

**MINISTÉRIO DA DEFESA
EXÉRCITO BRASILEIRO
DEPARTAMENTO DE CIÊNCIA E TECNOLOGIA
INSTITUTO MILITAR DE ENGENHARIA
CURSO DE MESTRADO EM SISTEMAS E COMPUTAÇÃO**

FELIPE ALVES DE OLIVEIRA

**MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO DE
MULTIRRELAÇÃO EM DATASETS NA WEB DE DADOS**

**Rio de Janeiro
2018**

INSTITUTO MILITAR DE ENGENHARIA

FELIPE ALVES DE OLIVEIRA

**MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO DE
MULTIRRELAÇÃO EM DATASETS NA WEB DE DADOS**

Dissertação de Mestrado apresentada ao Curso de Mestrado em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de Mestre em Ciências em Sistemas e Computação.

Orientadora: Prof^a. Maria Cláudia Reis Cavalcanti - D.Sc.

Co-Orientador: Prof. Ronaldo Ribeiro Goldschmidt - D.Sc.

Co-Orientadora: Prof^a. Raquel Lopes Costa - D.Sc.

Rio de Janeiro
2018

INSTITUTO MILITAR DE ENGENHARIA
Praça General Tibúrcio, 80 - Praia Vermelha
Rio de Janeiro - RJ CEP: 22290-270

Este exemplar é de propriedade do Instituto Militar de Engenharia, que poderá incluí-lo em base de dados, armazenar em computador, microfilmar ou adotar qualquer forma de arquivamento.

É permitida a menção, reprodução parcial ou integral e a transmissão entre bibliotecas deste trabalho, sem modificação de seu texto, em qualquer meio que esteja ou venha a ser fixado, para pesquisa acadêmica, comentários e citações, desde que sem finalidade comercial e que seja feita a referência bibliográfica completa.

Os conceitos expressos neste trabalho são de responsabilidade do(s) autor(es) e do(s) orientador(es).

005 Oliveira, Felipe Alves de
O48m Mineração de Regras de Associação de Multirrelação
em datasets na web de dados / Felipe Alves de Oliveira,
orientado por Maria Cláudia Reis Cavalcanti, Ronaldo
Ribeiro Goldschmidt e Raquel Lopes Costa - Rio de Ja-
neiro: Instituto Militar de Engenharia, 2018.

95p.: il.

Dissertação (Mestrado) - Instituto Militar de Enge-
nharia, Rio de Janeiro, 2018.

1. Curso de Sistemas e Computação - teses e disser-
tações. 2. Bancos de Dados. 3. Mineração de Dados. I.
Cavalcanti, Maria Cláudia Reis . II. Goldschmidt, Ro-
naldo Ribeiro. III. Costa, Raquel Lopes. IV. Instituto
Militar de Engenharia. V. Título.

INSTITUTO MILITAR DE ENGENHARIA

FELIPE ALVES DE OLIVEIRA

**MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO DE
MULTIRRELAÇÃO EM DATASETS NA WEB DE DADOS**

Dissertação de Mestrado apresentada ao Curso de Mestrado em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de Mestre em Ciências em Sistemas e Computação.

Orientadora: Prof^a. Maria Cláudia Reis Cavalcanti - D.Sc.

Co-Orientador: Prof. Ronaldo Ribeiro Goldschmidt - D.Sc.

Co-Orientadora: Prof^a. Raquel Lopes Costa - D.Sc.

Aprovada em 7 de Fevereiro de 2018 pela seguinte Banca Examinadora:

Prof^a. Maria Cláudia Reis Cavalcanti - D.Sc. do IME - Presidente

Prof. Ronaldo Ribeiro Goldschmidt - D.Sc. do IME

Prof^a. Raquel Lopes Costa - D.Sc. do INCA

Prof^a. Maria Luíza Machado Campos - Ph.D. da UFRJ

Prof. Luís Alexandre Estevão da Silva - D.Sc. do IPJBRJ

Rio de Janeiro
2018

Dedico este trabalho a Deus, meu Senhor e Salvador.
Aos meus pais Milton e Noêmia, pois mesmo sem compreenderem
este caminho, com muito amor, sempre me incentivaram.
Ao meu irmão Tiago, que foi meu amigo, sempre com palavras de
motivação.
À minha amada noiva Indira, que muito me apoiou e, nos momentos
críticos, ajudou-me a continuar.
A todos os meus amigos.

AGRADECIMENTOS

Agradeço a Deus, por ser meu amigo sempre presente, mostrando-me o caminho e ajudando-me durante toda a jornada até aqui.

À minha família, que sempre fez mais que o possível para que eu pudesse continuar; serei eternamente grato.

À minha noiva Indira que, por diversas vezes, com carinho, ajudou-me a repassar os textos e ouviu-me, treinando para as apresentações; amo-a.

À minha sogra Dilma Silva que, como professora de português, não exitou em socorrer um aluno, disponibilizando-se em rever o texto completo, ajudando-me nessa árdua tarefa.

Aos meus orientadores Maria Cláudia, Ronaldo e Raquel, pela disponibilidade e pela atenção, que foram os fatores decisivos em minha formação. Em especial, à professora Maria Cláudia, que foi além de orientar, mesmo com as “broncas”, nunca deixou de ser uma amiga com palavras de incentivo; só tenho a agradecer.

À professora Maria Luíza da UFRJ, por abrir um espaço de seu tempo para compor a banca avaliadora.

Ao professor Luís Alexandre, por fazer parte da banca avaliadora e que, desde a faculdade, já dedicava seu tempo a orientar os alunos que almejavam novas conquistas. Hoje, esse docente, além de colega de trabalho, tornou-se um amigo pessoal e um dos meus maiores incentivadores, com seus conselhos; só tenho a crescer.

Ao Jardim Botânico do Rio de Janeiro, por ceder os dados para os experimentos reais desta pesquisa. Em especial ao Núcleo de Computação Científica e Geoprocessamento (NCCG), pois somente com seu apoio que possível alcançar os objetivos pretendidos. Aos colegas de trabalho, pela compreensão nos momentos de ausência, apoio e as muitas palavras encorajadoras - em especial - ao Leonardo, sua ajuda foi fundamental em diversos momentos, grande amigo.

À pesquisadora Marinez Siqueira que, por diversas vezes, abriu espaço em sua corrida agenda para validar os resultados. Também a seu aluno Diogo Rocha, que me ajudou a obter os dados do IBGE e gerar o mapa de ocorrência dos espécimes coletados.

Aos obreiros e membros da Assembleia de Deus Ministério Primitivo Chama Viva, por suas orações e compreensão durante os momentos que estive ausente, em destaque, ao departamento jovem EL-Shalom que amo muito, espero que esta conquista motive o grupo a alcançar seus sonhos.

Ao Pastor e amigo Thiago Gabriel, que muito orou e impulsionou-me, como ex aluno

do IME, deu-me todas as dicas para ingressar na Instituição, isso mudou minha vida.

Ao Exército Brasileiro, Instituição que me recebeu pela segunda vez, já que nela prestei serviços por 6 anos, como recruta, soldado, aluno por 2 vezes e cabo, onde, com orgulho, galguei a graduação máxima que me foi permitida na época e, agora, para minha surpresa, na pós-graduação, consigo alcançar mais este degrau. A todos os amigos do EB, que sirva de motivação, pois, com mais de 3 anos de dispensa do serviço militar, não esperava receber novamente essa “Mão amiga”.

Ao Instituto Militar de Engenharia, onde sempre almejei estudar, ofereceu-me apoio para que eu superasse meus próprios limites, conduzindo-me adiante.

A todos os meus amigos do IME e externos, notavelmente, ao Major Arias, Mário, Vanessa e Erick, sem as muitas horas de estudos em grupo e sem o suporte de vocês eu não conseguiria. Ao Silas e à Yasmin, alunos das turmas anteriores, que transmitiram-me suas experiências e deram-me ótimos conselhos.

Por fim, a todos os profissionais e mestres do Departamento de Engenharia de Sistemas (SE/8) deste honrado Instituto Militar de Engenharia.

“Tende em vós este sentimento que houve também em Cristo Jesus, o qual, subsistindo em forma de Deus, não julgou que o ser igual a Deus fosse coisa de que não devesse abrir mão, mas esvaziou-se, tomando a forma de servo, feito semelhante aos homens e sendo reconhecido como homem, humilhou-se, tornando-se obediente até a morte, e morte de cruz. Por isso também Deus o exaltou soberanamente e lhe deu o nome que é sobre todo o nome, para que em o nome de Jesus se dobre todo o joelho dos que estão nos céus, na terra e debaixo da terra, e toda a língua confesse que Jesus Cristo é o Senhor para glória de Deus Pai.”

FILIPENSES 2:5-11

SUMÁRIO

LISTA DE ILUSTRAÇÕES	10
LISTA DE TABELAS	13
LISTA DE SIGLAS	15
1 INTRODUÇÃO	18
1.1 Motivação	18
1.2 Caracterização do Problema	20
1.3 Hipóteses / Questões de pesquisa	21
1.4 Objetivo	21
1.5 Contribuições Esperadas	21
1.6 Organização do Trabalho	22
2 FUNDAMENTAÇÃO TEÓRICA	23
2.1 Web de Dados	23
2.2 Sistemas de Bancos de Dados em Grafos	25
2.3 Mineração de dados	27
2.3.1 Mineração de regras de associação	28
2.4 Considerações finais	30
3 REGRAS DE ASSOCIAÇÃO DE MULTIRRELAÇÃO	31
3.1 Mineração de regras de associação de multirrelação	31
3.1.1 Implementação e validação do MRAR	35
3.1.1.1 Estudo de caso 1	35
3.1.1.2 Estudo de caso 2	36
3.2 Formalismo	38
3.3 Considerações Finais	40
4 ABORDAGEM/PROPOSTA	41
4.1 Visão Geral	41
4.2 Algoritmo <i>MRAR+</i>	46
4.3 Protótipo do <i>MRAR+</i>	48
4.3.1 Implementação	48
4.3.1.1 Interfaces	51

4.4	Considerações Finais	55
5	EXPERIMENTOS E RESULTADOS	57
5.1	Experimento 1 (<i>dataset</i> IME - Esportes)	57
5.2	Experimento 2 (<i>dataset</i> de Jogadores de futebol com recursos externos).....	61
5.3	Experimento 3 (<i>Dataset</i> JBRJ/IBGE)	64
5.4	Considerações Finais	75
6	TRABALHOS RELACIONADOS	78
6.1	Mineração de regras de associação entre rótulos de nós	78
6.2	Mineração de regras de associação para subgrafos frequentes em um único grafo grande	79
6.3	Mineração de regras de associação de multirrelação	80
6.4	Mineração de regras de associação multirrelação em grafos: direcionando o processo de busca	81
6.5	Comparação dos trabalhos relacionados	81
6.6	Considerações finais	83
7	CONCLUSÕES E CONSIDERAÇÕES FINAIS	84
7.1	Contribuições	85
7.2	Melhorias e trabalhos futuros	86
8	REFERÊNCIAS BIBLIOGRÁFICAS	88
9	ANEXOS	91
9.1	ANEXO 1: Resultados da mineração de dados com o Algoritmo MRAR sobre o <i>dataset</i> JabotG_IBGE.	92

LISTA DE ILUSTRAÇÕES

FIG.1.1	Diagrama em nuvem do Linked Open Data de 2017. Fonte: http://lod-cloud.net/versions/2017-08-22/lod.svg .	20
FIG.2.1	Exemplo de uma tripla RDF com links na <i>DBpedia</i> .	23
FIG.2.2	Exemplo de resultados da consulta SPARQL da página <i>DBpedia</i> com os predicados e objetos associados ao MIT.	25
FIG.2.3	Demonstração de resultados da consulta SPARQL da página <i>DBpedia</i> com os predicados e objetos associados ao IME.	26
FIG.2.4	Representação dos processos de KDD. Fonte: Fayyad et al. (1996).	28
FIG.2.5	Um exemplo clássico de uma base de dados composta de 4 itens e 5 transações.	29
FIG.3.1	Grafo direcionado com rótulo nas arestas. Fonte: Ramezani et al. (2014).	33
FIG.3.2	Fluxo de trabalho do algoritmo MRAR. Fonte: Ramezani et al. (2014).	33
FIG.3.3	<i>Dataset</i> IME (DtIME) com informações dos professores e alunos, juntamente com a localidade onde vivem e a instituição onde trabalham ou estudam.	37
FIG.4.1	Visão geral de todos os passos da proposta.	43
FIG.4.2	Visão geral passo 1. Exemplo de aplicação do algoritmo MRAR identificando uma regra de associação de multirrelação.	44
FIG.4.3	Visão geral passo 2, seleção dos recursos externos.	45
FIG.4.4	Visão geral passo 3, ampliando informações do <i>dataset</i> .	46
FIG.4.5	Visão geral passo 4, encontrando novas regras.	46
FIG.4.6	Visão geral passo 5, comparando as regras geradas ao aplicar o MRAR sobre os dados do DtA(R1) e DtA+(R2).	47
FIG.4.7	Visão geral passo 6, verificação opcional para descobrir se as regras (R1) geradas ao aplicar o MRAR sobre os dados do DtA estão contidas no conjunto de regras (R2), gerados após a análise feita sobre os dados do DtA+.	47
FIG.4.8	Tela de configuração para as variáveis de entrada do algoritmo <i>MRAR+</i> .	52

FIG.4.9	Tela de visualização das regras geradas, em formato numérico.	53
FIG.4.10	Tela de visualização das regras geradas, em formato de texto.	54
FIG.4.11	Tela de comparação das regras geradas pelo algoritmo MRAR e <i>MRAR+</i>	55
FIG.4.12	Figura da tabela contendo apenas as regras que são realmente no- vas, geradas após a comparação entre as regras encontradas com a mineração do DtIME e do DtIME+. Gerada após a seleção demonstrada na Figura 4.11.	55
FIG.5.1	<i>Dataset</i> Esportes com informações de pessoas que praticam espor- tes. Os nós são representados por retângulos e as arestas por setas. Os nós sombreados de cinza indicam os esportes praticados, já os nós em branco representam as pessoas.	58
FIG.5.2	<i>Datasets</i> demonstrando associações externas através da relação <i>Same as</i>	59
FIG.5.3	Resultado de uma consulta Sparql feita no endpoint da <i>DBpedia</i>	62
FIG.5.4	Subconjunto do esquema relacional do banco de dados Jabot.	65
FIG.5.5	Modelo em grafo do <i>dataset</i> JabotG. Os nós são representados em círculos ovais e as arestas em setas direcionadas, ambos apresen- tam um rótulo descritivo.	66
FIG.5.6	Modelo em grafo do <i>dataset</i> IBGE/INDE. Os nós são representa- dos em círculos ovais e as arestas em setas direcionadas, ambos apresentam um rótulo descritivo.	67
FIG.5.7	Schema de <i>datasets</i> interligados, através da relação <i>Same As</i>	68
FIG.5.8	Variações de clima para uma determinada família. Exemplo de consulta sobre as variações de clima para a família Bromeliacea. As setas vermelhas indicam junções entre os <i>datasets</i>	69
FIG.5.9	Bioma e tipo de vegetação das famílias mais frequentes, com número de ocorrências ≥ 200	70
FIG.5.10	Mapa de pontos das ocorrências (pontos em vermelho) dos registros selecionados no Jabot.	71
FIG.5.11	Novas regras geradas, após a mineração de dados estendida, com o algoritmo <i>MRAR+</i>	73
FIG.6.1	Exemplo de <i>dataset</i> em grafo fictício com nós rotulados, apresen- tando uma relação diferente mesmo que o nó se repita. Fonte:	

Hendrickx et al. (2015).	79
FIG.6.2 Exemplo de <i>dataset</i> em grafo com subgrafos frequentes. Fonte: Elseidy et al. (2014).	80
FIG.6.3 Exemplo do algoritmo $MRAR_m$ aplicando uma máscara de busca sobre os consequentes das regras. Fonte: Oliveira et al. (2017).	82

LISTA DE TABELAS

TAB.2.1	Lista de bancos de dados em grafos.	26
TAB.3.1	Exemplo de <i>ItemChains</i> gerados após a aplicação do <i>GenerateItemChains</i> , para os dados do grafo da Figura 3.1.	34
TAB.3.2	Exemplo de <i>LargeItemChain</i> , para os dados da Tabela 3.1.	34
TAB.3.3	Exemplo de Regras de associação de multirrelação gerado após a aplicação do algoritmo <i>Generating Multi-Relation Association Rules</i> , combinando os dados da Tabela 3.2.	35
TAB.3.4	Exemplo de regras encontradas após a aplicação do MRAR com o <i>dataset</i> visto na Figura 3.1.	36
TAB.3.5	Exemplo de regras encontradas após a aplicação do MRAR com o <i>dataset</i> IME, visto na Figura 3.3.	37
TAB.5.1	Exemplo de regras geradas aplicando o algoritmo <i>MRAR+</i> sobre o conjunto de dados estendidos (DtIME+).	60
TAB.5.2	Tabela comparativa dos resultados obtidos após a aplicação do MRAR e <i>MRAR+</i> sobre os dados do DtIME.	60
TAB.5.3	Tabela comparativa dos resultados obtidos após a aplicação do MRAR e <i>MRAR+</i> , sobre os dados do Dt_Neymar.	63
TAB.5.4	Novas regras geradas, após a mineração de dados estendida, com o algoritmo <i>MRAR+</i>	63
TAB.5.5	Tabela comparativa dos resultados obtidos após a aplicação do MRAR e <i>MRAR+</i> , sobre os dados do Dt_JabotG e Dt_JabotG_IBGE.	
	72	
TAB.5.6	Tabela com os gêneros taxonômicos de maior porcentagem associados aos recursos que deram suporte a cada regra. Os números das regras fazem referência as regras vistas na Figura 5.11.	74
TAB.5.7	Tabela contendo algumas das regras identificadas que possuem o bioma Mata Atlântica no antecedente das regras.	77
TAB.6.1	Tabela comparativa dos trabalhos relacionados.	82
TAB.9.1	Tabela de regras de associação de multirrelação, obtida após a mineração do <i>dataset</i> DtJabotG_IBGE. Contém apenas as 27 primeiras regras, ordenadas pelo valor de suporte.	92

TAB.9.2	Tabela de regras de associação de Multirrelação, obtida após a mineração do <i>dataset</i> DtJabotG_IBGE. Contém apenas as 27 primeiras regras, ordenadas pelo valor de suporte. (Continuação)	93
TAB.9.3	Tabela de regras de associação de Multirrelação, obtida após a mineração do <i>dataset</i> DtJabotG_IBGE. Contém apenas as 27 primeiras regras, ordenadas pelo valor de suporte. (Continuação)	94
TAB.9.4	Tabela de regras de associação de Multirrelação, obtida após a mineração do <i>dataset</i> DtJabotG_IBGE. Contém apenas as 27 primeiras regras, ordenadas pelo valor de suporte. (Continuação)	95

LISTA DE SIGLAS

ARM	<i>Association Rules Mining</i>
CSP	<i>Constraint Satisfaction Problem</i>
DBMS	<i>Database Management System</i>
GRAMI	<i>GRApH MIning</i>
HTML	<i>HyperText Markup Language</i>
IBGE	Instituto Brasileiro de Geografia e Estatística
IME	Instituto Militar de Engenharia
INDE	Infraestrutura Nacional de Dados Espaciais
IPJBRJ	Instituto de Pesquisas Jardim Botânico do Rio de Janeiro
JBRJ	Jardim Botânico do Rio de Janeiro
KDD	<i>Knowledge Discovery in Databases</i>
LNCC	Laboratório Nacional de Computação Científica
LOD	<i>Linked Open Data</i>
MRAR	<i>Multi-Relation Association Rules</i>
MIT	<i>Massachusetts Institute of Technology</i>
NCCG	Núcleo de Computação Científica e Geoprocessamento
PHP	<i>Hypertext Preprocessor</i>
RJ	Rio de Janeiro
RDF	<i>Resource Description Framework</i>
SPARQL	<i>SPARQL Protocol and RDF Query Language</i>
SGBDs	Sistemas Gerenciadores de Bases de Dados
SGBDRs	Sistemas Gerenciadores de Bases de Dados Relacionais
SGBDGs	Sistemas Gerenciadores de Bases de Dados em Grafos

RESUMO

A Web de Dados é hoje uma fonte extremamente rica e que contém diferentes tipos de informações. Extrair informações que levem ao conhecimento é um grande desafio para o avanço da área científica. Diversos algoritmos e ferramentas são desenvolvidos para auxiliar no processo de extração de conhecimento, valendo-se de diferentes estratégias de busca, como por exemplo a mineração de dados por meio da mineração de regras de associação. Entretanto, esses algoritmos são custosos em termos computacionais, e por isso aplicá-los no contexto na Web de Dados parece uma tarefa impossível devido à complexidade de manipular as diferentes fontes de dados. Os *datasets* disponíveis na Web de Dados, são representados em grafo, por meio dos recursos em RDF. É possível analisar os dados em grafos com a mineração de regras de associação de multirrelação fazendo uso do algoritmo MRAR. Neste trabalho, propomos uma forma de minerar dados de um *dataset* na Web de Dados, visando estender tal *dataset* em análise com informações de fontes externas (*datasets* conectados), possibilitando a geração de regras úteis para o usuário, além de apresentar uma formalização e extensão do algoritmo MRAR. Foram realizados experimentos que comprovaram a viabilidade e validade da abordagem proposta em três diferentes cenários, apresentando resultados promissores.

ABSTRACT

The Web of data is an extremely rich source containing different kinds of information. To extract information leading to knowledge is a major challenge. Several algorithms and tools were developed to assist in the process of knowledge extraction, using different search strategies, such as the data mining by means of mining association rules. However, these algorithms are computationally expensive, so applying them in the context of the Web of data seems an impossible task due to the complexity of the different data sources. The *datasets* available on the Web of data are represented as a graph, using RDF. It is possible to analyze data in graphs by mining multirelation association rules, making use of the MRAR algorithm. In this work, we propose an approach of mining datasets from the Web of data, in such a way that the analysis is performed on an extended *dataset*, with information from external sources (connected datasets), allowing the generation of useful rules for the user. In addition, we also present a formalization and extension of the MRAR algorithm. Three experiments were also reported and demonstrated the viability and validity of the proposed approach, showing promising results.

1 INTRODUÇÃO

Devido às facilidades que a era digital oferece, onde todos podem ter acesso às novas tecnologias, produzir dados e/ou acompanhar o que acontece no mundo ficou à distância de um simples clique. Com a inovação tecnológica e o volume crescente de dados, surgem ferramentas e motores que tentam organizar e navegar por essa nuvem de dados (PICKLER, 2007; VIEIRA et al., 2012). A internet moderna não é apenas mais um lugar onde se disponibiliza dados (SOUZA; ALVARENGA, 2004). Os novos provedores de conteúdo devem oferecer formas e padrões que ajudem a localização, acesso e processamento dos dados disponíveis. Assim, passa a ser possível desenvolver maneiras de encontrar o que realmente se busca.

O crescente volume de dados acabou tornando-se um problema na área da computação, devido à dificuldade de manipular e processar a grande quantidade de informação existente. A *Big Data*, como é chamada, pode ser definida como uma coleção de volumosas e complexas bases de dados, onde operações simples de inserção, ordenação e remoção tornam-se custosas para os SGBDs (Sistemas Gerenciadores de Bases de Dados) tradicionais (VIEIRA et al., 2012). Apesar de importante, a grandeza da base de dados não é o que mais importa para esse conceito. Existem outros atributos importantes em *Big Data*, a saber, variedade de dados e velocidade de dados. Sendo assim, os três V's da *Big Data* (volume, variedade e velocidade) constituem uma definição mais abrangente, que possibilita ir além de apenas ser uma base com grandes volumes de dados (RUSSOM et al., 2011).

1.1 MOTIVAÇÃO

Atualmente, o mundo tem produzido um volume muito grande de dados (REZENDE et al., 2003). O avanço tecnológico e a utilização das redes sociais permitiram a todos produzir conteúdos. Além disso, os sensores presentes nas novas tecnologias produzem uma vasta quantidade de dados e informações o tempo todo.

Esse crescimento avassalador levou ao surgimento de movimentos como:

- Dados abertos (do inglês, *Open data*): Esse movimento tem como objetivo tornar os dados disponíveis, abertos e processáveis, a fim de facilitar a pesquisa. No Brasil, a

lei de acesso à informação (BRASIL, 2011) tornou acessível os dados governamentais, que a partir de então passariam a estar disponíveis sem restrições (VAZ et al., 2010), disponibilizados através do PORTAL BRASILEIRO DE DADOS ABERTOS¹. Essa lei, regulamenta o direito constitucional do cidadão ao acesso às informações produzidas ou detidas pelo Governo, que só entrou em vigor no dia 16 de maio de 2012.

São considerados dados abertos aqueles que: estão disponíveis para qualquer pessoa ou máquina acessar, não importando o fim; e, também, aqueles que não têm nenhum custo para sua utilização. Eles têm como contrapartida apenas dar crédito a fontes onde dados foram obtidos.

- *Linked Open Data*² (LOD): Com os dados disponíveis, é preciso encontrar formas para entendê-los. A iniciativa LOD, além de incentivar a disponibilização de dados abertos e processáveis, tem a finalidade de ligá-los, possibilitando constatar que, se um dado é encontrado num *dataset*, possivelmente um outro *dataset* terá o mesmo dado ou alguma informação complementar sobre ele (BIZER et al., 2009). A Figura 1.1 mostra um exemplo de alguns *datasets* que estão conectados, o que indica a existência de relações e interligações entre eles. O formato RDF (*Resource Description Framework*) é utilizado como padrão de representação de dados no LOD. Esse formato é que permite a formação da Web de Dados. Nele, os dados são organizados na forma de triplas que se interceptam formando um grafo direcionado. As informações disponibilizadas nesse ambiente cobrem uma vasta gama de tópicos, tais como: localizações geográficas, pessoas, empresas, livros, publicações científicas, dados estatísticos, entre outros (BIZER et al., 2009).

Todos esses movimentos que trabalham e disponibilizam grandes fontes de dados, levam à seguinte pergunta: O que fazer com essa quantidade gigantesca de dados que estamos produzindo? Uma resposta possível é realizar análises, para obter conhecimento útil. Existem várias maneiras de efetuar análises, como visto em Goldschmidt et al. (2015), entre elas está a mineração de regras de associação. Esse tipo de análise é capaz de descobrir dados que coocorrem com frequência, por exemplo: um determinado ator pode atuar frequentemente em filmes de um determinado diretor. Além disso, como os *datasets* estão interligados, torna-se possível navegar por suas conexões e encontrar novas associações.

¹<http://dados.gov.br/>

²<http://lod-cloud.net/>

1.2 CARACTERIZAÇÃO DO PROBLEMA

Tavares et al. (2015) mostram que o crescimento do número de fontes da Web de Dados tem motivado o desenvolvimento de aplicações e ferramentas que consomem o conteúdo disponibilizado por essas fontes. Porém, essas aplicações, em geral, utilizam *datasets* específicos, o que limita a quantidade de informação, visto que só são obtidos resultados através de consultas.

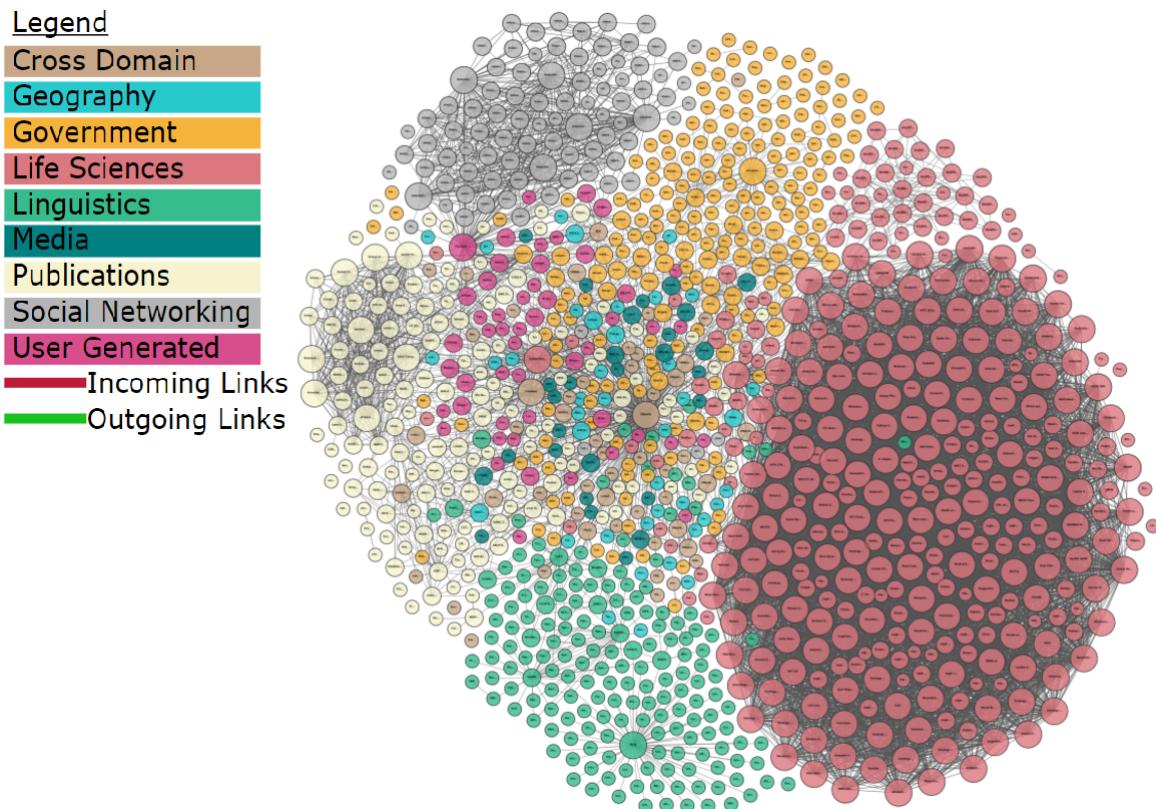


FIG. 1.1: Diagrama em nuvem do Linked Open Data de 2017. Fonte: <http://lod-cloud.net/versions/2017-08-22/lod.svg>.

Com o crescimento do número de *datasets*, é importante realizar análises sobre esses dados que estão espalhados na Web de Dados. Os algoritmos existentes são de alta complexidade e demandam um elevado poder de processamento, como os algoritmos apresentados em Ramezani et al. (2014) e Elseidy et al. (2014), tornando inviável sua aplicação na Web de Dados como um todo. Se fossem utilizados esses tipos de algoritmos, no contexto da Web de Dados, por causa do tráfego intenso de informações, na rede o processo de análise seria inviável. Mesmo que os *datasets* fossem replicados localmente para evitar os custos de comunicação, ainda assim, dependendo do tamanho, o processamento pode ser muito custoso.

Devido à imensa quantidade de dados existentes, cabe destacar o seguinte questionamento: como viabilizar a análise sobre a Web de Dados de modo a encontrar novas ligações e extrair conhecimento útil?

1.3 HIPÓTESES / QUESTÕES DE PESQUISA

De acordo com o que foi apresentado anteriormente, algumas questões surgiram com o objetivo de identificar maneiras de suprir a ausência de estudos, que fazem uso da mineração de regras de associação para reconhecer novos *datasets* e possibilitar a ampliação de seu conhecimento sobre os dados originais. Essas questões serão apresentadas a seguir:

Seria possível analisar um determinado *dataset* levando em consideração suas ligações com outros e descobrir novas ligações?

Se o *dataset* em questão tiver várias ligações com outros, como fazer para restringir/- selecionar dados de modo a evitar a inviabilização da análise? Qual o critério?

A mineração de dados em *dataset* local pode encontrar novos padrões. Será que ao ampliá-lo com dados provenientes de outros *datasets* podemos encontrar mais associações identificando novas regras úteis?

1.4 OBJETIVO

O objetivo principal da dissertação é desenvolver um método para viabilizar a análise sobre a Web de Dados. Esse método visa possibilitar a ampliação de um determinado *dataset* de maneira controlada, a partir de uma seleção de suas conexões com outros *datasets* e encontrar novas associações com o auxílio da mineração de regras de associação em grafos.

1.5 CONTRIBUIÇÕES ESPERADAS

As contribuições esperadas para este trabalho são:

- (i) Especificação e implementação de um método que torne possível a análise sobre *datasets* interligados na Web de Dados e encontrar novas regras úteis que demonstrem a validade dos resultados através da mineração de regras de associações.
- (ii) Utilização da técnica de mineração de regras de associação em grafos (multirrelação) como meio para identificação e seleção de um conjunto de *datasets* (*datasets* externos) conectados a um determinado *dataset* (*dataset* alvo) e, a partir desta seleção, esta técnica é também usada para a ampliação do *dataset* alvo.

- (iii) Ampliação do conhecimento sobre os dados de um *dataset* real (Jabot) sugerindo novas ligações com outros *datasets*.

1.6 ORGANIZAÇÃO DO TRABALHO

Este trabalho está dividido em sete capítulos. No capítulo atual, foi apresentada a introdução do trabalho, descrevendo a motivação, caracterização do problema, hipótese, objetivo e as contribuições esperadas. No capítulo seguinte, serão apresentadas as fundamentações teóricas pertinentes. No capítulo 3, discorreremos sobre regras de associação de multirrelação. No capítulo 4, falaremos sobre a abordagem proposta. No capítulo 5, apresentaremos a comparação de alguns dos resultados de nossos experimentos. Já no capítulo 6, serão apresentados os trabalhos relacionados ao tema desta dissertação e, por fim, no último capítulo será apresentada a conclusão do presente trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

A seguir, serão detalhados alguns conceitos importantes utilizados neste trabalho. Eles possuem relação com a importância da publicação dos dados interligados na Web de Dados e da utilização destes *datasets*. Definições como o modelo RDF, recurso externo, bem como o processo de descoberta de conhecimento (KDD), a mineração de regras de associação e a mineração de regras de associação de multirrelação serão apresentadas nesta seção.

2.1 WEB DE DADOS

A Web de Dados pode ser definida como uma rede de informações, onde os nós estejam semanticamente ligados, formando um grande grafo global, com informações advindas de várias fontes diferentes ao redor do planeta como mostram Eduardo e Segundo (2015). A Web de Dados é composta por *datasets*, que são conjuntos de dados expressos segundo o modelo RDF. Cada *dataset* versa sobre um domínio de conhecimento. No RDF é possível representar as informações ou recursos Web por meio de triplas, compostas por: Sujeito, Predicado e Objeto, sendo assim, elas podem ser organizadas como um grafo direcionado. Sujeito é o recurso descrito; o objeto pode ser um valor literal ou um recurso relacionado ao sujeito; e o predicado, também chamado de propriedade, indica a relação que existe entre o sujeito e o objeto. Os elementos da tripla são identificados por uma URI (TAVARES et al., 2015). A Figura 2.1 mostra uma tripla RDF indicando em qual cidade localiza-se a Universidade Federal do Rio de Janeiro.

<i>sujeito</i>	<i>predicado</i>	<i>objeto</i>
http://dbpedia.org/page/Federal_University_of_Rio_de_Janeiro	http://dbpedia.org/ontology/city	http://dbpedia.org/page/Rio_de_Janeiro

FIG. 2.1: Exemplo de uma tripla RDF com links na *DBpedia*.

Como exemplo de *dataset*, temos a *DBpedia*³ que, de acordo com sua página de informações, a versão em inglês descreve 4,58 milhão de coisas, das quais 4,22 milhões classificam-se em uma ontologia consistente⁴. Nessa base de dados podem ser encontra-

³<http://dbpedia.org/>

⁴<http://wiki.dbpedia.org/about>

das informações sobre: 1.445.000 pessoas, 735.000 lugares, 411.000 trabalhos criativos (incluindo 123.000 álbuns de música, 87.000 filmes e 19.000 jogos de vídeo), 241.000 organizações (incluindo 58.000 empresas e 49.000 instituições educacionais), 251.000 espécies e 6.000 doenças. Esse *dataset* tem o objetivo de extrair o conteúdo estruturado das informações da Wikipédia⁵ e, então, disponibilizar na Web de Dados. A *DBpedia* cria uma rede de ligações entre os dados permitindo aos usuários realizarem suas próprias consultas de forma simples, similar a uma consulta em um banco de dados. É um dos exemplos mais famosos da iniciativa *Linked Open Data* (BIZER et al., 2009).

Os *datasets* na Web de Dados permitem acesso ao seu conteúdo por meio de navegação (Web crawling), RDF *dump* ou via consultas SPARQL. Como esses estão interligados através dos chamados recursos externos, o usuário pode iniciar sua busca em um *dataset* e, logo em seguida, mover-se através desses recursos, podendo alcançar intermináveis *datasets* da Web de Dados (PALETTA et al., 2015). Por exemplo, ao navegar pelo *dataset* da *DBpedia*, na página (*site*) que descreve o recurso que representa o Instituto de Tecnologia de Massachusetts⁶ (MIT), encontra-se referência à seguinte tripla presente naquele *dataset*:

```
http://dbpedia.org/resource/Massachusetts_Institute_of_Technology
http://www.w3.org/2002/07/owl#sameAs
https://www.wikidata.org/wiki/Q49108.
```

Nota-se nessa tripla, que a propriedade “*sameAs*” pertence ao vocabulário da linguagem OWL, e indica uma similaridade entre o sujeito e o objeto da tripla. Além disso, observa-se que o objeto da tripla tem uma URI de outro *dataset*, o Wikidata. Assim, ao clicar neste objeto, o usuário passa a navegar pelos dados deste outro *dataset*.

Uma outra tripla, através da mesma propriedade (“*sameAs*”), leva a um recurso do *dataset* GeoNames⁷, agora com informações complementares, exibindo um mapa e sua geolocalização.

Os objetos presentes em um *dataset*, que apontam para outros *datasets*, passaremos a denominar “recursos externos”.

O SPARQL é uma linguagem de consulta e protocolo que permite realizar consultas a dados estruturados que estão no formato RDF. Ela funciona baseada na correspondência de padrões de triplas (TAVARES et al., 2015). O termo SPARQL vem do inglês e é um acrônimo recursivo que significa *SPARQL Protocol and RDF Query Language*. Uma query SPARQL consiste em uma estrutura simples, podendo ser composta por apenas

⁵<https://www.wikipedia.org/>

⁶<http://dbpedia.org/page/MIT>

⁷<http://www.geonames.org/4943351/massachusetts-institute-of-technology.html>

duas cláusulas, SELECT e WHERE. A cláusula SELECT identifica as variáveis que farão parte do resultado da consulta e a WHERE mostra o padrão básico do grafo a partir do qual se quer selecionar os dados. Outras cláusulas, como a cláusula FROM, também podem ser usadas. A FROM é usada caso seja necessário selecionar uma fonte de dados específica em uma consulta, mas seu uso não é obrigatório.

Ao acessar a página de consulta da *DBpedia*⁸, buscando as propriedades e objetos existentes sobre o MIT, seria obtido como resultado, por exemplo, uma lista das afiliações acadêmicas do MIT, como mostra Figura 2.2, utilizando a seguinte consulta expressa na linguagem SPARQL:

```
SELECT distinct ?predicado ?objeto
WHERE {
    <http://dbpedia.org/resource/Massachusetts_Institute_of_Technology> ?predicado ?objeto
}
```

predicado	objeto
http://dbpedia.org/resource/Massachusetts_Institute_of_Technology	http://dbpedia.org/resource/Association_of_American_Universities
http://dbpedia.org/resource/Massachusetts_Institute_of_Technology	http://dbpedia.org/resource/Association_of_Independent_Colleges_and_Universities_in_Massachusetts
http://dbpedia.org/resource/Massachusetts_Institute_of_Technology	http://dbpedia.org/resource/Association_of_Independent_Technological_Universities
http://dbpedia.org/resource/Massachusetts_Institute_of_Technology	http://dbpedia.org/resource/Association_of_Public_and_Land-Grant_Universities
http://dbpedia.org/resource/Massachusetts_Institute_of_Technology	http://dbpedia.org/resource/National_Association_of_Independent_Colleges_and_Universities
http://dbpedia.org/resource/Massachusetts_Institute_of_Technology	http://dbpedia.org/resource/The_Consortium_on_Financing_Higher_Education
http://dbpedia.org/resource/Massachusetts_Institute_of_Technology	http://dbpedia.org/resource/Universities_Research_Association

FIG. 2.2: Exemplo de resultados da consulta SPARQL da página *DBpedia* com os predicados e objetos associados ao MIT.

Como outro exemplo, é possível buscar todos os predicados e objetos associados ao Instituto Militar de Engenharia (IME). A Figura 2.3 mostra, em parte dos resultados, apenas as categorias que estão associadas a essa Instituição existentes na base da *DBpedia*. A consulta utilizada pode ser vista a seguir.

```
SELECT distinct ?predicado ?objeto
WHERE {
    <http://dbpedia.org/resource/Instituto_Militar_de_Engenharia> ?predicado ?objeto
}
```

2.2 SISTEMAS DE BANCOS DE DADOS EM GRAFOS

Os Sistemas Gerenciadores de Banco de Dados Relacionais (SGBDRs) dominaram os meios acadêmicos e empresariais por décadas, principalmente, por causa do uso de modelos e linguagens de consultas intuitivos e, também, por garantirem as propriedades ACID

⁸<http://dbpedia.org/sparql>

predicado	objeto
http://purl.org/dc/terms/subject	http://dbpedia.org/resource/Category:Educational_institutions_established_in_the_1790s
http://purl.org/dc/terms/subject	http://dbpedia.org/resource/Category:Engineering_universities_and_colleges_in_Brazil
http://purl.org/dc/terms/subject	http://dbpedia.org/resource/Category:1792_establishments_in_Brazil
http://purl.org/dc/terms/subject	http://dbpedia.org/resource/Category:Military_academies_of_Brazil
http://purl.org/dc/terms/subject	http://dbpedia.org/resource/Category:Brazilian_Army
http://purl.org/dc/terms/subject	http://dbpedia.org/resource/Category:Universities_and_colleges_in_Rio_de_Janeiro_(city)
http://purl.org/dc/terms/subject	http://dbpedia.org/resource/Category:Instituto_Militar_de_Engenharia

FIG. 2.3: Demonstração de resultados da consulta SPARQL da página *DBpedia* com os predicados e objetos associados ao IME.

(Atomicidade, Consistência, Isolamento e Durabilidade) nas diversas aplicações existentes. Mesmo com todos os benefícios de usar os SGBDRs, modelos de dados com alta complexidade podem enfrentar problemas ao transformá-los para o modelo relacional, principalmente quando se trata de aplicações com grande volume de dados (PENTEADO et al., 2014).

Os Sistemas Gerenciadores de Bancos de Dados em Grafos (SGBDGs) vieram como alternativa aos SBGBDRs, especialmente onde a interconectividade dos dados é um dos fatores importantes. Este sistema em grafos surgiu em torno dos anos 80 e 90 (KUNII, 1987; LECLUSE et al., 1988; GYSSENS et al., 1990) juntamente com o modelo orientado a objetos, porém logo perdeu espaço para os bancos de dados semiestruturados em virtude do crescimento do XML (ANGLES; GUTIERREZ, 2008). Entretanto, com o surgimento da Web de Dados e a popularização das redes sociais, o interesse pelos bancos de dados em grafos ressurgiu (LIPTCHINSKY et al., 2013).

Os bancos de dados em grafos têm em sua modelagem os vértices (nós) e arestas (relações) como em um grafo simples, o que facilita modelagens complexas. Além disso, eles podem definir naturalmente as relações existentes em uma base de dados. Atualmente, estão disponíveis diferentes bancos de dados em grafos. A Tabela 2.1 relaciona alguns deles.

TAB. 2.1: Lista de bancos de dados em grafos.

Nome	Versão	Linguagem
AllegroGraph ⁹	6.4.0 (12/2017)	C#, C, Common Lisp, Java, Python
InfiniteGraph ¹⁰	3.0 (01/2013)	Java
OrientDB ¹¹	2.2.24 (07/2017)	Java
Neo4J ¹²	3.3.1 (11/2017)	Java, .NET, JavaScript, Python, Ruby

O Neo4J foi lançado em fevereiro de 2010 pela empresa *Neo Technology*¹³, desen-

¹³<http://www.neotechnology.com/>

volvido em Java, e já se encontra na versão 3.3.1, podendo ser instalado nas principais plataformas (Windows, Linux e Mac). Possui duas versões de licenciamento para utilização, uma versão aberta e outra proprietária. A principal diferença existente é que a proprietária tem código fechado, oferece suporte aos usuários e possui uma variedade maior de recursos para garantir o desempenho de grandes aplicações. O Neo4J ainda oferece dois tipos de arquiteturas, uma centralizada e outra distribuída com suporte à replicação. Além de possibilitar consultas na linguagem SPARQL, ele possui uma linguagem própria, o Cypher. O armazenamento físico dos dados pode ocorrer tanto em memória quanto em disco e seu modelo físico se baseia em repositórios chave-valor (PENTEADO et al., 2014). Cypher é uma linguagem de consulta para grafos que, com o objetivo de se tornar o “SQL” para as consultas em grafos, passou a ser uma linguagem aberta. Ela fornece uma maneira familiar e legível para corresponder a padrões de nós e arestas dentro de um conjunto de dados em grafo. Como o SQL, a linguagem Cypher declarativa que permite aos usuários indicar quais comandos serão executados sobre seus dados em grafo, tal como match, insert, update ou delete, sem necessitar descrever/programar exatamente como fazê-lo. A seguir será apresentado um exemplo simples de consulta Cypher, que buscará o elenco dos filmes que começem com a letra “T”:

```

MATCH (actor :Person) -[:ACTED_IN]->(movie :Movie)
WHERE movie.title STARTS WITH "T"
RETURN movie.title AS title, collect(actor.name) AS cast
ORDER BY title ASC LIMIT 10;

```

2.3 MINERAÇÃO DE DADOS

A mineração de dados tornou-se um dos passos de maior importância no processo de extração de conhecimento KDD, do inglês, *Knowledge-Discovery in Databases*. O KDD é um processo para extrair informações de bases de dados, criando relações de interesse que normalmente não são observadas pelos pesquisadores do assunto, podendo ainda auxiliar na validação do conhecimento extraído. Esse tipo de análise sobre os dados vem auxiliando os tomadores de decisão e muitas das vezes é um fator importante para melhorar a eficiência de uma empresa. O problema de encontrar padrões frequentes e regras de associação transforma-se em um interesse comum entre os pesquisadores por fazer parte da vida cotidiana (FAYYAD et al., 1996). A Figura 2.4 destaca os passos aplicados para transformar dados em conhecimento. Todos os passos do processo são sequenciais, mas cada etapa pode ser reiniciada sempre que necessário, visando adequar os dados ao máximo, possibilitando alcançar o conhecimento verdadeiro.

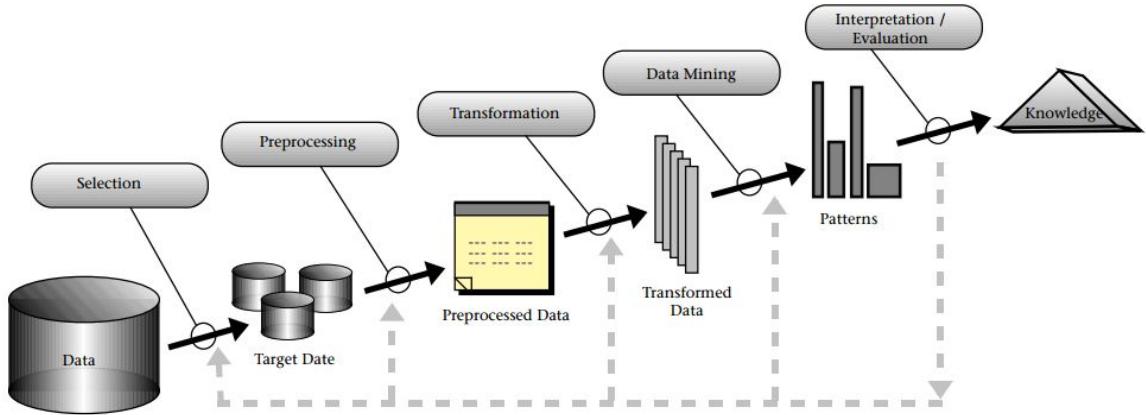


FIG. 2.4: Representação dos processos de KDD. Fonte: Fayyad et al. (1996).

2.3.1 MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO

Uma das principais tarefas na busca por novos conhecimentos em bases de dados é a chamada Mineração de Regras de Associação (do inglês, *Association Rules Mining* (ARM) (GOLDSCHMIDT et al., 2015). Em essência, essa tarefa consiste em identificar regras de associação frequentes e válidas em um conjunto de dados (AGRAWAL et al., 1993).

A utilização de padrões encontrados implica geração das regras de associação desejadas, se algum item ou evento ocorre em conjunto, alguns outros específicos também vão ocorrer com certa probabilidade, isso é conhecido como confiança. Dentre os objetivos da ARM, destaca-se o processo de encontrar padrões frequentes de dados existentes. Através desses padrões, são encontrados os itens que coocorrem frequentemente.

Uma regra de associação é uma implicação da forma $X \rightarrow Y$, onde X e Y são conjuntos de itens e $X \cap Y = \emptyset$. Um item é um predicado que pode assumir valor verdadeiro ou falso em função do registro de dados selecionado. Um exemplo de regra de associação pode ser $\{Sexo = M, JogaFutebol = S\} \rightarrow \{Saude = Boa\}$, onde $Sexo = M$, $JogaFutebol = S$ e $Saude = Boa$ são exemplos de itens.

O conceito de itens frequentes é apresentado por Agrawal et al. (1993), onde seus conjuntos de “*Frequent Itemset*” são criados respeitando um valor mínimo para cada item. Com os mesmos, é possível minerar regras de associação. Para cada regra gerada, é aplicado um suporte e confiança mínimos, respectivamente, chamados de (*MinSup*) e (*MinConf*), obtendo assim resultados válidos.

Uma regra de associação $X \rightarrow Y$ é dita frequente (resp. válida) se, e somente se, $|X \cup Y|/|D| \geq MinSup$ (resp. $|X \cup Y|/|X| \geq MinConf$). $|D|$ representa a quantidade total de registros de dados disponíveis no conjunto de dados D . $|I|$ é a quantidade de registros do conjunto de dados que satisfazem simultaneamente a todos os itens pertencentes ao

conjunto I . $MinSup$ (resp. $MinConf$) é um parâmetro definido pelo usuário que estabelece uma frequência (resp. confiança) mínima para que a regra seja considerada frequente (resp. válida) no conjunto de dados. Como exemplo, a regra $\{\text{leite}, \text{pão}\} \rightarrow \{\text{manteiga}\}$, extraída para os dados da Figura 2.5, tem uma confiança igual a $\frac{0.2}{0.4} = 0.5^{14}$ na base de dados, o que significa que para 50% das transações que contém leite e pão a regra está correta.

Trans	Leite	Pão	Manteiga	Cerveja
1	Sim	Sim	Não	Não
2	Não	Sim	Sim	Não
3	Não	Não	Não	Sim
4	Sim	Sim	Sim	Não
5	Não	Sim	Não	Não

FIG. 2.5: Um exemplo clássico de uma base de dados composta de 4 itens e 5 transações.

Em geral, o processo de mineração de regras de associação ocorre em duas etapas (AGRAWAL et al., 1993). A primeira, de maior custo computacional, busca por conjuntos de itens frequentes que ocorrem simultaneamente no conjunto de dados. A segunda consiste em identificar as regras válidas a partir de cada conjunto de itens considerado frequente na etapa anterior.

LIFT é um parâmetro estatístico que também pode ser utilizado em mineração de dados para definir o grau de interesse de uma regra. Em uma regra de associação, o *LIFT* é definido para indicar o quanto mais frequente se torna Y , quando X ocorre. Essa medida é computada da seguinte forma: $\text{Lift}(X \rightarrow Y) = \frac{\text{sup}(X \cup Y)}{\text{sup}(Y) * \text{sup}(X)}$. Sempre que $\text{Lift}(X \rightarrow Y)$ for 1, então X e Y serão independentes. Quando $\text{Lift}(X \rightarrow Y)$ for maior que 1, X e Y são positivamente dependentes, ou seja, a presença de X aumenta as probabilidades da ocorrência de Y . Entretanto, quando $\text{Lift}(X \rightarrow Y)$ for menor que 1, X e Y são negativamente dependentes, isso significa que a presença de X diminui as chances da ocorrência de Y . Essa medida pode ser interpretada da seguinte maneira: quanto maior for o valor do Lift, mais interessante é a regra (BÜRKLE, 2006). Por exemplo, a regra $\{\text{leite}, \text{pão}\} \rightarrow \{\text{manteiga}\}$ encontrada, ao analisar os dados da Figura 2.5, possui seguinte valor de LIFT $\frac{0.2}{0.4 * 0.4} = 1.25$.

Outro parâmetro que pode ser utilizado para avaliar uma regra de associação é a *Convicção*. A *convicção* pode ser definida da seguinte forma: $\text{Conv}(X \rightarrow Y) = \frac{1 - \text{sup}(Y)}{1 - \text{conf}(X \rightarrow Y)}$.

¹⁴Foi utilizado (.) como separador de decimal a fim de uniformizar o uso com o que é apresentado na ferramenta implementada.

Essa medida permite avaliar a regra de associação considerando o sentido das setas (implicação), significa que, a $Conv(X \rightarrow Y)$ é diferente da $Conv(Y \rightarrow X)$ (BÜRKLE, 2006). Seus valores podem variar entre 0 e ∞ , apresentando o valor 1 quando os conjuntos X e Y são independentes, e ∞ quando o valor da confiança for igual a 100%. Para a regra $\{\text{leite}, \text{pão}\} \rightarrow \{\text{manteiga}\}$, o resultado de *convicção* encontrado é $\frac{1-0.4}{1-0.5} = 1.2$. Pode ser interpretado como a razão da frequência esperada que X ocorre sem Y .

2.4 CONSIDERAÇÕES FINAIS

Nesta seção, foram apresentados os conceitos básicos relacionados ao tema do trabalho, entre eles, a Web de Dados, os sistemas de bancos de dados em grafos e a mineração de dados. Um outro conceito importante trata das regras de associação de multirrelação, que, por ser o foco deste trabalho, será apresentado e explorado mais profundamente na seção seguinte.

3 REGRAS DE ASSOCIAÇÃO DE MULTIRRELAÇÃO

Nesta seção serão tratados os assuntos relacionados à mineração de regras de associação de multirrelação. Tendo em vista que o foco deste trabalho está na descoberta de relações em dados representados em grafos (multirrelação), buscou-se por soluções de mineração de regras de associação de multirrelação. Até onde se sabe, somente um trabalho (RAMEZANI et al., 2014) versa sobre esse tema e apresenta um algoritmo (MRAR) específico para esse tipo de análise. Sendo assim, trataremos sobre a implementação do algoritmo MRAR, também serão apresentados alguns exemplos de uso para ilustrar a sua execução e, por fim, algumas definições serão apresentadas junto ao formalismo que foi concebido sobre esse tema.

3.1 MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO DE MULTIRRELAÇÃO

Ramezani et al. (2014) apresentam um novo algoritmo para minerar regras de associação, denominado MRAR (Mining Multi-Relation Association Rules). Diferente dos algoritmos clássicos, geralmente aplicados a banco de dados relacionais, são geradas regras de associações a partir de um grafo direcionado composto por nós e arestas. Tal algoritmo considera que cada item do grafo que compõe a regra representa uma entidade e várias relações, onde as arestas indicam um relacionamento indireto entre as entidades.

A mineração de regras de associação de multirrelação via MRAR é inspirada em princípios análogos aos utilizados na mineração de regras de associação tradicional. No caso do grafo direcionado, a ideia é encontrar caminhos frequentes que possam ocorrer, i.e., caminhos que percorram diferentes relações e cheguem em um mesmo nó. Para facilitar o entendimento do algoritmo MRAR, faz-se necessário apresentar as seguintes definições, de acordo com Ramezani et al. (2014):

- a) *MinSup*: O suporte absoluto do conjunto de itens S é o número de transações em D que contêm S . Dessa forma, o apoio relativo do S é a percentagem das transações em D que contêm S .

$$Sup_{abs}(S) = |S|$$

$$Sup_{rel}(S) = (|S|/|D|) * 100\%.$$

- b) *MinConf*: A confiança de uma regra de Associação $R = X \rightarrow Y$ é definida pelo

suporte do conjunto de todos os itens que fazem parte da regra, dividida pelo suporte do antecedente da regra. Elas só são consideradas válidas se sua confiança atinge ou ultrapassa um determinado valor (confiança mínima, definida pelo usuário)

$$Conf(R) = (Sup(\{X \cup Y\})/Sup(X)) * 100\%$$

- c) *Itemset* frequente: Conjunto de itens com suporte maior que o *MinSup* definido.
Na mineração, o objetivo é encontrar todos os conjuntos de itens frequentes.
- d) *ItemChains*: Cada *ItemChain* mostra que um conjunto de entidades está ligado a um *Endpoint* através de relações comuns.
- e) *MinLevel* e *MaxLevel*: O algoritmo MRAR gera *ItemChains* com varias relações. O *MinLevel* e *MaxLevel* é usado para definir o número mínimo e máximo de relações utilizadas.
- f) *Endpoints*: Vértice do grafo que inicia a busca do algoritmo.
- g) *List_EntityInfo*: Estrutura de dados que mostra onde cada vértice está conectado.

Por exemplo, tomando como *EndPoint* o nó “Humid” do grafo da Figura 3.1, há um caminho que se repete, formando a seguinte composição das relações “Live_In (Near_By (Climate_Type (Humid)))”, a partir dos nós “Hasan” e “Reza”. Nesse mesmo exemplo, temos que esses mesmos nós também formam caminhos até o nó “Good”, “Health_Condition (Good)”. Com base nesses caminhos frequentes, tem-se a geração de uma regra de associação de multirrelação: “Live_In (Near_By (Climate_Type (Humid))) → Health_Condition (Good)”, dizendo que quem vive perto de uma cidade com o clima úmido implica ter a condição de saúde boa, com o suporte de 11% e a confiança de 69%.

As regras de associação de multirrelação podem ser compostas por mais de um item no antecedente da regra, como é visto em Ramezani et al. (2014). Por exemplo, a regra “Live_In(Near_By(Climate_Type(Humid))), AgeLessThan(20) → Health_Condition(Good)” contém dois itens no antecedente da regra. O primeiro item apresenta três relações (*Live_In*, *near_by* e *climate_type*) e o segundo apenas a relação (*AgeLessThan*), já no consequente da regra é possível ver apenas uma relação (*HealthCondition*).

O algoritmo MRAR é subdividido em 3 algoritmos menores (*Generating 2-Large ItemsChains*, *Generating L-Large ItemsChains* e *Generating Multi-Relation Association Rules*), como mostra a Figura 3.2. O MRAR também é responsável por iniciar o processo e receber os parâmetros necessários para dar continuidade a todos os passos seguintes. Os parâmetros necessários para sua perfeita execução são informações como: um grafo

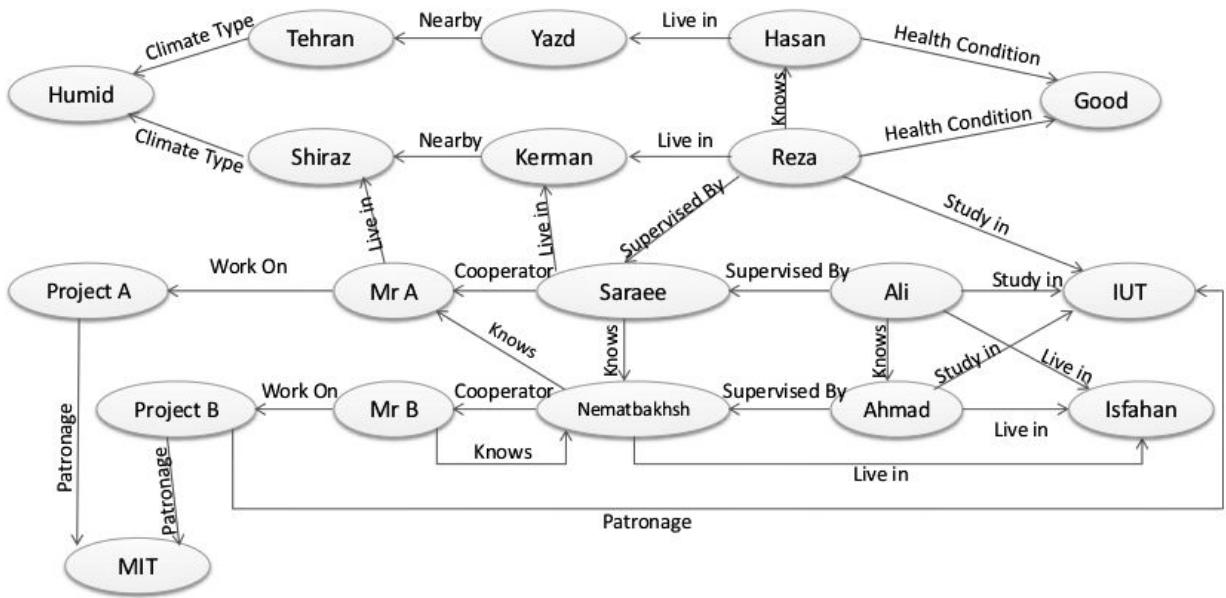


FIG. 3.1: Grafo direcionado com rótulo nas arestas. Fonte: Ramezani et al. (2014).

direcionado com rótulos nas arestas, o *MinSup*, a *MinConf*, o *MinLevel* e o *MaxLevel*. Antes de iniciar a execução do próximo algoritmo, é realizado um pré-processamento sobre os dados, assim, uma lista com informações das entidades é formada e chamada de *List_EntityInfo* onde os itens *Endpoints* são identificados, juntamente com as relações e nós associados.

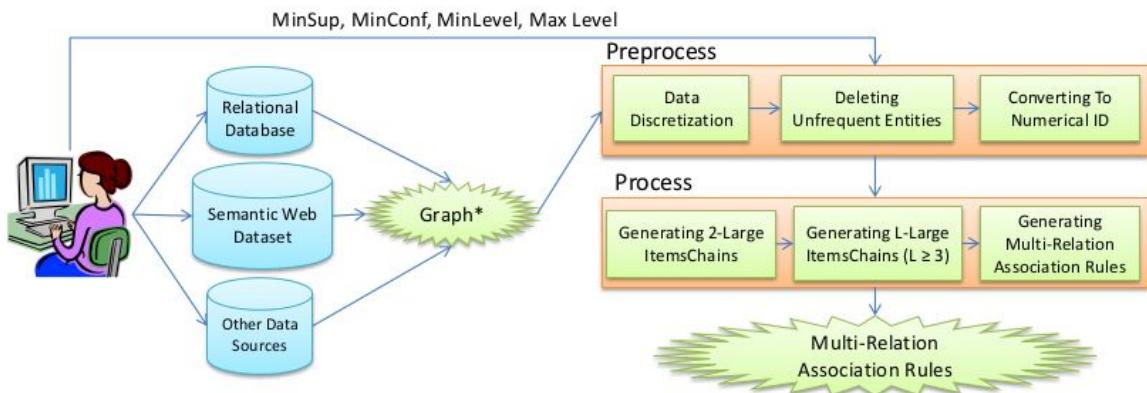


FIG. 3.2: Fluxo de trabalho do algoritmo MRAR. Fonte: Ramezani et al. (2014).

Na sequência, o primeiro sub-algoritmo (*Generating 2-Large ItemsChains*) é executado para gerar a lista de itens frequentes (*ItemChains*). Ela é composta pelos itens que estão conectados por relações comuns caminhando pelo grafo, respeitando a distância entre os valores mínimos e máximos de level (*MinLevel* e *MaxLevel*). Na Tabela 3.1 é possível ver um exemplo dos *ItemChains* que foram gerados a partir da aplicação do

algoritmo sobre os dados do grafo visto na Figura 3.1.

TAB. 3.1: Exemplo de *ItemChains* gerados após a aplicação do *GenerateItemChains*, para os dados do grafo da Figura 3.1.

<i>ChanID</i>	Entidades	Relações	EndPoint	Sup
1	Hasan, Reza	Health Condition	Good	2/19
2	Ali, Ahmad, Nematbakhsh	Live in	Isfahan	3/19
3	Ali, Ahmad, Reza	Study in	IUT	3/19
4	Yazd, Kerman	Near, ClimateType	Humid	2/19
5	Hasan, Reza	LiveIn, Near, ClimateType	Humid	2/19
6	Ali, Ahmad, Reza	Supervised By, Cooperator, Work On, Patronage	MIT	3/19

Em seguida, no segundo sub-algoritmo (*Generating L-Large ItemsChains*) é feita uma verificação combinando os itens da lista de *Itemchains*. Se os números de interseções existentes para as entidades forem maiores ou iguais ao suporte mínimo definido, uma nova lista é criada com os valores dessa combinação, chamada de *LargeItemChain*. Por exemplo, ao analisar as entidades das linhas 1 e 2 da Tabela 3.1, é possível perceber que a interseção entre suas entidades é nula (1 {Hasan, Reza} e 2 {Ali, Ahmad, Nematbakhsh}), pois as entidades são diferentes, sendo assim, elas não serão adicionados a Tabela 3.2. Entretanto, ao analisar as linhas 1 e 5, temos que a intercessão entre (1 {Hasan, Reza}) e (5 {Hasan, Reza}) é positiva (igual 2). Agora, a próxima verificação poderá ser feita para descobrir se o cálculo do suporte atende ao que foi definido pelo usuário. Alguns desses resultados podem ser vistos na Tabela 3.2.

TAB. 3.2: Exemplo de *LargeItemChain*, para os dados da Tabela 3.1.

<i>ChainIDs</i>	Valor da intercessão	Suporte
1,5	2 {Hasan, Reza}	2/19
2,3	2 {Ali, Ahmad}	2/19
2,6	2 {Ali, Ahmad}	2/19
3,6	2 {Ali, Ahmad, Reza}	3/19

Por fim, o último sub-algoritmo (*Generating Multi-Relation Association Rules*) é executando recebendo como parâmetro de entrada a lista de *LargeItemChain*. O algoritmo compara os elementos dessa lista identificando os antecedentes e consequentes, verifica se eles atendem ao critério da confiança mínima (*MinConf*). Por exemplo, se a combinação entre *ChainIDs* da lista de *LargeItemChain* for igual a {2,3,6}, o algoritmo separa os antecedentes e consequentes, calcula a confiança mínima e, então, gera uma lista com todas as regras de associação de multirrelação. A primeira linha da Tabela 3.3 codifica a seguinte

regra que, traduzindo para o português significa que: “Quem VIVE em ISFAHAN e ESTUDA em IUT → é SUPERVISIONADO por uma pessoa, que COOPERA com alguém, que TRABALHA em um projeto, que é PATROCINADO pelo MIT”.

TAB. 3.3: Exemplo de Regras de associação de multirrelação gerado após a aplicação do algoritmo *Generating Multi-Relation Association Rules*, combinando os dados da Tabela 3.2.

Antecedente	Consequente	Suporte	Confiança
2,3	6	2/19	1.00
2,6	3	2/19	1.00
3,6	2	2/19	0.66

3.1.1 IMPLEMENTAÇÃO E VALIDAÇÃO DO MRAR

Durante a etapa de revisão da literatura, foi possível perceber a necessidade de implementar o algoritmo MRAR. Ele foi desenvolvido para que fosse possível recriar o cenário apresentado por Ramezani et al. (2014). A implementação foi feita utilizando os seguintes softwares: PHP, JavaScript, JQuery e Bootstrap. Durante esse processo, encontramos algumas dificuldades relacionadas a falta de clareza do autor e a falta de um formalismo bem definido para os termos utilizados. Portanto, foi preciso criar um formalismo próprio, que será descrito nas próximas seções, sendo essa uma de nossas contribuições.

Com o entendimento e a implementação do algoritmo MRAR, permitiu-se extrair regras como as apresentadas em Ramezani et al. (2014), utilizando o mesmo conjunto de dados apresentado na Figura 3.1. Para demonstrar a utilização dessa ferramenta, alguns exemplos de uso serão ilustrados nas seções seguintes.

3.1.1.1 ESTUDO DE CASO 1

Como estudo de caso, neste exemplo, foi utilizado um grafo que representa uma rede de relacionamentos entre professores, alunos, instituições e cidades, construídos a partir do grafo visto na Figura 3.1 (19 nós e 31 arestas). Aplicou-se o algoritmo MRAR, com os valores de corte (*cut-off*) para o suporte mínimo ($MinSup = 10\%$) e para confiança ($MinConf = 70\%$), os mesmos valores utilizados nos exemplos vistos em Ramezani et al. (2014). Ocorreu que o resultado inicial obtido foi de um conjunto composto por 437 variações de regras de associação de multirrelação. A Tabela 3.4 tem como resultado algumas regras que foram geradas após a aplicação do algoritmo.

É importante destacar que os valores de suporte e confiança encontrados como resultados exibidos, nessa tabela, acabaram sendo valores parecidos. Uma vez que o grafo é pequeno e os valores atribuídos ao algoritmo como critérios necessários para criação das regras, fizeram com que as regras com valores diferentes, porém menores, fossem eliminadas dos resultados.

TAB. 3.4: Exemplo de regras encontradas após a aplicação do MRAR com o *dataset* visto na Figura 3.1.

Antecedente	Consequente	Conf	Sup
HealthCondition(Good)	LiveIn(NearBy(ClimateType(Humid)))	1.00	0.11
LiveIn(Kerman)	LiveIn(NearBy(Shiraz))	1.00	0.11
LiveIn(Kerman)	LiveIn(NearBy(ClimateType(Humid)))	1.00	0.11
SupervisedBy(LiveIn(Kerman))	SupervisedBy(LiveIn(NearBy(Shiraz)))	1.00	0.11
StudyIn(IUT)	SupervisedBy(Cooperator(WorkOn(Patronage(MIT))))	1.00	0.16
SupervisedBy(Cooperator(WorkOn(Patronage(MIT))))	StudyIn(IUT)	1.00	0.16

3.1.1.2 ESTUDO DE CASO 2

Para facilitar o entendimento e demonstrar um segundo exemplo de aplicação do MRAR, outro *dataset* foi criado de maneira análoga ao da Figura 3.1. Porém, utilizando informações do contexto do próprio IME, com dados de seus alunos e professores. Os nós desse *dataset* foram rotulados com os nomes dos professores, alunos, instituições e localidades, já as relações foram rotuladas a fim de demonstrar uma atividade exercida ou uma associação direta entre os itens do grafo, como mostra a Figura 3.3. Nessa figura os nós sombreados de cinza claro, identificam quem são os professores, os nós que estão com o tom de cinza escuro, indicam a Instituição e localidade, já os nós em branco representam os alunos. Após a aplicação do MRAR sobre os dados desse *dataset*, encontramos um total de 297 regras diferentes, em um tempo de execução inferior a 1 segundo, utilizando as mesmas configurações aplicadas anteriormente para os valores de *MinSup*, *MinConf*, *Min-level* e *MaxLevel*. A Tabela 3.5 mostra algumas das regras encontradas após a execução do algoritmo sobre os dados do *dataset* do IME.

Tendo em vista que as informações presentes nesse *dataset* eram apenas dos professores do IME, notou-se, então, uma anomalia quanto à porcentagem de confiança nas regras, pois não atingiam o valor que seria correto de 100%, justamente para a regra que dizia que estudar no IME implica ser supervisionado por alguém que trabalha no IME.

Após investigar o motivo do erro, foi descoberto que, por uma falha nos dados, um dos professores não estava relacionado à instituição, com isso, acabamos concluindo que as regras encontradas eram válidas, uma vez que, após a correção dos dados a regra passou a mostrar o resultado correto com 100% de confiança.

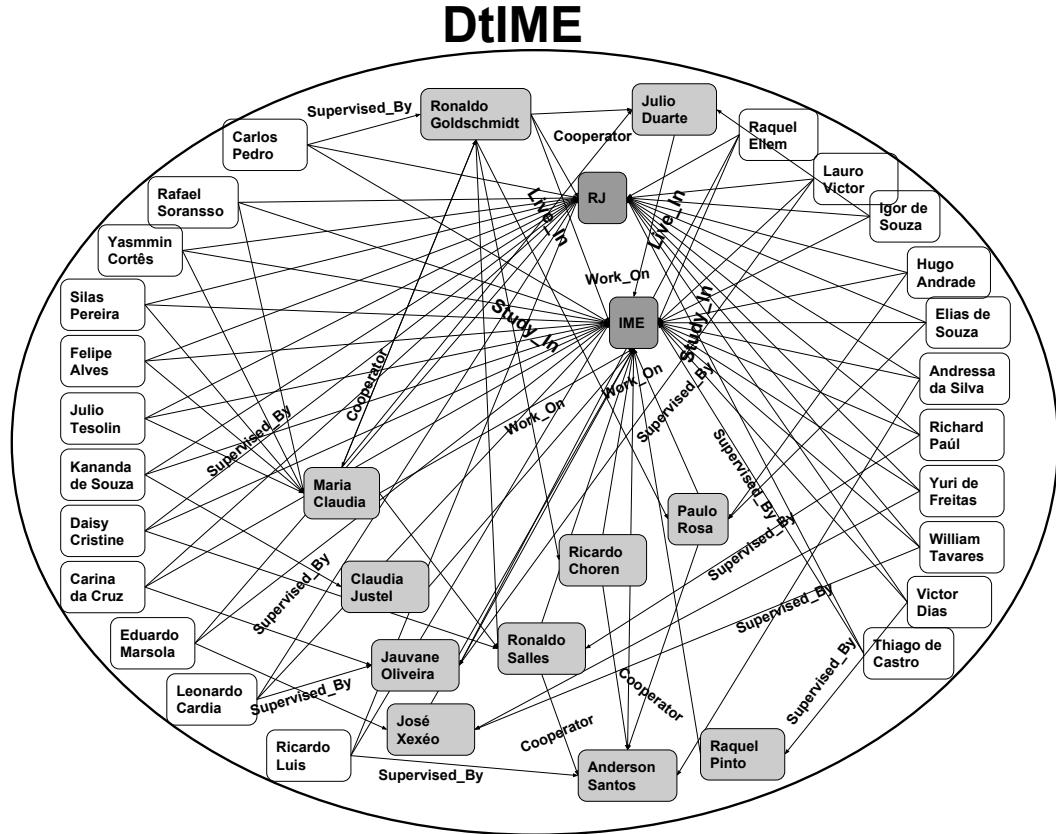


FIG. 3.3: *Dataset IME (DtIME)* com informações dos professores e alunos, juntamente com a localidade onde vivem e a instituição onde trabalham ou estudam.

TAB. 3.5: Exemplo de regras encontradas após a aplicação do MRAR com o *dataset IME*, visto na Figura 3.3.

Antecedente	Consequente	Conf	Sup
SupervisedBy(WorkOn(IME))	StudyIn(IME)	1.00	0.64
SupervisedBy(WorkOn(IME)), LiveIn(RJ)	StudyIn(IME)	1.00	0.64
SupervisedBy(Maria Claudia Cavalcanti)	SupervisedBy(Cooperator(Ronaldo Goldschmidt))	1.00	0.14
StudyIn(IME), LiveIn (RJ)	SupervisedBy(WorkOn(IME))	1.00	0.64

3.2 FORMALISMO

Para que possamos expressar a implementação realizada, algumas definições baseadas na funcionalidade do algoritmo MRAR precisaram ser formalizadas, sendo essa, também, uma contribuição importante.

Segundo a teoria dos grafos, um grafo dirigido $G = (V, A)$ é uma estrutura onde $V = \{v_1, v_2, \dots, v_n\}$ é um conjunto de vértices e $A = \{(v_i, v_j) / v_i, v_j \in V\}$ é um conjunto de arestas. Dada uma aresta e qualquer de A , representada genericamente por (v_i, v_j) , diz-se que e é parte do vértice v_i e chega ao vértice v_j .

Definição 1: Dado um grafo direcionado $G = (V, A)$, define-se $G_r = (R, P, PI)$ como um grafo em termos dos elementos do RDF, onde:

- R é o conjunto de recursos em G_r : $R = \{r / \exists v \in V \wedge r \subset v\}$.
- P é o conjunto de propriedades/predicados sobre os recursos $r_i \in R$ que podem formar arestas em G_r : $P = \{p^1, p^2, \dots, p^m\}$.
- PI é o conjunto de instâncias de propriedades/predicados que formam arestas direcionadas em G_r : $PI = \{p_y^x / \exists p^x \in P \wedge r_i, r_j \in R \wedge (r_i, r_j) \in A \wedge p_y^x = (r_i, r_j)\}$. De outra forma, pode-se dizer que $p_y^x = (r_i, r_j)$, ou ainda que $p_y^x(r_i) = r_j$.

Definição 2: Define-se $C(x, y)$ um caminho no grafo G_r , que leva um recurso x (chamado de origem do caminho) a um recurso y (chamado de destino do caminho), como um conjunto ordenado de instâncias de propriedades $p \in PI$ aplicadas sobre recursos de R , da seguinte forma: $C(x, y) = (p_1, \dots, p_k)$, onde $p_1, \dots, p_k \in PI$, e $p_1 = (x, r_j), p_2 = (r_j, r_{j+1}), \dots, p_k = (r_{j+k-1}, y)$, e $x, y, r_j, \dots, r_{j+k-1} \in R$. O caminho $C(x, y)$ também pode ser expresso como: $C(x, y) = x \xrightarrow{p_1} r_j \xrightarrow{p_2} r_{j+1} \xrightarrow{p_3} \dots r_{j+k-1} \xrightarrow{p_k} y$.

No grafo da Figura 3.1, um caminho entre os nós *Reza* (origem) e *Humid* (destino) pode ser expresso da seguinte forma: $C(Reza, Humid) = Reza \xrightarrow{\text{Live_in}} Kerman \xrightarrow{\text{Near_by}} Shiraz \xrightarrow{\text{Climate_Type}} Humid$.

Definição 3: Define-se uma cadeia $\mathcal{C}_{y,(p^1, \dots, p^k)}$ no grafo G_r como uma coleção de caminhos $C(r_j, y)$, da seguinte forma: $\mathcal{C}_{y,(p^1, \dots, p^k)} = \{C(r_j, y) / C(r_j, y) = (p_{i_1}^1, \dots, p_{i_k}^k)\}$

No exemplo da Figura 3.1, temos que os caminhos $Reza \xrightarrow{\text{Live_in}} Kerman \xrightarrow{\text{Near_by}} Shiraz \xrightarrow{\text{Climate_Type}} Humid$, e $Hasan \xrightarrow{\text{Live_in}} Yazd \xrightarrow{\text{Near_by}} Tehran \xrightarrow{\text{Climate_Type}} Humid$, são elementos da cadeia $\mathcal{C}_{Humid,(\text{Live_in}, \text{Near_by}, \text{Climate_type})}$.

Salvo casos onde haja ambiguidade, optaremos por simplificar a notação de cadeia $\mathcal{C}_{y,(p^1, \dots, p^k)}$ pela seguinte representação: $\mathcal{C}_{y,s}$, onde $s = (p^1, \dots, p^k)$.

Definição 4: Dada uma cadeia $\mathcal{C}_{y,s}$, define-se $\mathcal{I}(\mathcal{C}_{y,s})$ como a coleção formada por recursos que são origens em caminhos que pertençam a $\mathcal{C}_{y,s}$. Em termos formais:

$$\mathcal{I}(\mathcal{C}_{y,s}) = \{x / C(x, y) \in \mathcal{C}_{y,s}\}$$

No exemplo da Figura 3.1, temos que:

$$\mathcal{I}(C_{Humid, (Live_in, Near_by, Climate_type)}) = \{Reza, Hasan\}.$$

Definição 5: Sejam n cadeias $\mathcal{C}_{y_1,s_1}, \mathcal{C}_{y_2,s_2}, \dots, \mathcal{C}_{y_n,s_n}$. Diz-se que:

$$X = \{\mathcal{C}_{y_1,s_1}, \mathcal{C}_{y_2,s_2}, \dots, \mathcal{C}_{y_n,s_n}\}$$
 é um conjunto de cadeias (*ChainSet*).

No exemplo da Figura 3.1, temos que:

$$X = \{Live_In(Kerman), Near_By(Shiraz), Climate_Type(Humid)\}$$
 é um exemplo de conjunto de cadeias.

Definição 6: Seja $X = \{\mathcal{C}_{y_1,s_1}, \mathcal{C}_{y_2,s_2}, \dots, \mathcal{C}_{y_n,s_n}\}$ um conjunto de cadeias. Denomina-se $\mathcal{I}_T(X)$ ao conjunto dos recursos que são origem em algum caminho de cada cadeia que pertence a X . Formalmente: $\mathcal{I}_T(X) = \bigcap_{i=1}^n \mathcal{I}(\mathcal{C}_{y_i,s_i})$.

No exemplo da Figura 3.1, temos que:

$$X = \{Study_In(IUT), Live_In(Isfahan)\}$$
 é um conjunto de cadeias, onde $\mathcal{I}_T(X) = \{Ali, Ahmad\}$.

Definição 7: Define-se como regra de associação de multirrelação entre dois conjuntos de cadeias X e Y , a implicação $X \rightarrow Y$, onde $X \cap Y = \emptyset$ e $\mathcal{I}_T(X) \cap \mathcal{I}_T(Y) \neq \emptyset$.

Novamente considerando o exemplo da Figura 3.1, temos que:

$Study_In(IUT), Live_In(Isfahan) \rightarrow SupervisedBy(Cooperator(WorkOn(Patronage(MIT)))$) é uma regra de associação de multirrelação. Nesse exemplo os valores atribuídos seriam os seguintes: $X = \{Study_In(IUT), Live_In(Isfahan)\}$, $Y = \{SupervisedBy(Cooperator(WorkOn(Patronage(MIT))))\}$, $\mathcal{I}_T(X) = \{Ali, Ahmad\}$ e $\mathcal{I}_T(Y) = \{Ali, Ahmad\}$.

Definição 8: Seja $X \rightarrow Y$ uma regra de associação de multirrelação entre dois conjuntos de cadeias X e Y . Diz-se que $X \rightarrow Y$ é frequente (resp. válida) se, e somente se $|\mathcal{I}_T(X) \cap \mathcal{I}_T(Y)| / |R| \geq MinsuP$ (resp. $|\mathcal{I}_T(X) \cap \mathcal{I}_T(Y)| / |\mathcal{I}_T(X)| \geq MinConf$).

Considerando o exemplo da Figura 3.1, temos que:

$Health_Condition(Good) \rightarrow Live_In(Near_By(Climate_Type(Humid)))$ é uma regra de associação de multirrelação frequente, com o seguinte valor de suporte $2/19 = 0.10$.

3.3 CONSIDERAÇÕES FINAIS

Vale ressaltar que todo o formalismo foi desenvolvido especialmente para este trabalho, sendo esse uma de suas contribuições. Embora parcialmente baseado nas ideias apresentadas em um trabalho anterior (RAMEZANI et al., 2014), adaptações foram feitas respeitando os conceitos da abordagem do algoritmo Apriori. Com base no formalismo apresentado nesta seção, uma proposta de extensão do algoritmo MRAR será detalhada a seguir.

4 ABORDAGEM/PROPOSTA

Neste capítulo, será apresentada a abordagem proposta cujo objetivo é viabilizar análises sobre *datasets* na Web de Dados. Ficou organizado da seguinte forma: primeiro a visão geral será apresentada, em seguida, o novo algoritmo *MRAR+* será descrito e, por fim, alguns detalhes sobre a implementação serão expostos.

4.1 VISÃO GERAL

Como foi descrito anteriormente, o problema que se quer tratar neste trabalho caracteriza-se pela dificuldade de realizar análises em grandes volumes dados, organizados na forma de grafo, pois há vasta diversidade de tipos de nós e de relações para que se possa extrair informação útil. Esses dados estão dispostos em *datasets* interligados na Web de Dados, em formato RDF. Atualmente, sabe-se que é inviável tratar o conjunto completo destes *datasets*, que segundo as últimas estatísticas da LOD¹⁵, alcançou a marca de 10 mil *datasets* em 2017, contendo quase 150 bilhões de triplas.

Assim, o objetivo é apresentar uma possível abordagem para esse problema, através do enriquecimento de um determinado *dataset* com informações complementares, de maneira controlada, a fim de que, ao realizar uma análise, por meio da mineração de dados, novas regras de associação de multirrelação sejam encontradas, ampliando o conhecimento sobre os dados analisados. Tendo este *dataset* como alvo, a hipótese é que, ao realizar a mineração de regras de associação de multirrelação, entre os recursos que dão suporte às regras encontradas, seja possível identificar recursos externos, que apontam para outros *datasets* com área de domínio de conhecimento diferentes (complementares). Os *datasets* encontrados (chamados *datasets* externos) são uma seleção a partir dos quais, será possível enriquecer o *dataset* alvo com informações úteis (novas triplas), ampliando assim seu conhecimento. Dessa maneira, possibilita-se encontrar novas regras e evita-se ter que minerar todos os *datasets* externos por completo.

Para facilitar o entendimento e demonstrar que o processo de extensão do *dataset* alvo ocorre de maneira controlada, a fim de evitar a ampliação total, com a soma de todos os *datasets* envolvidos, daremos um exemplo hipotético. Supondo que o *dataset* alvo (*Dt_X*) seja um grafo com 100 nós, e que ele possua recursos externos presentes também

¹⁵<http://stats.lod2.eu>

no *dataset* externo (Dt_Y)¹⁶. Ao analisar o Dt_X , o algoritmo identifica 70 regras de associação de multirrelação. A partir dessas regras, são selecionados apenas os recursos externos que satisfazem algumas delas. No nosso exemplo, digamos que chegou-se ao total de 30 recursos externos. Para ampliar Dt_X , supondo que o Dt_Y possua um total de 500 nós, serão selecionados apenas parte desses nós. Explicando melhor, serão selecionados os equivalentes aos recursos identificados anteriormente, e além desses, para cada um deles, os recursos vizinhos (a uma dada distância pré-definida). Supondo que cada um desses nós, no grafo externo (Dt_Y), esteja ligado a apenas 2 nós e que eles sejam diferentes entre si, nesse caso, o algoritmo estaria selecionando apenas 60 nós juntamente com suas respectivas arestas, para serem acrescentados ao *dataset* alvo (Dt_X). Portanto, ao final desse processo, o *dataset* Dt_X ficará com o total de 160 ($100 + 60$) nós. Dessa forma, no lugar de somar os 2 *datasets* e obter um volume de dados muito grande (600 nós), foi realizada uma ampliação controlada que possibilitou identificar somente os recursos com potencial de gerar novas regras.

A Figura 4.1 mostra todos os passos da abordagem proposta aplicados para o processo de ampliação de conhecimento de um *dataset* hipotético chamado DtA . Inicia-se com a mineração do *dataset* alvo, DtA , (passo 1), passando pela seleção de recursos externos (passo 2). Uma vez identificados os recursos externos, passa-se, então, para o processo de ampliação do conhecimento existente do DtA (passo 3), com as informações encontradas no *dataset* dos recursos externos (DtB), gerando o *dataset* $DtA+$. Em seguida (passo 4), minera-se esse novo conjunto de dados ($DtA+$) para encontrar novas regras. No passo 5, compara-se as regras geradas sobre o *dataset* DtA com as geradas pelo conjunto $DtA+$. Por fim, no passo 6, o usuário tem a opção de reajustar o valor de *MinSup* para permitir que o valor utilizado seja suficiente tanto a geração das regras novas quanto das antigas, encontradas na primeira execução do MRAR.

O algoritmo MRAR foi utilizado nos passos 1 e 4. Para contemplar os passos 2 e 3, o algoritmo *MRAR+* foi desenvolvido. É importante destacar que o *MRAR+* também é responsável por dar início a execução do passo 4, passando os novos parâmetros para o MRAR. O passo 5 apenas compara e apresenta os resultados obtidos nos passos anteriores para interação com o usuário. Já no passo 6, é possível realizar uma verificação opcional, a fim de descobrir se as regras do primeiro conjunto fazem parte também do segundo e, caso o usuário ache interessante que as mesmas regras sejam geradas, ele pode fazer um ajuste no valor de *MinSup* e executar os passos 4 e 5 novamente. A seguir serão

¹⁶Neste exemplo, por simplificação, foi utilizado apenas um *dataset* externo, entretanto outros *datasets* externos poderiam ser utilizados.

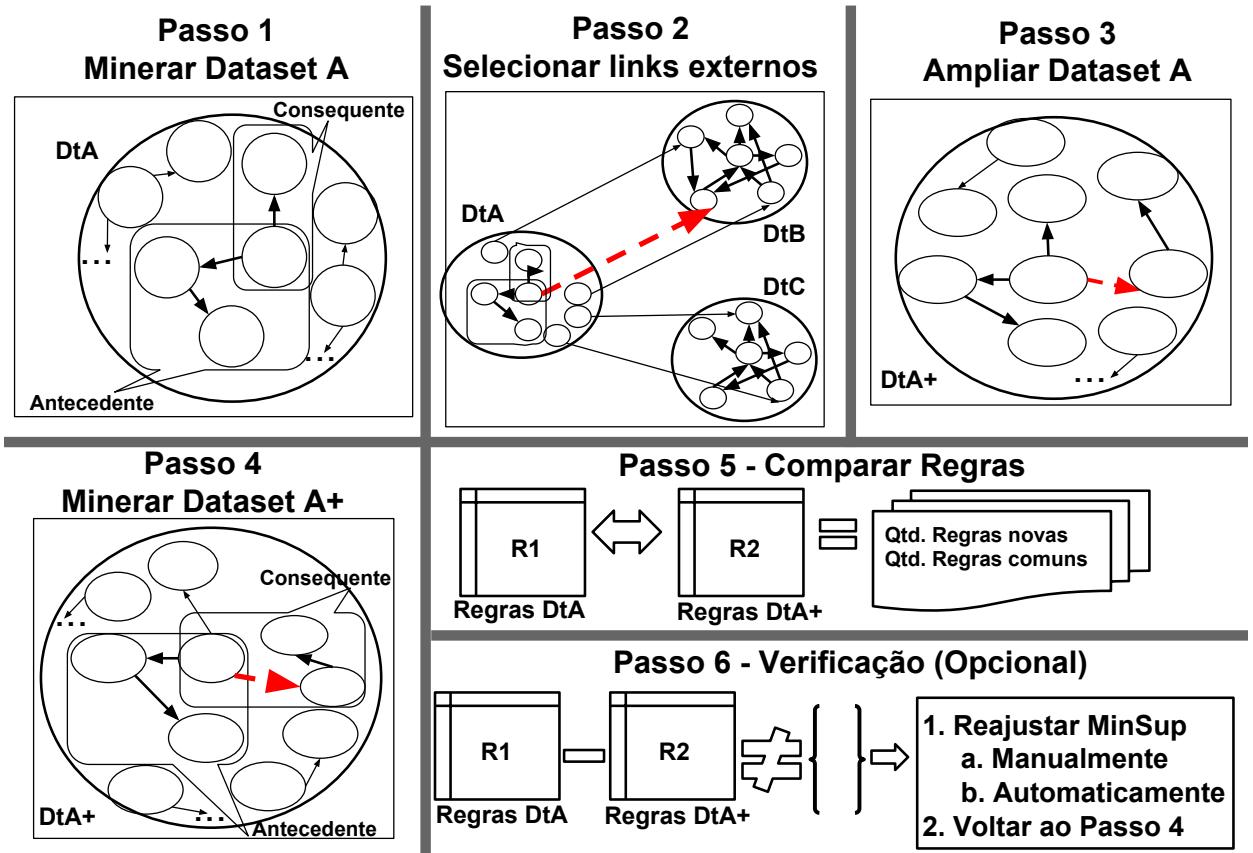


FIG. 4.1: Visão geral de todos os passos da proposta.

apresentados alguns esquemas meramente ilustrativos para detalhar todos os passos da abordagem proposta.

No passo 1, ao aplicar a mineração de dados com o algoritmo MRAR, é possível encontrar regras da forma $X \rightarrow Y$, onde X e Y contém multirrelações (conforme definição 7 da seção 3.2). As regras cujo valor de suporte não atenderem ao valor de suporte mínimo, serão eliminadas nesse passo. Como ilustrado na Figura 4.2, apenas uma regra foi encontrada: $L_2(L_3(DtA:a_4)) \rightarrow L_1(DtA:a_1)$. Em cada regra encontrada, um conjunto de recursos é selecionado para servir de entrada para o passo seguinte. Conforme as definições 6 e 8 da seção 3.2, analisa-se o conjunto $\mathcal{I}_T(X) \cap \mathcal{I}_T(Y)$, que representa os recursos que dão suporte à regra $X \rightarrow Y$.

Nesse exemplo, o nó que satisfaz a essa regra é o $DtB:b_4$ ¹⁷.

No passo 2 (Figura 4.3), é visto que o algoritmo *MRAR+* analisa o conjunto de recursos encontrado no passo anterior, selecionando somente os que são externos. Mesmo que o *dataset* alvo possua outros recursos externos, a ideia aqui é evitar trabalhar com

¹⁷Esta é uma simplificação, pois em situações reais, diversos nós deverão satisfazer cada regra identificada para atender o critério de suporte mínimo.

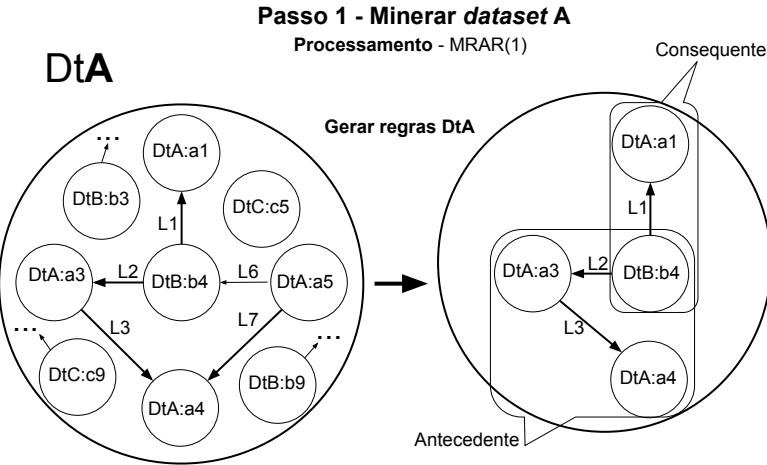


FIG. 4.2: Visão geral passo 1. Exemplo de aplicação do algoritmo MRAR identificando uma regra de associação de multirrelação.

todos eles, selecionando apenas os recursos que têm potencial para a descoberta de novas regras. Já que tais recursos foram suporte para encontrar regras no *dataset* alvo (sem expansão), estes têm maior chance de se tornar recursos que possam dar suporte para novas regras no *dataset* alvo estendido.

Nesse exemplo, apenas a aresta com o nome “link b4” é selecionada, tendo em vista que o recurso *DtB:b4* do *dataset* *DtA* é o único nó comum para o antecedente e consequente da regra selecionada. Também é possível observar que o *DtB:b4* do *dataset* *DtA* aponta para o recurso externo *DtB:b4* do *dataset* *DtB*, que por sua vez também está ligado a outros recursos no *dataset* *DtB*. No exemplo citado, apenas o *DtB:b4* do *dataset* *DtA* foi considerado como recurso potencial para extensão do *dataset* alvo e será utilizado no passo seguinte.

No passo 3, o *MRAR+* busca, nos *datasets* apontados pelos recursos externos selecionados no passo anterior, as novas triplas envolvendo esses recursos e seus vizinhos. Neste exemplo, limitou-se à vizinhança apenas de caminhos de comprimento 1 (apenas 1 aresta). Elas são adicionadas ao *dataset* alvo (*DtA*), ampliando dessa forma suas informações. No exemplo da Figura 4.4, nota-se que a tripla (*DtB:b4* L4 *DtB:b2*) foi adicionada. A junção entre os dados do *dataset* *DtA* e de um subconjunto dos dados extraídos do *dataset* *DtB* possibilita a criação de um novo conjunto de dados, que foi chamado de *dataset* *DtA+*.

No passo 4, o algoritmo MRAR é executado novamente, chamado a partir do algoritmo *MRAR+*, agora selecionando o novo *dataset* criado (*DtA+*) como entrada. Na Figura 4.5, nota-se que uma nova regra foi encontrada, essa fazendo uso dos nós e arestas trazidos do *dataset* externo (*DtB*). Nesse exemplo, a regra $L2(L3(DtA:a4)) \rightarrow \text{Link b4}(L4(DtB:b2))$

Passo 2 - Selecionar links externos

Processamento - MRAR+(1)

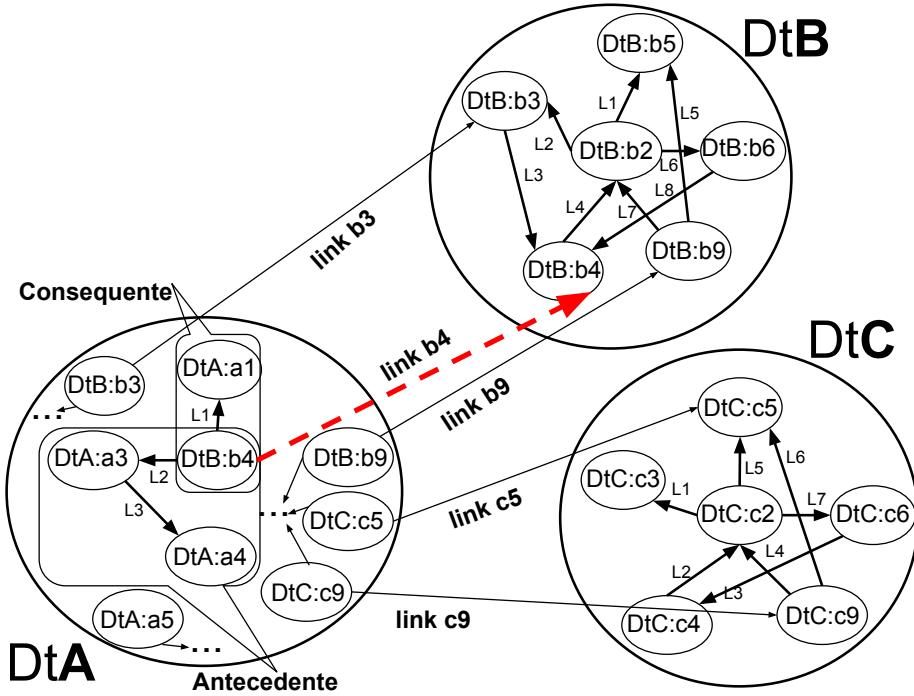


FIG. 4.3: Visão geral passo 2, seleção dos recursos externos.

é uma amostra que não poderia ser encontrada se apenas os dados do *DtA* fosse utilizado.

No passo 5, Figura 4.6, as regras obtidas com a mineração sobre os dados originais (R1 do *DtA*) são comparadas com as regras encontradas após o processo de extensão (R2 do *DtA+*). Essa comparação gera algumas estatísticas, demonstrando o total de regras que são novas e o total de regras comuns às duas execuções.

Por fim, no passo 6, Figura 4.7, o usuário tem a opção de verificar se todas as regras (R1) encontradas durante a primeira execução do algoritmo, sobre os dados do *DtA*, fazem parte do conjunto de regras (R2), obtidos após a análise do *dataset* estendido (*DtA+*). Funciona da seguinte maneira, se $R1 - R2 \neq \{\}$ então o usuário pode reajustar o *MinSup* manualmente ou selecionar a opção para o próprio *MRAR+* calcular o valor de suporte mínimo, tal que $R1$ esteja contido em $R2$. Com esse ajuste, os passos 4 e 5 são executados novamente.

Essa mineração de regras de associação de multirrelação nos dados estendidos possibilita ao algoritmo MRAR encontrar novas regras de multirrelação sem ter que analisar todos os *datasets* por completo.

Na próxima seção, o pseudo código do algoritmo *MRAR+* será apresentado e descrito para formalizar e detalhar ainda mais seu funcionamento.

Passo 3 - Ampliar dataset A

Processamento - MRAR+(2)

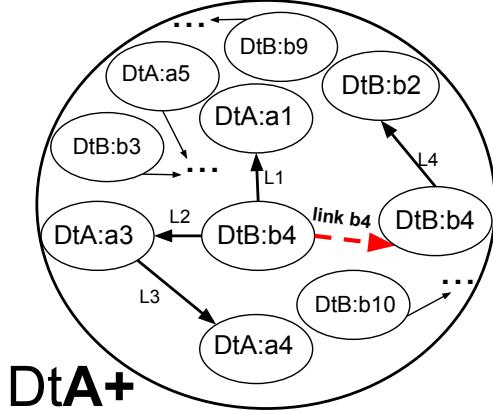


FIG. 4.4: Visão geral passo 3, ampliando informações do *dataset*.

Passo 4 - Minerar dataset A+

Processamento - MRAR(2)

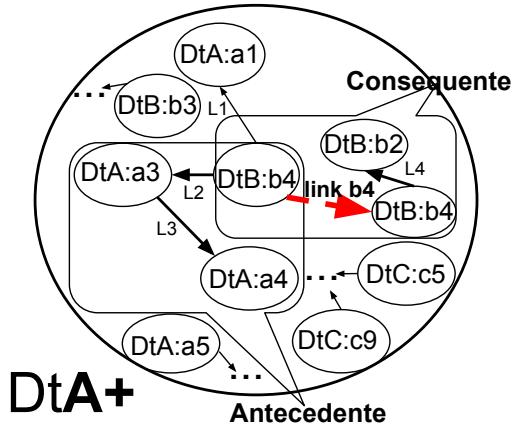


FIG. 4.5: Visão geral passo 4, encontrando novas regras.

4.2 ALGORITMO *MRAR+*

O algoritmo *MRAR+*, visto no pseudo código 1, é uma evolução do algoritmo MRAR. Como já foi mencionado, ele foi criado com o propósito de fazer uso dos recursos externos, normalmente presentes como nós de um grafo em RDF. Através destes recursos, é possível acessar outros *datasets* na Web de Dados e ampliar o *dataset* original de maneira controlada, possibilitando gerar novas regras de associação de multirrelação. Podemos afirmar, então, que o algoritmo *MRAR+* é uma extensão do MRAR com potencial de encontrar mais regras de associação de multirrelação, atravessando as fronteiras do *dataset* em análise.

A execução do *MRAR+* se inicia no passo 2 da abordagem. Ele recebe como entrada

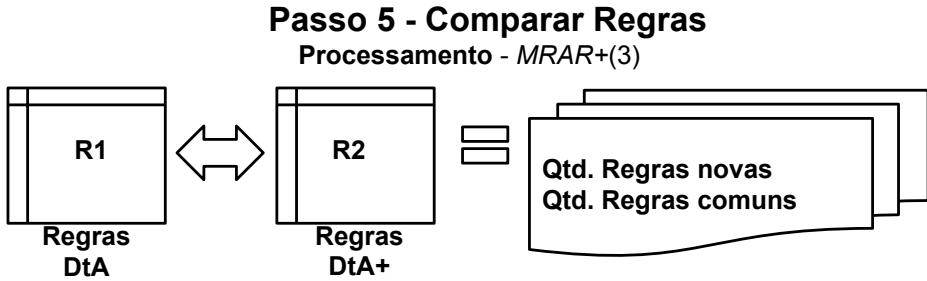


FIG. 4.6: Visão geral passo 5, comparando as regras geradas ao aplicar o MRAR sobre os dados do DtA(R1) e DtA+(R2).

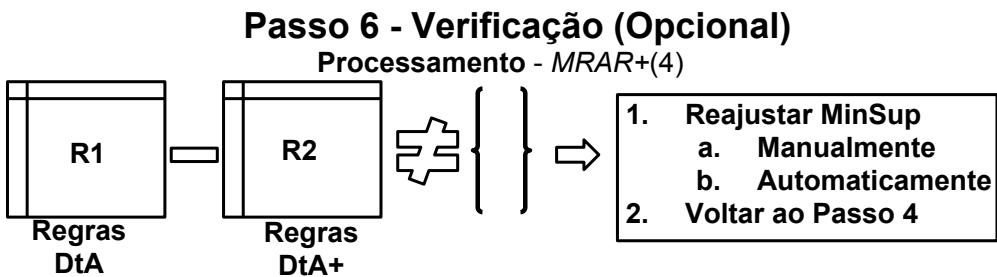


FIG. 4.7: Visão geral passo 6, verificação opcional para descobrir se as regras (R1) geradas ao aplicar o MRAR sobre os dados do DtA estão contidas no conjunto de regras (R2), gerados após a análise feita sobre os dados do DtA+.

o conjunto de regras de associação de multirrelação (Ψ), que resulta da execução do passo 1, isto é, da aplicação do algoritmo MRAR sobre um grafo direcionado $G_r(R, P, PI)$, que chamamos de *dataset* alvo (DS). Além disso, recebe o conjunto de recursos (Γ) que dão suporte a essas regras, onde $\Gamma \subset R$. O *MRAR+* recebe ainda o próprio *dataset* alvo (DS) como entrada, a partir do qual gera o novo *dataset* enriquecido (*NewDS*). A partir dos recursos externos presentes em Γ , o *MRAR+* trará mais recursos externos para enriquecer DS . Assim, o usuário deve definir como entrada do *MRAR+* o valor de comprimento máximo de caminho a ser percorrido no *dataset* externo (*MaxLength*). Outros parâmetros de entrada são usados para configuração da execução do *MRAR+*, e devem ser ajustados pelo usuário, tais como: o valor de suporte mínimo (*MinSup*), o valor de confiança mínima (*MinConf*), o valor mínimo e máximo de relações permitido para cada cadeia (*MinLevel* e *MaxLevel*). Estes parâmetros, definidos na seção 3, são usados para a seleção das regras de associação multirrelação geradas pelo MRAR sobre o novo *dataset* (*NewDataset*).

As variáveis de entrada do algoritmo são detalhadas entre as linhas 2 e 12, já entre as linhas 13 e 18 são as de saída do algoritmo. No primeiro passo do algoritmo (linha 19), são atribuídos os nós e arestas do grafo (DS) original à variável que receberá o novo grafo

NewDS. Na linha 20, o algoritmo inicia um loop selecionando cada recurso (γ) presente no conjunto de recursos (Γ). Logo após, na linha 21, ele verifica se o recurso selecionado corresponde a um recurso externo, através da função *ExtLink*. Na linha 22, um novo loop é iniciado para caminhar no grafo externo do ponto inicial (i) até o comprimento máximo (*MaxLength*) definido pelo usuário. Já na linha 23, o *MRAR+* adiciona ao grafo *NewDS* os novos nós e arestas selecionados do grafo externo na posição i , através da função *getData*. Após recuperar todos os recursos e arestas do grafo externo permitidos pelo (*MaxLength*), o algoritmo MRAR é executado novamente, agora sobre o novo grafo (*NewDS*). Isso permite encontrar novas regras (*NewRules*) que, antes, devido à falta de informação, não poderiam ser encontradas. Antes de finalizar, uma verificação é feita, na linha 28, para confirmar se a variável *NewRules* é diferente de NULL, garantindo que o novo conjunto de regras não está vazio. Caso contrário, nas linhas 29 e 30, novas variáveis são criadas, uma contendo apenas as regras que são novas (*AdditionalRules*), e outra contendo as regras que foram descartadas (*DiscardedRules*), isto é, são as regras que foram obtidas antes do enriquecimento do *DS* e que não fazem parte do novo resultado. Por fim, na linha 32, o algoritmo retorna, além do novo *dataset* *NewDS*, três conjuntos de regras: *NewRules*, *AdditionalRules* e *DiscardedRules*.

Vale notar que, embora o algoritmo especificado seja genérico e permita que quaisquer *datasets* externos sejam analisados, somente os recursos externos que dão suporte a alguma regra inicialmente encontrada sobre o *dataset* local, serão investigados, reduzindo, dessa forma, o número de *datasets* a investigar.

4.3 PROTÓTIPO DO *MRAR+*

Com o objetivo de realizar análises sobre alguns *datasets* de áreas de domínio de conhecimento diferentes, o protótipo do algoritmo *MRAR+* foi implementado. A seção seguinte descreve as linguagens de programação e as bibliotecas que foram utilizadas.

4.3.1 IMPLEMENTAÇÃO

O algoritmo *MRAR+* foi implementado em PHP, JavaScript, JQuery e Bootstrap. Essas linguagens de programação foram escolhidas por se tratar de linguagens do tipo *software* livre, de conhecimento amplo e fácil manipulação. Com o objetivo de acessar os *datasets* externos, na Web de Dados, utilizou-se a biblioteca EasyRdf¹⁸, também desenvolvida em PHP. Essa biblioteca tornou fácil a tarefa de consumir e produzir dados em RDF.

¹⁸<http://www.easyrdf.org/>

Algorithm 1 MRAR+ (Pseudo Código)

```
1: function MRAR+(\( \Gamma \),  $MaxLength$ ,  $DS$ ,  $MinSup$ ,  $MinConf$ ,  $MinLevel$ ,  
       $MaxLevel$ ,  $\Psi$  )  
2:   Inputs  
3:      $X_\psi$                                  $\triangleright$  Antecedentes das regras de  $\Psi$   
4:      $Y_\psi$                                  $\triangleright$  Consequentes das regras de  $\Psi$   
5:      $\Gamma = \{I_T(X_\psi) \cap I_T(Y_\psi) / \psi \in \Psi\}$      $\triangleright$  Conjunto de recursos que dão suporte às  
       regras frequentes e válidas, identificados pelo algoritmo MRAR  
6:      $MaxLength$      $\triangleright$  Tamanho máximo de caminho a percorrer no grafo externo  
7:      $DS$            $\triangleright$  Grafo direcionado com rótulos nas arestas  
8:      $MinSup$        $\triangleright$  Suporte Mínimo  
9:      $MinConf$      $\triangleright$  Confiança Mínima  
10:     $MinLevel, MaxLevel$      $\triangleright$  valor mínimo e máximo de relações em cada cadeia  
11:     $\Psi$            $\triangleright$  Conjunto de regras de associação de multirrelação  
12:  EndInpts  
13:  Output  
14:     $NewDS$          $\triangleright$  Novo dataset enriquecido com as triplas externas.  
15:     $NewRules$        $\triangleright$  Novo conjunto de regras de associação de multirrelação.  
16:     $AdditionalRules$      $\triangleright$  Conjuntos com apenas as regras que são novas.  
17:     $DiscardedRules$    $\triangleright$  Conjunto com apenas as regras que deixaram de ser  
       geradas.  
18:  EndOutput  
19:   $NewDS = DS$   
20:  while ( $\gamma$  in  $\Gamma$ ) do  
21:    if  $\exists \gamma \in \Gamma / ExtLink(\gamma)$  then  
22:      for  $i \leftarrow 1$  até  $MaxLength$  do  
23:         $NewDS.Add(\text{new getData}(\gamma, i))$   
24:      end for  
25:    end if  
26:  end while  
27:   $NewRules = MRAR(NewDS, MinSup, MinConf, MinLevel, MaxLevel)$   
28:  if  $NewRules \neq \text{NULL}$  then  
29:     $AdditionalRules = NewRules - \Psi$   
30:     $DiscardedRules = \Psi - NewRules$   
31:  end if  
32:  return  $NewDS, NewRules, AdditionalRules, DiscardedRules$   
33: end function
```

Durante o desenvolvimento do algoritmo, foram utilizados recursos nativos das linguagens para proporcionar economia de código e melhor desempenho na execução dos scripts. Ademais, o sistema foi desenvolvido em módulos, o que permitiu o crescimento das funcionalidades e facilitou sua evolução. Dessa maneira, a execução de um módulo gerou a entrada de outro de modo sequencial ou recursivo até que os recursos com potencial de extensão fossem identificados, estendidos e minerados. Esse processo possibilitou que novas regras de associação de multirrelação fossem geradas e apresentadas ao usuário.

Para facilitar os experimentos, a implementação tomou como base o formato JSON, ao invés de RDF, para servir como entrada no algoritmo. O JSON, acrônimo de “JavaScript Object Notation”, trata-se de um formato de padrão aberto para utilização de textos legíveis a humanos.

Os *datasets* locais foram criados utilizando arquivos no formato JSON. Os dados dos *datasets* foram estruturados em triplas. Por exemplo, uma tripla que afirma que alguém estuda no IME ficará estruturada da seguinte maneira em JSON: [{'Subject': 'Felipe', 'Predicado': 'Estuda', 'Object': 'IME'}]. Ao aplicar esse formato a todos os dados, será possível garantir que a estrutura do grafo estará sendo mantida. Logo abaixo, serão apresentadas as triplas em formato JSON, para o mesmo *dataset* visto na Figura 3.1.

```
{'dados': [
  {'Subject': 'Tehran', 'Predicado': 'Climate_Type', 'Object': 'Humid'},
  {'Subject': 'Shiraz', 'Predicado': 'Climate_Type', 'Object': 'Humid'},
  {'Subject': 'Yazd', 'Predicado': 'Near_By', 'Object': 'Tehran'},
  {'Subject': 'Kerman', 'Predicado': 'Near_By', 'Object': 'Shiraz'},
  {'Subject': 'Reza', 'Predicado': 'Live_In', 'Object': 'Kerman'},
  {'Subject': 'Hasan', 'Predicado': 'Live_In', 'Object': 'Yazd'},
  {'Subject': 'Hasan', 'Predicado': 'Health_Condition', 'Object': 'Good'},
  {'Subject': 'Reza', 'Predicado': 'Health_Condition', 'Object': 'Good'},
  {'Subject': 'Reza', 'Predicado': 'Study_In', 'Object': 'IUT'},
  {'Subject': 'Reza', 'Predicado': 'Study_In', 'Object': 'Saraee'},
  {'Subject': 'Reza', 'Predicado': 'Knows', 'Object': 'Hasan'},
  {'Subject': 'Ali', 'Predicado': 'Study_In', 'Object': 'IUT'},
  {'Subject': 'Ali', 'Predicado': 'Live_In', 'Object': 'Isfahan'},
  {'Subject': 'Ali', 'Predicado': 'Study_In', 'Object': 'Saraee'},
  {'Subject': 'Ali', 'Predicado': 'Knows', 'Object': 'Ahmad'},
  {'Subject': 'Ahmad', 'Predicado': 'Study_In', 'Object': 'IUT'},
  {'Subject': 'Ahmad', 'Predicado': 'Study_In', 'Object': 'Nematbakhsh'},
  {'Subject': 'Ahmad', 'Predicado': 'Live_In', 'Object': 'Isfahan'},
  {'Subject': 'Saraee', 'Predicado': 'Cooperator', 'Object': 'Mr_A'},
  {'Subject': 'Saraee', 'Predicado': 'Live_In', 'Object': 'Kerman'}
]}
```

```

{'Subject': 'Saraee', 'Predicado': 'Knows', 'Object': 'Nematbakhsh'},
{'Subject': 'Nematbakhsh', 'Predicado': 'Cooperator', 'Object': 'Mr_B'},
{'Subject': 'Nematbakhsh', 'Predicado': 'Knows', 'Object': 'Mr_A'},
{'Subject': 'Nematbakhsh', 'Predicado': 'Live_In', 'Object': 'Isfahan'},
{'Subject': 'Mr_A', 'Predicado': 'Work_On', 'Object': 'Project_A'},
{'Subject': 'Mr_A', 'Predicado': 'Live_In', 'Object': 'Shiraz'},
{'Subject': 'Mr_B', 'Predicado': 'Work_On', 'Object': 'Project_B'},
{'Subject': 'Mr_B', 'Predicado': 'Knows', 'Object': 'Nematbakhsh'},
{'Subject': 'Project_A', 'Predicado': 'Patronage', 'Object': 'Isfahan'},
{'Subject': 'Project_B', 'Predicado': 'Patronage', 'Object': 'Isfahan'},
{'Subject': 'Project_B', 'Predicado': 'Patronage', 'Object': 'IUT'}
]}

```

O código da implementação do algoritmo *MRAR+* está disponível no GitHub e pode ser acessado através da seguinte URL (https://github.com/feliperj629/MRAR_plus).

4.3.1.1 INTERFACES

Para iniciar a execução do algoritmo, o usuário precisa informar as configurações iniciais como: *MinLevel*, *MaxLevel*, *MinSup*, *MinConf*, selecionar o *dataset* que será analisado (*DS*) e informar o *endpoint* que será acessado para consumir as informações dos recursos externos. A Figura 4.8 mostra a tela que permite a inclusão dessas informações iniciais, que são essenciais para execução completa do *MRAR+*.

A tela do protótipo foi desenvolvida para permitir que o usuário possa ter a escolha de fazer a mineração de dados apenas com o algoritmo MRAR ou de fazer a mineração estendida com o algoritmo *MRAR+*. Funciona da seguinte maneira: para executar somente o MRAR, primeiro é necessário preencher as informações iniciais básicas, como foi detalhado anteriormente. Após o preenchimento, basta clicar no botão MRAR, como é visto na Figura 4.8. Entretanto, para executar o algoritmo *MRAR+*, além das informações básicas, é preciso informar no campo “Ext. Endpoint” o endereço do local que será acessado para buscar as informações externas. Após isso, basta clicar no botão *MRAR+*, visto na Figura 4.8.

Ao final da execução do *MRAR+*, um novo *dataset* é criado, em formato *Json*, contendo o conjunto de dados do *dataset* originais somados aos recursos e arestas que foram capturados do *dataset* externo, além de um cabeçalho contendo as informações básicas usadas para sua criação. Esse novo *dataset* será exibido junto à lista de seleção dos *datasets*, com o nome do *dataset* original mais a data de sua criação. Após a criação do novo conjunto de dados, basta preencher as informações iniciais, selecionar o novo *dataset* na

MRAR+ Dashboard

MinSup <input type="text" value="0.1"/>	MinConf <input type="text" value="0.6"/>
MinLevel <input type="text" value="1"/>	MaxLevel <input type="text" value="4"/>
Select dataset: <div style="border: 1px solid #ccc; padding: 5px; width: 100%;">Dt_Neymar</div>	
Ext. Endpoint <input type="text" value="http://dbpedia.org/sparql"/>	
Configuration	
<input style="background-color: #007bff; color: white; border: none; padding: 5px 10px; border-radius: 5px; font-weight: bold; width: fit-content; margin: auto;" type="button" value="MRAR"/>	<input style="background-color: #ffc107; color: black; border: none; padding: 5px 10px; border-radius: 5px; font-weight: bold; width: fit-content; margin: auto;" type="button" value="MRAR+"/>
<input style="background-color: #28a745; color: white; border: none; padding: 5px 10px; border-radius: 5px; font-weight: bold; width: fit-content; margin: auto;" type="button" value="Save Rules"/>	

FIG. 4.8: Tela de configuração para as variáveis de entrada do algoritmo *MRAR+*.

lista e clicar no botão MRAR, para realizar a mineração sobre o *dataset* enriquecido.

Sabendo que, acrescentar novos itens ao *dataset* original implicará alteração do suporte, com o qual as primeiras regras foram geradas. Achou-se necessário disponibilizar uma opção para auxiliar o usuário na hora de definir o valor de suporte mínimo. Visando garantir que as regras geradas, após a análise do primeiro *dataset*, pertençam ao segundo conjunto de regras, gerado ao analisar o *dataset* estendido, assim como foi demonstrado no passo 6 da visão geral, visto na seção 4.1. Para fazer esse uso, basta clicar no link do *dropdown* “*Configuration*”, visto na Figura 4.8, e marcar a opção “*Apply the best support*”. Ao iniciar a execução do algoritmo, um novo valor de suporte mínimo será calculado, visando utilizar um que além de permitir a geração das novas regras, possibilite também identificar as regras geradas com *dataset* original, antes do seu enriquecimento. A fórmula utilizada será detalhada adiante.

Objetivando garantir que o suporte escolhido seja adequado, a fim de permitir que as mesmas regras sejam geradas ao analisar os dois *datasets* (original e estendido), a fórmula $X = S * D / D'$ foi aplicada. Onde X é o novo valor de suporte que se deseja encontrar, D' é a quantidade de nós existentes no novo *dataset*, S é o suporte que foi utilizado

originalmente para analisar o *dataset* alvo e D é a quantidade de nós do grafo original. Por exemplo, em uma execução do algoritmo para um grafo com 19 nós, utilizando o suporte mínimo igual a 0.1, se após a extensão, o novo *dataset* criado passa a ter 25 nós, o cálculo é feito da seguinte forma: $X = 0.1 * 19/25$, $X = 1.9/25$, logo, $X = 0.076$, ou seja, o valor de suporte mínimo que melhor se aplica para possibilitar tanto a geração das regras originais quanto as novas será 0.076.

O sistema também mostra o tempo gasto na execução, o quanto de memória foi usada, além de abrir duas formas de visualização das regras que foram geradas. A primeira forma de visualização de regras, vista na Figura 4.9, mostra os antecedentes e consequentes em formato numérico, fazendo referências a cada cadeia de relações que foram geradas, além dos respectivos valores de suporte, confiança, lift e convicção de cada regra gerada. Já a segunda, vista na Figura 4.10, formata os valores dos antecedentes e consequentes das regras em formato de texto, proporcionando, assim, melhor entendimento e visualização das regras.

Rules						
Row	Ant.	Cons.	Sup.	Conf.	Lift	Conv.
1	4	3	0.11	1.00	0.63	0.944
2	3	4	0.11	0.69	0.63	0.000
3	7	3	0.11	1.00	0.63	0.944
4	3	7	0.11	0.69	0.63	0.000
5	9	3	0.11	1.00	0.63	0.944
6	3	9	0.11	0.69	0.63	0.000
7	7	4	0.11	1.00	0.91	1.000
8	4	7	0.11	1.00	0.91	1.000
9	6	5	0.11	1.00	0.91	1.000

FIG. 4.9: Tela de visualização das regras geradas, em formato numérico.

Após a execução do algoritmo, é possível salvar as regras que foram geradas. Na tela, Figura 4.8, é exibido o botão “*Save Rules*”. Ao selecioná-lo, o sistema captura as regras que estão sendo exibidas na tela, converte para o formato JSON e salva-as em um diretório.

Formatted Rules						
Row	Antecedent	Consequent	Sup.	Conf	Lift	Conv.
1	Live_In (Near_By (Shiraz)) →	Live_In (Near_By (Climate_Type (Humid)))	0.11	1.00	0.63	0.944
2	Live_In (Near_By (Climate_Type (Humid))) →	Live_In (Near_By (Shiraz))	0.11	0.69	0.63	0.000
3	Live_In (Kerman) →	Live_In (Near_By (Climate_Type (Humid)))	0.11	1.00	0.63	0.944
4	Live_In (Near_By (Climate_Type (Humid))) →	Live_In (Kerman)	0.11	0.69	0.63	0.000
5	Health_Condition (Good) →	Live_In (Near_By (Climate_Type (Humid)))	0.11	1.00	0.63	0.944
6	Live_In (Near_By (Climate_Type (Humid))) →	Health_Condition (Good)	0.11	0.69	0.63	0.000

FIG. 4.10: Tela de visualização das regras geradas, em formato de texto.

Com as regras salvas, o usuário tem a opção de compará-las. Para tanto, basta acessar o menu lateral, “*Dashboard*”, e ir na opção “*Compare Rules*”. Uma nova janela será carregada, trazendo duas opções de seleção. Na primeira, é preciso escolher um dos conjuntos de regras salvas, geradas com base nos *datasets* originais e, na segunda, é necessário selecionar o conjunto de regras geradas com o *dataset* estendido. Ao acionar o botão “*Compare Rules*”, a página será atualizada com as informações referentes à comparação feita, mostrando o número de regras que foram geradas com os dados do *dataset* original e do *dataset* estendido. As seguintes estatísticas: regras novas, regras comuns e regras que não foram geradas são exibidas após a comparação, como é possível ver na Figura 4.11. Além disso, uma tabela contendo somente as regras que são novas é exibida, no final da página, para a comparação feita entre os dados selecionados, como mostra a Figura 4.12.

Note que entre as novas regras, vistas na Figura 4.12, temos a regra “*Supervised_By (Cooperator (Claudia Justel)), Plays (Futebol) → Supervised_By(Maria Cláudia Cavalcanti)*”, que traduzindo para o português diz que: Quem é supervisionado por alguém que coopera com a professora Cláudia Justel e Joga Futebol é Supervisionado pela professora Maria Cláudia Cavalcanti, com o suporte de 10% e confiança de 100%. Sendo assim, é importante destacar que, apenas com os dados do *dataset* original, não é possível inferir esse tipo de regra. Esses resultados parciais foram apresentados apenas para demonstrar a execução do algoritmo. Novas análises sobre esse exemplo serão apresentadas no capítulo

seguinte.

Rules MRAR	Rules MRAR+	New Rules	Common Rules	Discarded Rules
297	400 (135%)	112 (38%)	288 (97%)	9 (3%)

FIG. 4.11: Tela de comparação das regras geradas pelo algoritmo MRAR e *MRAR+*.

New rules table						
Row	Antecedent	Consequent	Sup.	Conf	Lift	Conv.
23	Supervised_By (Cooperator (Claudia Justel)), Plays (Futebol) →	Supervised_By (Maria Claudia Cavalcanti)	0.10	1.00	0.77	1.000
24	Supervised_By (Cooperator (Claudia Justel)), Supervised_By (Maria Claudia Cavalcanti) →	Plays (Futebol)	0.10	0.77	0.33	0.003
25	Supervised_By (Cooperator (Cooperator (Anderson Santos))), Plays (Futebol) →	Supervised_By (Cooperator (Claudia Justel))	0.10	1.00	0.77	1.000
26	Supervised_By (Cooperator (Claudia Justel)), Plays (Futebol) →	Supervised_By (Cooperator (Cooperator (Anderson Santos)))	0.10	1.00	0.67	0.977
27	Supervised_By (Cooperator (Claudia Justel)), Supervised_By (Cooperator (Cooperator (Anderson Santos))) →	Plays (Futebol)	0.10	0.77	0.33	0.003
28	Supervised_By (Cooperator (Julio Duarte)), Plays (Futebol) →	Supervised_By (Cooperator (Claudia Justel))	0.10	1.00	0.77	1.000

FIG. 4.12: Figura da tabela contendo apenas as regras que são realmente novas, geradas após a comparação entre as regras encontradas com a mineração do DtIME e do DtIME+. Gerada após a seleção demonstrada na Figura 4.11.

4.4 CONSIDERAÇÕES FINAIS

Nesta seção, foi abordada uma forma de realizar a mineração de regras de associação na Web de Dados. Essa proposta possibilita identificar os elementos com potencial de extensão, que ao acessar os *datasets* externos, permite o enriquecimento do *dataset* local de maneira controlada, fazendo com que seja possível encontrar novas regras de associação de multirrelação. Uma visão geral da abordagem ilustrada, com exemplos, mostrou sua

funcionalidade que, em seguida, foi detalhada e expressa na forma de um algoritmo. Por fim, uma implementação da abordagem proposta também foi descrita. Esta ainda apresenta limitações como o acesso a somente um *dataset* externo. No entanto, mesmo com essa limitação, foi possível demonstrar sua funcionalidade, servindo como prova de conceito.

5 EXPERIMENTOS E RESULTADOS

Neste capítulo, serão apresentados alguns dos experimentos realizados, a fim de avaliar a abordagem proposta. Cada experimento inicia sua execução tendo como entrada dados de um determinado *dataset*, chamado *dataset* alvo. Um outro *dataset*, chamado de *dataset* externo, que é referenciado pelo *dataset* alvo também é preparado. Inicia-se o experimento com o passo 1 da abordagem, executando o MRAR sobre o *dataset* alvo. Nos passos subsequentes, um subconjunto de dados do *dataset* externo é localizado, e incluído no *dataset* alvo. Com o *dataset* alvo estendido, uma nova mineração de dados é feita, permitindo ao algoritmo encontrar novas regras de associação de multirrelação.

O experimento 1 teve como objetivo analisar um *dataset* menor com dados conhecidos. Assim, ao estendê-lo com informações complementares, os novos resultados obtidos seriam de fácil entendimento. O experimento 2 foi desenvolvido para analisar outro *dataset* menor, porém, consultando dados reais disponíveis na Web de Dados, por meio do *dataset* da *DBpedia*. Assim, demonstrou-se que é possível obter novos conhecimentos ao estender um *dataset* com dados oriundos da Web de Dados. Por fim, o experimento 3 teve como foco a análise de um *dataset* real de maior porte, esse contendo informações de uma banco de dados científico com dados botânicos, o Jabot. Ao estendê-lo com dados ambientais (clima, bioma, solo e relevo) disponíveis em outro *dataset*, foi possível obter novas regras, ampliando assim o seu conhecimento.

5.1 EXPERIMENTO 1 (*DATASET* IME - ESPORTES)

No primeiro experimento, utilizamos como alvo o *dataset* do IME (DtIME), visto na Figura 3.3, com 36 nós e 91 arestas. Este *dataset* é semelhante ao que foi usado em Ramezani et al. (2014), mostrado na Figura 3.1, e contém nós dos seguintes tipos: alunos, professores, instituições e cidades. As arestas também são de tipos diferentes: *estuda_em* (*studies_in*), *trabalha_para* (*works_on*), *orientado_por* (*supervised_by*), e *mora_em* (*lives_in*). Um outro *dataset* foi preparado, chamado DtEsportes, contendo um total de 32 nós e 34 arestas, para ser usado como *dataset* externo. Alguns dos alunos do DtIME são equivalentes a alunos presentes neste *dataset*, onde há informações sobre os esportes que os alunos praticam, como mostra a Figura 5.1.

Os dados do DtIME e do DtEsportes estão representados em forma de triplas (sujeito,

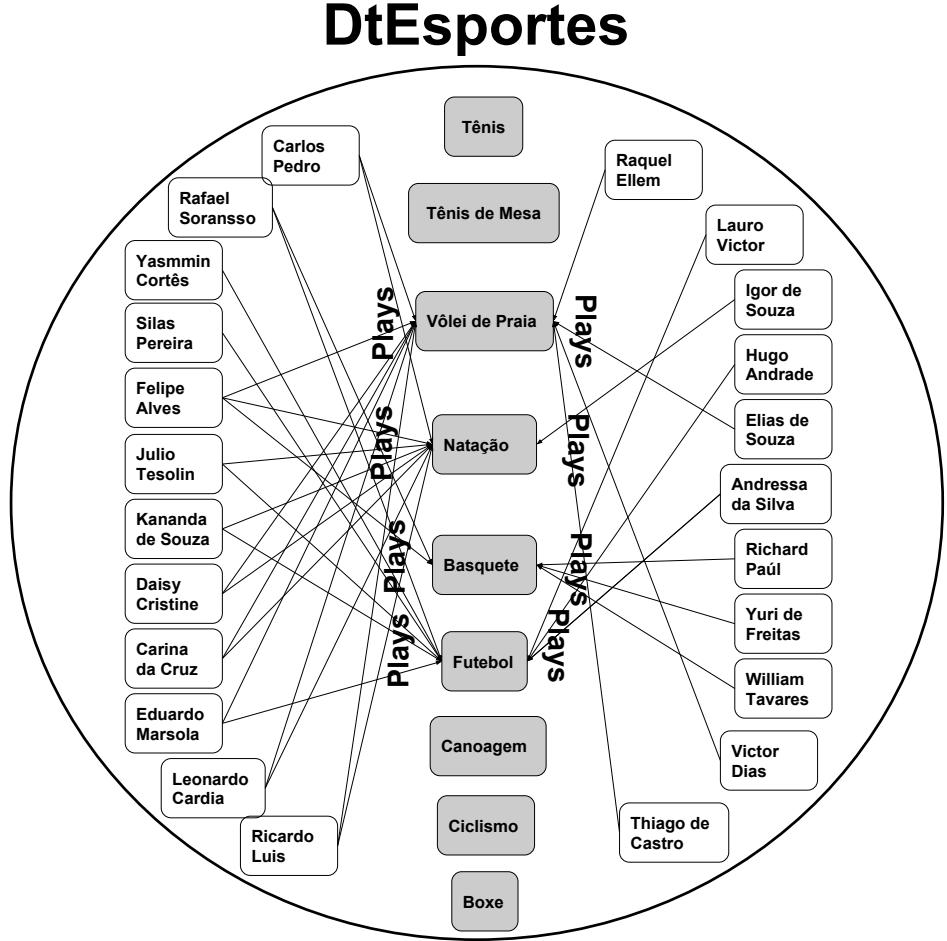


FIG. 5.1: *Dataset* Esportes com informações de pessoas que praticam esportes. Os nós são representados por retângulos e as arestas por setas. Os nós sombreados de cinza indicam os esportes praticados, já os nós em branco representam as pessoas.

predicado e objeto), no modelo RDF. Algumas URIs dos recursos (nós) existentes em DtIME apontam para o DtEsportes, através da relação dos recursos externos *Same_as*, que, por sua vez, possui informações complementares sobre os esportes praticados. Esse exemplo de associação com recursos externos pode ser visto na Figura 5.2.

Inicialmente, no passo 1, o algoritmo MRAR é executado sobre o DtIME, com a seguinte configuração de parâmetros: $MinSup = 10\%$, $MinConf = 70\%$, $MinLevel = 1$ e $MaxLevel = 4$. Foram encontradas 297 regras.

Em seguida, inicia-se o algoritmo *MRAR+*, que é executado com a mesma configuração de parâmetros. Com base nas regras encontradas no passo 1, e seus recursos (nós) de suporte, o algoritmo *MRAR+* identifica os recursos externos (passo 2), indicando qual *dataset* externo acessar, o que viabiliza o processo de extensão dos dados (passo 3). Neste passo o *dataset* alvo (DtIME) é estendido (DtIME+), e passa a ter um total de 40 nós e 126 arestas, o que representa um aumento de 11% em números de nós e 38% em relação

as arestas, comparando ao *dataset* original. É importante destacar que esse processo permitiu identificar os recursos externos com potencial de gerar novas regras, uma vez que os nós selecionados estavam diretamente ligados aos recursos que deram suporte às primeiras regras. Sendo assim, dos 32 nós presentes no *dataset* externo (DtEsportes), após remover as duplicações, foram selecionados apenas um conjunto de 4 nós, que representa apenas 12% do *dataset* externo.

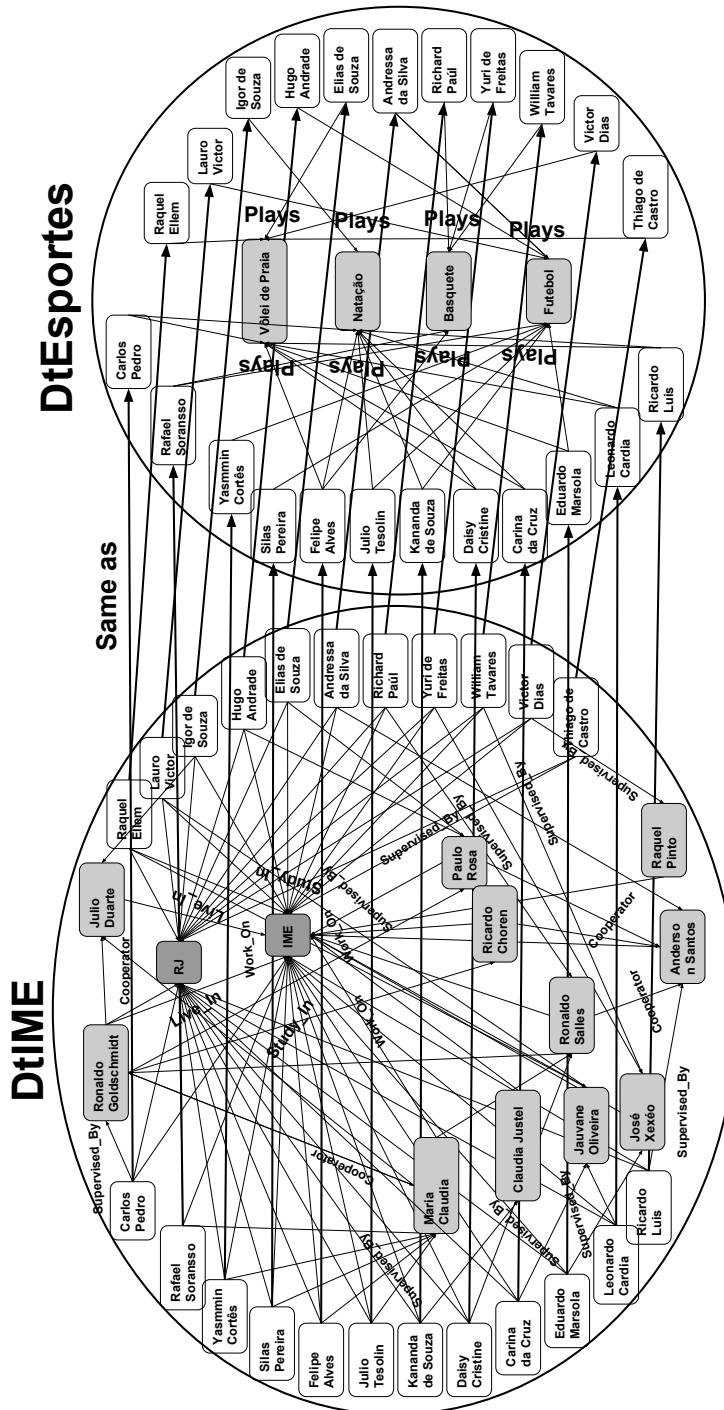


FIG. 5.2: *Datasets* demonstrando associações externas através da relação *Same as*.

A nova mineração sobre o DtIME+ encontrou um total de 400 regras de associação de multirrelação em menos de 1 segundo. A Tabela 5.1 mostra algumas das novas regras que foram geradas com as arestas identificadas e os esportes associados. É importante destacar que a regra “Supervised_By (Maria Cláudia Cavalcanti), Study_In(IME) → Plays (Futebol)” não seria encontrada sem o processo de extensão do *dataset* DtIME. Essa regra, traduzindo para o português, diz que: “Quem é supervisionado pela professora Maria Cláudia Cavalcanti e estuda no IME, joga futebol”, com o suporte de 10% e confiança de 77%. É possível observar, na Tabela 5.1, outras regras similares a essa relacionando outros professores. Isso pode sugerir que o IME oferece alguma facilidade para prática desse esporte em suas dependências.

Esse tipo de análise, em um *dataset* maior com várias instituições e alunos envolvidos, poderia levar à descoberta das instituições que possuem facilidade em apoiar a prática de esportes. Além disso, em alguns casos, é possível identificar aquelas que apoiam esportes menos populares como, por exemplo: esgrima, tênis, etc.

TAB. 5.1: Exemplo de regras geradas aplicando o algoritmo *MRAR+* sobre o conjunto de dados estendidos (DtIME+).

Antecedente	Consequente	Conf.	Sup.
Study_In(IME), Plays(Natação)	Supervised_By (Work_On(IME))	0.87	0.20
Supervised_By (Maria Claudia Cavalcanti), Study_In(IME)	Plays (Futebol)	0.77	0.10
Supervised_By (Ronaldo Goldschmidt), Study_In(IME)	Plays (Futebol)	0.77	0.10
Supervised_By (Claudia Justel), Study_In(IME)	Plays (Futebol)	0.77	0.10

Comparando os resultados obtidos com o *dataset* original, é possível ver que a análise realizada sobre o DtIME+ permitiu gerar o total de 103 regras a mais que a análise anterior. Isso representa um aumento de aproximadamente 35% na quantidade de regras geradas. Também foi possível analisar que, entre as novas regras, 112 são regras novas; 288 são comuns às duas abordagens. Para facilitar o entendimento, esses resultados são apresentados de maneira organizada na Tabela 5.2.

TAB. 5.2: Tabela comparativa dos resultados obtidos após a aplicação do MRAR e *MRAR+* sobre os dados do DtIME.

Total MRAR	Total <i>MRAR+</i>	Rg. Novas	Rg. Comuns
297	400	112	288

Como o *dataset* estendido (DtIME+) contém um número maior de nós que o *dataset* original (DtIME), mesmo que as regras encontradas possam ter valores parecidos no antecedente e consequente, em relação as regras que foram obtidas com a mineração do *dataset* original, ocorrerá que os valores de suporte atribuídos a cada regra serão diferentes. Portanto, se as regras encontradas com a mineração do *dataset* original possuírem valores de suporte próximos ao que foi definido como suporte mínimo, ao gerar o novo *dataset* com as informações complementares, isso fará que o valor de suporte para cada regra seja alterado. Como determinadas regras já possuíam um valor suporte muito próximo ao mínimo definido, a extensão realizada fará com que o novo suporte seja inferior, sendo assim, para que as mesmas regras sejam geradas após a extensão do *dataset* é necessário reajustar o valor de suporte mínimo.

Uma nova execução do algoritmo sobre os dados do *dataset* estendido foi feita, para garantir que as regras geradas após a mineração do *dataset* original, também sejam geradas ao analisar o *dataset* estendido. Para isso utilizamos a opção “Apply the best support” disponibilizada pelo sistema. Com essa funcionalidade ativa o próprio algoritmo ajusta o valor de suporte mínimo, com base nas informações salvas da execução anterior para o *dataset* original. Sendo assim, o algoritmo ajustou o valor de *MinSup* para 9%. Agora, após a execução com o novo *MinSup*, 427 regras foram geradas, sendo 130 novas e 297 comuns aos dois conjuntos analisados. As regras comuns são as mesmas que foram obtidas durante a primeira análise sobre o *dataset* alvo.

5.2 EXPERIMENTO 2 (DATASET DE JOGADORES DE FUTEBOL COM RECURSOS EXTERNOS)

Neste experimento, será apresentado o uso completo do algoritmo *MRAR+* desde a análise sobre os dados no *dataset* alvo, que passa pela identificação dos recursos com maior potencial de enriquecimento e acessa o *dataset* externo para buscar as informações complementares existentes, por último, realiza a mineração de regras de associação multirrelação.

O *dataset* alvo selecionado, chamado Dt_Neymar, para este experimento foi preparado a partir de uma adaptação do *dataset* visto na Figura 3.1, com um total de 19 nós e 31 arestas. Os recursos contendo o nome das pessoas foram trocados por URIs dos jogadores e treinadores de futebol, os quais foram encontrados na *DBpedia*. A ordem de substituição foi a seguinte: os nós (Reza, Hasan, Ali, Ahmad, Sarraee, Nematbakhsh, Mr A e Mr B) foram trocados por (dbr:Neymar, dbr:Lionel_Messi, dbr:Dani_Alves, dbr:Paulo_Henrique_Ganso, dbr:Ernesto_Valverde, dbr:Dorival_Júnior, dbr:Rogério_

Ceni e dbr:Kaká) respectivamente. Assim, o *dataset* externo nesse experimento foi o *DBpedia*, esse contendo aproximadamente 4,58 milhões de nós.

Ao iniciar a mineração de regras de associação sobre os dados do *dataset* Dt_Neymar, com a seguinte configuração mínima: $MinSup = 10\%$, $MinConf = 70\%$, $MinLevel = 1$ e $MaxLevel = 4$, foi encontrado um total de 437 regras em menos de 1 segundo. Ao acionar o *MRAR+*, foram identificados os recursos associados às regras (recursos(nós) que dão suporte às mesmas) e, para os que são externos, o *dataset* externo *DBpedia* foi consultado para buscar informações complementares sobre esses recursos, que são os jogadores de futebol.

Para ter acesso ao *dataset* externo, foi utilizado o *endpoint* de consulta do *DBpedia*¹⁹. Visando controlar o volume de informações vindas dessa fonte de dados, foi escrita uma consulta em *SPARQL* especificando, como filtro, apenas uma restrição para os predicados/relações. O filtro desta consulta foi utilizado para buscar apenas os clubes em que os referidos jogadores de futebol já fizeram parte. A consulta utilizada foi similar a *SELECT ?team WHERE {dbr:Neymar dbo:team ?team}*, sendo que, onde é visto a URI “dbr:Neymar” é, na verdade, uma variável que recebe os valores conforme cada interação do algoritmo, sobre o conjunto de recursos selecionados. Na Figura 5.3, tem-se o resultado para a consulta SPARQL vista anteriormente, com as URIs dos clubes onde o jogador de futebol Neymar já fez parte. Ao executar essa consulta para buscar o histórico de clubes para cada jogador, foram obtidos 31 nós, que representa apenas 0,0006% do *dataset* externo. Assim o Dt_Neymar foi então estendido, contendo agora um total de 50 nós e 76 arestas. Essa extensão representou um crescimento de 163% em relação aos nós do grafo original e 145% em relação as arestas.

team
http://dbpedia.org/resource/Brazil national under-23 football team
http://dbpedia.org/resource/FC Barcelona
http://dbpedia.org/resource/Brazil national under-17 football team
http://dbpedia.org/resource/Santos FC
http://dbpedia.org/resource/Brazil national under-20 football team
http://dbpedia.org/resource/Associação Atlética Portuguesa (Santos)

FIG. 5.3: Resultado de uma consulta Sparql feita no endpoint da *DBpedia*.

¹⁹<http://dbpedia.org/sparql/>

Ao aplicar novamente o MRAR (passo 4), percebemos que devido à extensão feita no *dataset*, o suporte escolhido estava muito rigoroso, a ponto de não permitir a geração de nenhuma regra, portanto foi necessário reduzir o valor do suporte mínimo. Assim, mantivemos as outras configurações e alteramos apenas o *MinSup* para 4%. Isso fez com que 39 regras fossem geradas, sendo 35 novas e 4 comuns. Para facilitar o entendimento, esses resultados são apresentados de maneira organizada na Tabela 5.3. Portanto, neste experimento, quase todas as regras geradas são novas e estão relacionadas aos clubes que os jogadores já atuaram, como mostra a Tabela 5.4.

TAB. 5.3: Tabela comparativa dos resultados obtidos após a aplicação do MRAR e *MRAR+*, sobre os dados do Dt_Neymar.

Total MRAR	Total <i>MRAR+</i>	Rg. Novas	Rg. Comuns
437	39	35	4

TAB. 5.4: Novas regras geradas, após a mineração de dados estendida, com o algoritmo *MRAR+*.

Antecedente	Consequente	Sup	Conf	Lift	Conv
dbo:team (dbr:FC_Barcelona)	Live_In (Near_By (Climate_Type (Humid)))	0.06	0.75	1.25	0.002
Live_In (Near_By (Climate_Type (Humid)))	dbo:team (dbr:FC_Barcelona)	0.06	1.00	1.25	0.979
Supervised_By (Cooperator (Work_On (Patronage (MIT))))	Study_In (IUT)	0.06	1.00	1.67	1.000
Study_In (IUT)	Supervised_By (Cooperator (Work_On (Patronage (MIT))))	0.06	1.00	1.67	1.000
dbo:team (dbr:Brazil_national_under- 20_football_team)	Study_In (IUT)	0.06	0.75	1.25	0.002

Vale notar que o processo de extensão do *dataset* alvo aumentou o número de nós existentes no grafo original. Sendo assim, quando uma nova mineração foi realizada o valor de suporte acabou sendo reajustado para cada regra. Por exemplo, a regra $X \rightarrow Y$ onde os recursos associados a ela são ‘a’ e ‘b’, em um grafo de 19 nós, faz com que seu suporte seja igual a 10% ($2/19=0.10$). Porém, quando o tamanho do grafo é aumentado para 50 nós, a mesma regra acabará ficando com suporte 4% ($2/50=0.04$). Sempre que esse suporte for inferior ao mínimo definido (*MinSup*) impedirá a geração das regras. Portanto, como nesse experimento o número de nós acrescentado ao *dataset* alvo foi superior à quantidade

já existente, acabou alterando o suporte das regras obtidas originalmente. Sendo assim, se o valor do *MinSup* for mantido igual ao da primeira execução, impedirá a identificação das mesmas regras. Por outro lado, novas regras foram identificadas, o que possibilitou ampliar o conhecimento sobre o *dataset* analisado.

É interessante destacar a regra “dbo:team (dbr:Brazil_national under-20_football_team) → Study_In (IUT)”, que traduzindo para o português diz que: “Quem joga na seleção brasileira de futebol sub-20 estuda em IUT (*Instituts Universitaires de Technologie*)”. Esta regra sugere que pode haver uma relação de cooperação entre a administração do time sub-20 e a IUT. Assim, novas relações podem surgir a partir dessa análise do *dataset* estendido. Esse é um bom exemplo de regra que não poderia ser encontrada sem o processo de extensão do *dataset* alvo.

5.3 EXPERIMENTO 3 (DATASET JBRJ/IBGE)

Neste experimento, utilizamos como alvo o *dataset* do Jardim Botânico do Rio de Janeiro (JBRJ) com seu banco de dados de coleções científicas, o Jabot²⁰, descrito por Silva et al. (2017). O Jabot foi desenvolvido em PostgreSQL DBMS, no modelo relacional, contendo 122 tabelas, 42 views, 19 triggers e 95 funções. Ele foi projetado para armazenar dados de 13 coleções científicas, como as de amostras de exsicatas (amostras de plantas desidratadas), lâminas de madeiras, sementes, DNA, carpoteca (frutos), entre outros. Para gerar uma primeira versão do grafo, escolhemos trabalhar com apenas a coleção principal, que são as exsicatas. O banco de dados completo tem mais de 790 mil tuplas, enquanto a coleção de exsicatas tem cerca de 690 mil tuplas. Para facilitar o entendimento sobre o Jabot, a Figura 5.4 mostra um subconjunto de sua modelagem no esquema relacional, que, futuramente, servirá como base para a conversão no modelo em grafos. Essa conversão será melhor descrita à frente.

Buscando representar o Jabot em grafo, que é uma etapa importante para esse experimento, foi necessário desenvolver uma modelagem para auxílio durante a conversão, vista na Figura 5.5, que passou a ser chamada de JabotG. Para obter os dados suficientes para a conversão, foi preciso realizar um conjunto de consultas em SQL e exportar os dados em formato CSV. O grafo gerado teve um total de 523.537 nós e 2.547.952 arestas com seus atributos e propriedades.

Em paralelo, uma segunda modelagem foi concebida para preparar o *dataset* externo, utilizando dados sobre localidade correlacionados ao seu bioma: clima, vegetação, relevo

²⁰<http://jabot.jbrj.gov.br/>

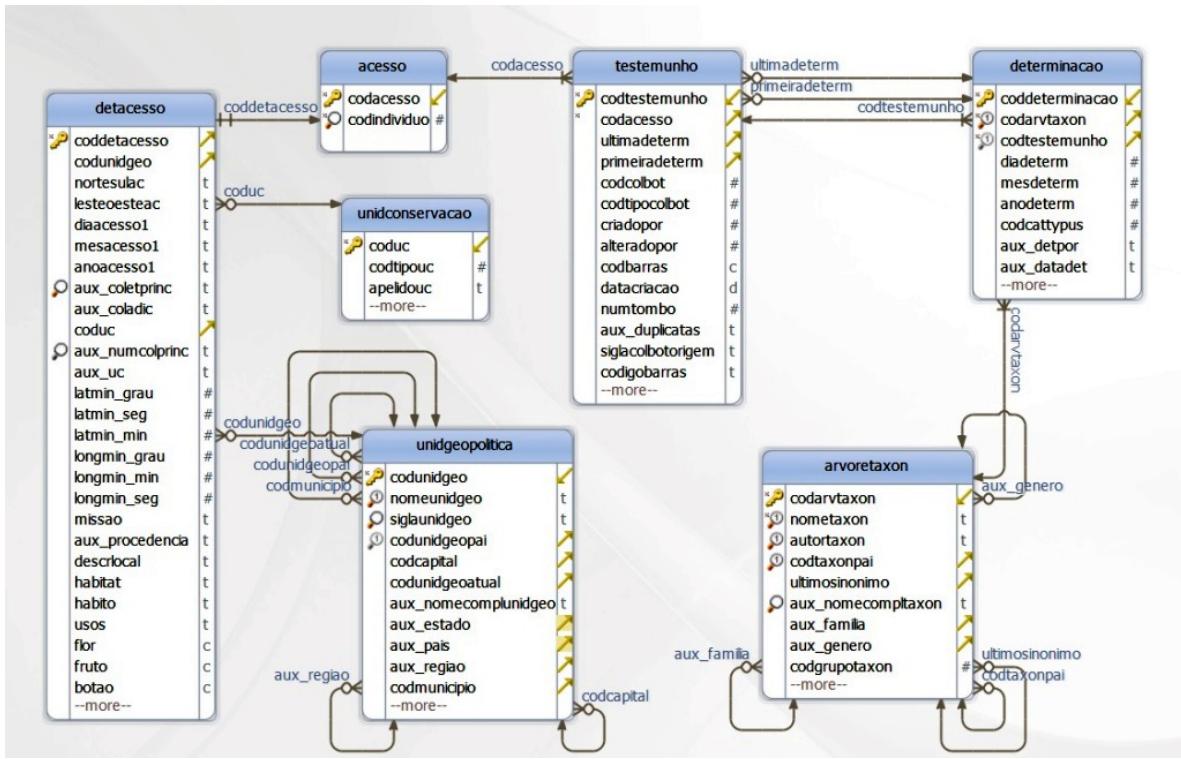


FIG. 5.4: Subconjunto do esquema relacional do banco de dados Jabot.

e/ou solo fornecidos pelo Instituto Brasileiro de Geografia e Estatística²¹ (IBGE) e pela Infraestrutura Nacional de Dados Espaciais²² (INDE). A Figura 5.6 mostra como ficou a modelagem deste *dataset*. Já a Figura 5.7 mostra que o JabotG referencia o *dataset* IBGE/INDE, através de seus nós de localização, onde cada nó aponta para outro que é semelhante a ele no conjunto de dados externos. Esse tipo de associação é comum ocorrer na Web de Dados.

É importante destacar que tanto na modelagem quanto no processo de conversão dos dados, foram utilizados alguns vocabulários, a fim de manter a semântica originalmente presente a cada área de conhecimento. O Darwin Core, representado como “dwc” nas Figuras 5.5 e 5.7, é o principal deles, pois inclui um glossário de termos destinados a facilitar o compartilhamento de informações sobre a biodiversidade, oferecendo definições, exemplos e comentários²³. O Vocabulário *Foaf*²⁴, que tem o objetivo de ligar pessoas a informações através da Web, foi utilizado para representar o nó *Person* e a relação *interest*, representados com “foaf”, vistos nas Figuras 5.5 e 5.7. Os termos representados com “ibge”, na Figura 5.6, fazem referência ao vocabulário utilizado no IBGE. Já os

²¹<https://www.ibge.gov.br/>

²²<http://www.inde.gov.br/>

²³<http://rs.tdwg.org/dwc/>

²⁴<http://xmlns.com/foaf/spec/>

termos representados com “jbг”, também vistos nas Figuras 5.5 e 5.7, fazem referência a um vocabulário próprio utilizado na criação do JabotG.

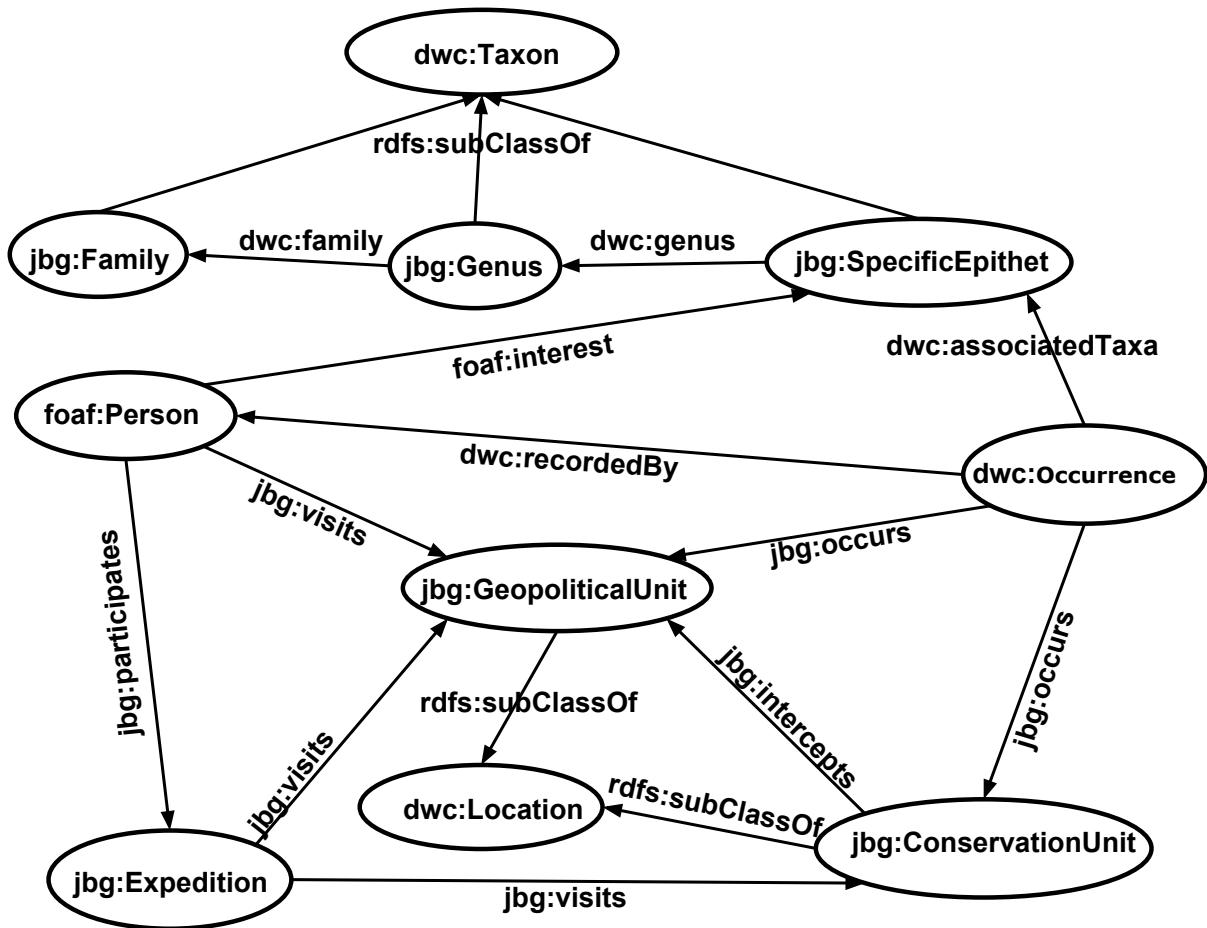


FIG. 5.5: Modelo em grafo do *dataset* JabotG. Os nós são representados em círculos ovais e as arestas em setas direcionadas, ambos apresentam um rótulo descritivo.

Para demonstrar a utilidade desse tipo de integração entre *datasets*, mostrando o resultado de consultas úteis que atravessam os dados dos dois *datasets*, foi utilizado o banco de dados em grafos Neo4j²⁵. Após o processo de importação, foram executadas algumas consultas na linguagem de Cypher, que percorreram os diferentes *datasets* para exemplificar a riqueza que os resultados dessa abordagem poderia trazer. A primeira delas, vista na Figura 5.8, apresenta quais são os diferentes climas onde uma determinada família (BROMELIACEAE) pode ocorrer. Nessa figura, as partes que descrevem e compõem a taxonomia são as seguintes: o círculo maior, em vermelho, representa a família; o círculo menor, em lilás, diz respeito ao gênero; já os amarelos representam as espécies associadas ao gênero. Os círculos verdes são as ocorrências representadas pelos códigos de barras de cada coleta. Em Rosa, são vistas as unidades geopolíticas, tanto do *dataset* alvo, quanto

²⁵<https://neo4j.com/>

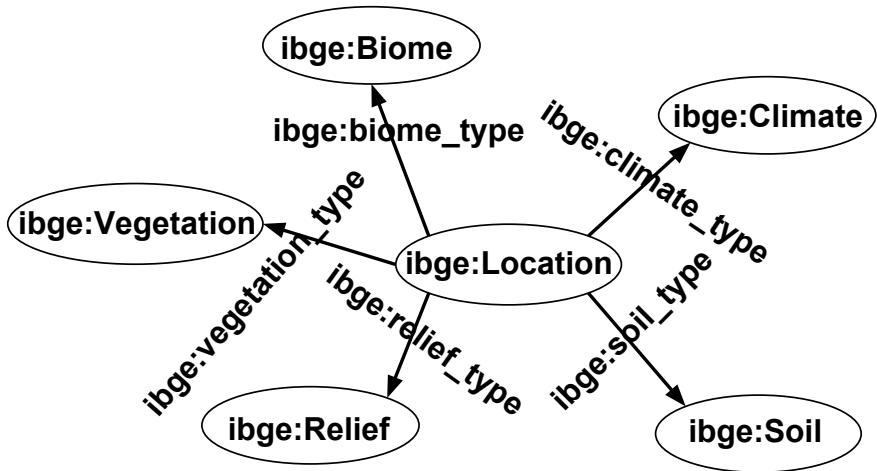


FIG. 5.6: Modelo em grafo do *dataset* IBGE/INDE. Os nós são representados em círculos ovais e as arestas em setas direcionadas, ambos apresentam um rótulo descritivo.

do externo. Observe que as setas vermelhas são as pontes entre os dois conjuntos de dados. Por fim, já no *dataset* externo, os círculos azuis são a representação dos tipos de climas associados a cada localidade. A segunda consulta, vista na Figura 5.9, apresenta uma agregação que indica a frequência com que cada família ocorre em determinado bioma e vegetação, por exemplo, a família ASTERACEAE foi encontrada 680 vezes no bioma “Cerrado” com a vegetação “Savana Arborizada com floresta-de-galeria” e 452 vezes no bioma “Mata Atlântica” com a vegetação “Floresta Ombrófila Densa Montana”.

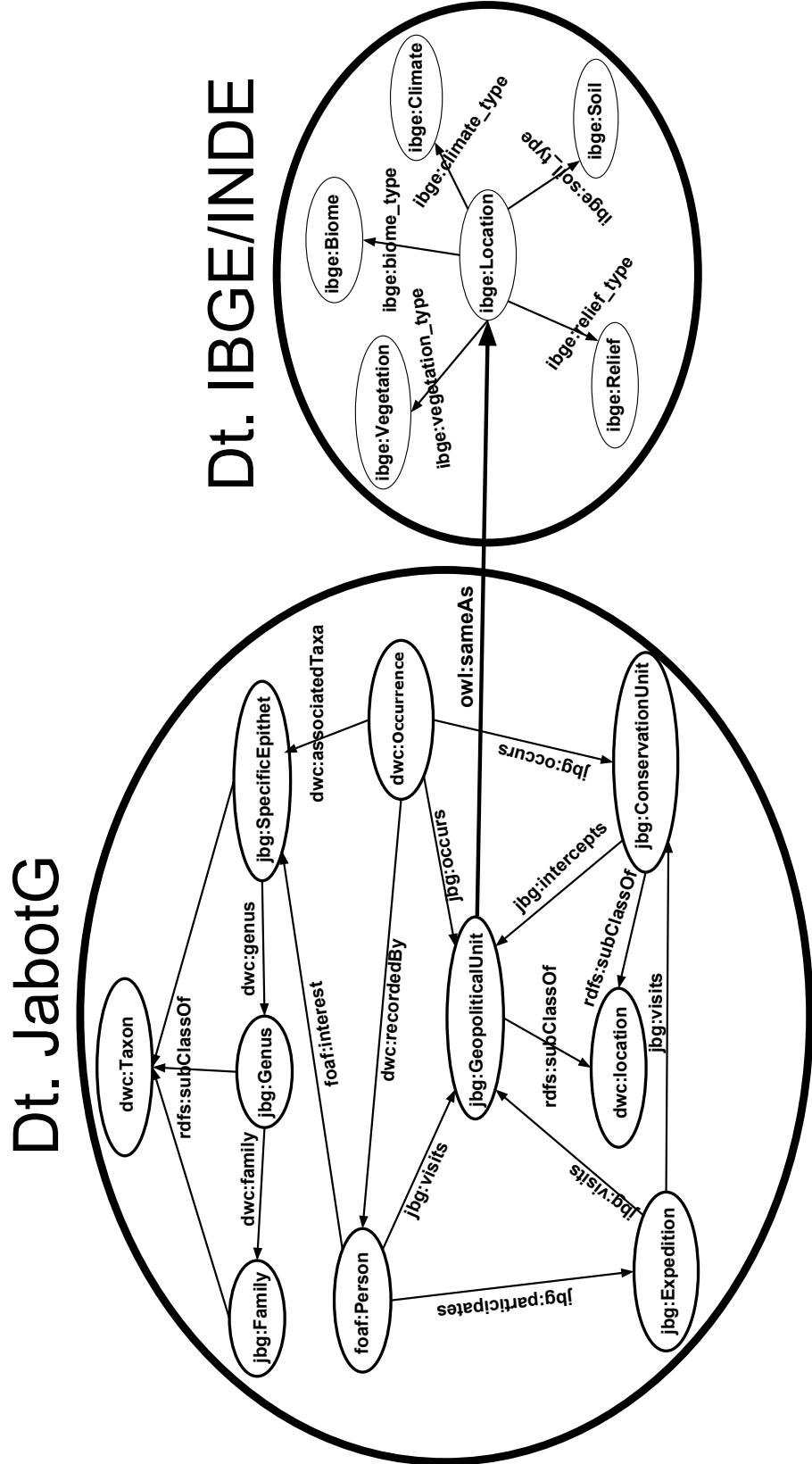


FIG. 5.7: Schema de *datasets* interligados, através da relação *Same As*.

```

MATCH (f:Family{family: "BROMELIACEAE"}) -[]-(g:Genus) -[]-
      (s:SpecificEpithet) -[]-(o:Occurrence) -[]-(geo:GeopoliticalUnit)
      -[:sameAs] ->(l:Location) -[]->(c:Climate)
RETURN c.clima_tp_umidade,c.clima_zona,c.clima_distr_umid
      ,c.clima_temperatur

```

FIG. 5.8: Variações de clima para uma determinada família. Exemplo de consulta sobre as variações de clima para a família Bromeliaceae. As setas vermelhas indicam junções entre os *datasets*.

```

MATCH (f:Family)-[]-(g:Genus)-[]-(s:SpecificEpithet)-[]-(o:Occurrence)
    -[]-(geo:GeopoliticalUnit)-[:sameAs]->(l:Location)-[]->
    (v:Vegetation), (l:Location)-[]->(b:Biome)
WITH f,v,b, count(o) as Total WHERE Total >= 200
RETURN DISTINCT f.family, b.biome, v.vegetacao_radam_nmuveg, Total
ORDER BY Total desc

```

f.family	b.biome	v.vegetacao_radam_nmuveg	Total
ASTERACEAE	Cerrado	Savana Arborizada com floresta-de-galeria	680
ASTERACEAE	Cerrado	Savana Arborizada sem floresta-de-galeria	623
ASTERACEAE	Mata Atlântica	Floresta Ombrófila Densa Montana	452
ASTERACEAE	Cerrado	Savana Parque com floresta-de-galeria	393
BROMELIACEAE	Mata Atlântica	Floresta Ombrófila Densa Montana	379
ASTERACEAE	Cerrado	Savana Florestada	333
ASTERACEAE	Cerrado	Savana Parque sem floresta-de-galeria	327
ASTERACEAE	Mata Atlântica	Savana Gramíneo-Lenhosa sem floresta-de-galeria	321
ASTERACEAE	Cerrado	Savana Gramíneo-Lenhosa sem floresta-de-galeria	319
ASTERACEAE	Cerrado	Savana Arborizada	315
ASTERACEAE	Cerrado	Savana Arborizada sem floresta-de-galeria	310
ASTERACEAE	Caatinga	Savana Arborizada sem floresta-de-galeria	280
ASTERACEAE	Amazônia	Savana Arborizada com floresta-de-galeria	278
ASTERACEAE	Mata Atlântica	Savana Arborizada	271
ASTERACEAE	Mata Atlântica	Floresta Ombrófila Densa Submontana	267
BROMELIACEAE	Mata Atlântica	Floresta Ombrófila Densa Submontana	264

Started streaming 22 records after 1301 ms and completed after 1301 ms.

FIG. 5.9: Bioma e tipo de vegetação das famílias mais frequentes, com número de ocorrências ≥ 200 .

Por se tratar de um banco de dados histórico, o Jabot possui coletas registradas a partir do ano 1768 (FORZZA et al., 2008); sendo assim, é de se esperar que parte dos dados (cerca de 71%) não possuam coordenadas geográficas para referenciar os pontos das coletas, como mostra a página de estatísticas do Jabot²⁶. Portanto, como critério de seleção dos dados, foi utilizada a base de municípios do IBGE, que foi atualizada em 2014, com quantitativo de 5570 municípios. Através de recursos da ferramenta PostGIS²⁷, foi carregado o *Shapefile* de municípios do IBGE e selecionados apenas os dados cujas coordenadas coincidiam com o município informado pelo pesquisador no momento da coleta. Outro filtro também foi aplicado para identificar os registros cujos dados possuíssem um correspondente para as variáveis ambientais no *dataset* do IBGE, visto na Figura 5.6. Esses filtros resultaram em um conjunto de dados com o total de 83.858 registros. A Figura

²⁶<http://jabot.jbrj.gov.br/v2/estatisticapublica.php>

²⁷<https://postgis.net/>

5.10 mostra como ficou a distribuição dos pontos para os registros que foram utilizados.

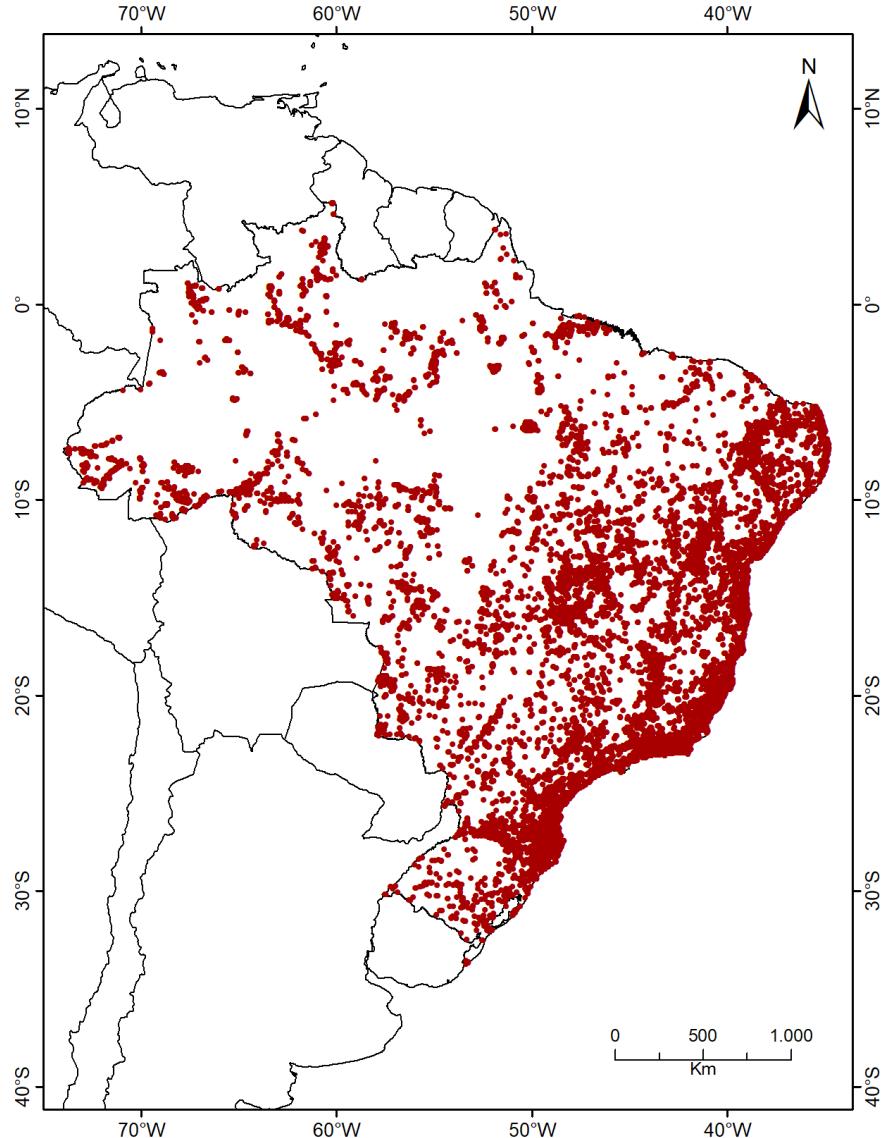


FIG. 5.10: Mapa de pontos das ocorrências (pontos em vermelho) dos registros selecionados no Jabot.

Ao executar o algoritmo MRAR sobre os dados do *dataset* Dt_JabotG, foram utilizados os seguintes valores para configuração inicial como entrada do algoritmo:

$MinSup = 40\%$, $MinConf = 60\%$, $MinLevel = 1$ e $MaxLevel = 4$. Isso permitiu que fosse gerado o total de 67 regras em menos de 1,66 minutos, para um grafo de 114895 nós e 366786 arestas.

A execução do algoritmo *MRAR+* permitiu identificar e capturar os recursos externos, em uma grafo contendo 3007 nós e 21226 arestas. Foram selecionados apenas 767 nós, que representam 25% do grafo externo. Ao adicionar os recursos externos ao Dt_JabotG um novo *dataset* foi criado, o JabotG_IBGE, contendo um total de 115662 nós e 388012

arestas, o que representou um aumento de 66% em relação aos nós e 5,78% em relação as arestas do *dataset* alvo.

Durante a análise inicial do *dataset* gerado (JabotG_IBGE), percebemos que, por ele possuir um volume grande de nós e arestas, caso fosse mantido o mesmo valor de suporte mínimo usado anteriormente, acarretaria a geração de um número muito grande de regras, aproximadamente 11 mil regras. Sendo assim, foi necessário fazer um ajuste no valor de suporte mínimo e aumentar o critério para a geração das regras. Dessa maneira, o algoritmo MRAR foi executado sobre esse novo conjunto de dados, agora com a seguinte configuração de entrada: $MinSup = 50\%$, $MinConf = 60\%$, $MinLevel = 1$ e $MaxLevel = 4$. Essa abordagem permitiu gerar um total de 552 regras em 3,37 minutos, para um grafo de 115622 nós e 388012 arestas. Ao comparar os resultados obtidos com a execução do MRAR sobre os dois *datasets*, obtivemos as seguintes números: 552 regras novas e nesse experimento não obtivemos regras comuns, como mostra a Tabela 5.5.

TAB. 5.5: Tabela comparativa dos resultados obtidos após a aplicação do MRAR e *MRAR+*, sobre os dados do Dt_JabotG e Dt_JabotG_IBGE.

Total MRAR	Total <i>MRAR+</i>	Rg. Novas	Rg. Comuns
67	552	552	0

Portanto, todas as regras geradas nesse experimento são novas e possuem relação com os dados estendidos. A Figura 5.11 mostra algumas delas e na Tabela 9.1 do Anexo 1 é possível ver um conjunto maior dessas regras. Ao analisar esse conjunto, duas das regras despertaram um maior interesse e serão detalhadas a seguir.

A regra nº 51, vista na Figura 5.11, “dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica)))) → dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))))”, traduzida para o português diz que: “As coletas realizadas por pesquisadores que visitaram localidades com o tipo de bioma Mata Atlântica, foram feitas por pesquisadores que visitaram localidades com o tipo de clima úmido”. Como essa regra, foi identificada com um suporte e confiança altos (63% e 97%), isso indica que as espécies coletadas nesse bioma, realmente, costumam ocorrer em regiões de clima úmido.

A regra nº 53, vista na Figura 5.11, “dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Argilossolo vermelho-amarelo Distrófico)))) → dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))))”, também traduzida para o português, diz que: “As coletas realizadas por pesquisadores em localidade com o tipo de solo Argilossolo vermelho-amarelo Distrófico foram feitas por pesquisadores que visitaram localidades com o tipo de clima úmido”. Semelhante ao que foi dito sobre a regra anterior, como essa também teve

Row	Antecedent	Consequent	Sup.	Conf	Lift	Conv.
51	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica)))) →	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))))	0.63	0.97	0.14	0.019
52	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido)))) →	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	0.63	0.91	0.14	0.005
54	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido)))) →	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Argilossolo vermelho-amarelo Distrófico))))	0.62	0.90	0.14	0.004
53	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Argilossolo vermelho-amarelo Distrófico)))) →	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))))	0.62	0.97	0.14	0.019
60	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido)))) →	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:super-úmido))))	0.62	0.90	0.14	0.004

FIG. 5.11: Novas regras geradas, após a mineração de dados estendida, com o algoritmo *MRAR+*.

um valor de suporte elevado (62%), indica que as espécies coletadas em regiões com esse tipo de solo ocorrem em regiões de clima úmido, com 97% de confiança. Portanto, assim como as regras em destaque, essa análise possibilitou encontrar regras que não seriam geradas sem o processo de extensão do *dataset* alvo. No Anexo 1 é possível ver mais algumas das regras encontradas.

Por se tratar de uma etapa importante, submetemos essas regras a especialistas do Jardim Botânico do Rio de Janeiro. Essa análise conjunta permitiu um esclarecimento ainda maior sobre os resultados. Uma investigação na base de dados levou a identificação de uma grande diversidade de gêneros taxonômicos que estavam associados aos recursos que deram suporte às regras. Para a regra nº 51: vista na Figura 5.11 e descrita anteriormente, dentre os gêneros encontrados, aplicou-se um filtro para selecionar os que teriam o maior número de ocorrências. Com essa análise chegou-se a lista de gêneros e seus percentuais, vista na Tabela 5.6, para o total geral de 72925 recursos associados à regra nº 51. Já para a regra nº 53, vista na Figura 5.11, também descrita anteriormente, dentre os gêneros encontrados também foi aplicado um filtro para selecionar os que teriam o maior número de ocorrências. Essa análise chegou a lista de gêneros e seus percentuais vista na Tabela 5.6, para o total geral de 71766 recursos associados à regra nº 53.

Essas informações são especialmente úteis para os pesquisadores que têm interesse nas espécies desses gêneros, pois eles podem selecionar as cidades com o clima úmido ou com o solo Argilossolo vermelho-amarelo Distrófico para suas expedições, mesmo que estejam fora do bioma Mata Atlântica, potencializando as chances de identificar as espécies de maior interesse. Outro dado importante que foi visto e que servirá para apoiar o pesquisador no momento de decidir qual localidade ele deve empregar o esforço de coleta,

TAB. 5.6: Tabela com os gêneros taxonômicos de maior porcentagem associados aos recursos que deram suporte a cada regra. Os números das regras fazem referência as regras vistas na Figura 5.11.

Regra Nº	Gêneros	%
51	PIPERACEAE Piper	2,50%
	MYRTACEAE Myrcia	1,99%
	MELASTOMATACEAE Miconia	1,82%
	MYRTACEAE Eugenia	1,49%
	SOLANACEAE Solanum	1,47%
	RUBIACEAE Psychotria	1,35%
	PIPERACEAE Peperomia	1,34%
	LEGUMINOSAE Mimosa	1,00%
	ASTERACEAE Baccharis	1,00%
53	PIPERACEAE Piper	2,51%
	MYRTACEAE Myrcia	1,99%
	MELASTOMATACEAE Miconia	1,75%
	SOLANACEAE Solanum	1,54%
	MYRTACEAE Eugenia	1,42%
	RUBIACEAE Psychotria	1,39%
	PIPERACEAE Peperomia	1,25%
	LEGUMINOSAE Mimosa	1,02%
	ASTERACEAE Baccharis	0,97%

foi que: de todos os municípios que fazem parte do bioma Mata Atlântica (3036), apenas 1354 foram visitados, que equivalem apenas a 44% dos municípios associados a esse bioma. Sendo assim, ao planejar uma nova expedição os pesquisadores podem escolher se concentrar nas localidades que ainda não foram visitadas ou que ainda não possuem o registro de ocorrência das espécies desejadas, uma vez que foi observado a existência de fortes indícios da ocorrência de suas espécies de interesse nessas regiões.

Ao analisar melhor os dados, percebeu-se a existência de uma grande similaridade entre as ocorrências dos gêneros para as duas regras descritas anteriormente. Como uma regra estava relacionada ao tipo de clima e bioma e a outra além de bioma estava relacionada ao tipo de solo, resolveu-se buscar no conjunto de regras gerado uma que estivesse relacionada ao clima, solo e bioma. Essa busca permitiu localizar a regra nº 27, vista na Tabela 9.4 do Anexo 1, “dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido)))), dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica)))) → dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Argilossolo vermelho-amarelo Distrófico))))”, gerada com o suporte de 58% e confiança de 92%. Traduzindo essa regra temos que “As coletas realizadas por pesquisadores que visitaram localidades com o tipo de clima úmido e bioma Mata Atlântica, foram feitas por pesquisado-

res que visitaram localidades com o tipo de solo Argilossolo vermelho-amarelo Distrófico”. Essa análise mostrou que as coletas foram feitas em localidades com o clima úmido, bioma Mata Atlântica e que também se verificam no tipo de solo Argilossolo vermelho-amarelo Distrófico. Dessa maneira, é possível que outras espécies do mesmo gênero pudessem ser encontradas nas mesmas localidades. Portanto, nesse mesmo sentido, sugere-se que pesquisadores com interesse em espécies distintas, porém do mesmo gênero, possam se juntar em futuras expedições a essas localidades que possuem as mesmas características ambientais identificadas pelas regras.

Fazendo analogia ao que já foi apresentado, temos várias outras formas de análises que comprovam a validade dessa abordagem. Como por exemplo, ao aplicar um filtro no antecedente do conjunto de regras e buscar apenas pelos caminhos similares que passam pelo bioma Mata Atlântica, é possível identificar todas as outras variáveis ambientais (solo, relevo, clima e vegetação) associadas aos gêneros identificados nesse bioma, como mostra a Tabela 5.7.

5.4 CONSIDERAÇÕES FINAIS

O primeiro experimento teve como objetivo analisar um *dataset* menor, visando realizar um estudo mais controlado. O objetivo do segundo foi explorar o acesso a um *endpoint* de um *dataset* real e de grande porte. Por fim, o terceiro buscou analisar um *dataset* real, o do Jardim Botânico do Rio de Janeiro (JabotG), explorando os benefícios das regras obtidas quando o JabotG foi estendido com os dados ambientais (clima, solo, relevo e vegetação). Assim sendo, os três experimentos comprovaram as hipóteses apresentadas, que afirmam:

- Seria possível analisar um determinado *dataset* levando em consideração suas ligações com outros e descobrir novas ligações? Sim, é possível, pois como foi demonstrado nos experimentos, analisar um *dataset* e considerar suas ligações externas permitem descobrir as ligações existentes em outro *dataset*.
- Se o *dataset* em questão tiver várias ligações com outros, como fazer para restringir/selecionar dados de modo a evitar a inviabilização da análise? Qual o critério? Para restringir/selecionar apenas os dados de maior relevância, bastou utilizar como critério, a seleção dos recursos do grafo que servem como suporte para as regras encontradas. Com esse critério, mesmo que um *dataset* possua muitos links externos, são utilizados apenas os que foram selecionados, o que viabiliza a análise em *datasets*

com infinitas conexões.

- Será que ao ampliá-lo com dados provenientes de outros *datasets* podemos encontrar mais associações identificando novas regras úteis? Como foi demonstrado anteriormente, os recursos selecionados do *dataset* externo foram primordiais para possibilitar a identificação de novas regras de associação de multirrelação.

Conforme foi possível observar nos experimentos, a abordagem proposta permitiu identificar dados que propiciaram o enriquecimento do *dataset* alvo com os dados externos; o que viabilizou a análise na Web de Dados. Novas regras de associações foram descobertas buscando apenas as informações sobre os recursos que davam suporte às regras.

No experimento 1, a ampliação do *dataset* que foi realizada, representou a captura de uma parcela de 12,5% do *dataset* externo (*DtEsportes*). Com essa ampliação foi possível identificar 112 regras novas, o que representou um crescimento de 37,71% em relação ao número de regras encontradas após a análise do *dataset* *DtIME*. No experimento 2, a ampliação realizada representou a captura de 0,0006% do *dataset* externo (*DBpedia*). Essa ampliação permitiu identificar 35 regras novas, representando uma crescimento de 8% em relação ao total de regras obtidos com a análise do *dataset* *Dt_Neymar*. Já no experimento 3, a ampliação representou a captura de 25,50% do *dataset* externo (*DtIBGE/INDE*). Essa ampliação fez com a análise descobrisse 552 novas regras, representando o total de 823% em relação ao número de regras obtidos após análise do *Dt_JabotG*.

Por derradeiro, novos experimentos poderiam validar ainda mais essa abordagem, entretanto, para isso, seria necessário realizar melhorias no protótipo. Por Exemplo, permitir que usuário possa explorar consultas mais avançadas em SPARQL para obter novos dados (aumentar o percentual de ampliação), e consequentemente, poder encontrar ainda mais novas regras.

TAB. 5.7: Tabela contendo algumas das regras identificadas que possuem o bioma Mata Atlântica no antecedente das regras.

Antecedente	Consequente	Sup	Conf	Lift	Conv
dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))))	0.63	0.97	0.14	0.019
dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Argilossolo vermelho-amarelo Distrófico))))	0.60	0.92	0.14	0.006
dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:super-úmido))))	0.60	0.92	0.15	0.006
dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:semi-úmido))))	0.59	0.91	0.14	0.005
dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Latossolo vermelho-amarelo Distrófico))))	0.59	0.91	0.14	0.005
dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))), dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica)))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:relief_type (ibge:Massa Dagua Costeira - Zona Econômica Exclusiva, 200 milhas))))	0.53	0.84	0.15	0.002

6 TRABALHOS RELACIONADOS

A medida que a importância de se realizar análises sobre *datasets* em grafos vem crescendo, no atual contexto da Web de Dados, sugeram trabalhos que buscam criar ferramentas e métodos que possibilitem extrair informações deste tipo de dado. Os trabalhos relacionados levantados apresentados abaixo resumiram-se aos que de fato aplicaram a mineração de regras de associação sobre *datasets* em grafo, uma vez que essa é uma condição necessária e que está diretamente ligada aos objetivos deste trabalho. Cabe ressaltar que a literatura sobre o tema é escassa. Foram encontrados poucos trabalhos e são recentes.

6.1 MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO ENTRE RÓTULOS DE NÓS

Hendrickx et al. (2015) propõem uma forma de minerar regras de associação entre rótulos de nós de um grafo, descobrindo assim informações adicionais sobre as interações existentes entre eles. Os itens frequentes são identificados através da análise sobre os rótulos. Com isso, são reconhecidos os conjuntos que são, em média, frequentes, mesmo que não estejam exatamente ligados da mesma maneira.

O algoritmo apresentado descobre regras que permitem afirmar que, se um conjunto de rótulos é encontrado em um grafo, há uma alta probabilidade de que algum outro conjunto de rótulos possa ser encontrado na sua proximidade. Descobrindo a regra que diz: “Se um nó com rótulo ‘X’ está conectado a algum outro nó no grafo de entrada, há uma alta probabilidade de que os nós com rótulos “Y” sejam encontrados nas proximidades”.

A Figura 6.1, apresenta um grafo fictício onde os nós e arestas são rotulados. Observa-se, que mesmo quando um nó se repete, suas relações (V_s) são sempre distintas. A análise realizada sobre esse tipo de grafo é diferente de outras abordagens (ELSEIDY et al., 2014; RAMEZANI et al., 2014; OLIVEIRA et al., 2017). Essa análise permite descobrir, por exemplo, uma regra que diz que: se um nó com rótulo “a” ocorre, um outro nó com rótulo “b” ocorre nas proximidades. No entanto, também é possível encontrar regras que dizem que: se um nó com rótulo “b” ocorre, um nó com rótulo “a” ocorre nas proximidades; o que não é sempre verdade, uma vez que alguns nós com rótulos “b” estão relativamente longe do nó mais próximo rotulado como “a”. Sendo assim, dada uma regra de associação do tipo $X \rightarrow Y$, foi definido como confiança a distância média de uma ocorrência de X

para a ocorrência mais próxima de Y.

Este trabalho, apesar de ser interessante, propõe-se a analisar apenas os rótulos dos nós, mas não leva em consideração a análise sobre a composição dos vários tipos de relações que estão presentes em muitas estruturas de grafos e redes atuais.

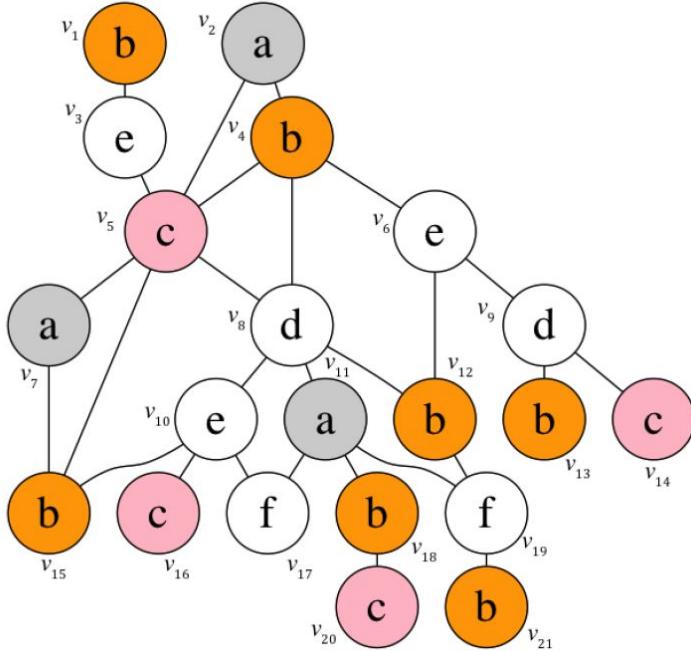


FIG. 6.1: Exemplo de *dataset* em grafo fictício com nós rotulados, apresentando uma relação diferente mesmo que o nó se repita. Fonte: Hendrickx et al. (2015).

6.2 MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO PARA SUBGRAFOS FREQUENTES EM UM ÚNICO GRAFO GRANDE

Elseidy et al. (2014) apresentam o GRAMI (GRAph MIning) para a mineração de subgrafos frequentes em um grafo grande. Essa abordagem encontra os conjuntos mínimos de casos para satisfazer o nível de frequência e evitar a enumeração custosa de todos os casos. O GRAMI faz a avaliação de modelos frequentes como um problema de satisfação de restrições (CSP – *Constraint Satisfaction Problem*). Em cada interação, ele resolve o CSP até encontrar o conjunto mínimo que é suficiente para avaliar a frequência do subgrafo, ignorando o conjunto restante.

A Figura 6.2 demonstra um exemplo da proposta do GRAMI para identificar subgrafos. A imagem identificada com (a) apresenta um grafo G de colaboração entre autores. Os nós correspondem aos autores, rotulados com seu campo de trabalho, as arestas representam coautoria rotuladas com o número de trabalhos publicados em colaboração. As

imagens (b) e (c) identificam os subgrafos S_1 e S_2 . Neste exemplo, é demonstrado que o algoritmo GRAMI identifica o número de trabalhos publicados em parceria, de acordo com os valores definidos nas relações do grafo. Com base nesses valores, são selecionados os subgrafos considerados frequentes, que atendem aos critérios especificados pelo interesse do usuário, mostrando apenas os resultados de maior relevância.

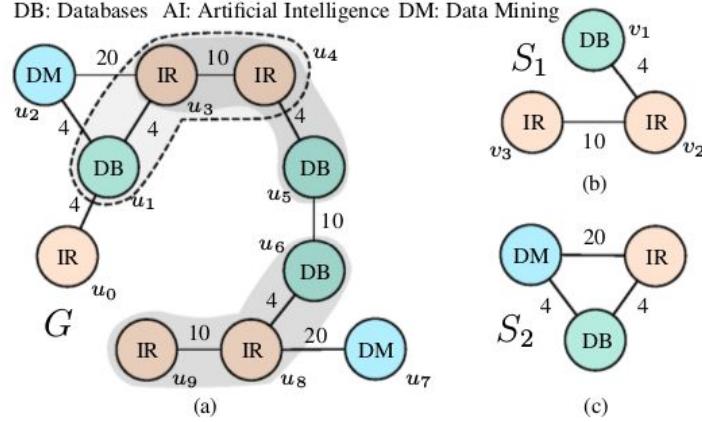


FIG. 6.2: Exemplo de *dataset* em grafo com subgrafos frequentes. Fonte: Elseidy et al. (2014).

Apesar do autor utilizar uma abordagem interessante para identificar os subgrafos de maior relevância, sua proposta não atende aos dados da Web de Dados, principalmente, quando pensamos nos vários tipos de composições de relações que estão presentes nos grafos conectados da Web de Dados.

6.3 MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO DE MULTIRRELAÇÃO

Ramezani et al. (2014) propõem o algoritmo MRAR para minerar regras de associação em banco de dados em grafo, como o exemplo da Figura 3.1, visando encontrar regras de associação de multirrelação.

Essa abordagem considera que, em um banco de dados em grafo, cada vértice tem um ou mais tipos de relacionamentos, e cada relação aponta para um outro vértice do grafo. Sendo assim, a composição dessas relações na direção em que os nós são apontados traz um entendimento maior sobre o conjunto de informação presente no grafo.

Como esse trabalho tratava de um tema similar ao desta dissertação, a seção 3.1 apresentou mais informações sobre ele. Cabe ressaltar que, durante o processo de implementação do MRAR, visto na seção 3.1.1, tivemos algumas dificuldades para entender os termos utilizados para referenciar os elementos do algoritmo, por exemplo, o que eles entendiam por *ItemChains*, sem mencionar a necessidade de um detalhamento maior para

algumas funções como a *UnionIncomingEdgesOf*. Essa é utilizada em um ponto importante do algoritmo, mas que não mostra nenhum exemplo do resultado após sua execução. Devido a essas e outras dificuldades, resolvemos dedicar parte do nosso tempo na concepção de um formalismo próprio, apresentado na seção 3.2.

Sendo assim, apesar de ser um trabalho com objetivos próximos a este, o MRAR se limita a analisar apenas um *dataset* local, diferentemente deste que considera os recursos externos para buscar novas informações sobre os dados em *datasets* externos na Web de Dados, o que permite que novas regras de associação de multirrelação sejam geradas.

6.4 MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO MULTIRRELAÇÃO EM GRAFOS: DIRECIONANDO O PROCESSO DE BUSCA

Recentemente, Oliveira et al. (2017) também apresentou um algoritmo para minerar regras de associação de multirrelação, chamado de *MRARm*, desenvolvido como uma extensão do MRAR. Diferentemente do apresentando em Ramezani et al. (2014), é possível atribuir uma máscara de busca durante o processo de mineração, isso faz com o algoritmo gere apenas as regras que estão relacionadas aos recursos de maior interesse do usuário, evitando o custo computacional para gerar todas as regras.

A Figura 6.3 mostra um exemplo da aplicação do algoritmo *MRARm*. Após atribuir os valores de suporte e confiança mínimos e os valores da máscara de busca, o algoritmo é executado e, ao invés de gerar todas as regras possíveis para aquele *dataset*, são geradas apenas as que estão associadas aos valores definidos na máscara.

Todavia, essa abordagem é diferente da que foi apresentada neste trabalho, uma vez que o algoritmo *MRARm* não leva em consideração os recursos externos para estender o *dataset* alvo e encontrar novas regras. Cabe ressaltar que o formalismo aplicado ao *MRARm* é semelhante ao que foi utilizando neste trabalho, porém, como só eram utilizados *datasets* locais, foi preciso fazer uma atualização no formalismo para que fosse possível trabalhar na Web de Dados.

6.5 COMPARAÇÃO DOS TRABALHOS RELACIONADOS

A partir das descrições acima, é possível observar que nenhum dos trabalhos leva em consideração as ligações externas com outros *datasets* da Web de Dados, o que, como foi mostrado em nossa abordagem, permitiu o enriquecimento dos dados analisados e a extração de novos conhecimentos sobre os *datasets* (alvos).

Formatted Rules			
	Humid,Good	Sup.	Conf
Live_In(Near_By(Climate_Type(Humid))) →	Health_Condition(Good)	0.11	0.69
Live_In(Kerman) →	Live_In(Near_By(Climate_Type(Humid)))	0.11	1.00
Live_In(Near_By(Shiraz)), Live_In(Kerman) →	Live_In(Near_By(Climate_Type(Humid)))	0.11	1.00
Health_Condition(Good) →	Live_In(Near_By(Climate_Type(Humid)))	0.11	1.00
Live_In(Near_By(Shiraz)) →	Live_In(Near_By(Climate_Type(Humid)))	0.11	1.00
Antecedent	Consequent	Sup	Conf

FIG. 6.3: Exemplo do algoritmo $MRAR_m$ aplicando uma máscara de busca sobre os consequentes das regras. Fonte: Oliveira et al. (2017).

Dentre os trabalhos estudados, apenas dois (RAMEZANI et al., 2014; OLIVEIRA et al., 2017) consideram relevante a composição de relações entre os nós do grafo para obtenção de conhecimento. Por outro lado, temos outros (ELSEIDY et al., 2014; HENDRICKX et al., 2015) que têm o objetivo de encontrar subgrafos frequentes e fazer análises sobre os rótulos dos nós. Entretanto, eles não consideram a possibilidade de extensão dos *datasets* com informações da Web de Dados. A Tabela 6.1 apresenta um resumo comparativo dos trabalhos relacionados.

TAB. 6.1: Tabela comparativa dos trabalhos relacionados.

Trabalho	Foco/objetivo	Análise sobre as Multirrelações
(RAMEZANI et al., 2014)	Minerar regras de associação de Multirrelação.	Sim
(ELSEIDY et al., 2014)	Minerar subgrafos frequentes.	Não
(HENDRICKX et al., 2015)	Mineração sobre os rótulos dos nós.	Não
(OLIVEIRA et al., 2017)	Utilização de máscara de busca para minerar de regras de associação de multirrelação	Sim
Mineração de regras de associação de multirrelação em datasets na Web de Dados	Minerar <i>datasets</i> na Web de Dados considerando os recursos externos para ampliar o <i>dataset</i> alvo de maneira controlada e encontrar novas regras de associação de multirrelação	Sim

6.6 CONSIDERAÇÕES FINAIS

Como já foi observado anteriormente, existem alguns trabalhos disponíveis na literatura que fazem uso da análise sobre grafos para extrair conhecimento. Os trabalhos de Elseidy et al. (2014); Hendrickx et al. (2015) apresentaram uma abordagem para identificar subgrafos frequentes, considerando os rótulos dos nós no grafo. Já Ramezani et al. (2014); Oliveira et al. (2017) realizaram análise sobre as múltiplas relações existentes em um grafo para obter conhecimento. Porém, todos eles distam do presente trabalho, pois limitam-se a analisar apenas um *dataset*; enquanto este oferece uma proposta enriquecida, mais avançada, em virtude de analisar mais de um *dataset* interligado na Web de Dados.

7 CONCLUSÕES E CONSIDERAÇÕES FINAIS

Atualmente, o mundo tem produzido um volume muito grande de dados, principalmente, devido às novas tecnologias que fazem uso de sensores e das mídias sociais - que permitem a todos produzir conteúdo. A Web de Dados vem disponibilizando esses dados por meios de formatos e padrões que possibilitem a algoritmos e ferramentas extraírem conhecimento útil. Entretanto, realizar análises sobre toda a Web de Dados é uma tarefa computacionalmente inviável devido ao grande volume de dados existentes.

Alguns movimentos sugeriram visando disponibilizar e ampliar o conhecimento sobre os dados existentes, a saber: *Open Data* e *Linked Open Data*. O *Open Data* busca tornar todos os dados abertos e permitir que eles sejam usados sem restrições. Já o *Linked Open Data* tem o objetivo de conectar os dados abertos, assim, um *dataset* de conhecimento específico, ao fazer uso de dados que não são do seu domínio deverá referenciar o *dataset* detentor destas informações. Sendo assim, o presente trabalho buscou fazer uso dos dados abertos e conectados para oferecer uma maneira de realizar análise e extrair conhecimento útil.

A mineração de dados é uma etapa importante no processo de extração do conhecimento (KDD). Nesse processo a mineração de regras de associação pode ser usada como um recurso importante permitindo identificar as regras existentes, que não são observadas facilmente. Devido à natureza dos dados oriundos da Web de Dados, foi necessário utilizar uma vertente desse recurso para realizar análise, a mineração de regras de associação de multirrelação.

Embora existam alguns trabalhos na literatura que fazem uso de algoritmos e ferramentas para extrair conhecimento, eles se limitam a analisar apenas um *dataset* por vez (RAMEZANI et al., 2014; ELSEIDY et al., 2014; HENDRICKX et al., 2015; OLIVEIRA et al., 2017). Nossa proposta visou possibilitar uma análise de maneira estendida, sem a necessidade de analisar todos os *datasets* na Web de Dados.

A partir da mineração de regras de associação de multirrelação em grafos sobre os *datasets* estendidos, foi possível confirmar a hipótese levantada, pois com nossa abordagem, foram descobertas novas regras de associação de multirrelação que antes não seriam geradas, ao analisar apenas um *dataset*.

Para possibilitar ampliar os dados do *dataset* alvo, o algoritmo *MRAR+* foi desenvol-

vido, como extensão do algoritmo MRAR, sendo ele uma das contribuições do presente trabalho.

Após a mineração de regras de associação de multirrelação sobre os dados do *dataset* original e do estendido, foi possível comparar os resultados e verificar algumas estatísticas sobre valores obtidos, a saber: o número de regras novas e regras comuns. As regras novas obtidas demonstraram que a validade desta abordagem possibilitou obter um conhecimento maior sobre os dados analisados.

7.1 CONTRIBUIÇÕES

As principais contribuições realizadas por esta dissertação são as seguintes:

- a) Uma maneira de realizar análise sobre os dados na Web de Dados sem necessitar analisar todos os *datasets*. Assim, passa a ser viável realizar análises sobre esse infinito mundo de dados, uma vez que a análise sobre um *dataset* leva a descoberta de informações complementares que permitem chegar a um novo conhecimento.
- b) Uma implementação do algoritmo MRAR, que permite encontrar regras de associação de multirrelação, e sua extensão (MRAR+) para ampliação e análise de mais de um *dataset* na Web de Dados.
- c) Proposta de representação de um esquema em grafos do *dataset* Jabot, chamado de JabotG. Dessa forma passou a ser possível disponibilizar os dados do banco de dados Jabot na Web de Dados, e com isso realizar novas análises sobre o *dataset* gerado e os *dataset*s eventualmente conectados a ele.
- d) A mecânica de conversão dos dados do Jabot para o modelo em grafo (JabotG). Essa mecânica mostrou que a conversão para o modelo em grafo oferece vantagens, além de permitir futuras adaptações visando atender diferentes tipos de análises que possam surgir.
- e) Formalismo dos conceitos envolvidos na geração de regras de associação de multirrelação em grafos. Esse formalismo foi fundamental para permitir o entendimento dos termos utilizados na mineração de regras de associação de multirrelação, pois originalmente seus conceitos conflitavam com os da mineração de regras de associação tradicional, que são vistos aplicados ao algoritmo Apriori.

Além dessas contribuições, vale mencionar que o presente trabalho gerou duas publicações. A primeira apresenta uma forma de direcionar o processo de análise sobre os

dados na Web de dados, através do uso de máscara de busca (OLIVEIRA et al., 2017), facilitando desta forma a seleção de regras mais relevantes, a serem consideradas. A outra publicação trata sobre a criação do *dataset* em grafos JabotG, a conversão dos dados relacionais para banco de dados em grafos e as possibilidades de análises que esse tipo de representação oferece (OLIVEIRA et al., 2017).

7.2 MELHORIAS E TRABALHOS FUTUROS

Nossa proposta demonstrou que acessar outros *datasets* na Web de Dados possibilita ao algoritmo enriquecer o *dataset* alvo com dados relevantes. Por meio desses dados, foi possível encontrar novas regras de associação de multirrelação. Entretanto, o presente trabalho limitou-se a acessar apenas um *dataset* externo específico por vez. Espera-se que, ao implementar uma alternativa que identifique qual *dataset* deve ser acessado, buscando dados dos múltiplos *datasets* externos disponíveis, seja possível encontrar ainda mais dados que permitam gerar novas regras de associação de multirrelação.

Além de outros experimentos com *datasets* reais, a exploração de extensões do *dataset* alvo, trazendo caminhos de maior comprimento provenientes do *dataset* externo também pode ser objeto de estudos futuros. Como já foi dito anteriormente, é um desafio encontrar uma forma de estender o *dataset* alvo, permitindo uma relação custo/benefício que valha a pena e que seja viável.

Em nossa abordagem, utilizamos somente um critério para selecionar e identificar os recursos externos dentro do *dataset* alvo, a partir dos quais a expansão é feita. Como explicado no passo 2 (seção 4.1), foi utilizado como critério a seleção dos recursos externos que dão suporte às regras. Outros critérios também poderão ser explorados, estendendo a abordagem original, como por exemplo, a utilização de outros recursos externos encontrados na composição dos *caminhos* que formam as *cadeias*, e que por sua vez formam as *regras*, conforme as definições 2, 3 e 7 da seção 3.2, respectivamente.

No método de expansão, consideramos como configurável (pré-definido) o comprimento do caminho usado para o conjunto de triplas a ser adicionado ao *dataset* alvo. No entanto, os experimentos foram realizados com comprimento = 1. Seria interessante realizar mais experimentos variando este parâmetro.

Outro trabalho futuro a considerar, é a incorporação das funcionalidades dos algoritmos MRAR e *MRAR+* a gerenciadores de banco de dados em grafo, mas, especialmente, os do tipo *Triple Store* como o AllegroGraph. Isso permitirá ao SGBDG realizar análises para descoberta de regras de associação multirrelação e ampliar seus *datasets* através da

busca por informações complementares em *datasets* externos.

Por fim, a Web de Dados é um grande desafio para as abordagens de análise em grafos e descoberta de conhecimento. Espera-se que este trabalho tenha contribuído nesta direção apresentando um caminho possível, e que inspire outros trabalhos futuros e relacionados.

8 REFERÊNCIAS BIBLIOGRÁFICAS

- AGRAWAL, R.; IMIELIŃSKI, T. ; SWAMI, A. Mining association rules between sets of items in large databases. **SIGMOD Rec.**, v. 22, n. 2, p. 207–216, 1993. Disponível em: <<http://doi.acm.org/10.1145/170036.170072>>. Acesso em: 25/08/2016.
- ANGLES, R.; GUTIERREZ, C. Survey of graph database models. **ACM Comput. Surv.**, v. 40, 2008. Disponível em: <<http://doi.acm.org/10.1145/1322432.1322433>>. Acesso em: 29/12/2017.
- BIZER, C.; HEATH, T. ; BERNERS-LEE, T. Linked data-the story so far. **International journal on Semantic Web and Information Systems**, v. 5, n. 3, p. 1–22, 2009. Disponível em: <<http://eprints.soton.ac.uk/271285/>>. Acesso em: 15/08/2016.
- BRASIL. Diário oficial da união. **Regula o acesso a informações previsto no inciso XXXIII do art. 5o, no inciso II do § 3o do art. 37 e no § 2o do art. 216 da Constituição Federal; altera a Lei no 8.112, de 11 de dezembro de 1990; revoga a Lei no 11.111, de 5 de maio de 2005, e dispositivos da Lei no 8.159, de 8 de janeiro de 1991; e dá outras provisões**, v. LEI Nº 12.527, DE 18 DE NOVEMBRO DE 2011, 2011. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Acesso em: 10/11/2015.
- BÜRKLE, P. Y. **Um Método de Pós-processamento de Regras de Associação com Base nas Relações de Dependência entre os Atributos**. 2006. 107 f. Dissertação (Mestrado em Computação) – Universidade Federal Fluminense, Niterói, Rio de Janeiro, 2006.
- EDUARDO, J.; SEGUNDO, S. WEB SEMANTICA, DADOS LIGADOS E DADOS ABERTOS: UMA VISÃO DOS DESAFIOS DO BRASIL FRENTE ÀS INICIATIVAS INTERNACIONAIS. **XVI Encontro Nacional de Pesquisa em Ciência da Informação (XVI ENANCIB)**, v. 8, n. 2177-3688, p. 219–239, 2015. Disponível em: <<http://www.ufpb.br/evento/lti/ocs/index.php/enancib2015/enancib2015/paper/view/3149>>. Acesso em: 20/08/2016.
- ELSEIDY, M.; ABDELHAMID, E. ; SKIADOPoulos, S. GRAMI: Frequent Subgraph and Pattern Mining in a Single Large Graph. **Proceedings of the VLDB**, v. 7, n. 7, p. 517–528, 2014.
- FAYYAD, U.; PIATETSKY-SHAPIRO, G. ; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.
- FORZZA, R.; MYNSSEN, C.; TAMAIO, N.; BARROS, C.; FRANCO, L. ; PEREIRA, M. As coleções do herbário. 200 anos do jardim botânico do rio de janeiro. In: LIRIO, A. (Org.). **Jardim Botânico do Rio de Janeiro: 1808 - 2008**. Rio de Janeiro: Instituto de Pesquisas Jardim Botânico do Rio de Janeiro, 2008. p. 45–55.

- GOLDSCHMIDT, R.; BEZERRA, E. ; PASSOS, E. **Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações.** [S.l.]: Rio de Janeiro: Elsevier, 2015.
- GYSSSENS, M.; PAREDAENS, J. ; GUCHT, D. V. A graph-oriented object model for database end-user interfaces. **SIGMOD Rec.**, v. 19, n. 2, p. 24–33, 1990. Disponível em: <<http://doi.acm.org/10.1145/93605.93616>>. Acesso em: 03/12/2017.
- HENDRICKX, T.; CULE, B.; MEYSMAN, P.; NAULAERTS, S.; LAUKENS, K. ; GOETHALS, B. Mining association rules in graphs based on frequent cohesive itemsets. In: CAO, T.; LIM, E.-P.; ZHOU, Z.-H.; HO, T.-B.; CHEUNG, D. ; MOTODA, H. (Org.). **Advances in Knowledge Discovery and Data Mining: 19th Pacific-Asia Conference, PAKDD 2015, Ho Chi Minh City, Vietnam, May 19-22, 2015, Proceedings, Part II.** Cham: Springer International Publishing, 2015. p. 637–648. ISBN 978-3-319-18032-8.
- KUNII, H. S. Dbms with graph data model for knowledge handling. In: PROCEEDINGS OF THE 1987 FALL JOINT COMPUTER CONFERENCE ON EXPLORING TECHNOLOGY: TODAY AND TOMORROW, s/n., ACM '87, s/n., 1987. **Anais...** Los Alamitos, CA, USA: IEEE Computer Society Press, 1987, p. 138–142. Disponível em: <<http://dl.acm.org/citation.cfm?id=42040.42071>>. Acesso em: 03/12/2017.
- LECLUSE, C.; RICHARD, P. ; VELEZ, F. O2, an object-oriented data model. **SIGMOD Rec.**, v. 17, n. 3, p. 424–433, 1988. Disponível em: <<http://doi.acm.org/10.1145/971701.50253>>. Acesso em: 03/12/2017.
- LIPTCHINSKY, V.; SATZGER, B.; ZABOLOTNYI, R. ; DUSTDAR, S. Expressive languages for selecting groups from graph-structured data. In: PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 22., WWW'13, 22., 2013. **Anais...** New York, NY, USA: ACM, 2013, p. 761–770. Disponível em: <<http://doi.acm.org/10.1145/2488388.2488455>>. Acesso em: 29/12/2017.
- OLIVEIRA, F. A.; MARTINS, Y. C.; ROCHA, D. S. B.; SIQUEIRA, M. F.; SILVA, L. A. E.; COSTA, R. L.; GOLDSCHMIDT, R. ; CAVALCANTI, M. C. Jabotg: Extending the herbarium dataset frontiers. In: INTERNATIONAL CONFERENCE ON METADATA AND SEMANTICS RESEARCH (MTSR'17), 11th., 2017. **Electronic proceedings...** Tallinn: Book of Abstracts and Posters, 2017, p. 45–53. Disponível em: <http://www.mtsr-conf.org/files/MTSR17_BOOK_OF_ABSTRACTS_AND_POSTERS_v5_final.pdf>. Acesso em: 10/12/2017.
- OLIVEIRA, F. A.; COSTA, R. L.; GOLDSCHMIDT, R. R. ; CAVALCANTI, M. C. Mineração de Regras de Associação Multirrelação em Grafos: Direcionando o Processo de Busca. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 32., SESSÃO TÉCNICA SBBD, 9., 2017. **Electronic proceedings...** Uberlândia: Simpósio Brasileiro de Banco de Dados, 2017, p. 270–275. Disponível em: <<http://sbbd.org.br/2017/wp-content/uploads/sites/3/2017/10/proceedings-sbbd-2017.pdf>>. Acesso em: 10/10/2017.
- PALETTA, F. C.; PAULO, U. D. S.; MUCHERONI, M. L. ; PAULO, U. D. S. O desenvolvimento da WEB 3 . 0 : Linked Data e DBPEDIA. **Revista PRISMA. COM**, v. 25, n. 25, p. 73–90, 2015.

PENTEADO, R. R.; SCHROEDER, R.; HOSS, D.; NANDE, J.; MAEDA, R. M.; COUTO, W. O. ; HARA, C. S. Um estudo sobre bancos de dados em grafos nativos. **ERBD-Escola Regional de Banco de Dados (ERBD'2014)**, v. 10, 2014. Disponível em: <<http://www.inf.ufpr.br/carmem/pub/erbd2014-artigo.pdf>>. Acesso em: 10/04/2016.

PICKLER, M. E. V. Web semântica: ontologias como ferramentas de representação do conhecimento. **Perspectivas em Ciência da Informação**, v. 12, n. 1, p. 65–83, 2007.

RAMEZANI, R.; SARAEE, M. ; NEMATBAKHSH, M. A. MRAR : Mining Multi-Relation Association Rules. **Journal of Computing and Security**, v. 1, n. 2, p. 133–158, 2014.

REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A. ; PAULA, M. D. Mineração de dados. **Sistemas inteligentes: fundamentos e aplicações**, v. 1, p. 307–335, 2003.

RUSSOM, P.; OTHERS. Big data analytics. **TDWI best practices report, fourth quarter**, v. 19, p. 40, 2011.

SILVA, L. A. E.; FRAGA, C. N.; ALMEIDA, T. M. H.; GONZALEZ, M.; LIMA, R. O.; ROCHA, M. S.; BELLON, E.; RIBEIRO, R. S.; OLIVEIRA, F. A.; CLEMENTE, L. S.; MAGDALENA, U. R.; VON SOHSTEN MEDEIROS, E. ; FORZZA, R. C. Jabot - botanical collections management system: the experience of a decade of development and advances. **Rodriguésia**, v. 68, n. 2, p. 391–410, 2017.

SOUZA, R. R.; ALVARENGA, L. A web semântica e suas contribuições para a ciência da informação. **Ciência da Informação**, v. 33, n. 1, p. 132–141, 2004.

TAVARES, A. T.; DE OLIVEIRA, H. R. ; LÓSCIO, B. F. Rdfmat—um serviço para criação de repositórios de dados rdf a partir de crawling na web de dados. **Revista da Escola Regional de Informática**, v. 1, n. 1, p. 6, 2015.

VAZ, J. C.; RIBEIRO, M. M. ; MATHEUS, R. Dados governamentais abertos e seus impactos sobre os conceitos e práticas de transparência no brasil. **Cadernos ppg-au/ufba**, v. 9, n. 1, p. 45–62, 2010.

VIEIRA, M. R.; FIGUEIREDO, J. M. D.; LIBERATTI, G. ; VIEBRANTZ, A. F. M. Bancos de dados nosql: conceitos, ferramentas, linguagens e estudos de casos no contexto de big data. **Simpósio Brasileiro de Bancos de Dados**, v. 27, p. 1–30, 2012. Disponível em: <http://data.ime.usp.br/sbbd2012/artigos/pdfs/sbbd_min_01.pdf>. Acesso em: 27/12/2017.

9 ANEXOS

ANEXO 1: RESULTADOS DA MINERAÇÃO DE DADOS COM O ALGORITMO MRAR SOBRE O *DATASET JABOTG_IBGE*.

TAB. 9.1: Tabela de regras de associação de multirrelação, obtida após a mineração do *dataset DtJabotG_IBGE*. Contém apenas as 27 primeiras regras, ordenadas pelo valor de suporte.

Nº	Antecedente	Consequente	Sup	Conf	Lift	Conv
1	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))))	0.63	0.97	0.14	0.019
2	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	0.63	0.91	0.14	0.005
3	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Argilossolo vermelho-amarelo Distrófico))))	0.62	0.90	0.14	0.004
4	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Argilossolo vermelho-amarelo Distrófico))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))))	0.62	0.97	0.14	0.019
5	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:super-úmido))))	0.62	0.90	0.14	0.004
6	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:super-úmido))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))))	0.62	0.98	0.14	0.031

TAB. 9.2: Tabela de regras de associação de Multirrelação, obtida após a mineração do dataset DtJabotG_IBGE. Contém apenas as 27 primeiras regras, ordenadas pelo valor de suporte. (Continuação)

Nº	Antecedente	Consequente	Sup	Conf	Lift	Conv
7	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Latossolo vermelho-amarelo Distrófico))))	0.61	0.88	0.14	0.003
8	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Latossolo vermelho-amarelo Distrófico))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))))	0.61	0.97	0.14	0.019
9	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:semi-úmido))))	0.6	0.87	0.14	0.002
10	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:semi-úmido))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))))	0.6	0.95	0.14	0.010
11	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Argilossolo vermelho-amarelo Distrófico))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	0.6	0.94	0.14	0.009
12	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Argilossolo vermelho-amarelo Distrófico))))	0.6	0.92	0.14	0.006
13	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:super-úmido))))	0.6	0.92	0.15	0.006
14	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:super-úmido))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	0.6	0.95	0.15	0.012

TAB. 9.3: Tabela de regras de associação de Multirrelação, obtida após a mineração do dataset DtJabotG_IBGE. Contém apenas as 27 primeiras regras, ordenadas pelo valor de suporte. (Continuação)

Nº	Antecedente	Consequente	Sup	Conf	Lift	Conv
15	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido)))), dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:super-úmido))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	0.59	0.95	0.15	0.012
16	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:semi-úmido))))	0.59	0.91	0.14	0.005
17	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:semi-úmido))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	0.59	0.94	0.14	0.009
18	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Latossolo vermelho-amarelo Distrófico))))	0.59	0.91	0.14	0.005
19	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Latossolo vermelho-amarelo Distrófico))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	0.59	0.94	0.14	0.009
20	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica)))), dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:super-úmido))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))))	0.59	0.98	0.14	0.031
21	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))), dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:super-úmido))))	0.59	0.94	0.15	0.01

TAB. 9.4: Tabela de regras de associação de Multirrelação, obtida após a mineração do dataset DtJabotG_IBGE. Contém apenas as 27 primeiras regras, ordenadas pelo valor de suporte. (Continuação)

Nº	Antecedente	Consequente	Sup	Conf	Lift	Conv
22	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Cerrado))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))))	0.58	0.97	0.14	0.019
23	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Latossolo vermelho-amarelo Distrófico))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:super-úmido))))	0.58	0.92	0.15	0.006
24	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Argilossolo vermelho-amarelo Distrófico))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:super-úmido))))	0.58	0.91	0.14	0.005
25	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido)))), dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:semi-úmido))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	0.58	0.97	0.15	0.022
26	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))), dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Latossolo vermelho-amarelo Distrófico))))	0.58	0.92	0.15	0.006
27	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:climate_type (ibge:úmido))), dwc:recordedBy (jbg:visits (owl:sameAs (ibge:biome_type (ibge:Mata Atlântica))))	dwc:recordedBy (jbg:visits (owl:sameAs (ibge:soil_type (ibge:Argilossolo vermelho-amarelo Distrófico))))	0.58	0.92	0.14	0.006