utilizar o *framework* OntoSDM. Ambos são algoritmos da categoria de agrupamento com base na densidade. As próximas seções detalham as características de cada algoritmo de prospecção espacial.

3.4.1. ALGORITMO *MRCLUSTERING*

O algoritmo *MRClustering* é uma extensão dos algoritmos *CLARANS* e *VDBSCAN* e seu fluxograma geral pode ser conferido na Figura 20.

Um dos diferenciais do algoritmo utilizado é a necessidade de entrada de apenas um parâmetro, grau de similaridade dos atributos não geográficos. Os demais parâmetros necessários para a execução do algoritmo são calculados de forma ótima e automaticamente pelo mesmo, evita-se assim que o usuário tenha que refinar esses valores manualmente. Todo esse processo é efetuado na etapa de inicialização dos parâmetros.

A próxima etapa é a de refinamento dos parâmetros, onde são removidos os pontos isolados, também referenciados como *noises* ou *outliers*.

Por fim, a etapa de agrupamento é iniciada com os parâmetros necessários, incluindo os definidos automaticamente nas duas etapas anteriores - k, número mínimo de objetos em um agrupamento e os valores de *Eps*, que são calculados automaticamente pelo algoritmo e

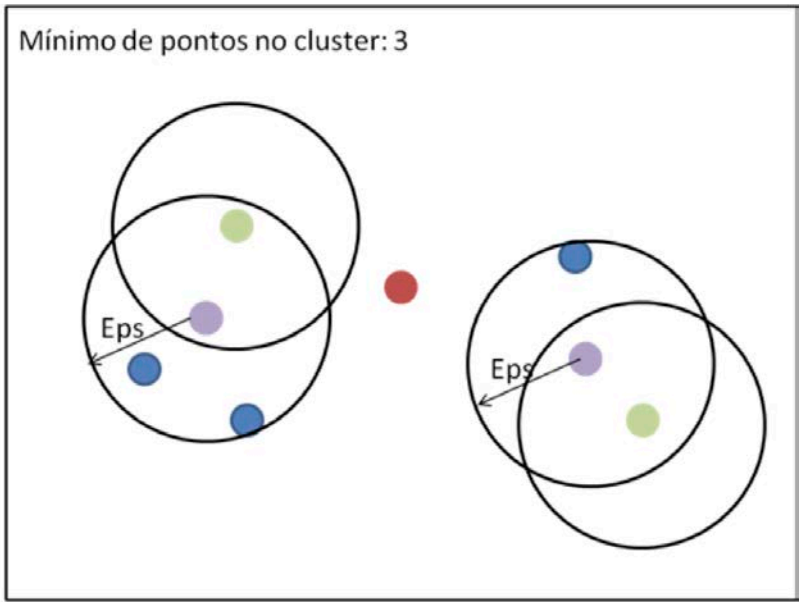
representam os níveis de densidade da base de dados, que determinarão se um ponto é próximo geograficamente de outro.

Figura 20: Fluxograma simplificado do *MRClustering*

Fonte: Retirada de (ICHIBA, 2013)

O algoritmo inicia a formação dos *clusters* com a mesma abordagem do *VDBSCAN*, seleciona-se um ponto central e verifica-se a existência de pontos que estão dentro do limite definido pelo *Eps*. Como diferencial do algoritmo, para cada ponto é analisado a similaridade dos atributos convencionais e só assim o ponto é adicionado no *cluster*. Observa-se na Figura 21 o processo da formação dos *clusters*, onde os pontos roxos representam os pontos centrais, os azuis os pontos alcançáveis, pois estão dentro do raio Eps, os pontos verdes são bordas do agrupamento e por fim os pontos vermelhos são *noises* pois não pertencem a nenhum agrupamento.

Figura 21: Criação de Agrupamentos

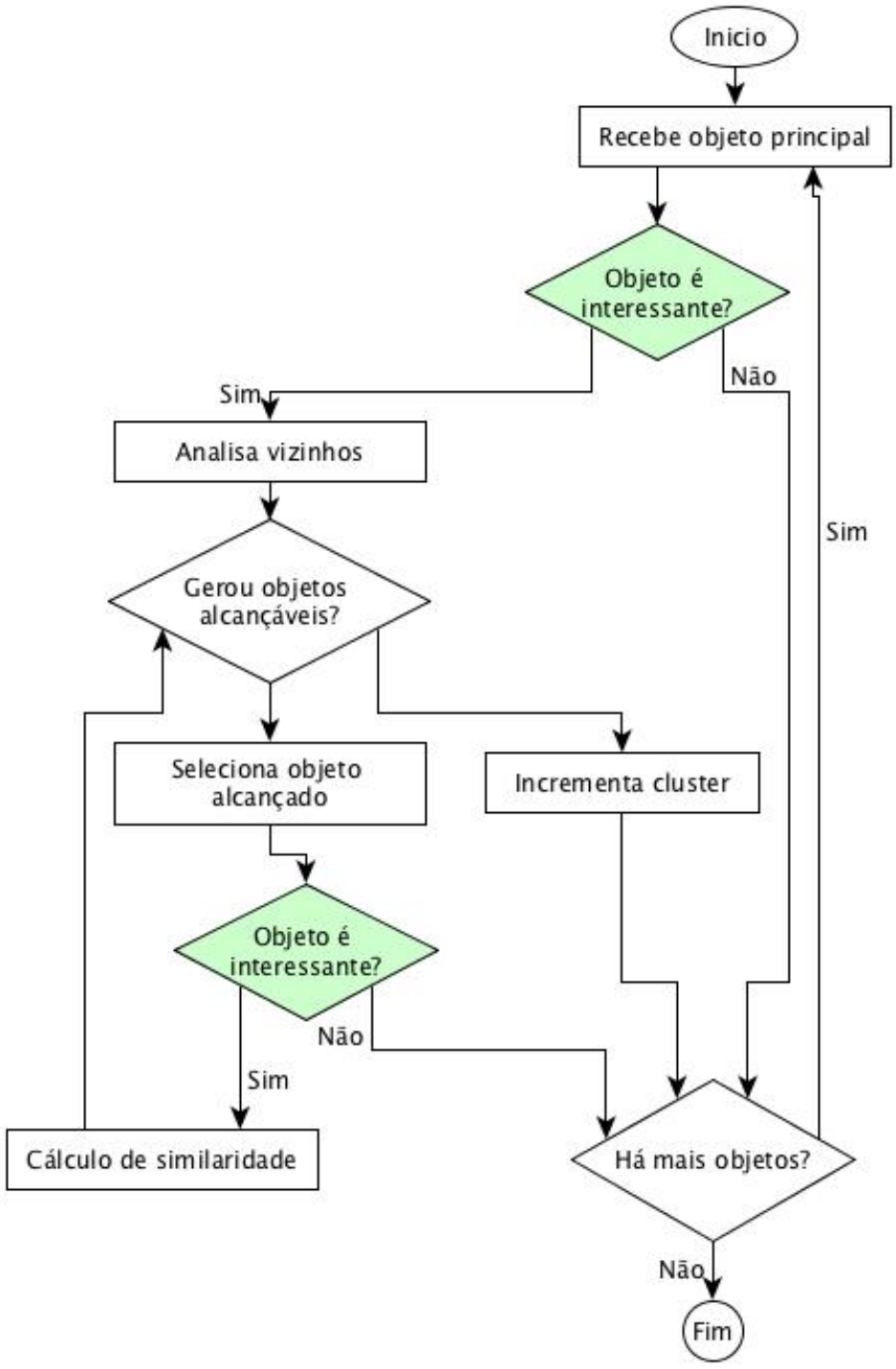


Fonte: Retirada de (ICHIBA, 2013)

No caso do algoritmo *MRClustering* foi decidido efetuar a abordagem OntoSDM no momento do cálculo de distância entre os pontos, visto que o objetivo principal é remover os agrupamentos já conhecidos. Essa abordagem pode reduzir o tempo de processamento, considerando que ao remover o ponto, o mesmo, não será mais verificado no decorrer do algoritmo, evitando assim a etapa de cálculo de similaridade e de densidade.

Como pode ser analisado no diagrama exibido na Figura 22, foi necessário alterar o algoritmo apenas para verificar se, no momento do cálculo de distância entre dois pontos, ambos tenham sua relevância verificada antes pela abordagem. Para isso basta utilizar o método disponibilizado pelo *framework* responsável por verificar o peso semântico do ponto com o corte definido pelo usuário, o mesmo irá retornar se o ponto é interessante para a região. A alteração realizada está destacada na Figura 22.

Figura 22: Fluxograma simplificado com a aplicação da abordagem OntoSDM no algoritmo MRClusting



Fonte: Adaptada de (ICHIBA, 2013)

Além das vantagens já apresentadas que justificam a escolha do algoritmo, o mesmo ainda contempla suporte a bases de dados multirrelacionais, garante-se assim que não ocorra perda de informação ou ainda que os dados possam ser interpretados de forma errônea e possui suporte a tecnologia *multi-thread*, dessa maneira otimiza-se de um modo geral o tempo

de processamento do algoritmo (ICHIBA, 2013).

Como dito anteriormente, caso o ponto não seja considerado interessante semanticamente para a região, será descartado e não será incluído em nenhum dos agrupamentos formados, com isso gera-se novos *clusters* como resultado final. Como os pontos não interessantes serão descartados, espera-se uma quantidade menor de agrupamentos, dessa maneira facilita-se a análise e interpretação dos dados.

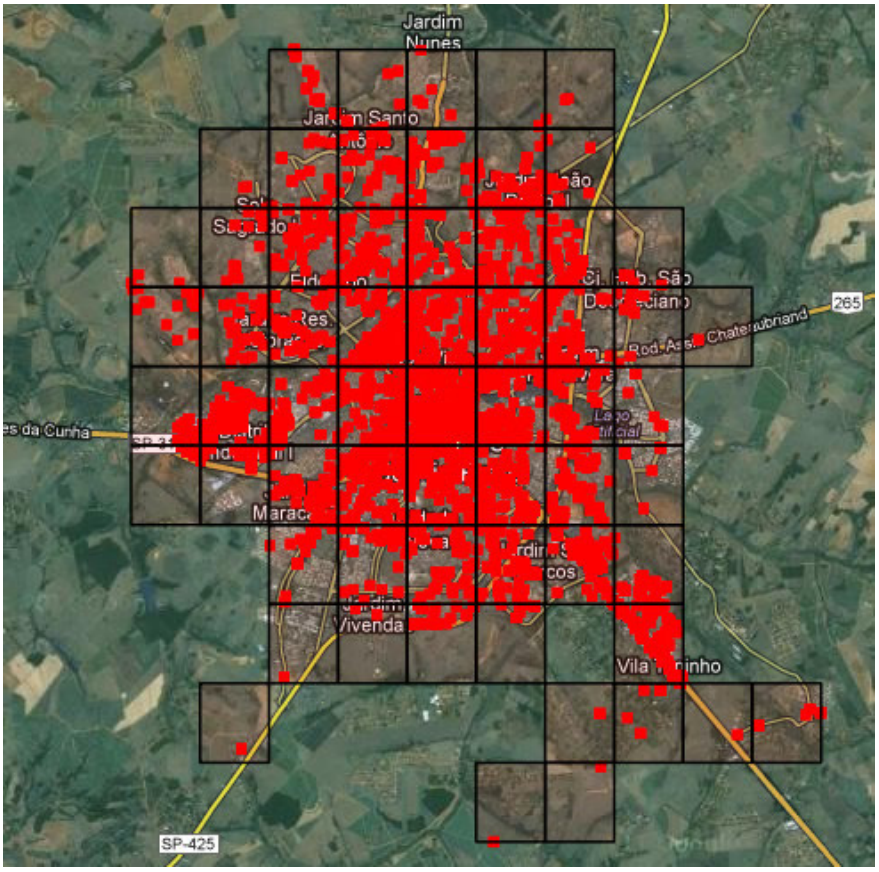
3.4.2. ALGORITMO CHSMST+

O algoritmo CHSMST+ fornece um aprimoramento dos resultados e melhora no desempenho em relação ao algoritmo CHSMST. O diferencial do algoritmo CHSMST é efetuar a divisão do conjunto de dados espaciais em sub-regiões, como está representado na Figura 23, para depois realizar a etapa de agrupamento. As regiões são subdividas de acordo com um parâmetro informado pelo usuário. Uma das melhorias do algoritmo CHSMST+ é a eliminação deste parâmetro por meio do cálculo de distância global e local de cada unidade de particionamento.

Após o cálculo das unidades, para cada, é criado o grafo de distâncias dos pontos contidos na mesma. O grafo possui como peso das arestas a distância que um ponto está de outro e o índice de similaridade entre os pontos.

Os agrupamentos são então formados por meio da poda das arestas do grafo onde são eliminados os pontos com maior distância e cuja a similaridade é inferior ao limite estipulado, formando assim os agrupamentos.

Figura 23: Divisão em sub-região CHSMST



Fonte: Retirada de (MEDEIROS, 2014)

Para adaptar o algoritmo ao *framework* é necessário efetuar o mesmo procedimento, bastando estender a classe “ponto” do algoritmo para utilizar a classe referente ao ponto semântico disponibilizada pelo *framework*. Após isso, basta decidir qual será a estratégia utilizada e fazer as alterações no algoritmo.

3.5. AMBIENTE DE GERENCIAMENTO DE ALGORITMOS

Além do *framework* desenvolvido, como um subproduto do trabalho desenvolvido, foi confeccionada uma ferramenta responsável por gerenciar os algoritmos de prospecção de dados desenvolvidos.

O objetivo da ferramenta, é disponibilizar um ambiente único e centralizado capaz de suportar e gerenciar os algoritmos desenvolvidos até o momento no grupo de pesquisa e também os algoritmos futuros. Com isso os próximos trabalhos poderão reutilizar componentes da ferramenta desenvolvida, permitindo que esforços despendidos para realizar

tarefas triviais e comuns entre os algoritmos não sejam refeitos, como por exemplo, conexão com a base de dados, seleção de atributos e consolidação dos resultados.

A ferramenta necessita de um arquivo de configuração que irá definir os algoritmos e suas etapas, dentre elas, etapa de conexão com a base de dados, conexão com ontologia, seleção de atributos, execução do algoritmo e consolidação de resultados. Novas etapas podem ser criadas e definidas sem dificuldade. A figura 24 contem um exemplo do arquivo de configuração.

Figura 24: Arquivo de configuração

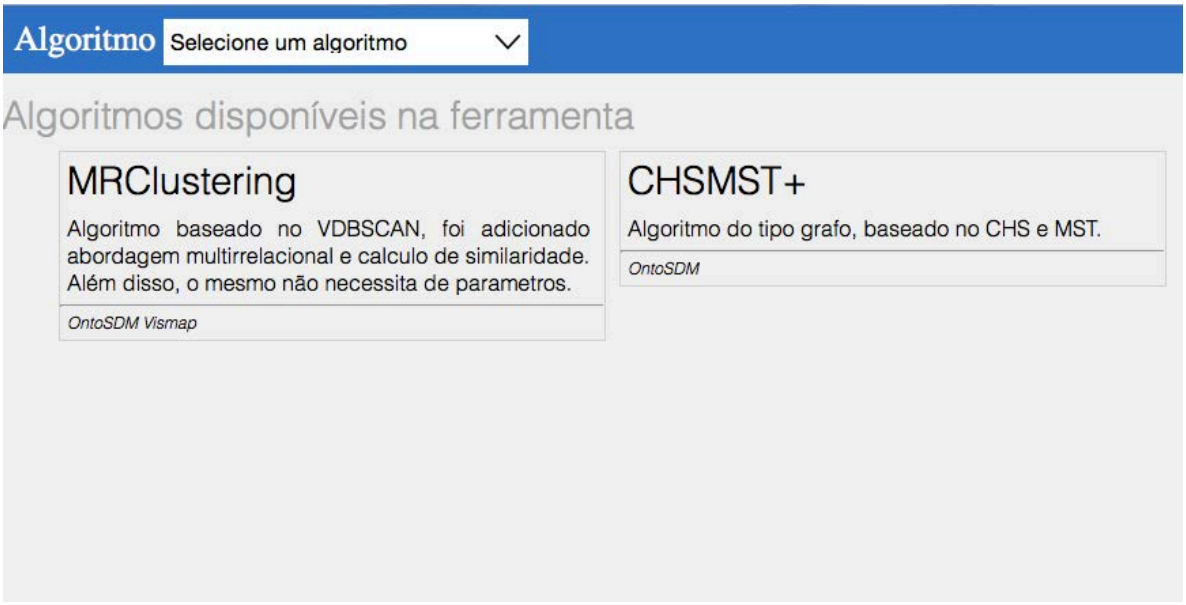
```
<?xml version="1.0" encoding="UTF-8"?>
<algoritmos>
    <algoritmo classe="com.unesp.gbd.Algoritmo.MRClustering">
        <nome>MRClustering</nome>
        <descricao>Algoritmo baseado no VDBSCAN, foi adicionado abordagem multirrelacional e
calculo de similaridade. Além disso, o mesmo não necessita de parâmetros.</descricao>
        <etapas>
            <etapa view="conectarDB" ontologia="true">Conexão com Base e Ontologia</etapa>
            <etapa view="selecaoAtributos">Seleção de Atributos</etapa>
            <etapa view="executarAlgoritmo">Execução do algoritmo</etapa>
            <etapa view="consolidaResultado">Consolidar Resultados</etapa>
        </etapas>
        <suporta>
            <func>OntoSDM</func>
            <func>Vismap</func>
        </suporta>
    </algoritmo>

    <algoritmo classe="com.unesp.gbd.Algoritmo.CHSMST+">
        <nome>CHSMST+</nome>
        <descricao>Algoritmo do tipo grafo, baseado no CHS e MST.</descricao>
        <etapas>
            <etapa view="conectarDB">Conexão com base de dados</etapa>
            <etapa view="naoImplementado">Definição de Parâmetros</etapa>
            <etapa view="consolidaResultado">Consolidar Resultados</etapa>
        </etapas>
        <suporta>
            <func>OntoSDM</func>

        </suporta>
    </algoritmo>
```

Após adicionar as informações do algoritmo no arquivo de configuração o mesmo estará disponível para seleção e será configurado corretamente de acordo com cada etapa que o mesmo possui, como pode ser visto na figura 25.

Figura 25: Seleção de algoritmo



Selecionado um algoritmo, a ferramenta será configurada para respeitar as etapas definidas, podendo reaproveitar etapas que já foram implementadas anteriormente, como por exemplo conexão com a base e consolidação dos resultados, que são demostradas nas Figuras 26 e 27, respectivamente.

Figura 26: Conexão com base de dados e ontologia

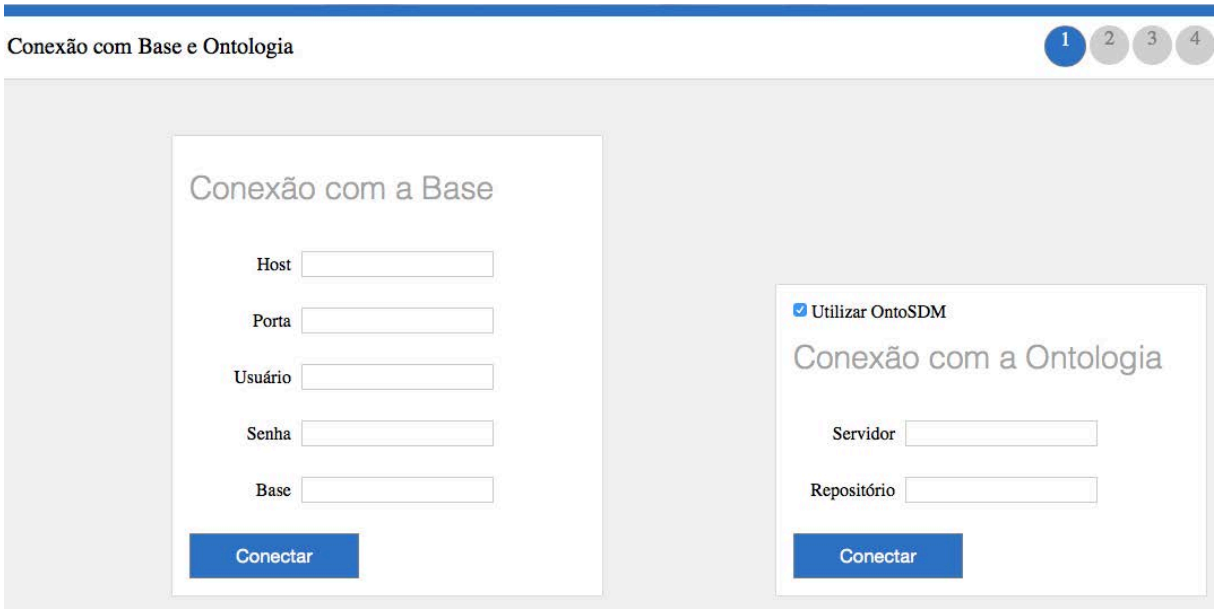


Figura 27: Etapa de consolidação de dados

3.6. CONSIDERAÇÕES FINAIS

Neste capítulo foi apresentado o algoritmo de prospecção de dados espaciais utilizado para validação da abordagem, foi explicado as principais etapas do mesmo e os motivos que justificam sua escolha.

Além disto, foi detalhada a arquitetura desenvolvida deixando claro sua particularidade de ser configurável e portável para outros algoritmos de prospecção de dados e também a qualquer ontologia, desde que algumas adaptações sejam realizadas.

Para finalizar, cada uma das principais etapas da abordagem foram detalhadas, exemplificando cada parte do algoritmo, que incluem a configuração dos parâmetros e a aplicação da lógica do mesmo.

4. Testes e Resultados

4.1. CONSIDERAÇÕES INICIAIS

Neste capítulo são apresentados os experimentos e resultados obtidos por meio da aplicação da abordagem desenvolvida por meio do *framework* OntoSDM. Para validar a abordagem a mesma foi implementada em dois algoritmos de agrupamento de dados, o *MRClustering* e *CHSMST+*.

A intenção é comprovar sua eficiência em relação aos resultados obtidos quando comparados com os algoritmos originais e, além disto, demostrar que a abordagem é configurável e adaptável.

4.2. AMBIENTE DE TESTE

Para validar a eficácia da abordagem proposta foram realizados uma série de experimentos e testes com o intuito de verificar dois pontos importantes: qualidade do agrupamento formado e impacto de desempenho no algoritmo original.

Os testes foram realizados em bases de dados reais. A execução em bases reais é importante pois os resultados obtidos serão qualitativos e ao utilizar dados reais a análise se torna mais prática.

A primeira base refere-se ao Sistema de Vigilância de Acidentes de Trabalho, *SIVAT*, o qual é responsável por armazenar casos de acidentes de trabalho, sejam de trajeto, na empresa ou em locais de terceiros. O sistema contêm dados de acidentes de trabalhos de São José do Rio Preto e região, totalizando mais de 100 municípios atendidas. A base de dados contêm mais de 100 mil ocorrências de acidentes de trabalho, sendo dessas 30 mil georreferenciadas.

A base de dados do SIVAT possui diversas informações georreferenciados, como o local do acidente, empresa de trabalho e domicílio do acidentado. Foi considerado nos testes apenas o local do acidente e alguns atributos não espaciais, como a máquina causadora do acidente, ocupação e escolaridade do acidentado.

Foram criados 4 conjuntos de testes, variando a quantidade de pontos pertencente em cada. Os pontos foram adicionados de forma aleatório e selecionados apenas os acidentes que ocorreram na cidade de São José do Rio Preto. A quantidade de pontos distribuída em cada conjunto pode ser consultada na Tabela 3.

Tabela 3: Conjuntos de pontos

	C1	C2	C3	C4
Quantidade de Pontos	3 mil	5 mil	10 mil	25 mil

Para a análise dos resultados é importante ressaltar que um mesmo ponto no mapa pode representar várias ocorrências de acidentes, isso pode ocorrer pois caso o local do acidente seja reincidente um ponto ficará sobreposto a outro.

A outra base utilizada para testes refere-se a ao sistema Gerenciador de Recursos Hídricos - GRH. O sistema é responsável por coletar informações referente a recursos hídricos em propriedades na região do rio Marinheirinho. São coletadas informações referentes a qualidade do recurso hídrico, seja um poço, um barramento ou um fio d'agua, assim como informações referentes a propriedade, como se existe cultivo, quantos moradores, renda familiar etc.

A base do GRH possui cerca de 650 recursos hídricos georreferenciados e todos foram considerados durante a etapa de teste. Diferente da base do SIVAT cada ponto representa exclusivamente um recurso hídrico, pois não é possível que dois recursos estejam em uma mesma coordenada.

Em relação as ontologias, foram definidas duas ontologias de domínio para cada uma das bases de dados. Para armazenar as ontologias foi utilizado o OpenRDF Sesame, onde foi configurado um repositório OWLIM-lite de modo a permitir realizar inferências e ser considerado um repositório semântico robusto.

As ontologias foram confeccionadas com intuito de apenas validar a abordagem proposta e não representam totalmente o domínio. A ontologia referente ao SIVAT, relacionada aos acidentes de trabalhos, possuí a seguinte estrutura:

- 4 regiões (Região comercial, residencial, industrial e hospitais);
- 50 máquinas causadoras;
- 50 ocupações;
- 2 níveis de escolaridade;
- 3 propriedades de objeto relacionando as regiões com as demais classes da

ontologia. Sendo responsáveis por informar a frequência em que ocorrem, podendo ser frequente, comum ou raro.

Por sua vez, a ontologia referente ao GRH, recursos hídricos, possui a seguinte estrutura:

- 2 regiões (Área urbana e rural);
- 5 níveis de escolaridade;
- 6 tipos de situação administrativa;
- 2 propriedades de objeto relacionando as regiões com as demais classes. As propriedades são responsáveis por indicar a ocorrência ou não dos itens para uma determinada região.

Para finalizar, em relação aos testes quanto à qualidade do agrupamento, serão aplicados métodos intrínsecos para mensurar a compactação de cada agrupamento e sua dissimilaridade em relação aos demais agrupamentos (HAN; KAMBER, 2011).

Existem diversos métodos intrínsecos na literatura, porém um dos mais utilizados é o coeficiente de silhueta que consiste em calcular o coeficiente entre a média da distância de um objeto para os demais pertencentes ao mesmo agrupamento e a distância média mínima do objeto para todos os outros agrupamentos. Portanto, o coeficiente **s(o)** será calculado a partir de um objeto **o**, para isso, deve-se calcular primeiramente o valor de **a(o**), que representa a média da distância de o para os demais objetos pertencentes ao agrupamento $\mathbf{C_i}$, $\mathbf{C_i} \in \mathbf{C} = \{\mathbf{C_1}, \mathbf{C_2}, \mathbf{C_3}, ..., \mathbf{C_k}\}$. Posteriormente é calculado a dissimilaridade do agrupamento $\mathbf{C_i}$ para os demais através de **b(o),** onde as fórmulas estão dispostas a seguir.

$$\boldsymbol{a(o)} = \frac{\sum_{o' \in Ci\ o \neq o'} \boldsymbol{dist(o,o')}}{|\boldsymbol{C_i}| - \mathbf{1}} \tag{4}$$

$$\boldsymbol{b(o)} = min_{C_j : 1 \leq j \leq k, i \neq j} \left\{ \frac{\sum_{o' \in Cj} \boldsymbol{dist(o,o')}}{|\boldsymbol{C_j}|} \right\} \tag{5}$$

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}} \tag{6}$$

O coeficiente da silhuetada dado pela equação (6) é responsável por determinar a qualidade do agrupamento. Os valores do coeficiente variam entre -1 e 1. Quando o valor de **s(o)** é negativo significa que o agrupamento $\mathbf{C_i}$ é similar aos demais agrupamentos obtidos, enquanto que se o valor de **s(o)** é próximo de 1 significa que o agrupamento $\mathbf{C_i}$ é compacto e dissimilar em relação aos outros agrupamentos e, portanto, a qualidade do agrupamento neste caso é alta (HAN; KAMBER, 2011).

4.3. EXPERIMENTOS MRCLUSTERING

A próxima seção apresenta os experimentos que foram realizados no algoritmo de prospecção de dados espaciais MRClustering.

4.3.1. QUANTIDADE DE AGRUPAMENTOS

Os primeiros experimentos foram executados com o intuito de verificar a quantidade de agrupamentos que são desconsiderados ao utilizar a abordagem desenvolvida. Para isso, o algoritmo *MRClustering* foi executado em cada um dos conjuntos de dados definidos na Tabela 3. O mesmo foi executado sem utilizar e utilizando a abordagem e depois a quantidade total de agrupamentos foi comparada.

Devido a estratégia utilizada, a de desconsiderar os pontos não relevantes, esperava-se uma redução na quantidade de agrupamentos gerados. Essa redução propicia melhores resultados finais facilitando a análise posterior do resultado por um analista. Os primeiros experimentos em relação a redução de agrupamentos foi feita ao utilizar apenas os atributos não espaciais referente a Máquina Causadora e Ocupação.

A redução de agrupamentos alcançou índices de quase 25% com uma média de 19% para todos os conjuntos de dados, no qual a quantidade exata de agrupamentos esta disponibilizada na Tabela 4.

Tabela 4: Comparação da quantidade de agrupamentos gerados - 2 Atributos

	C1	C2	C3	C4
MRClustering	142	198	492	1247
MRClustering & OntoSDM	107	150	409	1075

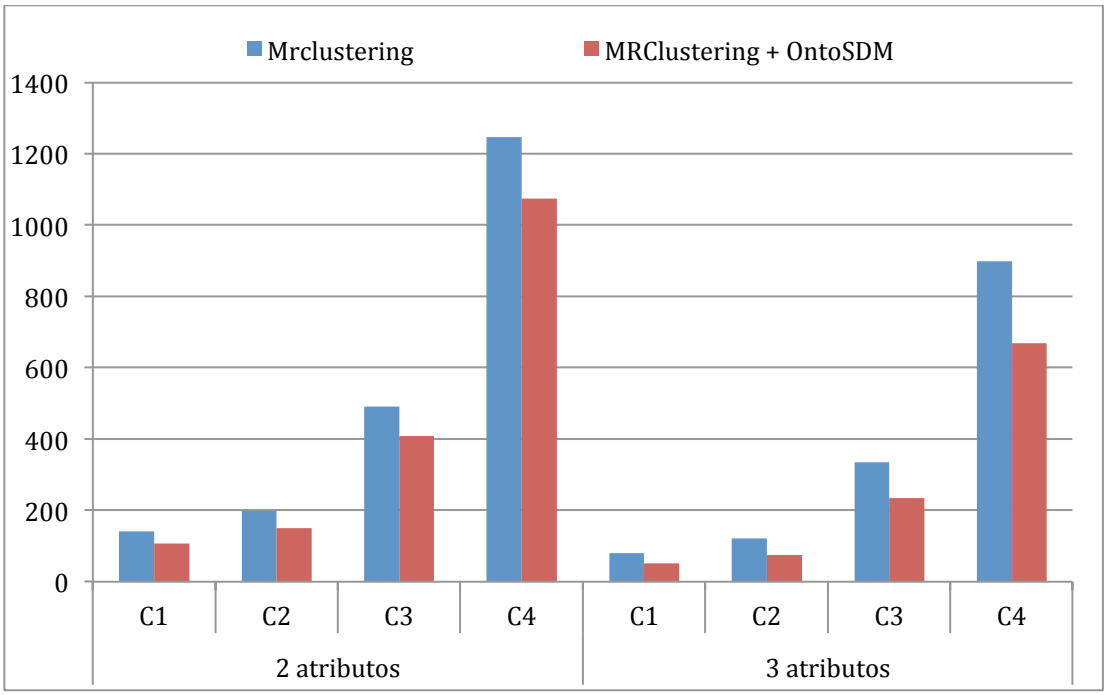
Ao adicionar um terceiro item para a análise, o nível de escolaridade, foi possível perceber um aumento ainda maior na redução da quantidade de agrupamentos, obteve-se uma redução máxima 38% e em média de 32%. A quantidade de agrupamentos gerados ao considerar os três itens (máquina, ocupação e escolaridade) pode ser observado na Tabela 5.

Tabela 5: Comparação da quantidade de agrupamentos gerados - 3 Atributos

	C1	C2	C3	C4
MRClustering	80	121	335	899
MRClustering & OntoSDM	52	74	234	669

Com isso evita-se um trabalho extra do analista, que necessitaria verificar cada um desses agrupamentos para que apenas posteriormente os mesmos pudessem ser removidos por não serem relevantes para a análise. O gráfico exibido na Figura 28 ilustra a diferença dos resultados obtidos em ambos os experimentos.

Figura 28: Gráfico com comparação de agrupamentos gerados



Fonte: Do autor

Além disso, como a estratégia utilizada desconsidera o ponto e não o agrupamento como um todo, agrupamentos diferentes são gerados, evita-se assim que esses pontos não relevantes interfiram no resultado, o que pode levar a uma compreensão equivocada dos agrupamentos.

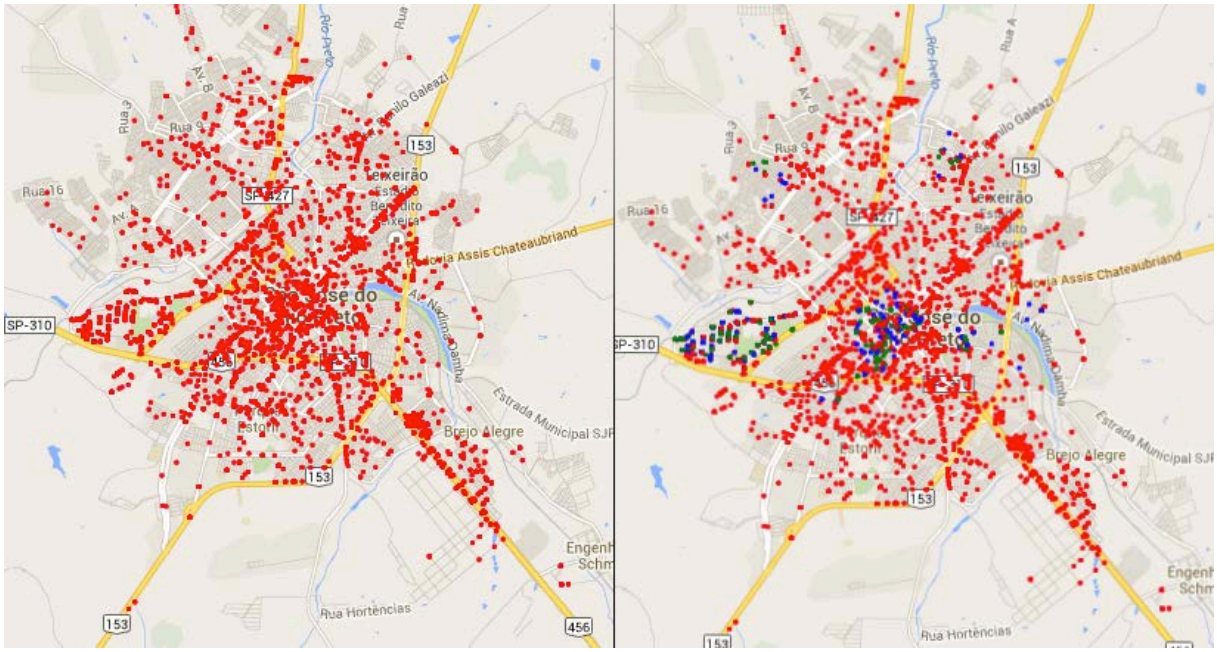
4.3.2. QUALIDADE DOS AGRUPAMENTOS

Os próximos experimentos tem como objetivo demonstrar o ganho de qualidade nos agrupamentos gerados e em melhorias que podem ser adicionadas aos algoritmos para aprimorar e facilitar a compreensão dos dados.

Para estes experimentos foi utilizado o conjunto de dados 3 (C3, ver Tabela 3). A distribuição dos pontos pode ser observada na Figura 29. A esquerda é exibido o resultado obtido sem a abordagem e a direita reflete a aplicação com a abordagem OntoSDM. Como é possível perceber na imagem, ao adicionar a relevância semântica nos pontos é possível enriquecer o resultado destacando aqueles pontos que são mais ou menos relevantes.

Na Figura 29 os pontos vermelhos representam ocorrências onde a relevância é considerada normal, os pontos azuis são os pontos irrelevantes e os pontos verdes são considerados com relevância alta.

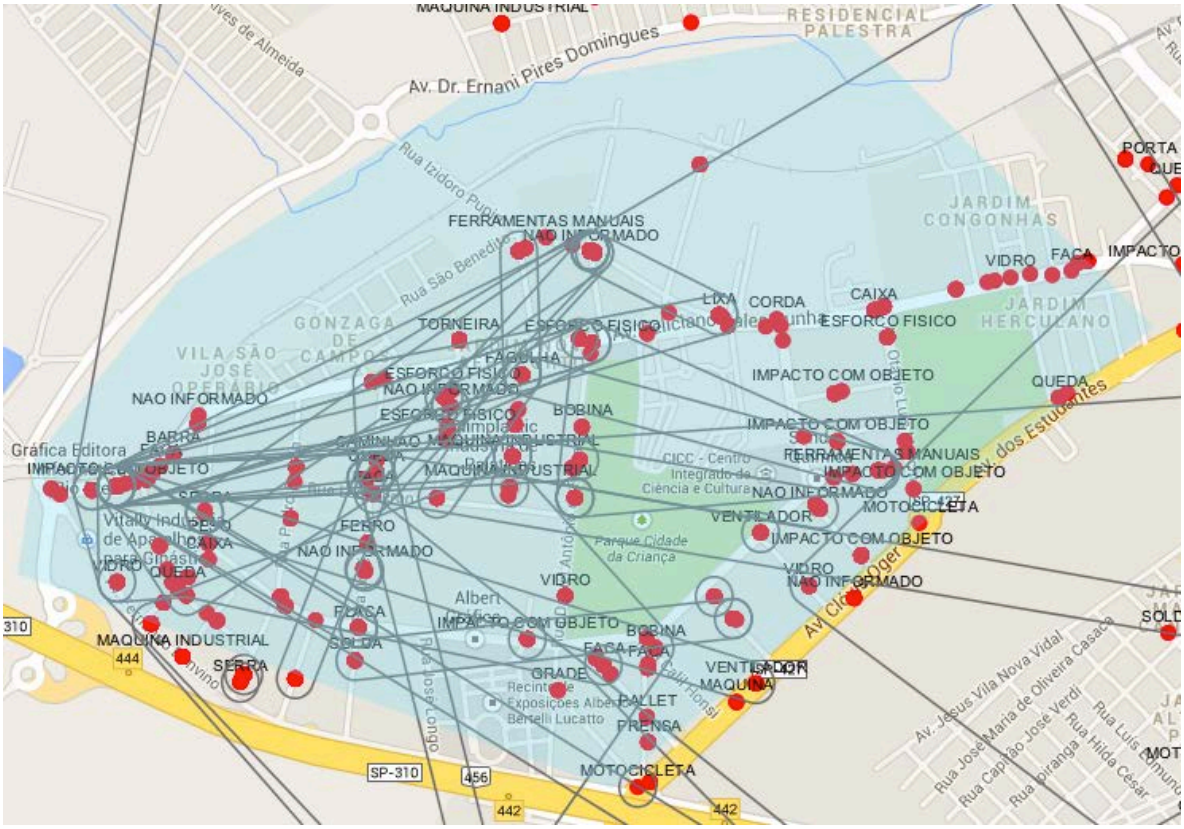
Figura 29: Comparação da distribuição dos pontos sem e com a abordagem



Fonte: Do autor

Ao analisar a região industrial que está destacada na Figura 30, pode-se notar o alto número de agrupamentos gerados, totalizou-se 151 agrupamentos apenas nesta região. Porém ao analisar os agrupamentos gerados utilizando a abordagem OntoSDM, Figura 31, na mesma região, é possível notar uma redução significativa dos agrupamentos, com isso, facilita-se a análise do especialista.

Figura 30: Agrupamentos da Região Industrial



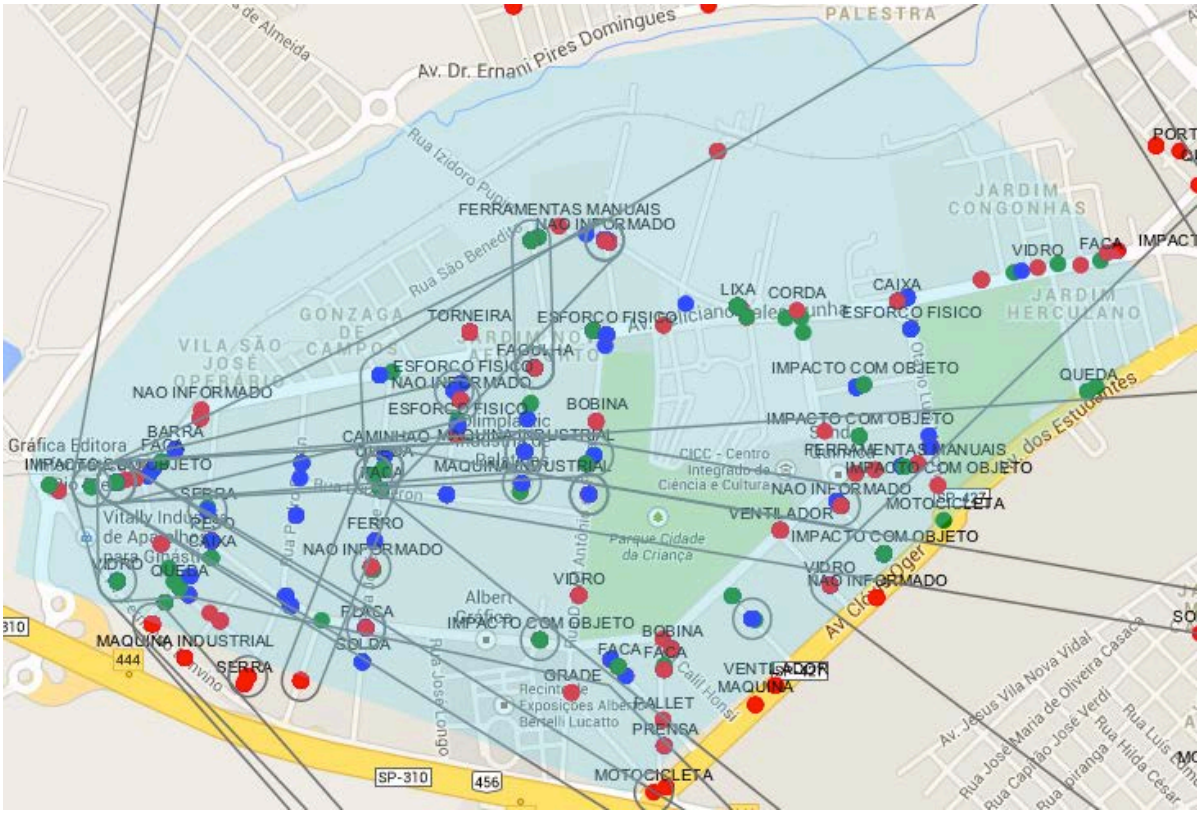
Fonte: Do autor

Após a aplicação da abordagem foi gerado um total de 84 agrupamentos. Com o intuito de validar que apenas agrupamentos considerados triviais foram desconsiderados, foi realizada uma comparação entre as características dos pontos de cada agrupamento. Com isso, foi possível perceber os seguintes fatos:

- 9 dos agrupamentos removidos envolviam a máquina causadora “Máquina Industrial”;
- 3 agrupamentos continham a máquina causadora “Serra”;
- 5 agrupamentos onde a ocupação é “Operador de Máquina Pesada”;
- 12 agrupamentos a ocupação é “Soldador”.

Todas essas características são consideradas comuns para a região industrial, o que causa um processamento desnecessário durante a criação dos agrupamentos.

Figura 31: Agrupamentos da Região Industrial utilizando abordagem OntoSDM



Fonte: Do autor

Ao analisar o coeficiente de silhueta dos agrupamentos gerados tanto com e sem a abordagem, notou-se que não houve grandes mudanças nos valores gerados, um resultado melhor foi apresentado ao utilizar a abordagem, porém a diferença não pode ser considerada relevante. Este fato pode ser explicado pois os cálculos utilizados também não consideram a semântica do agrupamento gerado, ao analisar apenas a distância entre os pontos e a similaridade dos atributos entre os agrupamentos.

No caso deste experimento, considerou-se os agrupamentos da região industrial, obteve-se os seguintes valores descritos na Tabela 6.

Tabela 6: Comparativo com os coeficientes de silhueta - Região Industrial

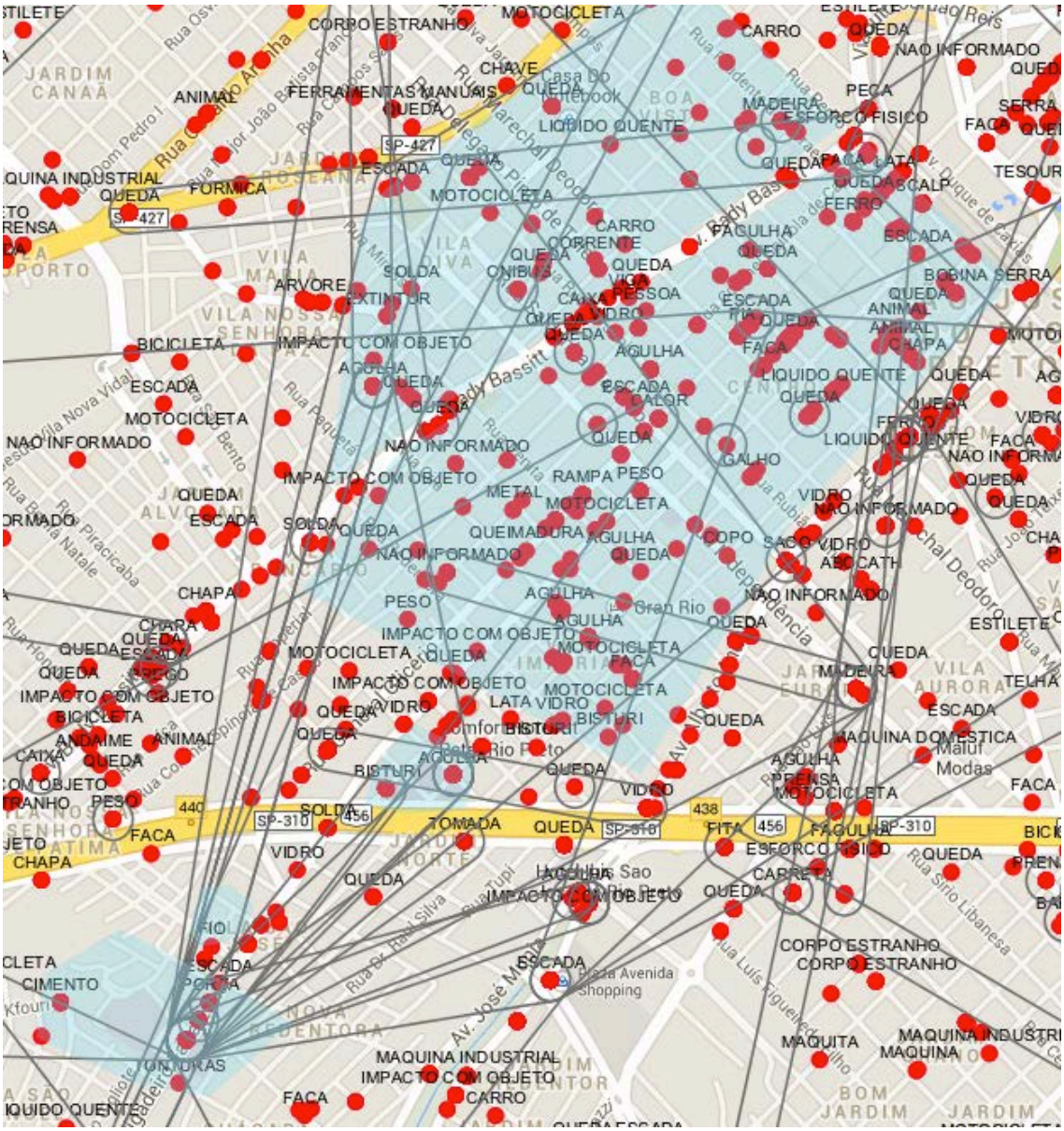
	Similaridade		Distância	
	Menor	Média	Menor	Média
MRClustering	0.85	0.96	0.98	0.99
MRClustering & OntoSDM	0.91	0.97	0.99	0.99

Outra região também analisada refere-se a região central da cidade de São José do Rio Preto, onde existem regiões definidas na ontologia que se referem a região comercial e região de hospitais, que estão destacadas de azul. Na Figura 32 são exibidos os agrupamentos da

região sem considerar a semântica implícita nas regiões.

Como é possível perceber, existe uma grande quantidade de agrupamentos que dificultam a compreensão por parte do usuário. Por sua vez, quando se analisa a Figura 33 nota-se uma quantidade reduzida de agrupamentos

Figura 32: Agrupamentos gerados na região de hospitais



Fonte: Do autor

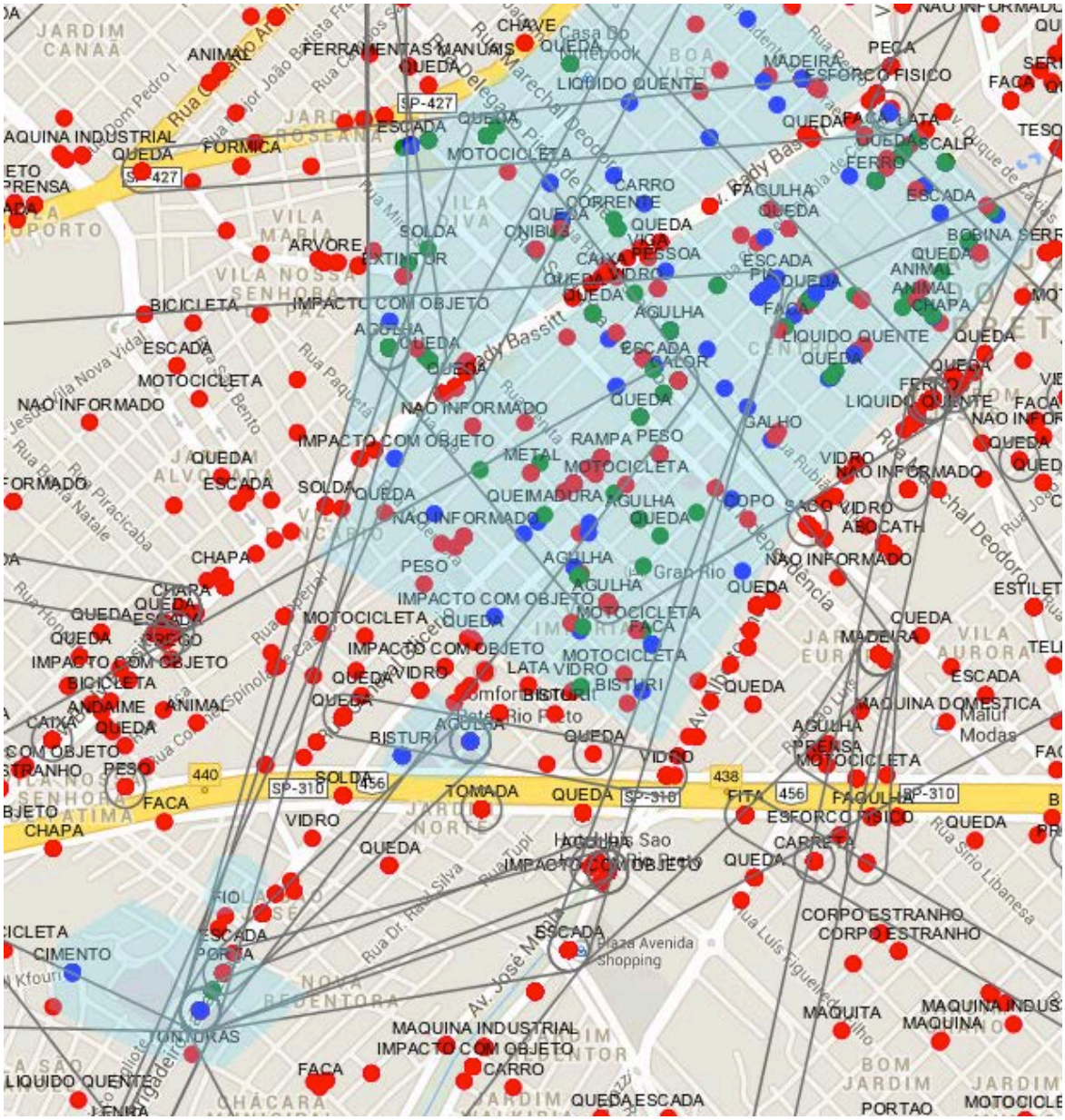
Ao analisar as informações dos agrupamentos que foram removidos é possível perceber que grande parte dos pontos desconsiderados estavam na região de hospital. Com a aplicação da abordagem de 91 agrupamentos existentes na região, houve uma redução para 59. Os agrupamentos que foram desconsiderados eram formados basicamente pela característica

“Auxiliar de Enfermagem”. Também foram desconsiderados pontos que possuíam como máquina causadora os itens, “Sangue” e “Agulha”.

Vale ressaltar que não é suficiente que um ponto possua uma dessas características para ser desconsiderado. O conjunto de atributos não espaciais do mesmo interfere no cálculo do peso semântico.

Ao considerar a região de hospital, haviam 23 agrupamentos formados com a característica de ocupação como “Auxiliar de Enfermagem” e após a aplicação da abordagem OntoSDM foram reduzidos para 20. Portanto, 3 outros agrupamentos ainda foram considerados relevantes pois suas outras características influenciaram no peso semântico, mantendo estes pontos nos agrupamentos.

Figura 33: Aplicação da abordagem OntoSDM na região de hospitais



Fonte: Do autor

Assim como o experimento anterior, foi calculado o coeficiente de silhueta para cada agrupamento. Novamente, o cálculo da silhueta não apresentou resultados discrepantes entre as duas aplicações, percebeu-se dessa vez um resultado melhor sem utilizar a abordagem. Os valores obtidos podem ser consultados na Tabela 7.

Tabela 7: Comparativo com os coeficientes semânticos

	Similaridade		Distância	
	Menor	Média	Menor	Média
MRClustering	0.93	0.97	0.99	0.99
MRClustering & OntoSDM	0.90	0.96	0.99	0.99

Porém, como ressaltado anteriormente, a diferença nos valores pode ser justificada pelo fato das medidas utilizadas não considerarem a semântica da região. No caso, uma métrica mais apurada seria necessária para mensurar a qualidade dos agrupamentos gerados.

4.3.3. APLICAÇÃO PARA DIFERENTES DOMÍNIOS

A fim de demonstrar que é possível adaptar a abordagem a diferentes aplicações, foram realizados experimentos utilizando outra base de dados, dessa vez, com informações de recursos hídricos na região do rio Marinheiro.

Devido a estrutura do *framework* toda a definição de regras e conhecimento fica descrito na ontologia, dessa maneira é possível aplicar a abordagem para diferentes domínios, bastando alterar a ontologia que está sendo utilizada.

Da mesma forma como é possível realizar a aplicação para diferentes domínios ao alterar a ontologia, também é possível obter resultados melhores aumentando a qualidade e os termos definidos da mesma. Portanto quanto melhor for a definição da ontologia melhores resultados serão obtidos.

Como detalhado no início do capítulo a ontologia referente a recursos hídricos possui duas regiões, “Área Urbana” e “Área Rural” e foram definidas regras que especificam os tipos de recursos hídricos em cada área.

Figura 34: Pontos desconsiderados dos agrupamentos

id integer	geom geometry	escolaridade text	sit_administrativa text	regiao text
242	010100002(	NAO_INFORMADO	sem_autorizacao	Região Rural
258	010100002(	primario_incompleto	sem_autorizacao	Região Rural
259	010100002(	ginasial_completo	sem_autorizacao	Região Rural
260	010100002(	NAO_INFORMADO	sem_autorizacao	Região Rural
261	010100002(	Fundamental_incompleto	cadastrado	Região Rural
263	010100002(	primario_incompleto	sem_autorizacao	Região Rural
271	010100002(	fundamental_incompleto	NAO_INFORMADO	Região Rural
283	010100002(	primario_completo	sem_autorizacao	Região Rural
291	010100002(	NAO_INFORMADO	sem_autorizacao	Região Rural
294	010100002(	primario_incompleto	sem_autorizacao	Região Rural

Fonte: Do autor

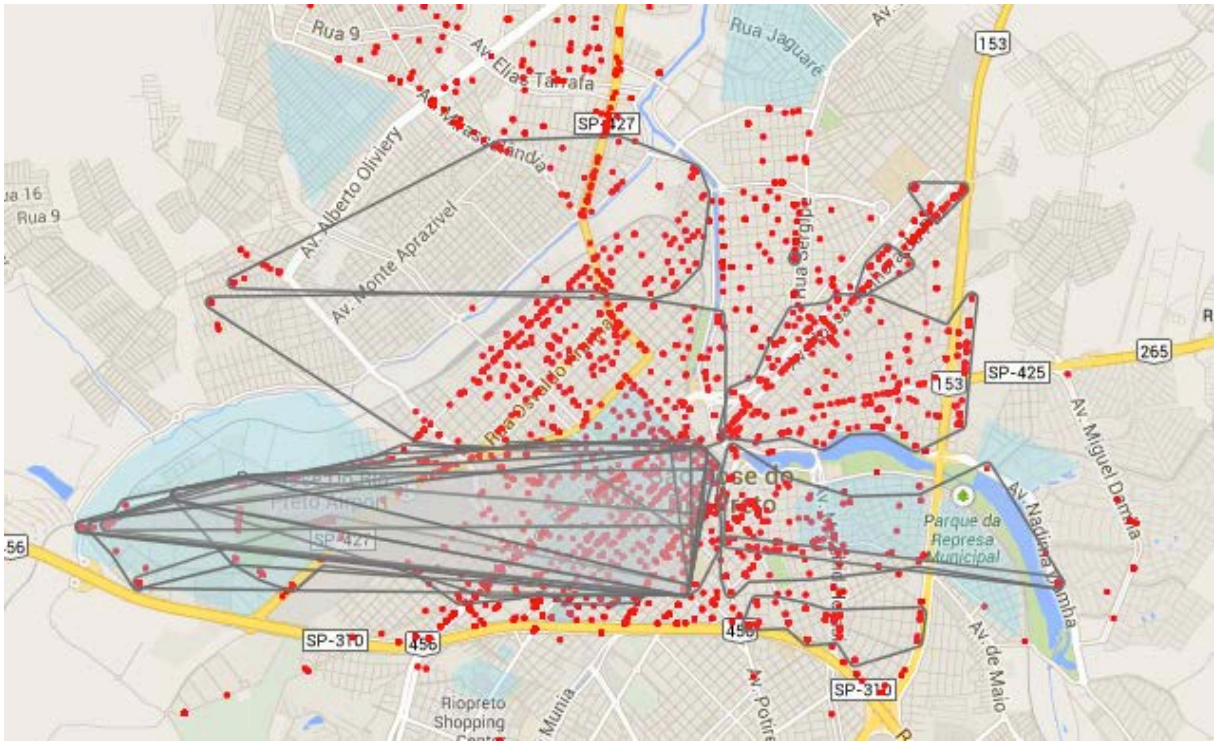
A Figura 34 exibe pontos que foram desconsiderados durante o agrupamento. Como é possível perceber são pontos que apresentam características comuns para as regiões que o mesmo se encontra, ocasionando uma redução na quantidade de agrupamentos gerados ao eliminar aqueles que são considerados triviais. Vale lembrar que como o *framework* depende diretamente da qualidade da ontologia, é possível obter sempre resultados melhores, a medida que a ontologia é refinada e adequada para um determinado domínio.

4.4. EXPERIMENTOS CHSMST+

Os experimentos com o algoritmo CHSMST+ tem como objetivo principal demostrar que a abordagem pode ser implementada em outros algoritmos de agrupamento de dados espaciais de modo trivial. Para isso, a abordagem foi adaptada ao algoritmo CHSMST+ e o mesmo executado na base de dados do SIVAT.

Na Figura 35 são exibidos os agrupamentos obtidos ao aplicar o algoritmo em um conjunto de dados com 10 mil pontos. Como visto anteriormente, o CHSMST+ possui uma abordagem distinta da utilizada pelo MRClustering. Porém mesmo com um funcionamento diferente é possível utilizar a abordagem OntoSDM.

Figura 35: Resultado da aplicação do algoritmo CHSMST+

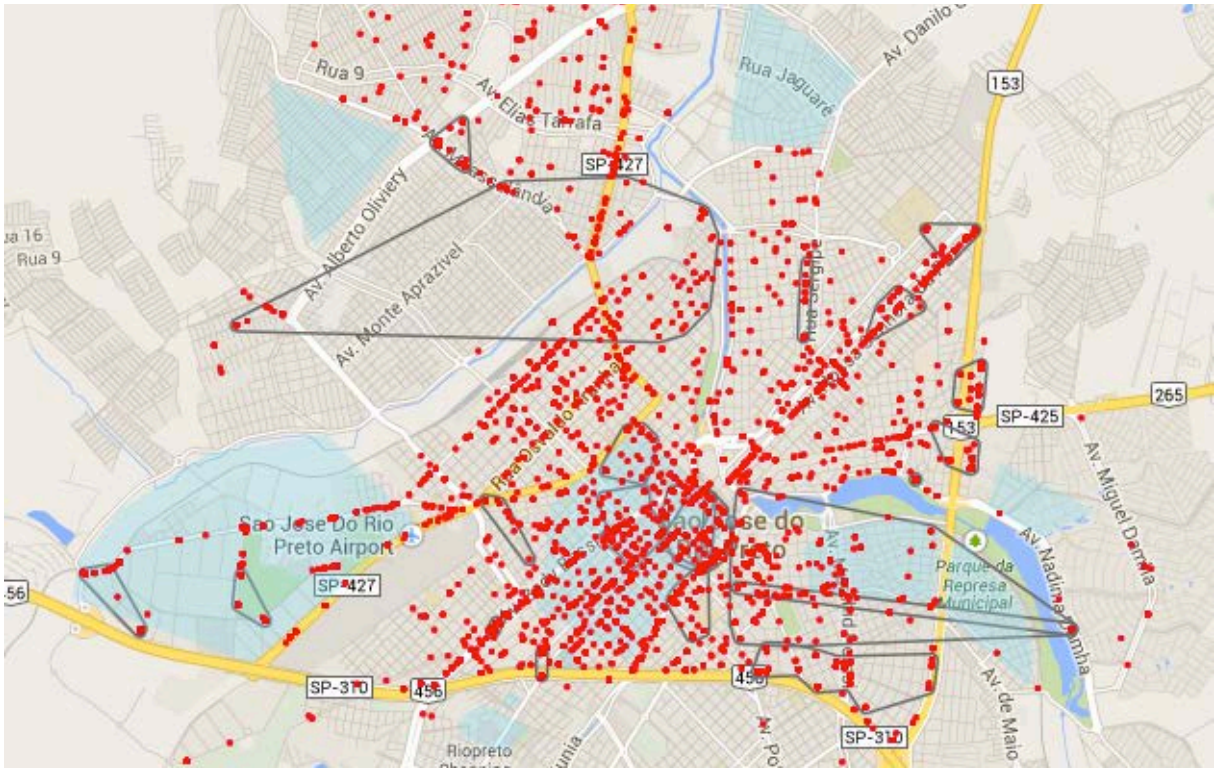


Fonte: Do autor

Ao analisar a Figura 36 é possível notar uma melhoria similar a obtida nos demais experimentos. Resulta-se assim agrupamentos mais relevantes ao remover aqueles que não são relevantes para o usuário final, dessa forma facilita-se a análise do mesmo.

Ao comparar as duas imagens, percebe-se que o algoritmo não desconsidera um agrupamento por completo, mas sim elimina os pontos não relevantes, ao gerar novos agrupamentos. Nota-se na Figura 35 que a maioria dos agrupamentos cruzam a cidade, ocupou-se assim uma grande parte e não gerou resultados muito interessantes. Ao considerar a semântica da região, os pontos que estão na região industrial passam a ser desconsiderados, gerando assim resultados mais refinados.

Figura 36: Resultado da aplicação do algoritmo CHSMST+ utilizando a abordagem OntoSDM



Fonte: Do autor

Além disso, na figura 35 foram gerados 87 agrupamentos, ao aplicar o suporte semântico, como pode ser observado na figura 36, há uma redução de agrupamentos, o que facilita a interpretação dos resultados resultando em uma redução de 23 agrupamentos gerados o que representa aproximadamente 25% dos agrupamentos.

4.5. DESEMPENHO

Com o intuito de verificar qual o impacto no desempenho do algoritmo causado pela aplicação da abordagem, o tempo total de execução do algoritmo foi mensurado para cada um dos conjuntos de dados da base de dados do SIVAT e aplicando a abordagem no algoritmo *MRClustering*.

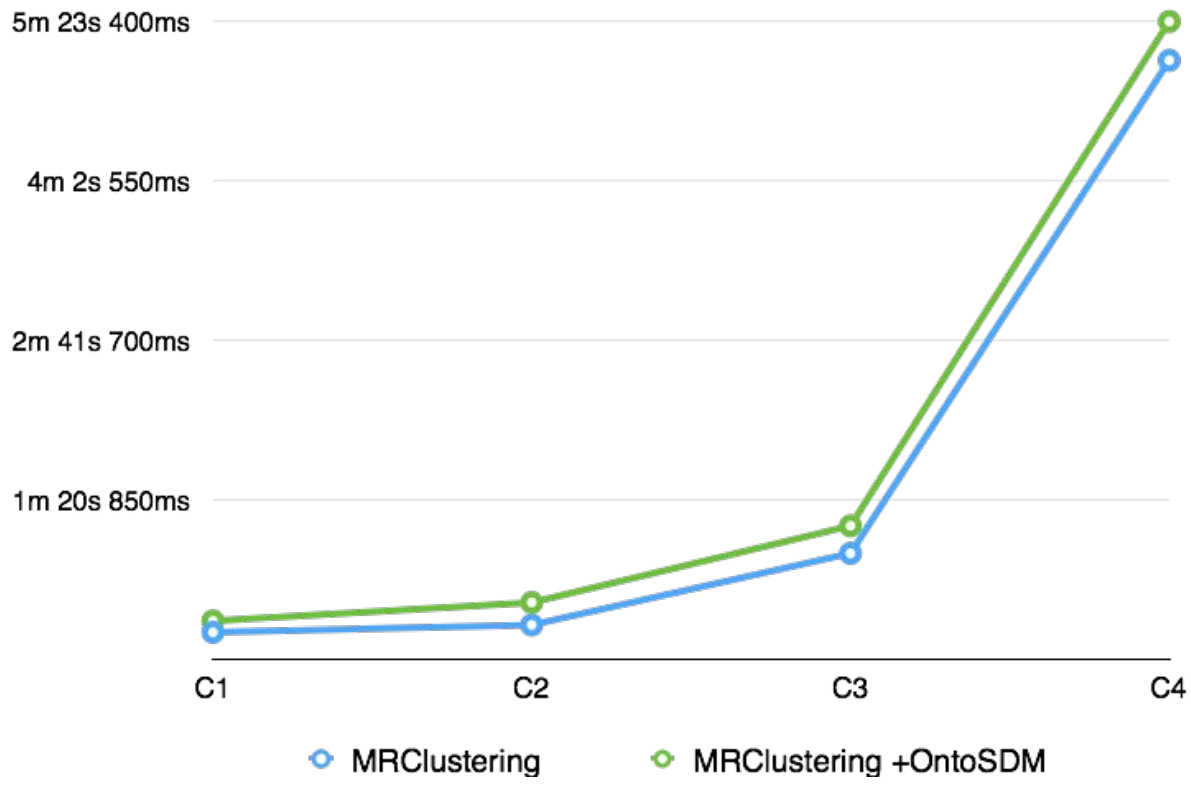
O tempo de execução com e sem a abordagem pode ser conferido na Tabela 8 e no gráfico ilustrado na

Figura 37

Tabela 8: Tempo total de execução do algoritmo

	C1	C2	C3	C4
MRClustering	13s 89ms	17s 60ms	53s 88ms	5m 3s 631ms
MRClustering & OntoSDM	19s 66ms	28s 96ms	67s 82ms	5m 23s 300ms

Figura 37: Gráfico com tempo total de execução do algoritmo



Fonte: Do autor

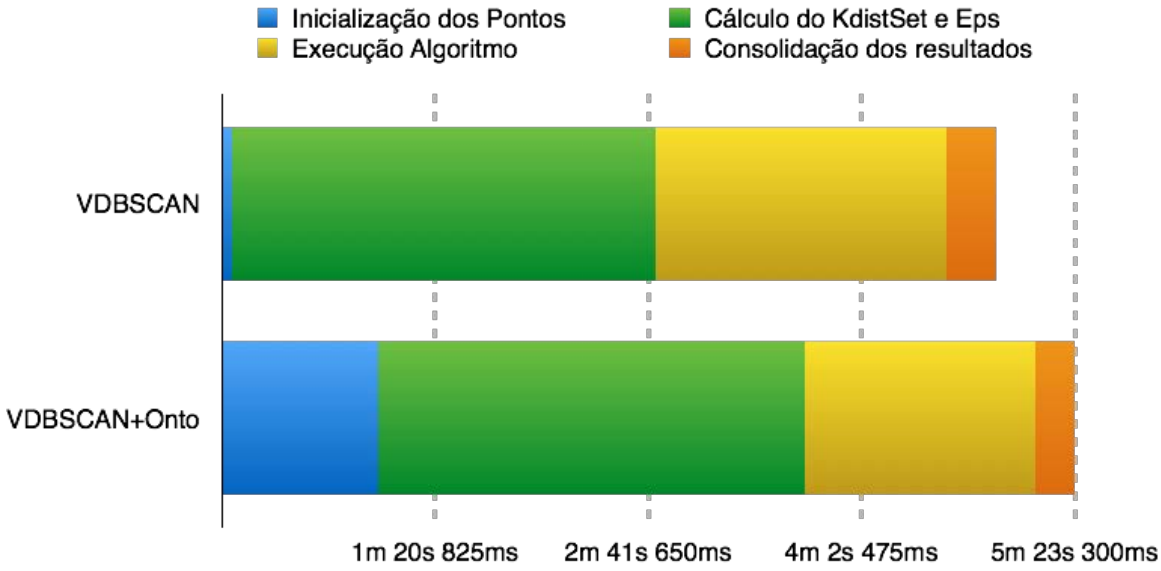
Ao utilizar a abordagem é possível perceber um impacto no tempo total de execução do algoritmo. Para analisar melhor o que causa o aumento no tempo, o conjunto 4 (C4) foi observado com mais detalhes.

Para isso, a execução do algoritmo foi dividida em 4 etapas: “Inicialização dos Pontos”, “Cálculo do KDistSet e Eps”, “Execução do Algoritmo” e “Consolidação dos Resultados”. Os tempos para cada etapa podem ser analisados na Tabela 9 e no gráfico da Figura 38.

Tabela 9: Tempos parciais de execução do algoritmo

	MRClustering	MRClustering + OntoSDM
Inicialização dos Pontos	3s 85ms	59s 49ms
Cálculo do KdistSet e Eps	2m 40s 27ms	2m 41s 52ms
Execução do Algoritmo	1m 50s 71ms	1m 27s 35ms
Consolidação dos Resultados	18s 792ms	14s 88ms

Figura 38: Gráfico com tempos parciais de execução do algoritmo



Fonte: Do autor

Como é possível verificar, o impacto da abordagem se concentra na etapa de inicialização dos pontos. Isso ocorre pois é durante essa etapa que a ontologia é acessada e verificada para cada um dos pontos contidos no conjunto de dados. Como já descrito anteriormente, os repositórios de ontologias não são tão otimizados como os SGBDs o que explica a queda de desempenho.

Em compensação, como a abordagem utilizada desconsidera os pontos não relevantes, os mesmos não precisam ser analisados pelas outras etapas do algoritmo, como o cálculo de distância e cálculo de similaridade. Além disso, outra etapa onde é possível obter uma redução do tempo é referente a consolidação dos resultados. Isso ocorre pois menos agrupamentos são gerados o que evita que o algoritmo dispenda tempo para gerar e gravar resultados que não são relevantes.

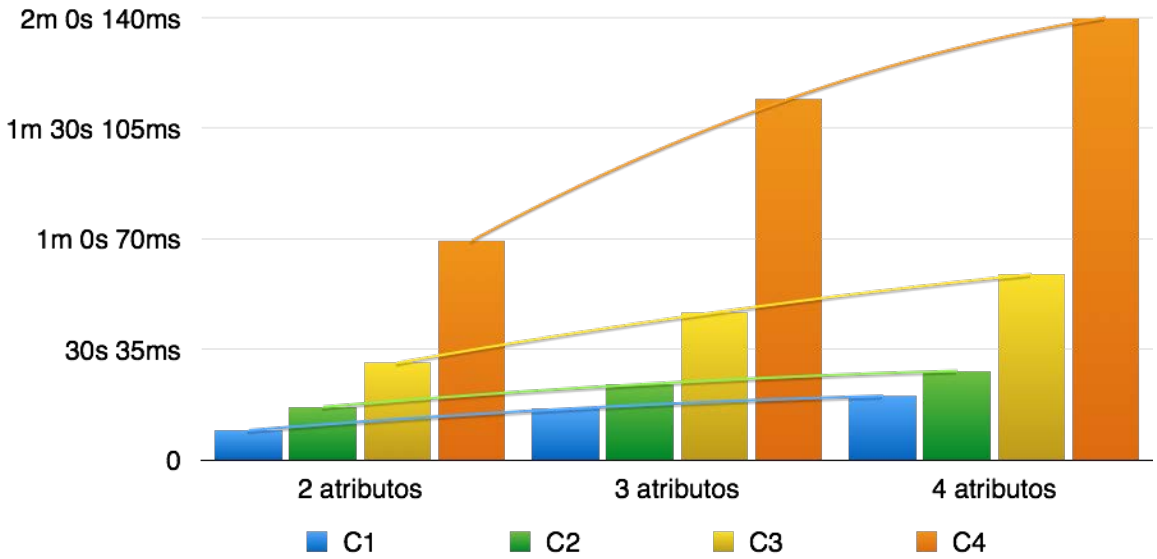
Outros experimentos foram realizados a fim de verificar o comportamento da abordagem em outros cenários. Para isso, foi mensurado o tempo de execução variando a quantidade de atributos não espaciais considerados. Os tempos para cada um dos conjuntos de

dados pode ser conferido na Tabela 10.

Tabela 10: Tempo de execução variando quantidade de atributos

	C1	C2	C3	C4
2 atributos	8s 73ms	16s 41ms	28s 37ms	59s 49ms
3 atributos	13s 88ms	20s 54ms	40s 56ms	1m 38s 66ms
4 atributos	17s 39ms	24s 21ms	50s 423ms	2m 0s 12ms

Figura 39: Gráfico com o tempo de execução ao variar a quantidade de atributos



Fonte: Do autor

Ao analisar o gráfico exibido na Figura 39, nota-se um aumento no tempo de execução considerável na etapa de “Inicialização dos Pontos” para cada um dos conjuntos de dados. Isso é explicado pois cada novo atributo deve ser consultado na ontologia, fazendo com que o tempo de execução aumente sempre que aumentar a quantidade de atributos não espaciais considerados.

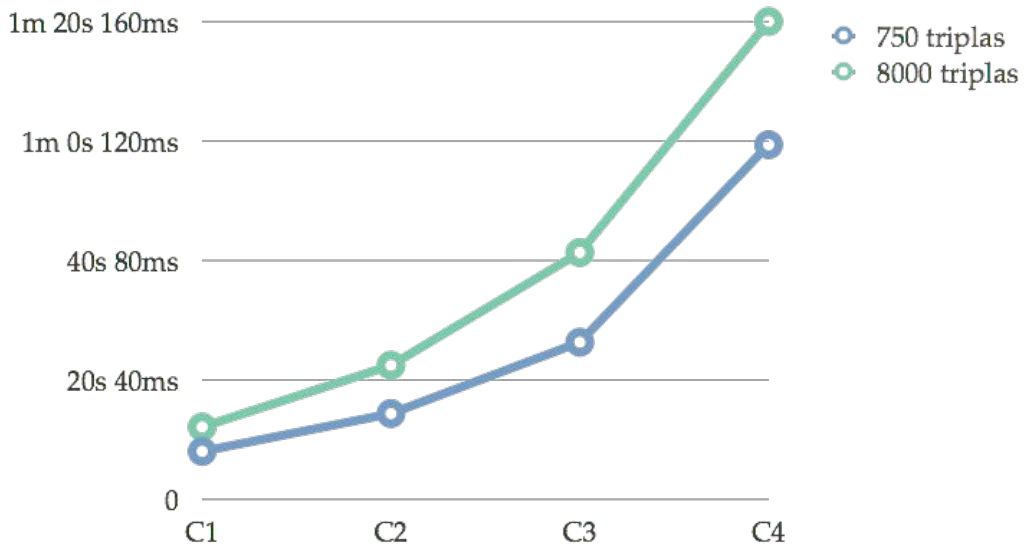
Por fim, foi analisado o impacto em que o aumento de triplas na ontologia proporciona na abordagem desenvolvida. O experimento foi realizado com uma ontologia contendo 750 triplas e outra com 8000 triplas. Os tempos podem ser analisados na Tabela 11 e no gráfico da Figura 40.

Tabela 11: Tempo de execução variando quantidade de tuplas contidas na ontologia

	C1	C2	C3	C4
750 triplas	8s 73ms	16s 41ms	28s 37ms	59s 49ms
8000 triplas	12s 13ms	22s 50ms	41s 42ms	1m 20s 15ms

A medida que o tamanho da ontologia aumenta maior é o tempo gasto para a execução da etapa de “Inicialização dos pontos”. Apesar do aumento obtido, não considera-se o mesmo alto, visto que o tamanho da ontologia aumentou praticamente dez vezes.

Figura 40: Gráfico com o tempo de execução do algoritmo de acordo com a quantidade de tuplas



Fonte: Do autor

4.6. CONSIDERAÇÕES FINAIS

Este capítulo apresentou uma conjunto de experimentos que foram executados com o propósito de validar os quesitos buscados pelos objetivos deste trabalho: a eficiência e a utilidade ao considerar o suporte semântico durante o processo de prospecção de dados espaciais.

Com os resultados obtidos foi possível observar uma redução significativa de agrupamentos considerados irrelevantes e que seriam descartados por uma analista na etapa posterior de análise dos resultados. Portanto, efetuando essa análise durante a etapa de prospecção garante-se um resultado melhor, além de facilitar a etapa seguinte, ou seja, de análise manual dos dados.

5. Conclusão

5.1. CONSIDERAÇÕES

Dados espaciais estão cada vez mais frequentes e com isso cada vez mais dados devem ser analisados e verificados. Algoritmos de prospecção de dados espaciais tradicionais consideram normalmente apenas a densidade, ou seja, a distância que um ponto está de seus vizinhos e o grau de similaridade entre os pontos durante a etapa de criação dos agrupamentos.

Diversos algoritmos na área vem sendo desenvolvidos com o intuito de melhorar cada vez mais os resultados gerados, porém nenhum considera um aspecto importante para a análise, as informações semânticas implícitas na posição geográfica do ponto. O trabalho desenvolvido permite que algoritmos de prospecção consigam se comunicar com uma ontologia de domínio definida de modo a gerar melhores resultados.

Para isso, a abordagem apresenta um novo atributo para um ponto geográfico, o seu valor semântico, que é calculado pelo *framework* desenvolvido ao considerar o conhecimento definido na ontologia e os parâmetros definidos pelo usuário.

Com o intuito de comprovar a eficácia do *framework* e a possibilidade de utilização em algoritmos de prospecção de dados espaciais que apresentam diferentes abordagens, o mesmo foi implementado em dois algoritmos, o *MRClustering* e *CHSMST+*. Foram então executados diversos experimentos ao utilizar os dois algoritmos adaptados e comparar os resultados obtidos com e sem o uso da abordagem.

Os experimentos foram aplicados em duas bases de bases de dados reais, demonstra-se assim que o mesmo pode se adaptar a diferentes domínios. Foi selecionado diversos conjuntos de teste com dados entre 5mil e 25mil pontos georreferenciados. Ao analisar os resultados obtidos foi possível perceber uma melhora nos agrupamentos formados, pois uma vez que pontos não relevantes são desconsiderados, menos agrupamentos são gerados, permanecendo apenas aqueles que possuem informação realmente relevante, facilita-se assim a análise de um especialista posteriormente.

Com os experimentos realizados, conclui-se que a abordagem apresenta uma nova possibilidade para algoritmos de prospecção de dados espaciais, ao possibilitar gerar resultados mais interessantes para qualquer domínio no qual deseja aplicar.

5.2. TRABALHOS FUTUROS

O trabalho apresentou uma abordagem inicial, que permite utilizar ontologias para obter melhores resultados. Porém existem vários pontos que podem ser estudados e melhorados em trabalhos futuros.

Uma frente que pode ser abordada é a otimização do algoritmo por meio do uso de técnicas de *multi-thread* ou do *MapReduce*, principalmente, na etapa de criação dos pontos, onde a ontologia é consultada diversas vezes gerando um alto custo computacional. Pode-se também trabalhar em otimização direta na consulta da ontologia, pois como demonstrado nos testes esta foi a etapa que consumiu maior tempo.

Outro ponto que pode ser melhorado é referente à informação da região na ontologia, onde se pode utilizar uma extensão do *SPARQL*, conhecida como *GeoSPARQL*[3], garantindo assim, que a identificação da classe, com a informação georrefenciada seja feita de forma automática, além disso, permite verificar diretamente por meio da ontologia todos os pontos que estão dentro de uma região, sem a necessidade de utilizar o *PostGIS*.

Por fim, pode-se trabalhar em diversos outras técnicas para melhorar o *framework*, como, por exemplo, adicionar o tratamento de múltiplas regiões para um ponto, regras de inclusão de agrupamentos, utilização de propriedades de objetos entre diferentes classes da ontologia entre outras técnicas.

[3] GeoSPARQL – Disponível em <http://www.opengeospatial.org/standards/geosparql>. Acessado em 01/08/2013

Referências Bibliográficas

AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A. **Mining association rules between sets of items in large databases**. New York, New York, USA: ACM Press, 1993. p. 207–216

ALLEMANG, D.; HENDLER, J. **Semantic web for the working ontologist: effective modeling in RDFS and OWL**. [s.l.] Elsevier, 2011.

ALMEIDA, M. B.; BAX, M. P. Uma visão geral sobre ontologias : pesquisa sobre definições , tipos , aplicações , métodos de avaliação e de construção. **SciELO Brasil**, p. 7–20, 2003.

BAE, D.-H. et al. SD-Miner: A spatial data mining system. **2009 IEEE International Conference on Network Infrastructure and Digital Content**, p. 803–807, nov. 2009.

BECKETT, D. et al. **RDF 1.1 Turtle**. Disponível em: <http://www.w3.org/TR/turtle/>. Acesso em: 20 jan. 2014.

BERMAN, J. **Principles of Big Data: preparing, sharing, and analyzing complex information**. [s.l.] Elsevier, 2013.

BHATIA, C. S.; JAIN, S. Semantic Web Mining: Using Ontology Learning and Grammatical Rule Inference Technique. **2011 International Conference on Process Automation, Control and Computing**, p. 1–6, jul. 2011.

BRAUN, M.; SCHERP, A.; STAAB, S. **Collaborative creation of semantic points of interest as linked data on the mobile phone**Extended Semantic Web Conf. **Anais**...Springer, 2010Disponível em: <http://kola.opus.hbz-nrw.de/volltexte/2010/504/>. Acesso em: 4 maio. 2014

BRAVO, C. D. O. **Geração Automática de Ontologias para a Web Semântica**. Brasília: Universidade de Brasília, 2010.

CHANDRASEKARAN, B.; JOSEPHSON, J. R.; BENJAMINS, V. R. What are ontologies, and why do we need them? **IEEE Intelligent Systems**, v. 14, n. 1, p. 20–26, jan. 1999.

COELHO, E. DE M. P. **Ontologias Difusas no Suporte à Mineração de Dados : aplicações na Secretaria de Finanças da Prefeitura Municipal de Belo Horizonte**. [s.l.] Universidade Federal de Minas Gerais, 2012.

CONTRERAS, J.; CORCHO, O.; GÓMEZ-PÉREZ, A. Six Challenges for the Semantic Web. n. i, p. 1–15, 2009.

CUGLER, D. C.; YAGUINUMA, C. A.; SANTOS, M. T. P. WOntoVLab: A Virtual Laboratory Authorship Process Based on Workflow and Ontologies. **2010 10th IEEE International Conference on Advanced Learning Technologies**, p. 690–694, jul. 2010.

DOMINGOS, P. Prospects and challenges for multi-relational data mining. **ACM SIGKDD explorations newsletter**, v. 5, n. 1, p. 80–83, 2003.

DOU, W. et al. Distributed multi-relational data mining based on genetic algorithm. **2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)**, p. 744–750, jun. 2008.

DUCHARME, B. **Learning SPARQL**. Second ed. Sebastopol: O'Reilly Media Inc., 2013. p. 357

ELSAYED, A. et al. **Applying data mining for ontology building.**Proc. of ISSR. **Anais**...Cairo: 2007

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v. 39, n. 11, p. 27–34, 1996.

FERNÁNDEZ, M.; GÓMEZ-PÉREZ, A.; JURISTO, N. **METHONTOLOGY : From Ontological Art Towards Ontological Engineering**Proceedings of the AAAI97 Spring Symposium Series on Ontological Engineering. **Anais**...Stanford, USA: 1997

GAZOLA, A.; FURTADO, A. L. **Bancos de Dados Geográficos Inteligentes**. Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro, 2007.

GUIZZARDI, G. **Desenvolvimento para e com Reuso: Um Estudo de Caso no Domínio de Vídeo sob Demanda**. Vitória: Centro Tecnológico da Universidade Federal do Espírito Santo, 2000.

HAN, J.; KAMBER, M. **Data mining: concepts and techniques**. Third ed. [s.l.] Morgan Kaufmann, 2011. p. 744

HAN, J.; KAMBER, M.; TUNG, A. K. H. Spatial Clustering Methods in Data Mining: A Survey. **Geographic Data Mining and Knowledge Discovery**, p. 1–29, 2001.

HE, Q.; ZHAO, W.; SHI, Z. CHSMST: a clustering algorithm based on hyper surface and minimum spanning tree. **Soft Computing**, v. v. 15, p. 1097–1103, 2011.

HENDLER, J. Web 3.0 Emerging. **IEEE Computer Society**, n. January, p. 111–113, 2009.

HEPP, M. Ontologies: State of the art, business potential, and grand challenges. **Ontology Management**, p. 3–22, 2008.

HONGFEI, C.; XIAOYAN, W. Research on GIS-based spatial data mining. **2011 International Conference on Business Management and Electronic Information**, p. 351–354, maio 2011.

HORRIDGE, M. et al. **A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools**. Manchester: [s.n.].

HWANG, S. Using formal ontology for integrated spatial data mining. **Computational Science and Its Applications–ICCSA**, v. 3044, p. 1026–1035, 2004.

ICHIBA, F. T. **Algoritmo para prospecção multirrelacional de dados espaciais**. São José do Rio Preto: Universidade Estadual Paulista, 2013.

JI, M. et al. **Mine geological hazard multi-dimensional spatial data warehouse construction research**18th International Conference on Geoinformatics. **Anais**...IEEE, jun. 2010

JIEHAI, C.; WEI, L. Research on the Storage and Management of Mine Spatial Data Based on PostgreSQL. **Proceedings of 2010 International Conference on Computer Application and System Modeling**, p. 493–496, 2010.

JIN, H.; MIAO, B. **The research progress of spatial data mining technique**2010 3rd International Conference on Computer Science and Information Technology. **Anais**...IEEE, jul. 2010

KAUFMAN, L.; ROUSSEEUW, P. . Finding groups in data: an introduction to cluster analysis. 1990.

KNOBBE, A.; BLOCKEEL, H. Multi-relational data mining. **Technical Report CWI**, p. 11, 1999.

KUO, Y.-T. et al. **Domain ontology driven data mining**Proceedings of the 2007 international workshop on Domain driven data mining - DDDM '07. **Anais**...New York, New York, USA: ACM Press, 2007

LE, W. et al. Scalable Multi-query Optimization for SPARQL. **2012 IEEE 28th International Conference on Data Engineering**, p. 666–677, abr. 2012.

LIANG, Y. et al. A method for OWL ontology module partition. **2010 IEEE 2nd Symposium on Web Society**, p. 372–377, ago. 2010.

LIAO, S.-H.; CHEN, J.-L.; HSU, T.-Y. Ontology-based data mining approach implemented for sport marketing. **Expert Systems with Applications**, v. 36, n. 8, p. 11045–11056, out. 2009.

LIU, P.; ZHOU, D.; WU, N. **VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise.**INTERNATIONAL CONFERENCE ON SERVICE SYSTEMS AND SERVICE MANAGEMENT. **Anais**...New York: 2007

LLOYD, S. Least squares quantization in PCM. **IEEE Transactions on Information Theory**, v. v. 28, p. 129–137, 1982.

MEDEIROS, C. A. DE. **Extração de conhecimento em bases de dados espaciais: Algoritmo CHSMST+**. [s.l.] Universidade Estadual Paulista - UNESP, 2014.

MENNIS, J.; GUO, D. Spatial data mining and geographic knowledge discovery—An introduction. **Computers, Environment and Urban Systems**, v. 33, n. 6, p. 403–408, nov. 2009.

MILLLER, H. J.; HAN, J. **Geographic Data Mining and Knowledge Discovery**. Second ed. Illinois: CRC Press, 2009. p. 486

MIZOGUCHI, R.; VANWELKENHUYSEN, J.; IKEDA, M. Task Ontology for Reuse of Problem Solving Knowledge. **Towards Very Large Knowledge Bases**, p. 46–59, 1995.

MORAIS, E.; AMBRÓSIO, A. **Ontologias: conceitos, usos, tipos, metodologias, ferramentas e linguagens**. Jataí: Universidade Federal de Goiáis, 2007.

ONGENAE, F. et al. A probabilistic ontology-based platform for self-learning context-aware healthcare applications. **Expert Systems with Applications**, v. 40, n. 18, p. 7629–7646, dez. 2013.

OSEI-BRYSON, K.-M. A context-aware data mining process model based framework for supporting evaluation of data mining results. **Expert Systems with Applications**, v. 39, n. 1, p. 1156–1164, jan. 2012.

PANOV, P.; DZEROSKI, S.; SOLDATOVA, L. **OntoDM: An Ontology of Data Mining**IEEE International Conference on Data Mining Workshops. **Anais**...IEEE, dez. 2008

PARREIRAS, F. S. **Semantic Web and Model-driven Engineering**. Piscataway: IEEE Press, 2012. p. 187

SANTOS, M. Y.; AMARAL, L. Mining geo-referenced data with qualitative spatial reasoning strategies. **Elsevier Computers and graphics**, 2004.

SHVAIKO, P.; EUZENAT, J. Ten challenges for ontology matching. **On the Move to Meaningful Internet Systems: OTM …**, p. 1163–1181, 2008.

STAAB, S.; STUDER, R. Knowledge Processes and Ontologies. **Intelligent Systems, IEEE**, 2001.

USCHOLD, M.; GRUNINGER, M. Ontologies and semantics for seamless connectivity. **ACM SIGMOD Record**, v. 33, n. 4, p. 58, 1 dez. 2004.

VAARANDI, R. A Data Clustering Algorithm for Mining Patterns From Event Logs A Data Clustering Algorithm for Mining Patterns From Event Logs. 2003.

VALÊNCIO, C. R. et al. Multi-relational Algorithm for Mining Association Rules in Large Databases. **2011 12th International Conference on Parallel and Distributed Computing, Applications and Technologies**, p. 269–274, out. 2011.

VAN HEIJST, G.; SCHREIBER, A. T.; WIELINGA, B. J. Using explicit ontologies in KBS development. **International Journal of Human-Computer Studies**, v. 46, n. 2-3, p. 183–292, fev. 1997.

VANDECASTEELE, A.; NAPOLI, A. Spatial ontologies for detecting abnormal maritime behaviour. **OCEANS, 2012-Yeosu**, 2012.

VENANCIO, L. R.; FILETO, R.; MEDEIROS, C. B. **Aplicando Ontologias de Objetos Geográficos para Facilitar Navegação em GIS**SIMPÓSIO BRASILEIRO DE GEOINFORMÁTICA. **Anais**...2003

WANG, J. et al. Research of GIS-based Spatial Data Mining Model. **2009 Second International Workshop on Knowledge Discovery and Data Mining**, p. 159–162, jan. 2009.

YASMIN, M. Dynamic referencing rules creation using intelligent agent in geo spatial data mining. **Computer & Information Science (ICCIS), 2012 …**, v. 5, p. 4–8, 2012.

YONG-GUI, W.; ZHEN, J. Research on semantic web mining. **International Conference On Computer Design and Applications**, v. 1, p. 67–70, 2010.

YUE, P. et al. Intelligent services for discovery of complex geospatial features from remote sensing imagery. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 83, p. 151–164, set. 2013.

YUESHUN, H. A study of spatial data mining architecture and technology. **2009 2nd IEEE International Conference on Computer Science and Information Technology**, p. 163–166, 2009.