Web semântica em mineração de dados e descoberta de conhecimento: Uma pesquisa abrangente

Petar Ristoski ⋆ **Heiko Paulheim**

Data and Web Science Group, Universidade de Mannheim, B6, 26, 68159 Mannheim

Resumo

Mineração de dados e descoberta de conhecimento em bancos de dados (KDD) é um campo de pesquisa preocupado em obter informações de nível superior a partir de dados. As tarefas executadas nesse campo são intensivas em conhecimento e geralmente podem se beneficiar do uso de conhecimento adicional de várias fontes. Portanto, muitas abordagens foram propostas nessa área que combinam dados da Web Semântica com o processo de mineração de dados e descoberta de conhecimento. Este artigo de pesquisa fornece uma visão abrangente dessas abordagens em di ff estágios diferentes do processo de descoberta do conhecimento. Como exemplo, mostramos como o Linked Open Data pode ser usado em vários estágios para criar sistemas de recomendação baseados em conteúdo. A pesquisa mostra que, embora existam numerosos trabalhos de pesquisa interessantes realizados, o potencial total da Web Semântica e dos Dados Abertos Vinculados para mineração de dados e KDD ainda está para ser desbloqueado.

Palavras-chave: Dados abertos vinculados, Web semântica, mineração de dados, descoberta de conhecimento

1. Introdução

A mineração de dados é definida como "um processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e, em última análise, compreensíveis nos dados" [1], ou "o análise de conjuntos de dados observacionais (geralmente grandes) para encontrar relacionamentos insuspeitados e resumir os dados de maneiras novas, que sejam compreensíveis e úteis para o proprietário dos dados ”[2]. Como tal, a mineração de dados e a descoberta de conhecimento são normalmente consideradas tarefas tensas. Assim, o conhecimento desempenha um papel crucial aqui. O conhecimento pode estar (a) nos próprios dados primários, de onde é descoberto usando ferramentas e algoritmos apropriados, (b) em dados externos, que devem ser incluídos primeiro no problema (como estatísticas de segundo plano ou estatísticas). dados do arquivo ainda não vinculados aos dados primários) ou (c) apenas na mente do analista de dados.

Os dois últimos casos são oportunidades interessantes para aumentar o valor dos processos de descoberta de conhecimento. Considere o seguinte caso: um conjunto de dados consiste países da Europa e alguns fatores econômicos e sociais indicadores. Há, com certeza, alguns padrões interessantes que podem ser descobertos nos dados. No entanto, um analista que lida regularmente com esses dados saberá que alguns dos países fazem parte do Euro- União Europeia, enquanto outros não. Assim, ela pode adicionar um variável adicional Membro da UE ao conjunto de dados, o que pode levar a novas idéias (por exemplo, certos padrões válidos apenas para os estados membros da UE).

Nesse exemplo, o conhecimento foi adicionado ao dados da mente do analista, mas poderia igualmente estar contido em alguma fonte externa de conhecimento, como o Linked Open Data.

O Linked Open Data (LOD) é uma coleção aberta e interconectada de conjuntos de dados em formato interpretável por máquina, vários domínios, das ciências da vida aos dados governamentais [3, 4]. Portanto, deve ser possível fazer uso desse cofre de conhecimento em uma determinada mineração de dados, em várias etapas do processo de descoberta de conhecimento.

Muitas abordagens foram propostas nos recentes passado por usar LOD em processos de mineração de dados, por vários finalidades, como a criação de variáveis adicionais, como no exemplo acima. Neste artigo, fornecemos uma pesquisa estruturada de tais abordagens. Seguindo o conhecido modelo de processo de mineração de dados proposto por Fayyad et al. [1], discutimos como os dados semânticos são trançado na di ff estágios diferentes do modelo de mineração de dados. Além disso, analisamos como ff características diferentes

⋆ autor correspondente

Endereço de e-mail: petar.ristoski (Dinformatik.uni-mannheim.de (Petar Ristoski), heiko (Dinformatik.uni-mannheim.de (Heiko Paulheim)

de dados abertos vinculados, como a presença de interlinks entre conjuntos de dados e o uso de ontologias como esquemas para os dados, são explorados pelo di ff abordagens diferentes. O restante deste artigo está estruturado da seguinte forma. A Seção 2 define o escopo desta pesquisa e coloca-a no contexto de outras pesquisas em áreas semelhantes. A seção 3 descreve o processo de descoberta de conhecimento de acordo com Fayyad et al. Na seção 4, apresentamos um modelo geral de mineração de dados usando o Linked Open Data, seguido de uma descrição das abordagens usando dados da Web Semântica nos di ff estágios diferentes do processo de descoberta de conhecimento nas seções 5 a 9. Na seção 10, damos uma exemplo de caso de uso do processo KDD ativado para LOD no domínio dos sistemas de recomendação. Concluímos com um resumo de nossas descobertas e identificamos várias direções promissoras para pesquisas futuras.

2) Escopo desta Pesquisa

Na última década, uma grande quantidade de abordagens foi proposta, combinando métodos de mineração de dados e descoberta de conhecimento com dados da Web Semântica. O objetivo dessas abordagens é apoiar di ff tarefas de mineração de dados diferentes ou para melhorar a própria Web Semântica.

Todas essas abordagens podem ser divididas em três categorias:

- Usando abordagens baseadas na Web Semântica, Tecnologias da Web Semântica e Dados Abertos Vinculados para apoiar o processo de descoberta de conhecimento.
- Usando técnicas de mineração de dados para explorar a Web Semântica, também chamada *Mineração da Web Semântica*.
- Usando técnicas de aprendizado de máquina para criar e melhorar dados da Web Semântica.

Stumme et al. [5] forneceram uma pesquisa inicial de todos três categorias, mais tarde focando mais na segunda categoria. Datado de 2006, esta pesquisa não reflete trabalhos e tendências recentes de pesquisa, como o advento e o crescimento de Dados Abertos Vinculados. Pesquisas mais recentes sobre a segunda categoria, ou seja, Semantic Web Mining, têm publicado por Sridevi et al [6], Quoboa et al. [7], Sivakumar et al. [8] e Dou et al. [9]

Tresp et al. [10] fornecem uma visão geral dos desafios e oportunidades da terceira categoria, isto é, aprendizado de máquina na Web Semântica e uso de aprendizado de máquina. abordagens para apoiar a Web Semântica. O trabalho foi estendido em [11].

Em contraste com essas pesquisas, a primeira categoria - ou seja, o uso da Web Semântica e dos Dados Abertos Vinculados para apoiar e melhorar a mineração de dados e a disseminação de conhecimento covery - não foi objeto de uma pesquisa recente. Assim, nesta pesquisa, focamos nessa área. O objetivo desta pesquisa é fazer uma pesquisa no campo o mais amplo possível, ou seja, capturar o máximo de ff direções de pesquisa diferentes quanto possível. Como conseqüência, nem sempre é possível uma comparação direta de abordagens, pois elas podem ter sido desenvolvidas com ff objetivos diferentes, adaptados a casos de uso e / ou conjuntos de dados específicos, etc. No entanto, tentamos formular pelo menos comparações e recomendações de granularidade grossa, sempre que possível.

3. O processo de descoberta de conhecimento

Em seu artigo seminal de 1996, Fayyad et al. introduziu um modelo de processo para processos de descoberta de conhecimento. O modelo compreende cinco etapas, que levam de dados brutos a conhecimentos e insights acionáveis que são de valor imediato para o usuário. Todo o processo é mostrado na Figura 1. Ele compreende cinco etapas:

1. *Seleção* A primeira etapa é desenvolver um entendimento do domínio do aplicativo, capturando conhecimentos prévios relevantes e identificando a meta de mineração de dados da perspectiva do usuário final. Com base nesse entendimento, os dados de destino usados no processo de descoberta de conhecimento podem ser escolhidos, ou seja, selecionando amostras de dados apropriadas e um subconjunto relevante de variáveis.

2. *Pré-processamento* Nesta etapa, os dados selecionados são processados de forma a permitir uma análise subsequente. As ações típicas executadas nesta etapa incluem a manipulação de valores ausentes, a identificação (e potencialmente correção) de ruídos e erros nos dados, a eliminação de duplicatas, assim como a correspondência, fusão e resolução de conflitos para dados extraídos de ff fontes diferentes.

3. *Transformação* A terceira etapa produz uma projeção dos dados para um formulário no qual os algoritmos de mineração de dados podem trabalhar - na maioria dos casos, isso significa transformar os dados em um formulário proposicional, em que cada instância é representada por um vetor de recurso. Para melhorar o desempenho dos algoritmos de mineração de dados subsequentes, os métodos de redução de dimensionalidade também podem ser aplicados nesta etapa para reduzir o ff número efetivo de variáveis em consideração.

4. *Mineração de Dados* Depois que os dados estão presentes em um formato útil, o objetivo inicial do processo é correspondido a um método específico, como classificação, regressão ou cluster. Esta etapa inclui decidir quais modelos e parâmetros podem ser apropriados (por exemplo, modelos para dados categóricos

são di ff diferentes dos modelos para dados numéricos), e combinar um método específico de mineração de dados com os critérios gerais do processo KDD (por exemplo, o usuário final pode estar mais interessado em um modelo interpretável, mas menos preciso do que um modelo muito preciso, mas difícil de interpretar). Uma vez que os dados

Quando o método e o algoritmo de mineração são selecionados, a mineração de dados ocorre: procurando padrões de interesse em uma forma representacional específica ou em um conjunto dessas representações, como conjuntos de regras ou árvores.

5) *Avaliação e Interpretação* Na última etapa, o padrões e modelos derivados do (s) algoritmo (s) de mineração de dados são examinados em relação à sua validade. Além disso, o usuário avalia a utilidade do conhecimento encontrado para o aplicativo especificado. Esta etapa também pode envolver a visualização dos padrões e modelos tratados ou visualização dos dados usando os modelos extraídos.

A qualidade dos padrões encontrados depende dos métodos empregados em cada uma dessas etapas, bem como de suas interdependências. Assim, o modelo de processo prevê a possibilidade de voltar a cada etapa anterior e revisar as decisões tomadas nessa etapa, conforme ilustrado na Figura 1. Isso significa que o processo geral geralmente é repetido após o ajuste da parametrização ou mesmo a troca dos métodos em qualquer uma dessas etapas até a etapa qualidade dos resultados é su ffi cient.

4. Mineração de Dados usando Dados Abertos Vinculados

Como um meio de expressar conhecimento sobre um domínio na Web Semântica, *ontologias* foram introduzidos no início dos anos 90 como "especificações formais explícitas do conceitos e relações entre eles que podem existir em um determinado domínio "[12]. Para a área de descoberta de conhecimento e mineração de dados, Nigoro et al. [13] dividem as ontologias usadas nesta área em três categorias:

- *Ontologias de domínio:* Expresse conhecimento de fundo vantagem sobre o domínio do aplicativo, ou seja, o domínio dos dados em questão no qual o KDD e a mineração de dados são executados.
- *Ontologias para o processo de mineração de dados:* Definir o conhecimento sobre o processo de mineração de dados, suas etapas e algoritmos e seus possíveis parâmetros.
- *Ontologias de metadados:* Descreva o meta conhecimento sobre os dados disponíveis, como informações de proveniência, por exemplo, os processos usados para construir determinados conjuntos de dados.

Já foi demonstrado que ontologias para o processo de mineração de dados e ontologias de metadados podem ser usadas em cada etapa do processo KDD. No entanto, queremos enfatizar mais o uso de LOD (Linked Open Data) no processo de descoberta de conhecimento, que representa uma coleção interconectada publicamente disponível de conjuntos de dados de vários domínios tópicos [3, 4]. A Figura 2 fornece uma visão geral do pipeline de descoberta de conhecimento vinculado a Dados Abertos. Dado um conjunto de dados locais (como um banco de dados relacional), a primeira etapa é vincular os dados aos conceitos LOD correspondentes do conjunto de dados LOD escolhido (consulte a seção 5).[1] Depois que os dados locais são vinculados a um conjunto de dados LOD, podemos explorar os links existentes no conjunto de dados apontando para as entidades relacionadas em outros conjuntos de dados LOD. Na próxima etapa, várias técnicas para consolidação de dados, pré-processamento e limpeza são aplicadas, por exemplo, correspondência de esquema, fusão de dados, normalização de valores, tratamento de valores ausentes e outliers, etc. (cf. seção 6). Próximo, alguns As transformações nos dados coletados precisam ser realizadas formados para representar os dados de uma maneira que possam ser processados com qualquer algoritmo arbitrário de análise de dados (consulte a seção 7). Como a maioria dos algoritmos exige uma forma proposicional dos dados de entrada, isso geralmente inclui uma transformação dos dados LOD baseados em gráficos em uma forma proposicional canônica. Após a transformação dos dados, um algoritmo de mineração de dados adequado é selecionado e aplicado aos dados (consulte a seção 8). Na etapa final, os resultados do processo de mineração de dados são apresentados ao usuário. Aqui, facilite a interpretação e avaliação dos resultados do processo de mineração de dados, Web Semântica e LOD também pode ser usado (consulte a seção 9). Para a pesquisa apresentada na seção a seguir, compilamos uma lista de abordagens que atendem aos seguintes critérios:

1. Eles são projetados e adequados para melhorar o processo KDD em pelo menos uma etapa
2. Eles usam um ou mais conjuntos de dados na Web Semântica

Cada uma das abordagens é avaliada usando um número de critério:

1. A abordagem é independente do domínio ou adaptada a um domínio específico?

[1] Devemos observar que os dados podem ser vinculados aos conjuntos de dados LOD em di ff Nos estágios diferentes do processo KDD, por exemplo, em algumas abordagens, apenas os resultados e os padrões descobertos do processo de mineração de dados são vinculados a um determinado conjunto de dados LOD para facilitar a interpretação deles. Por uma questão de simplicidade, descrevemos o processo de vinculação como o primeiro passo, que também é mostrado na Figura 2.

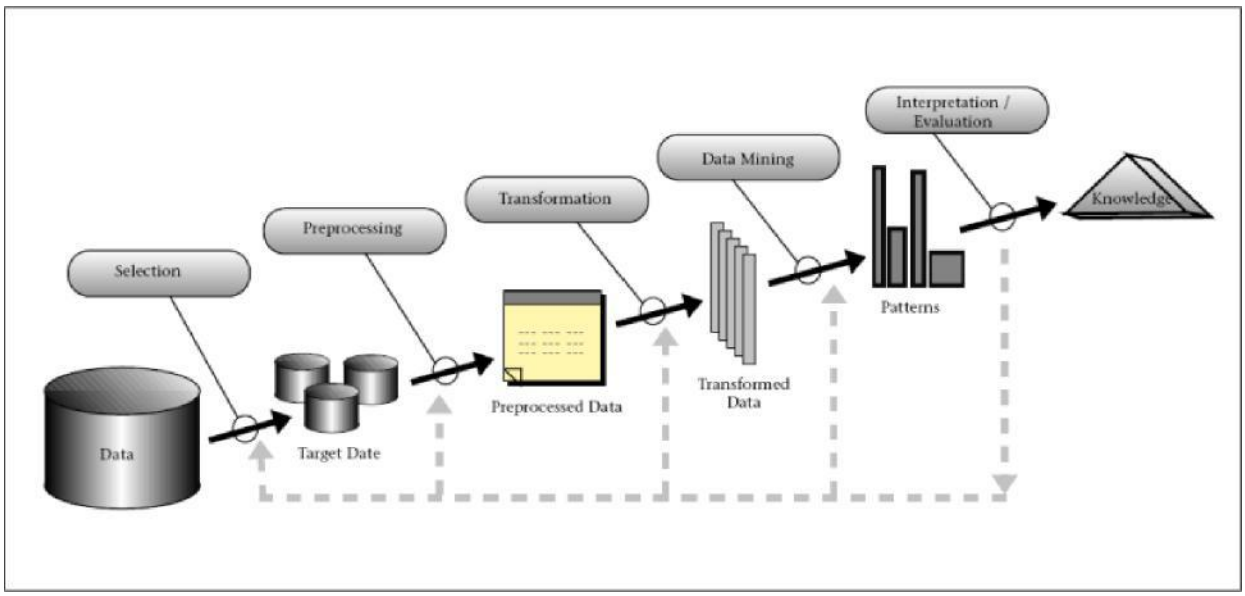

Figura 1: Visão geral das etapas que compõem o processo KDD

2. A abordagem é adaptada a uma técnica específica de mineração de dados (por exemplo, indução de regras)?
3. Utiliza uma ontologia complexa ou apenas uma axiomatizada fracamente (como uma hierarquia)?
4) Existe algum raciocínio envolvido?
5) Existem links para outros conjuntos de dados (um ingrediente central do Linked Open Data) usado?
6. A semântica dos dados (isto é, a ontologia) é explorada?

Além disso, analisamos quais conjuntos de dados da Web Semântica são usados nos artigos, para entender quais são os mais usados com destaque.

Nas seções a seguir, a pesquisa apresenta e discute as abordagens individuais [2] Uma pequena caixa no final de cada seção fornece um breve resumo, uma comparação simplificada e algumas diretrizes para os profissionais de mineração de dados que desejam usar as abordagens projetos individuais.

5. Seleção

Para desenvolver um bom entendimento do domínio do aplicativo e dos métodos de mineração de dados apropriados para os dados fornecidos, um entendimento mais profundo do dados são necessários. Primeiro, o usuário precisa entender o que é o domínio dos dados, qual conhecimento é capturado nos dados e qual é o possível conhecimento adicional que pode ser extraído dos dados. Em seguida, o usuário pode identificar a meta de mineração de dados com mais facilidade e selecione uma amostra dos dados que seriam apropriados para alcançar esse objetivo.

No entanto, a etapa de entender os dados geralmente não é trivial. Em muitos casos, o usuário precisa ter conhecimento específico do domínio para entender com êxito os dados. Além disso, os dados disponíveis são frequentemente representados em uma estrutura bastante complexa que contém relações ocultas.

Para superar esse problema, várias abordagens propõem o uso de técnicas da Web Semântica para melhor representação apresentação e exploração dos dados, explorando principais ontologias específicas e dados abertos vinculados. Esta é a primeira etapa do pipeline KDD aprimorado da Web Semântica, chamado *vinculação*. Nesta etapa, um *ligação,* ou *mapeamento,* ontologias existentes, e os conjuntos de dados LOD são formado nos dados locais.

Depois que a vinculação for concluída, o conhecimento adicional dos dados locais poderá ser extraído automaticamente. Isso permite estruturar formalmente os conceitos e informações de domínio sobre os dados, definindo tipos comuns e relações entre conceitos. Usando back-conhecimento básico em muitos casos, os usuários podem entender facilmente o domínio de dados, sem a necessidade de contratar especialistas em domínio.

Além disso, muitas ferramentas para visualização e exploração

[2] Devemos observar que algumas das abordagens podem ser aplicáveis em várias etapas do pipeline KDD habilitado para LOD. No entanto, em quase todos os casos, há uma etapa que está particularmente no foco desse trabalho, e categorizamos essas obras nessa etapa.

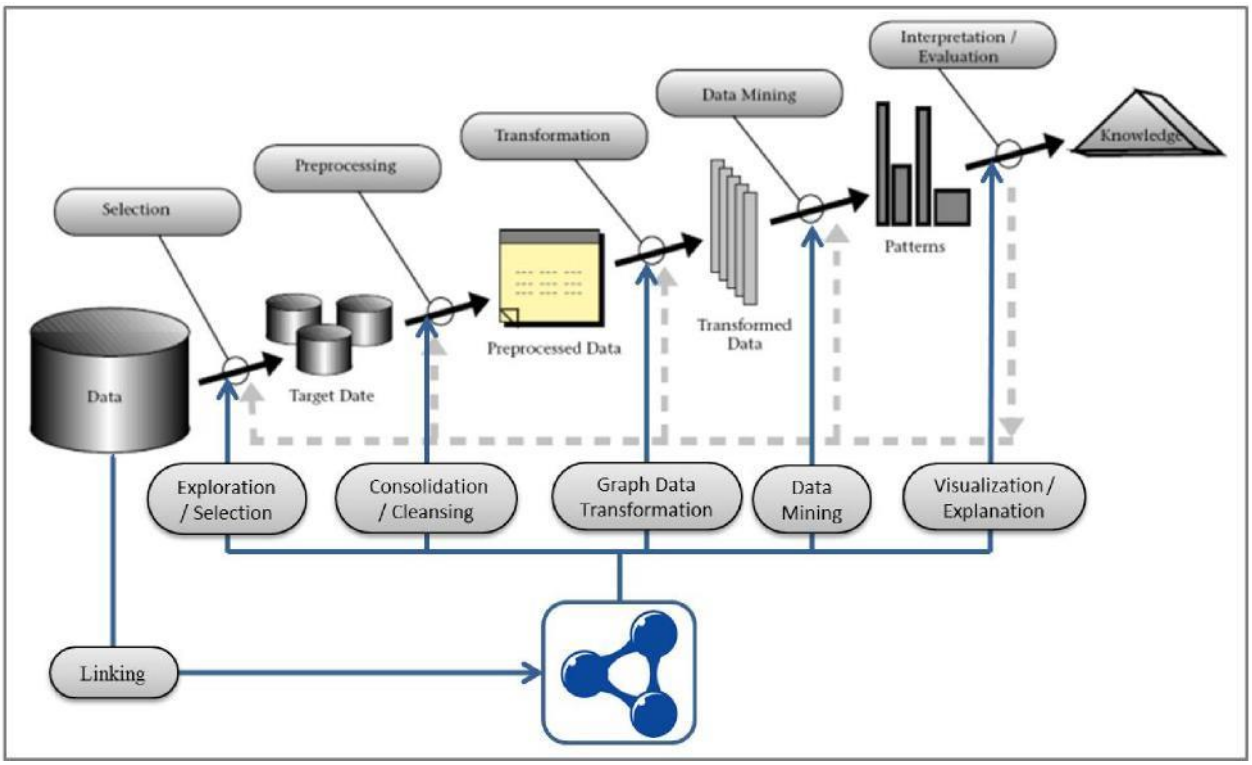

Figura 2: Uma visão geral das etapas do pipeline KDD ativado por dados abertos vinculados.

existe uma relação de dados de LOD que permitiria uma compreensão mais fácil e profunda dos dados. Uma visão geral das ferramentas e abordagens para visualização e exploração de LOD é fornecida na pesquisa de Dadzie et al. [14] Os autores primeiro estabelecem os requisitos ou o que é esperado de as ferramentas para visualização ou navegação no LOD: (i) a capacidade de gerar uma visão geral dos dados subjacentes, (ii) suporte para filtrar dados menos importantes para se concentrar em regiões de interesse selecionadas (ROI) e (iii) suporte para visualizar os detalhes nas ROIs. Mais longe Além disso, todas essas ferramentas devem permitir ao usuário intuitivamente navegue pelos dados, explore entidades e relações entre eles, explore anomalias nos dados, execute consultas avançadas e extração de dados para reutilização. Eles dividiram os navegadores analisados entre os apresentando uma apresentação baseada em texto, como o Disco [3] e Sig.ma [15] e Piggy Bank [16], e aqueles com opções de visualização, como Fenfire [17], IsaViz [4] e RelFinder [5]) A análise das abordagens mostra que a maioria dos navegadores baseados em texto fornece funcionalistas para apoiar o usuários de tecnologia, enquanto o navegador baseado em visualização focado principalmente nos usuários não técnicos. Embora os autores concluam que há apenas um número limitado de navegadores SW disponíveis, ainda podemos utilizá-los para entender melhor os dados e selecionar os que atendem às necessidades do analista de dados. A categorização das abordagens na pesquisa de Dadzie et al. foi ampliada por Peña et al. [18], com base nos tipos de dados visualizados e na funcionalidade necessária pelos analistas. Os autores listam algumas abordagens mais recentes para visualização e exploração avançadas de LOD, como CODE [19], LDVizWiz [20], LODVisualization [21] e Payola [22].

As abordagens para vincular dados locais ao LOD podem ser divididas em três categorias mais amplas, com base na representação estrutural inicial dos dados locais:

5.1 Usando LOD para interpretar bancos de dados relacionais

Os bancos de dados relacionais são considerados uma das soluções de armazenamento mais populares para vários tipos de dados e são amplamente utilizados. Os dados representados nos bancos de dados relacionais geralmente são apoiados por um esquema, que define formalmente as entidades e relações entre eles. Na maioria dos casos, o esquema é específico para cada banco de dados, o que não permite a inserção automática de dados.

[3] http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco
[4] http://www.w3.org/2001/11/IsaViz/
[5] http://www.visualdataweb.org/relfinder.php

integração de vários bancos de dados. Para facilitar e domínio-
integração e extensão de dados automáticos, uma definição de esquema compartilhado global deve ser usada nos bancos de dados.

Para superar esse problema, muitas abordagens para o mapeamento de bancos de dados relacionais para ontologias globais e conjuntos de dados LOD foram propostas. Em pesquisas recentes [23, 24, 25] as abordagens foram categorizadas em várias categorias mais amplas, com base em três critérios: existência de uma ontologia, domínio da ontologia gerada e aplicação da engenharia reversa de banco de dados. Além disso, [25] fornece uma lista dos existentes ferramentas e estruturas para o mapeamento de bancos de dados relacionais para o LOD, dos quais o mais popular e o mais usado é a ferramenta D2RQ [26]. D2RQ é uma linguagem declarativa para descrever mapeamentos entre esquemas de bancos de dados relacionais específicos de aplicativos e ontologias RDF-S / OWL.

Usando o D2RQ, os aplicativos da Web Semântica podem consultar um banco de dados não RDF usando RDQL, publique o conteúdo de um banco de dados não RDF na Web Semântica usando a API de rede RDF [6] fazer inferências RDFS e OWL sobre o conteúdo de um banco de dados não RDF usando a ontologia Jena API [7] e acessar informações em um banco de dados não RDF usando a API do modelo Jena [8]) O D2RQ é implementado como um gráfico Jena, o objeto básico de representação de informações na estrutura do Jena. Um gráfico D2RQ agrupa um ou mais bancos de dados relacionais locais em um ambiente virtual, gráfico RDF somente leitura. O D2RQ reescreve consultas RDQL e chamadas da API Jena em consultas SQL específicas do modelo de dados do aplicativo. Os conjuntos de resultados dessas consultas SQL são transformados em triplos RDF que são passados para as camadas mais altas da estrutura Jena.

5.2 Usando LOD para interpretar dados semiestruturados

Em muitos casos, os dados disponíveis são representados em uma representação semiestruturada, o que significa que os dados podem ser facilmente compreendidos por seres humanos, mas não podem ser processados automaticamente por máquinas, porque não são apoiado por um esquema ou qualquer outra representação formal. Uma das representações semiestruturadas de dados mais utilizadas é a representação tabular, encontrada em documentos, planilhas, na Web ou em bancos de dados. Essa representação geralmente segue uma estrutura simples e, ao contrário bancos de dados internacionais, não há representação explícita de um esquema.

Evidências para a semântica de dados semiestruturados podem ser encontradas, por exemplo, nos cabeçalhos das colunas, nos valores das células, nas relações implícitas entre as colunas e nas legendas.

[6] http://wifo5-03.informatik.uni-mannheim.de/ bizer / rdfapi / tutorial / netapi.html
[7] https://jena.apache.org/documentation/ontology/
[8] https://jena.apache.org/tutorials/rdf_api.html

e texto circundante. Contudo, general e É necessário conhecimento prévio específico para interpretar o significado da tabela.

Muitas abordagens foram propostas para extrair o esquema das tabelas e mapeá-lo para os existentes ontologias e LOD. Mulwad et al. fizeram significativa contribuição para a interpretação de dados tabulares usando LOD, provenientes de domínios independentes [27, 28, 29, 30, 31, 32]. Eles propuseram várias abordagens que usam conhecimento de base do Linked Open Data nuvem, como Wikitology [33], DBpedia [34], YAGO [35], Freebase [36] e WordNet [37], para inferir a semântica dos cabeçalhos das colunas, os valores das células da tabela e as relações entre as colunas e representar o significado inferido como gráfico de triplos RDF. O significado de uma tabela é, portanto, mapeado colunas para classes em uma ontologia apropriada, vinculando valores de células a constantes literais, medições implícitas ou entidades na nuvem LOD e identificando relações entre colunas. Seus métodos variam desde a simples pesquisa de índice de uma fonte de LOD até técnicas baseadas em modelos gráficos e probabilísticos raciocínio para inferir o significado associado a uma tabela [32], aplicável em ff diferentes tipos de tabelas. ou seja, tabelas relacionais, tabelas quase-relacionais (Web) e tabelas de planilhas.

Liu et al. [38] propõem uma semântica baseada na aprendizagem algoritmo de pesquisa para sugerir termos e ontologias da Web Semântica apropriados para os dados fornecidos. A abordagem combina várias medidas para semelhança semântica de documentos para criar uma semântica ponderada baseada em recursos modelo de pesquisa, que é capaz de encontrar as ontologias Os pesos são aprendidos com os dados de treinamento, usando o método de descida do subgradiente e a regressão logística.

Limaye et al. [39] propõem um novo gráfico probabilístico modelo físico para escolher simultaneamente entidades para células, tipos para colunas e relações para pares de colunas, usando o YAGO como base de conhecimento em segundo plano. Para a construção dos modelos gráficos, foram utilizados vários tipos de recursos, como texto da célula e rótulo da entidade, tipo de coluna e rótulo de tipo, tipo de coluna e entidade de célula, relação e par de tipos de coluna, pares de relação e entidade. Os experimentos mostraram que abordar os três subproblemas coletivamente e em uma estrutura de inferência gráfica unificada leva a uma maior precisão em comparação com a produção local. decisões.

Ventis et al. [40] associam vários rótulos de classe (ou conceitos) a colunas em uma tabela e identificam relações entre a coluna "assunto" e o restante das colunas da tabela. Tanto a identificação do conceito para colunas e identificação de relações são baseadas em hipótese de probabilidade mínima, ou seja, o melhor rótulo de classe

(ou relação) é aquele que maximiza a probabilidade dos valores, conforme o rótulo da classe (ou relação) da coluna. As evidências para as relações e para o classes são recuperadas de um isA extraído anteriormente banco de dados, descrevendo as classes das entidades, e banco de dados de relações, que contém relações entre as entidades. As experiências mostram que a abordagem pode obter rótulos significativos para tabelas que raramente existem em as próprias tabelas, considerando que os recuperados

a semântica leva à pesquisa de alta precisão com pouca perda de recuperação de tabelas em comparação com abordagens baseadas em documentos.

Wang et al. [41] propõem um algoritmo multifásico que usando a taxonomia probabilística universal chamada *Probase [* 42] é capaz de entender os direitos, atributos e valores em muitas tabelas na Web. A abordagem começa identificando uma única "coluna de entidade" em uma tabela e, com base em seus valores e no restante da coluna cabeçalhos, associa um conceito do conhecimento Probase base da borda com a mesa.

Zhang et al. [43, 44] propõem uma abordagem incremental de inicialização que aprende a rotular colunas da tabela usando dados parciais na coluna e usa um recurso genérico modelo capaz de usar vários tipos de contexto de tabela na aprendizagem. O trabalho foi estendido em [45], onde o autor mostra que, usando técnicas de seleção de amostras, é possível anotar semanticamente as tabelas da Web de maneira mais ffi maneira eficiente.

Da mesma forma, uma abordagem para interpretar dados de formulários da Web usando LOD foi proposta [46]. A abordagem começa extraindo as pars de valor-atributo do formulário, o que é feito usando métodos de análise. Em seguida, os dados extraídos dos formulários da Web são representados como RDF gráfico RDF triplo ou completo. Para enriquecer o gráfico com semântica, ele é alinhado com uma grande ontologia de referência, como o YAGO, usando abordagens de alinhamento de ontologias.

Um caso específico são as tabelas na Wikipedia, que seguem uma certa estrutura e, com links para outros Páginas da Wikipedia, podem ser mais facilmente vinculadas às existentes Fontes de LOD, como DBpedia. Portanto, várias abordagens para interpretar tabelas da Wikipedia com LOD foram propostas. Munoz et al. [47, 48] propõem métodos para triplicar as tabelas da Wikipedia, chamadas WikiTables, usando as bases de conhecimento existentes de LOD, como DBpedia e YAGO. Seguindo a idéia das abordagens anteriores, essa abordagem começa extraindo entidades das tabelas e descobrindo as relações existentes entre elas. Da mesma forma, uma abordagem de aprendizado de máquina foi proposto por Bhagavatula et al. [49], onde não

A base de conhecimento de LOD é usada, mas apenas metadados para os tipos de entidades e relações entre eles são adicionados. Da mesma forma, foram propostas abordagens para pretendendo dados tabulares em planilhas [50, 51], CSV [52], e XML [53].

5.3 Usando LOD para interpretar dados não estruturados

A mineração de texto é o processo de análise de informações não estruturadas, geralmente contidas em um texto em linguagem natural, para descobrir novos padrões. Mais comum As tarefas de mineração de texto incluem categorização de texto, agrupamento de texto, análise de sentimentos e outros. Na maioria dos casos, os documentos de texto contêm entidades nomeadas que podem ser identificadas no mundo real e mais informações podem ser extraídas sobre elas. Várias abordagens e APIs têm foi proposto para extrair entidades nomeadas do texto documentos e vinculá-los ao LOD. Uma das APIs mais usadas é o DBpedia Spotlight [54, 55], que permite anotar automaticamente documentos de texto com URIs do DBpedia. Essa ferramenta é usada em vários LOD habilitados abordagens de mineração de dados, por exemplo, [56, 57, 58, 59]. De várias Existem APIs para extrair riqueza semântica do texto, como Alchemy API [9] API do OpenCalais [10], API Textwise SemanticHacker [11]) Todas essas APIs são capazes de anotar entidades nomeadas com conceitos de vários conhecimentos bases, como DBpedia, YAGO e Freebase. Essas ferramentas e APIs foram avaliadas na estrutura NERD, implementada por Rizzo et al. [60] Além disso, o Linked Open Data também é muito usado para melhor entendimento das mídias sociais, o que notícias de autoria e outros conteúdos textuais da Web, redes sociais os dados da mídia apresentam uma série de novos desafios para as tecnologias semânticas, devido à sua natureza em larga escala, barulhenta, irregular e social. Uma visão geral de ferramentas e abordagens para representação semântica de meios sociais dia streams é dado em [61]. Esta pesquisa discute cinco questões-chave de pesquisa: (i) Quais ontologias e recursos da Web de Dados podem ser usados para representar e raciocinar sobre a semântica dos fluxos de mídia social? Por exemplo, FOAF [12] ontologia GUMO [62] para descrição pessoas e redes sociais, SIOC [13] e ontologia DLPO [63] para modelagem e interligação de mídias sociais, ontologia MOAT [64] para modelagem de semântica de tags (ii) Como os métodos de anotação semântica capturam a rica semântica implícita nas mídias sociais? Para a prova- extração de frase-chave [65, 66], pesquisa baseada em ontologia reconhecimento de cidades, detecção de eventos [67] e detecção de sentimentos citegangemi2014frame, sentilo. (iii) como pode

[9] http://www.alchemyapi.com/api/
[10] http://www.opencalais.com/documentation/ documentação-opencalais
[11] http://textwise.com/api
[12] http://xmlns.com/foaf/spec/
13http: //sioc-project.org/

extraímos informações confiáveis desses fluxos de conteúdo barulhentos e dinâmicos? (iv) Como podemos modelar usuários identidade digital e atividades de mídia social? Por exemplo, descobrindo dados demográficos do usuário [68], derivando interesses do usuário [69] e capturando o comportamento do usuário [70]. (v) Quais métodos de acesso à informação baseados em semântica podem ajudar a resolver o comportamento complexo de busca de informações nas mídias sociais? Por exemplo, a busca semântica nas mídias sociais [71] e a mídia social transmite recomendações [72].

Uma vez que o usuário tenha desenvolvido um su ffi entendimento eficiente do domínio e a tarefa de mineração de dados é definida, eles precisam selecionar uma amostra de dados apropriada. Se os dados já foram mapeados para ontologias específicas de domínio apropriadas ou vinculados a dados abertos vinculados externos, os usuários podem selecionar mais facilmente um representante amostra e / ou significativo subpopulação dos dados para a tarefa de mineração de dados fornecida. Por exemplo, para uma coleção de textos, o usuário pode decidir selecionar aqueles que mencionam um político *depois de* os dados foram vinculados à web semântica, para que essa seleção seja possível.

A Tabela 1 fornece uma visão geral das abordagens discutidas nesta seção. [14] Pode-se observar que, na etapa de seleção, os links entre os conjuntos de dados desempenham apenas um papel menor, e o raciocínio é escassamente usado. Na maioria dos casos, bases de conhecimento de uso geral, como DBpedia ou YAGO, são usados como fontes de conhecimento.

A seleção de conjuntos de dados semânticos relevantes da web geralmente é feita por *interligação* um conjunto de dados em mãos com dados do Linked Open Data. Existem estratégias e ferramentas para di ff tipos diferentes de dados: os bancos de dados relacionais geralmente são mapeados para a web semântica usando regras e ferramentas de mapeamento, como o D2R. Nesses casos, as regras de mapeamento geralmente são escritas manualmente, o que é facilmente possível porque o esquema de um banco de dados relacional é geralmente definido explicitamente.

Dados semiestruturados, como tabelas da Web, geralmente vêm sem semântica explícita e em grandes quantidades. Aqui, di ff Muitas vezes, abordagens heurísticas e de aprendizado de máquina são aplicadas para vinculá-las às fontes de LOD. Nesse caso, foi demonstrado que a combinação de abordagens que executam a correspondência de esquema e instância de uma maneira holística geralmente supera as abordagens que lidam com ambas as tarefas isoladamente. Para dados não estruturados, ou seja, conteúdo textual, a interligação é normalmente feita vinculando entidades nomeadas no texto a fontes LOD com ferramentas como o DBpedia Spotlight.

Após a interligação, as técnicas de visualização e resumo de dados podem se beneficiar de conhecimento adicional contido nos conjuntos de dados interligados.

6. Pré-processamento

Depois que os dados são mapeados para o conhecimento específico do domínio, as restrições expressas nas ontologias podem ser usado para executar verificações de validade e limpeza de dados. As ontologias podem ser usadas para detectar discrepâncias e ruídos, bem como para lidar com valores ausentes e violações de intervalo de dados e restrições, além de orientar os usuários através de etapas personalizadas de pré-processamento.

Ontologias são freqüentemente usadas em muitas aplicações de pesquisa. abordagens para o uso de limpeza e pré-processamento de dados. Nomeadamente, há duas aplicações de ontologias nesse estágio: ontologias independentes de domínio usadas para gerenciamento de qualidade de dados e ontologias de domínio.

A primeira categoria de ontologias geralmente contém especificações instruções para executar operações de limpeza e pré-processamento. Nessas abordagens, a ontologia é geralmente usada para orientar o usuário no processo de limpeza e validação de dados, sugerindo possíveis operações para ser executado sobre os dados. A segunda categoria de As tecnologias fornecem conhecimento específico do domínio necessário para validar e limpar dados, geralmente de maneira automática.

6.1 Abordagens independentes de domínio

Uma das primeiras abordagens que usa uma qualidade de dados ontologia é proposta por Wang et al. [74] Eles

[14] As tabelas usadas para resumir as abordagens no final de cada seção estão estruturadas da seguinte maneira: A segunda coluna da tabela indica o domínio do problema no qual a abordagem é aplicada. A terceira coluna indica a tarefa / domínio de mineração de dados que foi usada na abordagem. As próximas duas colunas capturam as características das onologias utilizadas na abordagem, ou seja, o nível de complexidade da ontologia e se o raciocínio é aplicado à ontologia. Com base em uma categorização prévia de ontologias apresentada em [73], distinguimos dois graus de complexidade ontológica: ontologias de baixa complexidade que consistem em hierarquias de classe e relações de subclasse (marcadas com *EU),* e onotologias com alta complexidade, que também contêm outras relações além das subclasses, além de outras restrições, regras etc. (marcadas com *H)* A sexta coluna indica se os links (como *owl: sameAs)* a outras fontes LOD foram seguidas para extrair informações adicionais. A próxima coluna indica se as informações semânticas explícitas foram usadas de uma determinada fonte de LOD. As duas colunas finais listam as fontes LOD usadas e ontologias compartilhadas, respectivamente. Se uma fonte LOD for usada, a respectiva ontologia também será usada, sem declarar explicitamente isso na tabela.

Tabela 1: Resumo das abordagens usadas na etapa de seleção.

	Domínio		Ontologia				dados usados de ontologia	
Aproximação	Problema	Mineração de dados	Complexidade	Raciocínio			LOD	Conjuntos de
[27, 28, 29, 30, 31, 32]	organizações	*Eu*			não	sim		Wikitologia
	pessoas, lugares, organizações		LH		não	não	WordNet *Eu*	
	comércio biologia medicina, publicações			sim	sim	sim	YAGO, Base livre, DBpedia DBpedia, ClinicTrials.gov , BibBase uma	
	pessoas, lugares, organizações			não	Sim Sim	não	não Links	
	pessoas, lugares,			não				
[51] [50]	geografia, música, filmes			não	não	não Semântica	DBpedia, YAGO, WordNet	
[40] [39] [53]	organizações,			não			Banco de dados gratuito Probase, DBpedia YAGO,	
[43, 44, 45] [41]	pessoas, lugares,			não sim	não sim	sim sim LOD		
				não			YAGO, DBpedia, MusicBrainz b	
[46]	livros	////////	HHHHHHH	não	não	sim sim	YAGO	////////

uma http://data.bibbase.org/

b http://linkedbrainz.org/

propor uma estrutura chamada *OntoClean* [15] para limpeza de dados baseada em ontologia. O componente principal da estrutura é o componente de ontologia de limpeza de dados, usado na identificação do problema de limpeza e dados relevantes. Nesse componente, a ontologia da tarefa especifica os métodos possíveis que podem ser adequados para atender às metas do usuário, e a ontologia do domínio inclui todas as classes, instâncias e axiomas em um domínio específico, que fornece conhecimento de domínio, como como restrições de atributo para verificar valores inválidos durante executar as tarefas de limpeza.

Uma abordagem semelhante é proposta por Perez et al. [75] com o *OntoDataClean* estrutura, capaz de orientar o processo de limpeza de dados em um ambiente distribuído ambiente. A estrutura usa uma ontologia de pré-processamento para armazenar as informações sobre as transformações necessárias. Primeiro, o processo de identificação e armazenamento das etapas de pré-processamento necessárias deve ser realizado por um especialista em domínio. Então, essas transformações são necessárias para mogenizar e integrar os registros para que possam ser corretamente analisados ou unificados com outras fontes. Finalmente, as informações necessárias são armazenadas na ontologia de pré-processamento e as transformações de dados podem ser realizadas automaticamente. A abordagem foi testada em quatro bases de dados no domínio da biomedicina, mostrando que, usando a ontologia, os dados podem ser corretamente pré-processados e transformados de acordo com as necessidades.

6.2 Abordagens específicas de domínio

Uma das primeiras abordagens para usar um domínio específico ontologia é proposta por Philips et al. [76] A abordagem usa ontologias para organizar e representar o conhecimento sobre atributos e suas restrições nos bancos de dados relacionais. A abordagem é capaz de identificar automática ou semi-automaticamente, com a assistência do usuário, a identificação os domínios dos atributos, relações entre os atributos, atributos duplicados e entradas duplicadas no banco de dados.

Kedad et al. [77] propõem um método para lidar com a heterogeneidade semântica durante o processo de limpeza de dados ao integrar dados de várias fontes, o que é diferente ff diferenças nas terminologias. A solução proposta é baseada no conhecimento linguístico fornecido por um domínio é uma ontologia. A idéia principal é gerar automaticamente asserções de correspondência entre instâncias de objetos com base na hierarquia is-a, onde o usuário pode especificar o nível de precisão expresso usando a ontologia de domínio. Depois que o usuário especificar o nível de precisão, dois conceitos serão considerados iguais se houver uma relação de subsunção entre eles ou ambos pertencerem a a mesma classe. Usando essa abordagem, o número de resultados pode ser aumentado ao consultar os dados, por exemplo, para a consulta "Os carros vermelhos sofrem mais acidentes do que outros?" o sistema não apenas procurará *carros vermelhos,* mas também para carros com cores *rubi, vermelhão,* e *sevilha,* qual são subclasses da cor vermelha.

Milano et al. introduzir a estrutura OXC [78] que permite a limpeza de dados em documentos XML com base em uma representação uniforme do conhecimento do domínio por meio de uma ontologia, coletada a partir da análise de domínio atividades e pelas DTDs dos documentos. o A estrutura compreende uma metodologia para avaliação e limpeza da qualidade dos dados com base na ontologia de referência e uma arquitetura para limpeza de dados XML com base nessa metodologia. Dada uma ontologia de domínio, um mapa relação entre a DTD e a ontologia é determinada multado, que é usado para definir dimensões de qualidade (precisão, integridade, consistência e moeda) e executar a melhoria da qualidade dos dados, contando com a semântica codificada pela ontologia.

Brueggemann et al. [79] propõem uma combinação de ontologias específicas de domínio e ontologias de gerenciamento de qualidade de dados, anotando ontologias de domínio com metadados específicos de gerenciamento de qualidade de dados. Os autores mostraram que essa abordagem híbrida é adequada para verificação de consistência, detecção duplicada e gerenciamento de metadados. A abordagem foi estendida em [80], onde sugestões de correção estão sendo geradas para cada inconsistência detectada. A abordagem usa a estrutura hierárquica da ontologia para oferecer ff er o usuário correção semântica relacionada ao contexto sugere ções. Além disso, a estrutura usa várias medidas de distâncias semânticas em ontologias para encontrar as correções mais adequadas para as inconsistências identificadas. Com base nessas métricas, o sistema pode ff er set- sugestões gerais para correções de valor, ou seja, valor da próxima irmão, primeiro filho e pai. A abordagem foi aplicada aos dados do registro de câncer da Baixa Saxônia [16] mostrando que ele pode suportar com êxito especialistas em domínio.

Wang et al. [81] apresentam um índice externo discrepante baseado em densidade método de detecção usando a ontologia de domínio, denominada *ODSDDO (* Detecção de Outlier para Documentos Curtos usando Ontologia de Domínio). O algoritmo é baseado no *fator outlier local* algoritmo e usa ontologia de domínio calcular a distância semântica entre documentos curtos, o que melhora a precisão da detecção de valores extremos.

[15] Não deve ser confundido com o método de engenharia de ontologia de Guarino e Welty.

[16] http://www.krebsregister-niedersachsen.de

Para calcular a semelhança semântica entre dois documentos, primeiro cada palavra de cada documento é mapeada para o conceito correspondente na ontologia. Então, usando árvore conceitual da ontologia, a semelhança entre cada par de conceitos é calculado. A distância entre dois documentos é então simplesmente calculada como média da soma das semelhanças máximas entre os pares de conceitos. Os documentos que têm segregação pequena ou nula semântica semântica com outros documentos no conjunto de dados são considerados outliers.

Lukaszewski [82] propõe uma abordagem para admitir e utilizar dados ruidosos, permitindo modelar di ff diferentes níveis de granularidade do conhecimento, tanto em treinamento quanto em testes amplos. Os autores argumentam que erros ou falta de Os valores de tributo podem ser introduzidos pelos usuários de um sistema que são obrigados a fornecer valores muito específicos, mas o nível de seu conhecimento do domínio é muito geral para descrever com precisão a observação pelo valor de um atributo. Portanto, eles propõem conhecimento representação de borda que usa hierarquias de conjuntos de valores de atributos, derivadas de hierarquias de subsunção de conceitos de uma ontologia, o que diminui o nível de ruído de atributo nos dados.

Fëuber e Hepp [83, 84, 85, 86] propõem abordagens por usar tecnologias da Web Semântica e Dados Abertos Vinculados para reduzir o ff esforço para gerenciamento da qualidade dos dados em bancos de dados relacionais. Eles mostram que o uso de dados de referência LOD pode ajudar a identificar valores ausentes, valores de gal e violações de dependência funcional. No seu primeiro trabalho [83], os autores descrevem como identificar e classificar problemas de qualidade de dados em bancos de dados relacionais, por meio do uso da SPARQL Inferencing Notation (SPIN) [17] SPIN é um vocabulário da Web Semântica estrutura de processamento que facilita a representação implementação de regras baseadas na sintaxe do protocolo SPARQL e da linguagem de consulta RDF. Para aplicar a abordagem em bancos de dados relacionais, a ferramenta D2RQ [26] é usada para extrair dados de bancos de dados relacionais em um representante RDF. ressentimento. A estrutura permite que especialistas em domínio definir requisitos de dados para seus dados com base em formulários como parte do processo de gerenciamento da qualidade dos dados. A estrutura SPIN identifica automaticamente violações de requisitos em instâncias de dados, como erros sintáticos, valores ausentes, violações de valores exclusivos, fora ou intervalo valores e violações de dependência funcional. Essa abordagem é estendida em [85] para avaliar o estado de qualidade dos dados em dimensões adicionais.

Em um trabalho adicional [84], em vez de definir manualmente Com as regras de validação de dados, os autores nos propõem a criação de dados abertos vinculados como uma base de conhecimento confiável que já contém informações sobre as dependências de dados. Foi demonstrado que essa abordagem reduz significativamente o ff orientação para gerenciamento da qualidade dos dados, quando dados de referência estão disponíveis na nuvem LOD. A abordagem foi avaliada com base em uma base de conhecimento local que continha dados de endereço criados manualmente. Usando GeoNames como um conjunto de dados LOD de referência, a abordagem conseguiu identificar entradas de cidade inválidas e relações cidade-país inválidas. Uma abordagem semelhante usando SPIN, foi desenvolvida por Moss et al. [87] para avaliar dados médicos. O sistema compromete um conjunto de ontologias que suportam o raciocínio em um domínio médico, como psicologia humana, domínio médico e dados do paciente. Para executar a limpeza de dados, foram usadas várias regras para verificar pontos de dados ausentes e verificação de valor. A abordagem é avaliada em dados da rede Brain-IT [18], mostrando que é capaz de identificar valores inválidos nos dados. Ontologias são frequentemente usadas no domínio da saúde para gerenciamento de qualidade e limpeza de dados. A revisão da literatura de tais artigos é apresentada em [88].

Em [89], desenvolvemos uma abordagem para preencher valores ausentes em uma tabela local usando LOD, que é implementado em um sistema chamado *Mannheim Search Join Engine* [19] O sistema conta com um grande corpus de dados, rastreado de mais de um milhão de di ff diferentes sites. Além de dois grandes conjuntos de dados quase-relacionais, o corpus de dados inclui o *Conjunto de dados Billion Triples Cahllenge 2014* [20] [90] e o *Conjunto de dados de microdados WebDataCommons* [21] [91] Para uma determinada tabela local, o mecanismo procura no corpus de dados dados adicionais para os atributos das entidades na tabela de entrada. Para executar a pesquisa, o mecanismo usa as informações existentes na tabela, ou seja, os rótulos das entidades, os cabeçalhos dos atributos e os tipos de dados dos atributos. Os dados descobertos geralmente são recuperados de várias fontes, portanto, os novos dados são consolidados primeiro usando métodos de correspondência de esquema e fusão de dados. Em seguida, os dados descobertos são usados para preencher os valores ausentes na tabela local. Além disso, a mesma abordagem pode ser usada para validar os dados existentes na tabela fornecida, isto é, detecção externa, detecção e correção de ruído. A Tabela 2 fornece uma visão geral das abordagens discutidas nesta seção. Podemos observar que, embora ontologias sejam freqüentemente usadas para limpeza de dados, conjuntos de dados LOD conhecidos como o DBpedia, dificilmente são explorados. Além disso, muitas abordagens foram adotadas

[17] http://spinrdf.org/

[18] http://www.brain-it.eu/

[19] http://searchjoins.webdatacommons.org/

[20] http://km.aifb.kit.edu/projects/btc-2014/

[21] http://webdatacommons.org/structureddata/

lamentadas e avaliadas no domínio médico, provavelmente porque existem muitas ontologias sofisticadas nesse domínio. Ontologias e dados da Web Semântica ajudam no pré-processamento dos dados, principalmente para aumentar a qualidade dos dados. Existem várias dimensões de qualidade de dados que podem ser tratadas. Valores extremos e valores falsos podem ser encontrados identificando pontos de dados e valores que violam as restrições definidas nessas ontologias. Hierarquias de sub-suposição e relações semânticas ajudam a unificar sinônimos e detectar inter-relações entre atributos. Finalmente, os valores ausentes podem ser inferidos e / ou preenchidos a partir de conjuntos de dados LOD.

7) **Transformação**

Nesta fase, a geração de melhores dados para os dados processo de mineração está preparado. A etapa de transformação inclui redução de dimensionalidade, geração e seleção de recursos, amostragem de instância e transformação de atributos, como discretização de dados numéricos, agregação, transformações funcionais etc.

texto de mineração de dados semântica habilitada para a Web, *geração de recursos* e *seleção de recursos* são particularmente relevantes.

7.1 Geração de Recursos

O Open Data Linked foi reconhecido como uma fonte valiosa de conhecimento em segundo plano em muitos data mining tarefas. Aumentar um conjunto de dados com recursos extraídos do Linked Open Data pode, em muitos casos, melhorar os resultados de um problema de mineração de dados, enquanto externaliza o custo de criar e manter esse conhecimento em segundo plano [92].

A maioria dos algoritmos de mineração de dados trabalha com uma proposição *vetor de recurso* representação dos dados, ou seja, cada instância é representada como um vetor de recursos (f_1 f_2 $\ldots$, f_n) onde os recursos são binários (ou seja, $f_i \in \{$ *verdadeiro falso*$\}$), numérico (ou seja, $f_i \in R$), ou nomeado inal (ou seja, $f_i \in S$, Onde S é um conjunto finito de símbolos) [93]. Os dados abertos vinculados, no entanto, vêm na forma de *gráficos* conectando recursos com tipos e relações, apoiados por um esquema ou ontologia.

Assim, para acessar dados abertos vinculados com Para ferramentas de mineração de dados, é necessário realizar transformações, que criam recursos proposicionais a partir dos gráficos no Linked Open Data, ou seja, um processo chamado *proposicionalização* [94] Geralmente, recursos binários (por exemplo, verdadeiro se um tipo ou relação existir, falso caso contrário) ou recursos numéricos (por exemplo, contar o número de relações de um determinado tipo) são usados. Além disso, recursos numéricos ou nominais elementares (como a população de uma cidade ou o estúdio de produção de um filme) podem ser adicionado [95]. Outras variantes, por exemplo, computando as frações possíveis relações de um certo tipo, mas raramente são usava.

No passado recente, foram propostas algumas abordagens para proposicionalizar dados abertos vinculados para fins de mineração de dados. Muitas dessas abordagens são super- vistos, ou seja, eles permitem que o usuário formule consultas SPARQL, o que significa que eles deixam a estratégia de proposicionalização para o usuário e uma geração totalmente automática de recursos não é possível. Geralmente, os recursos resultantes são agregados binários ou numéricos usando SPARQL COUNT construções.

LiDDM [96] é um sistema integrado para mineração de dados na Web Semântica. A ferramenta permite que os usuários declarem consultas SPARQL para recuperar recursos do LOD que podem ser usados em di ff tecnologia de aprendizado de máquina técnicas, como agrupamento e classificação. Mais longe- mais a ferramenta ff operadores para integrar dados de várias fontes, filtragem e segmentação de dados, que são transportados manualmente pelo usuário. A utilidade da ferramenta foi apresentada em dois casos de uso, usando DBpedia, World FactBook [22] e LinkedMDB [23], na aplicação da análise de correlações e aprendizado de regras.

Uma abordagem semelhante foi usada no Rapid-Miner [24] semweb plugin [97], que processa previamente o RDF dados de forma que possam ser processados posteriormente por um ferramenta de mineração, o RapidMiner nesse caso. Novamente, o usuário precisa especificar uma consulta SPARQL para selecionar os dados de interesse, que são convertidos em vetores de recursos. Os autores propõem dois métodos para lidar com valores definidos dados, mapeando-os em um vetor N-dimensional espaço. O primeiro é *FastMap,* que incorpora pontos em um espaço N-dimensional com base em uma métrica de distância, bem como o Escala Multidimensional (MDS). O segundo é a Análise de Correspondência (CA), que mapeia valores para um novo espaço com base em sua coocorrência com valores de outros atributos. As abordagens foram avaliadas em dados do IMDB [25], mostrando que as funções de mapeamento podem melhorar os resultados sobre a linha de base. Cheng et al. [98] propõem uma abordagem para automação geração de recursos após o usuário especificar o tipo de recursos. Para fazer isso, os usuários precisam especificar a consulta SPARQL, que torna essa abordagem supervisionado. A abordagem foi avaliada no domínio de sistemas de recomendação (domínio de filmes) e classificação de texto.

[22] ttp: //wifo5-03.informatik.uni-mannheim.de/ factbook /

[23] http://www.linkedmdb.org/

[24] http://www.rapidminer.com/ 25http: //www.imdb.com/

Tabela 2: Resumo das abordagens usadas na etapa de pré-processamento.

Aproximação	Domínio		Ontologia		LOD		Conjuntos de dados usados	
	Problema	detecção de outlier	Complexidade	Raciocínio			LOD	Ontologia
[7 4]	geografia			não	não	não		Ontologia OntoClean
[7 5]				não	não	não		Ontologia OntoDataClean
[77]				não	não			ontologia personalizada ontologia personalizada
	medicamento		HHH	não	não	Links semânticos não		
[78]	Eu biomedicina			não	não	não		
[79, 80]	medicamento		HH	sim	não não	não	///	
		Mineração de dados de		não	não	não	///	ontologia personalizada ontologia personalizada
[83, 84, 85, 86] [81] [77]	geografia	////	HH	sim		sim		
[89]	social medicina geografia, empresas, filmes, livros, música, pessoas, drogas	///	H	não	sim	não sim	GeoNames uma microdados DBpedia BTC 2014, Conjunto de dados de WebDataCommons	// ontologia personalizada
[87]	medicina mídia	//	HH	não	não	não	//	ontologia personalizada

uma http: // sws geonames org /

(classificação de tweets). Os resultados mostram que o uso de recursos semânticos pode melhorar os resultados dos modelos de aprendizado em comparação ao uso apenas de recursos padrão. Mynarz et al. [99] consideraram usar consultas SPARQL especificadas pelo usuário em combinação com SPARQL agregados, incluindo CONTAGEM, SOMA, MIN, MÁX. Kaup-pinen et al. desenvolveram o pacote SPARQL para R [100, 101], que permite importar dados LOD em um ambiente muito conhecido para computação e gráficos estatísticos *R*. Em suas pesquisas, eles usam a ferramenta para realizar formar análise estatística e visualização dos dados vinculados da floresta amazônica brasileira. A mesma ferramenta foi usada em [102] para análise estatística em dados de relatórios de ataques de pirataria. Além disso, eles usam a ferramenta para importar dados RDF de várias fontes LOD no ambiente de *R,* o que lhes permite analisar, interpretar e visualize os padrões descobertos nos dados.

*FeGeLOD [*95] foi a primeira abordagem totalmente automática para enriquecer dados com recursos derivados do LOD. Nesse trabalho, propusemos seis di ff características estratégias de geração de estruturas, permitindo recursos binários e agregados numéricos simples. As duas primeiras estratégias dizem respeito apenas às próprias instâncias, ou seja, à recuperação das propriedades dos dados de cada entidade e dos tipos da entidade. As quatro outras estratégias considere a relação das instâncias com outros recursos no gráfico, ou seja, relações de entrada e saída, e *relações qualificadas,* isto é, agrega o tipo de relação e da entidade relacionada. O trabalho foi continuado nos dados abertos vinculados do RapidMiner, tensão [27] [103, 104]. Atualmente, a extensão RapidMiner LOD suporta o usuário em todas as etapas do processo de descoberta de conhecimento ativado por LOD. ou seja, vincular, combinar dados de várias fontes de LOD, pré-processamento e limpeza, transformação, análise de dados e pretação dos resultados da mineração de dados.

O FeGeLOD e a extensão RapidMiner LOD foram utilizados em ff aplicações de mineração de dados diferentes, ou seja, classificação de texto [58, 57, 56, 105], explicação de estatísticas [106, 107, 108], detecção de erros de ligação [109] e sistemas de recomendação [110, 111]. Além de usar uma representação binária e numérica simples dos recursos, propusemos o uso de versões adaptadas de medidas baseadas em TF-IDF. Em [112], realizamos uma comparação inicial de di ff estratégias de proposicionalização diferentes (isto é, binário, contagem, contagem relativa e TF-IDF) para gerar recursos a partir de tipos e relações do Linked Open Dados.

Um problema semelhante à geração de recursos é solucionado por *Funções do kernel,* que calculam a distância entre entre duas instâncias de dados. A similaridade é calculada contando subestruturas comuns nos gráficos das instâncias, por exemplo, passeios, caminhos e três. Os kernels gráficos são usados nos algoritmos de mineração de dados e aprendizado de máquina baseados em kernel, mais comumente suportam vec- SVMs, mas também pode ser explorado para tarefas como armazenamento em cluster. No passado, muitos kernels gráficos foram propostos para aplicações específicas [113, 114, 115] ou para representações semânticas específicas [116, 117, 118, 119]. Mas apenas algumas abordagens são gerais o suficiente para serem aplicadas em qualquer Dados RDF, independentemente da tarefa de mineração de dados. Lösch et al. [120] introduz dois núcleos gráficos RDF gerais, baseados em gráficos de interseção e árvores de interseção. Primeiro, eles propõem o uso de grãos de caminhada e de caminho, que contam o número de passeios e trilhas na região gráficos selecionados. Em seguida, eles propõem o kernel completo da subárvore, que conta o número de subárvores completas da árvore de interseção.

O kernel do caminho da árvore de interseção introduzido por L¨osch et al., foi modificado e simplificado por Vries et al. [121, 122, 123, 124], que também permite o cálculo explícito dos vetores de recursos das instâncias, em vez de similaridades em pares. A computação dos vetores de recursos melhora significativamente o tempo de computação e permite usando qualquer método arbitrário de aprendizado de máquina. Eles desenvolveram dois tipos de kernels sobre dados RDF, kernel RDF walk count e kernel subárvore RDF WL. O kernel do RDF walk count conta os di ff percorre os subgráficos (até a profundidade do gráfico fornecida) ao redor os nós das instâncias. O kernel da subárvore RDF WL conta os di ff subárvores completas nos subgráficos (até a profundidade do gráfico fornecida) ao redor dos nós das instâncias, usando o algoritmo Weisfeiler-Lehman [125]. As abordagens desenvolvidas por Lëosch et al. e por Vries et al. foram avaliados em duas aprendizagens relacionais comuns tarefas: classificação da entidade e previsão de links.

7.2 Seleção de Recursos

Mostramos que existem várias abordagens que geram vetores de características proposicionais do Linked Dados abertos. Geralmente, os espaços de recursos resultantes podem ter uma dimensionalidade muito alta, que leva a problemas tanto no desempenho quanto na precisão dos algoritmos de aprendizado. Portanto, é necessário aplicar algumas abordagens de seleção de recursos para reduzir a possibilidade de espaço adequado. Além disso, para conjuntos de dados que já possuem alta dimensionalidade, o conhecimento de base do LOD

[26] http://linkedscience.org/tools/sparql-package - for-r /

[27] http://dws.informatik.uni-mannheim.de/en / pesquisa / extensão rápida-mineiro /

recursos lingüísticos ou linguísticos, como o WordNet, podem ajudar a reduzir o espaço de recursos melhor do que as técnicas padrão que não explore esse conhecimento prévio.

A seleção de características é um problema muito importante e bem estudado na literatura. O objetivo é identificar recursos correlacionados ou preditivos do rótulo da classe. Geralmente, todos os métodos de seleção de recursos podem ser dividido em duas categorias mais amplas: métodos de wrapper e métodos de filtro (John et al. [126] e Blum et al. [127]).

Nos vetores de características gerados a partir do conhecimento externo, geralmente podemos observar relações entre as características. No Em muitos casos, essas relações são hierárquicas ou podemos dizer que os recursos se complementam e carregam informação semântica semelhante. Essas relações hierárquicas podem ser facilmente recuperadas da ontologia ou esquema usado para publicar o LOD e podem ser usadas para executar melhor seleção de recursos.

Introduzimos uma abordagem [128] que explora hierarquias para seleção de recursos em combinação com métricas padrão, como *ganho de informação* ou *correlação.* A idéia central da abordagem é identificar recursos com relevância semelhante e selecione o valor mais recursos abstratos aceitáveis, ou seja, recursos dos níveis mais altos possíveis da hierarquia, sem perder o poder preditivo e, assim, encontrar uma negociação ideal ff entre o poder preditivo e a generalidade de uma característica em ou evitar excesso de ajuste. Para medir a semelhança de relevância entre dois nós, usamos a correlação padrão e a medida de ganho de informações. A abordagem funciona em duas etapas, ou seja, uma seleção inicial e uma etapa de poda adicional.

Jeong et al. [129] propõe a *TSEL* método usando uma hierarquia semântica de recursos com base nas relações do WordNet. O algoritmo apresentado tenta encontrar o mais representativo e o mais ff recursos efetivos do espaço completo de recursos. Para fazer isso, eles selecionam um representante característica de cada caminho na árvore, onde o caminho é o conjunto de nós entre cada nó folha e a raiz, com base no *lift* medir e usar χ^2 para selecionar o mais ff recursos efetivos do espaço reduzido de recursos.

Wang et al. [130] propõe uma *escalada de baixo para cima* algoritmo de busca para encontrar um subconjunto ideal de conceitos para representação de documentos. Para cada recurso no espaço inicial, eles usam um classificador kNN para detectar o *k* vizinhos mais próximos de cada instância no conjunto de dados de treinamento e use a pureza dessas instâncias como sinal pontuações para recursos.

Lu et al. [131] descrevem um *ganancioso de cima para baixo* estratégia de pesquisa para seleção de recursos em um espaço hierárquico de recursos. O algoritmo começa com a definição de todos os caminhos possíveis de cada nó folha para o nó raiz da hierarquia. archy. Os nós de cada caminho são classificados em ordem decrescente com base na taxa de ganho de informações dos nós. Em seguida, uma estratégia baseada em ganancioso é usada para remover as listas classificadas. Especificamente, ele remove iterativamente o primeiro elemento da lista e o adiciona à lista de recursos selecionados. Em seguida, remove todos os ascendentes e descendentes deste elemento na lista classificada. Portanto, a lista de recursos selecionados pode ser interpretada como uma mistura de conceitos de di ff diferentes níveis da hierarquia.

Ao criar recursos de várias fontes LOD, geralmente um único recurso semântico pode ser encontrado em várias fontes LOD representadas por di ff propriedades diferentes. Por exemplo, a área de um país na DBpedia é representada com *db: areaTotal,* e com *yago: hasArea* em YAGO. O problema de alinhar propriedades, bem como instâncias e classes, em ontologias é abordado por *correspondência de ontologia* técnicas [132]. Embora exista uma grande quantidade de trabalho na área de correspondência ontológica, a maioria das abordagens para gerar recursos a partir de Dados Abertos Vinculados não está abordando explicitamente esse problema. A extensão RapidMiner LOD oferece um operador para propriedades correspondentes extraídas de várias fontes LOD, que posteriormente são fundidas em um único recurso. O operador é baseado no algoritmo probabilístico para ontologia correspondente a PARIS [133]. Diferentemente da maioria dos outros sistemas, o PARIS é capaz de alinhar entidades e relações. Isso é feito através da inicialização de um alinhamento dos literais correspondentes e da propagação de evidências com base nas funcionalidades da relação. Em [104], mostramos que, por exemplo, o valor para a população de um país pode ser encontrado em 10 dias ff fontes diferentes na nuvem LOD, que usando o operador de correspondência e fusão de extensão RapidMiner LOD foram mescladas em um único recurso. Essa fusão pode fornecer um recurso que mitiga valores ausentes e erros únicos para fontes individuais, levando a apenas um recurso de alto valor.

Na mineração de padrões e na mineração de regras de associação, as ontologias de domínio geralmente são usadas para reduzir o espaço de recursos, a fim de obter padrões mais significativos e interessantes. Na abordagem proposta por Bellandi et al. [134] várias restrições específicas do domínio e definidas pelo usuário são usadas, isto é, restrições de remoção, usadas para filtrar itens desinteressantes e restrições de abstração que permitem a generalização de itens para conceitos de ontologia. Os dados são pré-processados primeiro de acordo com as restrições extraídas da ontologia e, em seguida, a etapa de mineração de dados ocorre. A aplicação das restrições de remoção exclui as informações nas quais o usuário não está interessado antes de aplicar a abordagem de mineração de dados. O Onto4AR é um algoritmo baseado em restrições para mineração de associação proposto por Antunes [135] e revisado posteriormente em [136], onde taxonômicos e não-taxonômicos

restrições são definidas sobre uma ontologia de item. Essa abordagem é interessante na maneira como a ontologia ff oferece um alto nível de expressão para as restrições, o que permite realizar a descoberta de conhecimento na melhor nível de abstração normal, sem a necessidade de entrada do usuário. Garcia et al. desenvolveu uma técnica chamada *Coesão do conhecimento [137, 138]* extrair regras de associação mais significativas. A métrica proposta é baseada na distância semântica, que mede a proximidade de dois itens. com base na ontologia, onde cada tipo de relação é ponderada di ff constantemente.

7.3 De outros

Zeman et al. [*139]* Apresente a ferramenta Ferda DataMiner, focada na etapa de transformação de dados. No Nessa abordagem, as ontologias são usadas para dois propósitos: construção de categorização adequada de atributos e identificação e exploração de atributos semanticamente relacionados. Os autores afirmam que ontologias podem ser eficientemente usadas para categorizar atributos como semântica de nível pode ser atribuída a valores individuais. Por exemplo, para pressão arterial, existem valores predefinidos que dividem o domínio de maneira significativa: digamos, pressão arterial acima *140/90* mm Hg é considerado hipertensão. Para o segundo objetivo, ontologias são usado para descobrir a relação entre os atributos, que pode ser explorada para organizar de maneira significativa os atributos de dados correspondentes na fase de transformação de dados.

Mesa *3* fornece uma visão geral dos aplicativos discutidos abordagens nesta seção. Pode-se observar que, nesta fase do processo de mineração de dados, muitas abordagens também exploram os links entre os conjuntos de dados LOD para identificar mais recursos. Por outro lado, os recursos são gerados com mais freqüência sem considerar o esquema dos dados, que é, na maioria dos casos, usado para pós-processamento dos recursos, por exemplo, para seleção de recursos. Da mesma forma, o raciocínio é pouco usado.

A maioria dos algoritmos e ferramentas de mineração de dados exige uma *proposicional* representação, isto é, vetores de características para instâncias. As abordagens típicas para a proposição proposicional são, por exemplo, adicionar todas as propriedades de tipos de dados numéricos como recursos numéricos ou adicionar todos os tipos diretos como recursos binários. Existem métodos não supervisionados e supervisionados, nos quais, para o último, o usuário especifica uma consulta para os recursos a serem gerados - esses são úteis se o usuário conhece o conjunto de dados LOD em mãos e / ou tem uma idéia de quais recursos podem ser valiosos. Embora esses métodos clássicos de proposicionalização crio humano características interpretáveis e, portanto, também são aplicáveis a *descritivo* mineração de dados, os métodos do kernel geralmente oferecem melhores *preditivo* resultados, mas ao preço de perder a interpretabilidade desses resultados. Um problema crucial ao criar recursos explícitos a partir do Linked Open Data é a escalabilidade e o número de recursos gerados. Como apenas poucas abordagens se concentram na identificação de recursos de alto valor já na etapa de geração, é claramente recomendável combinar a geração de recursos com a seleção de subconjuntos de recursos. As informações do esquema para as fontes LOD, como hierarquias de tipos, podem ser exploradas para redução do espaço de recursos. Existem alguns algoritmos que exploram o esquema, que geralmente oferecem uma troca melhor ff entre redução de espaço de recurso e desempenho preditivo do que abordagens independentes de esquema.

8. Mineração de Dados

Depois que os dados são selecionados, pré-processados e transformados na representação mais adequada, a próxima etapa é escolher a tarefa de mineração de dados e o algoritmo de mineração de dados apropriados. Dependendo dos objetivos do KDD e das etapas anteriores do processo, os usuários precisam decidir que tipo de mineração de dados usar, como classificação, regressão, clustering, sumarização, ou detecção externa. A compreensão do domínio ajudará a determinar que tipo de informação é necessária no processo KDD, o que facilita a decisão dos usuários. Existem duas categorias mais amplas de objetivos na mineração de dados: previsão e descrição. A previsão geralmente é chamada de mineração de dados supervisionada, que tenta prever os possíveis valores futuros ou desconhecidos dos elementos de dados. Por outro lado, a mineração de dados descritiva é chamada de mineração de dados não supervisionada, que busca descobrir padrões interpretáveis nos dados. Após a seleção da estratégia, o algoritmo de mineração de dados mais apropriado deve ser selecionado. Esta etapa inclui a seleção de métodos para pesquisar padrões nos dados e a decisão sobre modelos e parâmetros específicos dos métodos.

Tabela 3: Resumo das abordagens usadas na etapa de transformação.

Aproximação	Domínio		Ontologia				dados usados de ontologia	
	Problema		Complexidade	Raciocínio			LOD	Conjuntos de
[9 6]				não	sim não		LinkedMDB no	
	filmes	Mining			sim	Links		
	economia, filmes	de associação Data		não	sim	não		
[97]	mídia social geografia governo,	sistemas de recomendação classificação		não	sim	não Semântica	FactBook uma,	
101] [98]	filmes econômicos, publicações	correlações de de classificação de	LH	não	sim	sim LOD	DBpedia, World	/ /
[9 5] [100,	biologia, sociologia,	correlações, correlação			sim		DBpedia, LinkedMDB DBpedia YAGO	/ /
[104]	economia,	classificação de	HHHH	não	sim	sim sim	DBpedia, YAGO, LinkedGeoData b Brasileira vinculada, Eurostat c GeoNames, OMS d Floresta Amazônica Factbook DBpedia vinculada e OpenCycf, World	/ /

17

uma http://wifo5-03.informatik.uni-mannheim.de/factbook/

b http://linkedgeodata.org

c http://eurostat.linked-statistics.org/ e http://wifo5-03.informatik.unimannheim.de/eurostat/

d http://gho.aksw.org/

e ht tp: /en.openei.org/lod/

fhttp: //sw.opencyc.org/

Abordagem	Domínio		Ontologia		LOD		Conjuntos de dados usados	
	Problem	Mineração de dados	H	Raciocínio	Ligações		LOD	NDF-RTf
[56]	news	análise de sentimentos	Complexidade	não		não	DBpedia	Eu
[109]	música, filmes, livros		H	não			DBpedia, DBTropes uma Descasca Sessões b	Eu
[121, 122]	publicações, geologia	ligação detecção de erro de	H	não	não	semântica não		SWRC d
[123]	publicações bio-medicina,	valor, predição de link propriedade predição de classificação de classificação de texto		não	não	sim não	MUTAG, ENZIMAS, Geological Survey c Survey, British British Geological Se- maluco Rede de dados personalizado Conferência Corpus e Conjunto	Eu UMLSg
[1 1 2 9]	notícia		HH	não	não	não	WordNet	Eu
[1 1 3 0 0]	biomedicina	classificação de texto	HH	não	não	sim sim sim não	///	UMLS
[1 1 3 1 1]	farmacologia	clássica		não	não	não sim		
[134]	comércio	regra aprendizagem	H	não	não	não		
[135, 136]	medicina	regra de associação	H	não	não	não	//	ontologia personalizada
[139]	filmes de	mineração de associação	H	sim			Ontologia	produtos ontologia
[137, 138]	relatórios de acidentes	mineração de texto, aprendizado de regras	H	não	não	não	Eu	ontologia personalizada

uma http://skipforward.opendfki.de/wiki/DBTropes

b http://dbtune.org/bbc/peel/

c http://www.bgs.ac.uk/opengeoscience/

d http://ontoware.org/swrc/

e http://data.semanticweb.org/ f http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/

g http://www.nlm.nih.gov/research/umls/

Depois que o método e o algoritmo de mineração de dados são selecionados selecionado, a mineração de dados ocorre.

Até onde sabemos, raramente existem abordagens na literatura que incorporem dados publicados como Dados Abertos Vinculados nos próprios algoritmos de mineração de dados. No entanto, muitas abordagens somos nós ontologias para o processo de mineração de dados, não apenas para oferecer suporte ao usuário no estágio de seleção dos métodos de mineração de dados, mas também para orientá-lo por todo o processo de descoberta de conhecimento.

8.1 Abordagens independentes de domínio

Embora ainda não exista uma ontologia de mineração de dados estabelecida universalmente, existem várias ontologias de mineração de dados atualmente em desenvolvimento, como a Ontologia de descoberta de conhecimento (KD) [140], a KD-DONTO Ontologia [141], a ontologia de fluxo de trabalho de mineração de dados (DMWF) [28] [142], a ontologia de otimização de mineração de dados (DMOP) [29] por Hilario [143, 144], On-toDM [30] [145, 146] e seus módulos de sub ontologia OntoDT [31], OntoDM-core [32] [147] e OntoDM-KDD [33] [148]

Uma visão geral dos assistentes inteligentes existentes para análise de dados que usam ontologias é apresentada em [149]. Nesta pesquisa, todas as abordagens são categorizadas por vários critérios. Primeiro, quais tipos de suporte os assistentes inteligentes o ff ao analista de dados. Segundo, ele examina os tipos de conhecimento de base em que os IDAs se baseiam para fornecer o suporte. Por fim, realiza uma comparação completa dos IDAs à luz das dimensões definidas e a identificação de limitações e recursos ausentes.

Uma das primeiras abordagens, *CHAMALOTE,* foi proposto por Suyama et al. [150], que usa duas ontologias leves de entidades de aprendizado de máquina para apoiar a composição automática de sistemas de aprendizado indutivo.

Entre os primeiros protótipos está o *Assistente de descoberta inteligente* proposto por Bernstein et al. [151], que fornece aos usuários enumerações sistemáticas de sequências válidas de operadores de mineração de dados. A ferramenta é capaz de determinar as características dos dados e dos dados desejados. resultado da mineração e usa uma ontologia para procurar e enumerar os processos KDD válidos para produzir o resultado desejado a partir dos dados fornecidos. Além disso, a ferramenta auxilia o usuário na seleção dos processos a serem executados, classificação dos processos de acordo com o que é importante para o usuário. Uma ontologia leve é usada que contém apenas uma hierarquia de operadores de mineração de dados divididos em três classes principais: operadores de pré-processamento, algoritmos de indução e operadores de pós-processamento. Muitas abordagens estão usando tecnologias da Web Semântica para ajudar o usuário a criar fluxos de trabalho complexos de mineração de dados. ˙ Zákováet al. [152, 140] propuseram uma abordagem para geração de fluxo de trabalho semiautomático que requer apenas a entrada do usuário e a saída desejada do usuário para gerar fluxos de trabalho completos de mineração de dados. Para implementar a abordagem, os autores desenvolveram a ontologia de descoberta de conhecimento, que fornece uma representação formal de tipos de conhecimento e algoritmos de mineração de dados. Segundo, é implementado um algoritmo de planejamento que monta fluxos de trabalho com base nas descrições de tarefas de planejamento extraídas da ontologia de descoberta de conhecimento e nos requisitos de tarefa de entrada e saída de usuários. Nesse ambiente semiautomático, o usuário não precisa conhecer as inúmeras propriedades da ampla gama de algoritmos relevantes de mineração de dados. Em seus trabalhos posteriores, a metodologia é implementada no ambiente Orange4WS para mineração de dados orientada a serviços [153, 154].

Diamantini et al. [155] introduzem uma estrutura baseada em semântica, orientada a serviços, para compartilhamento e reutilização de ferramentas, fornecendo suporte avançado para o enriquecimento semântico através da anotação semântica de ferramentas KDD, implantação de ferramentas como serviços da Web e descoberta e uso de tais serviços. Para suportar o sistema, é utilizada uma ontologia chamada KDDONTO [141], que representa uma ontologia formal que descreve o domínio dos algoritmos KDD. A ontologia fornece as informações exigidas pelo compositor do KDD para ajudá-lo a escolher os algoritmos adequados para alcançar seu objetivo a partir dos dados disponíveis e para compor corretamente os algoritmos para formar um processo válido [156].

Kietz et al. [142, 157] apresentaram uma ontologia de mineração de dados para o planejamento do fluxo de trabalho, capaz de ff organizar de maneira criativa centenas de operadores, que é a base para verificar a correção dos fluxos de trabalho do KDD e um componente de planejamento da Hierarchical Task Network capaz de enumerar efetivamente os fluxos de trabalho úteis do KDD. Isso inclui os objetos manipulados pelo sistema, os metadados necessários, os operadores utilizados e uma descrição do objetivo. O gerador de fluxo de trabalho está fortemente associado a um meta-minerador, cuja função é classificar os fluxos de trabalho e é baseado na ontologia do DMOP. Além disso, os autores introduziram a ferramenta eProPlan [158], que representa um ambiente baseado em ontologia para o planejamento de fluxos de trabalho KDD. Posteriormente, a ferramenta é usada para anotar semanticamente todos os operadores nos minimos dados bem conhecidos.

[28] http://www.e-lico.eu/dmwf.html
[29] http://www.e-lico.eu/DMOP.html
[30] http://www.ontodm.com/doku.php
[31] http://www.ontodm.com/doku.php?id=ontodt
[32] http://www.ontodm.com/doku.php?id=ontodm-core
[33] http://www.ontodm.com/doku.php?id=ontodm-kdd

ferramenta RapidMiner. Isso permite que os usuários construam mais facilmente e com mais rapidez ffi fluxos de trabalho suficientes do KDD no RapidMiner [159]. Sua avaliação mostrou que o uso de

As tecnologias da Web semântica podem acelerar o tempo de design do fluxo de trabalho em até 80%. Isso é obtido por sugestão automática de possíveis operações em todas as fases do processo KDD.

Além disso, Hilario et al. [143] apresentam os dados ontologia de otimização de mineração, que fornece um estrutura conceitual para analisar tarefas de mineração de dados, algoritmos, modelos, conjuntos de dados, fluxos de trabalho e métricas de desempenho, bem como seus relacionamentos. Um dos principais objetivos da ontologia é apoiar a meta-mineração de experimentos completos de mineração de dados para extrair padrões de fluxo de trabalho [144]. Além disso, os autores desenvolveram uma base de conhecimento definindo instâncias da ontologia do DMOP. A ontologia do DMOP não se baseia em nenhuma ontologia de nível superior e usa uma grande conjunto de relações personalizadas para fins especiais. Panov et al. [145, 146] propõem uma ontologia de mineração de dados *OntoDM* inclui definições formais de entidades básicas de mineração de dados, como *tipo de dados* e *conjunto de dados, tarefa de mineração de dados* e *algoritmo de mineração de dados,* qual baseia-se na proposta de um quadro geral para mineração de dados proposto por D˘zeroski [160]. A ontologia é uma das primeiras ontologias pesadas / pesadas para mineração de dados. Para permitir a representação de dados estruturados de mineração, os autores desenvolveram um módulo de tecnologia, denominado OntoDT, por representar o conhecimento sobre tipos de dados. Para representar entidades de mineração de dados principais e ser suficientemente geral para representar a mineração de dados estruturados, os autores introduziram o segundo módulo de ontologia chamado OntoDM-core [147].

O terceiro e final módulo da ontologia é o OntoDM-KDD, que é usado para representar investigações de mineração de dados [148].

Gabriel et al. [161] propõem o uso de informações semânticas sobre os atributos contidos em um conjunto de dados para regras de classificação de aprendizado potencialmente melhor compreensíveis. Eles usam *coerência semântica,* isto é, a proximidade semântica dos atributos usados em uma regra, como critério alvo para aumentar a compreensibilidade de uma regra. Em seu artigo, eles mostram que usar o WordNet como fonte de conhecimento e adaptando um algoritmo padrão de aprendizado de regras de separação e conquista [162], eles podem aumentar significativamente a coerência semântica em uma regra sem diminuir a precisão da classificação.

8.2 Abordagens específicas de domínio

Santos et al. [163] descreve uma pesquisa de uma abordagem ontológica para alavancar o conteúdo semântico de ontologias para melhorar a descoberta de conhecimento em bancos de dados.

Os autores dividem o processo KDD em três operações principais e tentam dar suporte a cada um deles usando ontologias. Primeiro, nas fases de entendimento e preparação de dados, as ontologias podem facilitar a integração de dados heterogêneos e orientar a seleção de dados relevantes a serem minerados, em relação aos objetivos do domínio. Segundo, durante a fase de modelagem, o conhecimento do domínio permite a especificação de restrições para orientar os algoritmos de mineração de dados, restringindo o espaço de pesquisa. Finalmente, na fase de interpretação e avaliação, o conhecimento do domínio ajuda os especialistas a validar e classificar os padrões extraídos.

Ding et al. [164, 165] introduzem outra estrutura baseada em ontologia para incorporar o conhecimento do domínio no processo de mineração de dados. A estrutura é capaz de suportar o processo de mineração de dados em várias etapas do pipeline: exploração de dados, definição de metas de mineração, seleção de dados, pré-processamento de dados e seleção de recursos, transformação de dados, seleção de parâmetros do algoritmo de mineração de dados e avaliação de resultados de mineração de dados.

- Ce˘spivováet al. [166] mostraram como as ontologias e o conhecimento prévio podem ajudar em cada etapa do processo KDD. Eles executam a mineração de associação usando a ferramenta LISp-Miner, sobre o conjunto de dados médicos STULONG. Para dar suporte à mineração de dados, eles usam ontologias UMLS [34] para mapear os dados para conceitos semânticos. O mapeamento ajudou os autores a entender melhor o domínio. Eles foram capazes de identificar e filtrar atributos redundantes e desnecessários e agrupar atributos relacionados semanticamente, analisando os relacionamentos dentro da ontologia. Além disso, eles usam ontologias para interpretar e dar uma melhor explicação dos resultados da mineração de dados.

A Tabela 4 fornece uma visão geral das abordagens discutidas nesta seção. Isso mostra que, embora os conjuntos de dados baseados em dados abertos vinculados tenham um papel menor nesta etapa, ontologias e raciocínios pesados são usados com bastante frequência. Além disso, a maioria das ontologias é independente do domínio, enquanto os desenvolvimentos específicos do domínio nesta etapa são bastante raros.

[34] http://www.nlm.nih.gov/research/umls/.

Ontologias são frequentemente usadas para apoiar o usuário na criação de um processo de mineração de dados adequado. Eles podem ser usados para representar fontes de dados, algoritmos etc. nos dados processos de mineração e auxiliar o usuário na construção de processos razoáveis de mineração de dados, por exemplo, garantindo que um algoritmo escolhido seja capaz de manipular os dados fornecidos. Por exemplo, a plataforma *RapidMiner* internamente usa descrições semânticas de operadores para ajudar o usuário a evitar erros, por exemplo, ao combinar operadores de pré-processamento de dados e aprendizado de máquina. Aqui, o raciocínio não apenas verifica a validade de um processo, mas também propõe soluções para corrigir um processo inválido.

Abordagens que usam informações semânticas diretamente em um algoritmo para influenciar o resultado desse algoritmo são bastante raras. Existem algumas instruções para usar o conhecimento semântico de fundo em algoritmos de mineração de dados, por exemplo, para encontrar padrões mais fáceis de consumir por um usuário final.

9. Interpretação

Após a etapa de mineração de dados ter sido aplicada, esperamos descobrir alguns padrões ocultos dos dados. Para serem interpretados e compreendidos, esses padrões geralmente exigem o uso de algum conhecimento prévio, o que não é sempre fácil de encontrar. Na maioria dos contextos do mundo real, o fornecimento de conhecimento prévio é comprometido com os especialistas, cujo trabalho é analisar os resultados de um processo de mineração de dados, dar-lhes um significado e refiná-los. A interpretação acaba sendo uma intensa e processo demorado, em que parte do conhecimento pode permanecer não revelada ou inexplicável.

O Explain-a-LOD [106] é uma das primeiras abordagens na literatura para gerar automaticamente hipóteses para explicar estatísticas usando LOD. A ferramenta usa O FeGeLOD (descrito na Seção 7.1) para aprimorar conjuntos de dados estatísticos com informações de base da DBpedia e usa análise de correlação e aprendizado de regras para produzir hipóteses que são apresentadas ao usuário. A ferramenta foi usada posteriormente para encontrar e explicar hipóteses. pela qualidade de vida em cidades do mundo [107], e taxas de desemprego na França [108], entre outros. Por exemplo, em [107] a ferramenta foi capaz de descobrir automaticamente hipóteses como "Cidades onde muitas coisas acontecem têm uma qualidade de vida alta". e "europeu capitais da cultura têm uma alta qualidade de vida. ". Enquanto em [108] onde os dados da DBpedia, Euro-stat e LinkedGeoData foram usados, a ferramenta descobriu hipóteses como “Regiões na França que têm alta consumo de energia tem baixa taxa de desemprego ”. e "As regiões francesas que estão fora da Europa, as ilhas africanas francesas e as ilhas francesas no Oceano Índico têm uma alta taxa de desemprego". Além disso, a abordagem é estendida em [167], que permite análise de correlação automática e visualização de dados estatísticos em mapas usando Dados Abertos Vinculados. A ferramenta permite que os usuários importem dados estatísticos de planilhas locais ou dat-acubes RDF, realizem análises de correlação e visualizem automaticamente as descobertas em um mapa.

Dados abertos vinculados não podem apenas adicionar informações categóricas, o que permite uma exploração mais fácil dos resultados, mas também pistas visuais adicionais. Em [108, 167], mostramos que dados de polígonos para entidades geográficas publicadas como LOD, como GADM [35] pode ser explorado para criar uma visualização baseada em mapa dos resultados da mineração de dados. Além disso, o GADM o ff modelam dados de entidades geográficas em ff administrativos diferentes, que podem ser acessados através da DBpedia, seguindo *owl: sameAs* links.

d'Aquin et al. [168] propuseram um método que explora informações externas disponíveis como LOD para apoiar a interpretação dos resultados da mineração de dados, através da construção automática de uma estrutura de exploração de navegação nos resultados de um tipo específico de mineração de dados, neste caso mineração de padrão sequencial, ferramenta com base nas dimensões de dados escolhidas pelo analista. Para fazer isso, os autores primeiro representam os resultados da mineração de dados de maneira compatível com uma representação de LOD e os vinculam às fontes existentes de LOD. Em seguida, o analista pode explorar facilmente os resultados extraídos com dimensão adicional. Além disso, para organizar os resultados enriquecidos em uma hierarquia, os autores usam a análise formal de conceitos para construir uma estrutura conceitual. Isso pode permitir que o analista faça uma pesquisa detalhada dos detalhes de um subconjunto dos padrões e veja como eles se relacionam com os dados originais. Uma abordagem semelhante é usada em [169] para interpretar padrões sequenciais nos dados do paciente. Dados vinculados são usados para apoiar a interpretação de padrões extraídos das trajetórias de atendimento ao paciente. Os dados vinculados expostos pelo sistema BioPortal são usados para criar uma estrutura de navegação dentro dos padrões obtidos da mineração sequencial de padrões. A abordagem fornece uma maneira flexível de explorar dados sobre trajetórias de diagnósticos e tratamentos de acordo com di ff diferentes classificações médicas. Tiddi [170] propõe uma abordagem em três etapas para interpretar os resultados da mineração de dados, ou seja, agrupamentos, regras de associação e padrões de sequência. Na primeira etapa, informações adicionais para os resultados dos padrões são extraídas

[35] http: // gadm org / geovocab /

Tabela 4: Resumo das abordagens usadas na etapa de mineração de dados.

Abordagem	de classificação		Ontologia		LOD		dados usados do WordNet	
		Mineração de dados	H	Raciocínio	links	semântica	LOD	
[151]			–	não	não	não		personalizada
[152, 140, 153, 154]	genômica, engenharia,	classificação redes neurais regressão, regressão,		sim	não	não	*Eu*	KD
[155, 141, 156] [142, 157, 158, 159]	Comércio de conhecimento	classificação de agrupamento, classificação,	HH – –	sim Sim	não não	não não	//	KDDONTO DMO, DMWF, DMOP DMO, DMOP
[143, 144]	medicamento	classificação	Complexidade	sim	não	não	//	Ontologia ontologia OntoDM, OntoDT,
[145, 146, 148, 147]	química, farmacologia		– –	sim	não	não		OntoDMcore, OntoDMKDD
[161]	*Èu*	domínio de	eu	não	não	não	*Èu*	Conjuntos de
[163]	*Eu*	*Eu*	ΣHH	não	sem	sem	*Eu*	ontologia personalizada
[164, 165]	//	//	LL	não	não	não	//	ontologia personalizada

da nuvem LOD. Usando programação lógica indutiva, novas hipóteses são geradas a partir dos resultados da mineração de padrões e do conhecimento extraído do LOD.

Na última etapa, as hipóteses são avaliadas usando estratégias de classificação, como Precisão Relativa Ponderada e Medida F de Recuperação de Informações. A mesma abordagem foi usada em [171] para explicar por que grupos de livros, ob
provenientes de um processo de agrupamento, foram emprestados pelos mesmos alunos. A análise foi feita no conjunto de dados de uso de livros de Huddersfields 36, usando a bibliografia nacional britânica 37. e Biblioteca do Congresso 38. como conjuntos de dados LOD. Os experimentos levam a interessantes
uma hipótese para explicar os agrupamentos, por exemplo, “livros emprestados por estudantes da Music Technologies estão agrupados porque falam sobre música ”.

O trabalho foi continuado em [172, 173], introduzindo *Dedalo,* estrutura que atravessa dinamicamente o Linked
Dados para encontrar semelhanças que formam explicações opções para itens de um cluster. O Dedalo usa uma abordagem iterativa para percorrer os gráficos LOD, onde as raízes são os itens dos clusters. A suposição subjacente é que os itens que pertencem a um cluster compartilham mais
caminhos comuns no gráfico LOD, que os itens externos o cluster. Os autores conseguiram extrair explicações interessantes e representativas para os clusters, no entanto, o número de regras atômicas resultantes é bastante grande e precisa ser agregado em uma etapa de pós-processamento. o
Uma estratégia típica para superar esses problemas é fornecida fornecer os padrões a especialistas humanos, cujo papel consiste em analisar os resultados, descobrir os interessantes e, ao mesmo tempo, explicar, remover ou refinar os que não estão claros. Para lidar com um profissional tão árduo e demorado
os autores em seu próximo trabalho [174] propuseram uma abordagem que está usando o modelo de rede neural para prever se duas regras, se combinadas, podem levar à criação de uma nova regra aprimorada (ou seja, uma nova regra, com uma melhor medida). A abordagem foi aplicada no
domínio da educação e publicações.

Lavra˜c et al. fizeram um trabalho de pesquisa notável no campo da descoberta de subgrupos semânticos. A tarefa de descoberta de subgrupos é definida da seguinte forma: “Dada uma população de indivíduos e uma propriedade desses indivíduos
nos quais estamos interessados, encontre subgrupos populacionais que são estatisticamente mais interessantes, por exemplo, são as maiores possíveis e têm as características estatísticas (de distribuição) mais incomuns em relação à propriedade de interesse ”[175]. Os autores definem semântica

descoberta de subgrupos táticos como parte de mintologias de dados semânticos e dados empíricos anotados por domínio on-

definido como: “Dado um conjunto de domínios on-line
termos de tecnologia, pode-se encontrar uma hipótese (um modelo preditivo ou um conjunto de padrões descritivos), expressa por termos de ontologia de domínio, explicando os dados empíricos dados ”. A descoberta de subgrupos semânticos foi implementada pela primeira vez no sistema SEGS [176]. O SEGS usa como dados de conhecimento de segundo plano de três repositórios de dados biológicos anotados semanticamente e publicamente anotados. Com base no conhecimento prévio, ele formula automaticamente hipóteses biológicas: regras que definem grupos de ff genes expressamente identificados. Finalmente, estima a relevância (ou significância) de

o automaticamente
hipóteses formuladas sobre dados experimentais de raios-micro. O sistema foi estendido no sistema SegMine, que permite a análise exploratória de dados de microarranjos, realizada através da descoberta de subgrupos semânticos por SEGS [177], seguida pela descoberta e visualização de links por Biomine [178],

a bioinformática anotada integrada
recurso informativo de dados interligados. O sistema SEGS foi posteriormente estendido a dois sistemas gerais de descoberta de subgrupos semânticos, SDM-SEGS e SDMAleph [179, 180, 181]. Finalmente, os autores introduziram o sistema Hedwig [182], que supera algumas das limitações dos sistemas anteriores. As conclusões desta série de trabalhos foram concluídas em [183, 184]. Um problema semelhante é abordado em [185]. Em vez de identificar subgrupos, buscamos encontrar características especiais de uma determinada instância, dado um conjunto de contraste. Para esse fim, os dados sobre a instância em questão, bem como seu conjunto de contraste, são recuperados do DBpedia. A detecção de discrepância de atributos, que calcula pontuações discrepantes para valores de atributo único [186], é explorada para identificar os valores de atributo da instância que são significativamente diferentes. ff diferente dos das outras instâncias. Muitas abordagens estão usando ontologias para pós-mineração de padrões e interpretação dos resultados. O conhecimento do domínio e a especificação de metadados armazenados na ontologia são usados no estágio de interpretação para remover e filtrar os padrões descobertos. Ontologias são comumente usadas para filtrar padrões redundantes e padrões muito específicos sem perder informações semânticas. Uma das primeiras abordagens que utiliza ontologias de domínio para esse fim é o trabalho de Srikant [187], que introduziu o conceito de regras de associação generalizadas. Da mesma forma, Zhou et al. [188] introduzem o conceito de criação. Levantar é a operação de generalizar regras de mineração de dados para aumentar o suporte e manter a confiança alta o suficiente. Isso é feito com a generalização das entidades, elevando-as para um nível especificado na ontologia. Os autores utilizam uma ontologia que consiste em dois

36. http://library.hud.ac.uk/data/usagedata/readme.html
37. http://bnb.data.bl.uk/
38. http://id.loc.gov/

taxonomias, uma das quais descreve di ff classificações de clientes diferentes, enquanto a outra contém uma grande hierarquia, baseada no Yahoo, que contém interesses. Nas experiências, os valores de suporte dos conjuntos de regras foram bastante aumentado, até 40 vezes. O GART é uma abordagem muito semelhante [189], que usa várias taxonomias sobre atributos para generalizar iterativamente regras e, em seguida, remover regras redundantes a cada etapa. Os experimentos foram realizados utilizando um banco de dados de vendas de uma supermarca brasileira ket. Os experimentos mostram taxas de redução dos conjuntos de regras de associação variando de 14,61% a 50,11%. Marninica et al. [190] apresenta uma estrutura interativa de pós-processamento, chamada ARIPSO (pós-processamento interativo de regras de associação usando Schemas e Ontolo- gies). A estrutura auxilia o usuário durante toda a tarefa de análise a remover e filtrar as regras descobertas. O sistema permite formalizar o conhecimento e os objetivos do usuário, que são usados posteriormente para aplicar iterativamente um conjunto de filtros sobre regras extraídas, a fim de extrair interesses. regras: filtro de restrição de aprimoramento mínimo, filtro de relação ao item, filtros / remoção do esquema de regras. Os experimentos foram realizados nos dados do Nantes Habitat [39], lidar com a satisfação dos clientes em relação à acomodação, para a qual foi estabelecida uma ontologia correspondente desenvolvido pelos autores. Os resultados mostraram que o número de regras pode ser reduzido significativamente ao usar o esquema, resultando em regras mais descritivas. Huang et al. [191] use LOD para interpretar os resultados da mineração de texto. A abordagem começa com a extração de entidades laços e relações semânticas entre eles a partir de documentos de texto, resultando em gráficos semânticos. Em seguida, um algoritmo de descoberta frequente de sub-gráficos é aplicado nos gráficos de texto para encontrar padrões frequentes. Para interpretar os subgráficos descobertos, um algoritmo é proposto para tra- verso Gráficos de dados vinculados para relações usadas para anotar os vértices e as arestas dos subgráficos frequentes. A abordagem é aplicada em um conjunto de dados militar, em que o DBpedia é usado como um conjunto de dados LOD. Outra abordagem que usa ontologias em regras de regras é a ferramenta 4ft-Miner [192]. A ferramenta é usada em quatro etapas do processo KDD: entendimento de dados, mineração de dados, interpretação e disseminação de resultados. Na etapa de entendimento dos dados, foi realizado um mapeamento de dados para ontologia, que resultou na descoberta de atributos dundantes. No estágio de mineração de dados do processo KDD, a ontologia foi usada para decompor a tarefa de mineração de dados em tarefas mais específicas, que podem ser executadas mais rapidamente, resultando em resultados mais homogêneos e, portanto, facilmente interpretáveis. Na etapa de interpretação, os dados Os mapeamentos de ontologia são usados para corresponder a alguns dos associações descobertas às relações semânticas correspondentes ou suas cadeias mais complexas da ontologia, que podem ser consideradas como explicação potencial das associações descobertas. A abordagem foi usada para interpretar associações em aplicações de clima médico e social. No domínio médico, o conjunto de dados STU-LONG [40]. é usado, que contém dados de risco cardiovascular. Como uma ontologia é usada a ontologia UMLS. Usando a abordagem, os autores foram capazes de descobrir hipóteses como “Pacientes que não são fisicamente ativos no trabalho, nem após o trabalho (Antecedente), terão maior pressão arterial (Sucessente)” e “Aumento do tabagismo leva ao aumento da cardio- doenças cardiovasculares ”.

A Tabela 9 fornece uma visão geral das abordagens discutidas nesta seção. Observamos que, nesta etapa, o raciocínio não desempenha papel crucial. Os conjuntos de dados explorados são bastante mistos, conjuntos de dados de uso geral, como o DBpedia, são frequentemente usados, mas também conjuntos de dados altamente específicos podem ser explorados. Aproximadamente metade das abordagens também faça uso dos interlinks entre esses conjuntos de dados.

Os dados semânticos da Web podem ajudar na interpretação dos padrões encontrados, principalmente para tarefas descritivas. Geralmente, eles incluem subgrupos ou clusters encontrados ou modelos de regras que são usados para descrever um conjunto de dados.

As informações usadas nos conjuntos de dados e / ou ontologias do LOD podem ajudar a analisar melhor essas descobertas, por exemplo, explicando os recursos típicos das instâncias em um subgrupo ou cluster, assim, eles podem explicar o agrupamento escolhido por um algoritmo de mineração de dados. Além disso, as regras podem ser mais refinadas e / ou generalizadas, o que melhora sua interpretabilidade.

10. Exemplo de Caso de Uso

Os sistemas de recomendação mudaram a maneira como as pessoas encontram e compram produtos e serviços. À medida que a Web cresce ao longo do tempo e o número de produtos e serviços, os sistemas de recomendação representam um método poderoso para os usuários filtrarem essas informações e espaço de produtos. Com a introdução dos sistemas de recomendação de dados abertos vinculados, estão surgindo uma área de pesquisa que usa extensivamente dados abertos vinculados como conhecimento de base para extrair recursos úteis de mineração de dados que poderiam melhorar os resultados das recomendações. Foi demonstrado que os Dados Abertos Vinculados podem melhorar os sistemas de recomendação para uma melhor compreensão e representação das preferências do usuário, recursos de itens e

[39] http: // www nantes-habitat fr /

[40] http: // euromise vse cz / stulong-pt /

Tabela 5: Resumo das abordagens utilizadas na etapa de interpretação.

Aproximação	Domínio		Ontologia				dados usados de ontologia	
	Problema		Complexidade	Raciocínio			LOD	Conjuntos de
	sociologia, economia	dados			não		DBpedia yes LinkedGeoData, GADM	
[107, 108, 167] [106]	Estatisticas	padrão mineração de		não	sim	Links	DBpedia, Eurostat,	
		padrão mineração padrão mineração		não	sim sim	Semântica LOD	Catálogo de cursos da Open Universitys *uma*, ICD10, CCAM Bio Ontologia *b*	/
171] [168, 169]	livros, publicações de publicação, medicina	associação regras		não	sim	não sim sim	Biblioteca do Congresso *d*	/ /
[172, 173, 174] [170,	educação, estudantes	clustering clustering,	HHHHH	não	sim	não	DBpedia, UIS *e* Britânica do Congresso *c* Bibliografia nacional britânica, Biblioteca do Bibliografia Nacional	/ /

uma http://data.open.ac.uk

b http://sparql.bioontology.org/sparql/

c http://bnb.data.bl.uk/

d http://id.loc.gov/

e http://uis.270a.info/.html

Aproximação	militar	Mineração de dados	Ontologia		LOD		usados da ontologia personalizada	
					Links		LOD	Ontologia
[176, 177, 178, 179, 180, 181, 184]	biomedicina	aprendizagem, descoberta de subgrupos subgrupo de detecção de		não	não	não	*Eu*	[b] Entrez
[185]	*Eu*	regra de descoberta de	HH	não	sim não	não	DBpedia	//
[182]	finança comércio sociologia		HH	não raciocínio		não semântica	GeoNames	produtos GO [uma], KEGG
[1 1 8 8]	associação	aprendizagem de regras	H	não	não sim	não	*Eu*	interesse em- tecnologia (do Yahoo)
[1 1 8 9]	comércio	aprendizagem de regras	H	não	não	não	*Eu*	taxonomia de
[1 1 9 0 0]		descoberta de subgrupo de	Complexidade	não não	não	não sim	*Eu*	conjuntos de dados
[1 1 9 1 1]	problema da	aprendizado de regras					DBpedia	*Eu*
[1 1 9 2]	medicina, sociologia	domínio de mineração de	HH	sim	não	não		UMLS, Ontologia do clima social

[uma] http://geneontology.org/

[b] http://www.genome.jp/kegg/

sinais contextuais com os quais lidam. O LOD foi usado em técnicas baseadas em conteúdo, colaborativas e híbridas, em várias tarefas de recomendação, como previsão de classificação, Recomendações Top-N, recomendação entre domínios e diversidade nas recomendações baseadas em conteúdo. Portanto, nesta seção, mostramos um exemplo de caso de processo de descoberta de conhecimento habilitado para LOD no domínio de sistemas de recomendação. Através desta Como exemplo, descreveremos cada etapa do processo KDD ativado para LOD, ou seja, vincular os dados locais ao conjunto de dados LOD, combinando dados de vários conjuntos de dados LOD, transformação dos dados, construção de sistema de recomendação e interpretação dos resultados. Neste exemplo, usaremos o conjunto de dados usado no Desafio de sistemas de recomendação recomendados vinculados a dados abertos na ESWC 2014 [193].

10.1 Vinculando dados locais ao LOD

A primeira etapa do pipeline KDD habilitado para LOD é vincular os dados aos conceitos correspondentes de LOD do conjunto de dados LOD escolhido (consulte a seção 5) O conjunto de dados inicial **continha uma lista de livros recuperados do conjunto de dados LibraryThing [41] juntamente** com classificações de usuários para livros. Para poder criar recomendadores habilitados para LOD, os conjuntos de dados foram vinculados à DBpedia. Para isso, o rótulo e o ano de produção dos livros é usado para encontrar a entidade do livro correspondente na DBpedia, usando a seguinte consulta SPARQL [194]:

SELECIONE DISTINTA? Filme? Etiqueta? Ano ONDE {? Filme rdf: tipo dbpedia-owl: Film.
? movie rdfs: label? label. dcterms do filme: sujeito? cat. ? cat rdfs: label? ano.

FILTRO langMatches (lang (etiqueta?), "EN"). Regex FILTER (? Ano, "^ [0-9] {4} filme", "i")
}
Etiqueta ORDENAR POR?

Isso resulta em um conjunto de dados de livros com os URIs da DBpedia correspondentes e classificações do usuário. Vemos aqui que, em vez de uma ferramenta de uso geral, usamos uma ferramenta artesanal regra de ligação que explora uma certa quantidade de conhecimento de domínio **não-formalizado (por exemplo, pode haver ff filmes diferentes com o mesmo título,** mas somos capazes de diferenciá-los pelo ano de produção). Conforme explicado na seção 5, essa é uma estratégia comum para o conhecimento estruturado. fontes de ponta, como bancos de dados relacionais.

[41] **http://www.macle.nl/tud/LT**

10.2 Combinando vários conjuntos de dados LOD

A segunda etapa é explorar os links iniciais para extrair dados adicionais de outros conjuntos de dados LOD que podem ser úteis para a tarefa especificada (consulte a seção 6). Além do DB-pedia, existem vários outros conjuntos de dados na nuvem LOD que contêm informações sobre livros. Para extrair os URIs de entidade correspondentes dos outros conjuntos de dados, podemos **seguir os *owl: sameAs* links na DBpedia. Por exemplo, podemos extrair os URIs** das entidades correspondentes no YAGO e no Freebase. Além disso, podemos encontrar os URIs correspondentes a outros conjuntos de dados para os quais um

***owl: sameAs* não existe, como dbTropes, usando o título do livro e o ano de** produção. Em [110], usamos o ISBN e o título do livro para vincular os livros **ao conjunto de dados RDF Book Mashup [42], que fornece a pontuação média** atribuída a um livro na Amazon.

Na terceira etapa do pipeline, os dados coletados de di ff fontes diferentes precisam ser consolidadas em um conjunto de dados limpo. **No entanto, ao combinar dados de di ff fontes diferentes de LOD, essas usualmente usam ff esquemas diferentes. Por exemplo, o autor do livro no** DBpe-dia está listado sob o *dbpprop: autor* enquanto na YAGO as **mesmas informações estão sob o *criada* propriedade. Para usar esses dados mais e ff efetivamente, esses atributos podem ser mesclados em** um aplicando a correspondência de esquema. Por exemplo, em [104] **para esse fim, usamos o *PARIS* abordagem de ontologia. O conjunto de** dados resultante conterá informações abrangentes e de alta qualidade sobre os livros.

Conforme discutido na seção 6, isso mostra como os dados da Web Semântica podem ajudar a criar dados mais valiosos, por exemplo, fundindo informações semelhantes de várias fontes para aumentar a cobertura e reduzir a redundância de atributos no conjunto de dados.

10.3 Construindo um sistema de recomendação baseado em LOD

Na quarta etapa, os dados do gráfico precisam ser transformados para a forma proposicional, para que possam ser usados em um sistema de recomendação padrão (consulte a seção 7). Para esse propósito em [110], usamos a extensão RapidMiner LOD. Nesta abordagem, desenvolvemos um sistema híbrido de recomendação baseado em conteúdo de várias estratégias. Essa abordagem se baseia no treinamento de recomendações individuais de base e no uso de índices de popularidade global como recomendadores genéricos. Os resultados de recomendações individuais são combinados usando regressão de empilhamento e agregação de classificação.

[42] **http://wifo5-03.informatik.uni-mannheim.de/ bizer / bookmashup /**

Para criar os recomendadores baseados em conteúdo, usamos os seguintes conjuntos de recursos para descrever um livro
recuperados do DBpedia e do conjunto de dados RDF Book Mashup:

- Todos *tipos diretos*, ou seja, rdf: tipo, de um livro [43]
- Todos *categorias de um livro*
- Todos *categorias de um livro, incluindo categorias mais amplas* [44]
- Todos *categorias do (s) autor (es) de um livro*
- Todos *categorias do (s) autor (es) de um livro e de todos os outros livros* pelos autores do livro
- Todos *gêneros de um livro* e de todos os outros livros dos autores do livro
- Todos *autores que influenciaram ou foram influenciados* de autores do livro
- Um conjunto de palavras criadas a partir do *resumo* do livro na DBpedia. Esse pacote de palavras é pré-processado por tokenização, originando, removendo para kens com menos de três caracteres e remover todos os tokens com menos de 3% ou mais de 80%.

Essa estratégia de criação de recursos é uma mistura de geração automática e manual de recursos. Por um lado, nós auto
crie automaticamente todos os tipos diretos, sem se preocupar se são úteis para a tarefa em questão ou não. A maioria dos outros recursos, no entanto, é guiada por conhecimentos e suposições de domínio, por exemplo, que as categorias e os gêneros de um livro podem ser relevantes para um livro.

sistema ommender. Conforme discutido na seção 7, a geração de recursos totalmente automática abrange todos os recursos possíveis, com o risco de criar um espaço de recursos dimensionais muito altos com muitos recursos irrelevantes. Portanto, combinações de estratégias de geração de recursos automáticas e artesanais, como neste exemplo, são bastante comuns na prática.

O sistema de recomendação baseado em conteúdo é baseado no algoritmo k-NN, onde usamos $k = 80$ e semelhança de cosseno para os recomendadores de base. A lógica do uso da semelhança de cosseno é que, diferentemente, por exemplo, da distância euclidiana, apenas características comuns influenciam a semelhança,

mas não ausência comum de recursos (por exemplo, dois livros *não* sendo American Thriller Novels).

[43] Isso inclui tipos na ontologia YAGO, que podem ser bastante específicos (por exemplo, *Romances de suspense americanos)*

[44] O motivo para não incluir categorias mais amplas por padrão é que o gráfico de categorias não é uma árvore sem ciclo, com algumas sub-subposições sendo bastante questionáveis.

10.4 Interpretação dos Resultados do Recomendador

A etapa final do pipeline KDD habilitado para LOD é a avaliação e interpretação do modelo de mineração de dados desenvolvido (consulte a seção 9). No caso de sistemas de recomendação, além de poder ffi produzir recomendações precisas, a capacidade de ff explicar de forma eficaz as recomendações aos usuários é outro aspecto importante de um sistema de recomendação. Para esse objetivo, o Linked Open Data desempenha um papel importante, pois facilita o cálculo de uma explicação compreensível pelo ser humano, pois permite ao usuário explorar o espaço de resultados após di ff dimensões diferentes, ou seja, listar explicitamente, para cada propriedade, os valores comuns entre os filmes no perfil do usuário e os sugeridos [194].

Essa abordagem é particularmente interessante se, ao contrário do caso de uso acima, a recomendação se basear puramente em métodos estatísticos como filtragem colaborativa. Por exemplo, para um usuário que já gostou do livro "O Senhor dos Anéis", um sistema de recomendação pode recomendar o livro "O Hobit". O sistema pode facilmente dar uma explicação do motivo pelo qual o livro foi recomendado ao usuário, exibindo as relações compartilhadas mais importantes para esses dois livros, por exemplo, os dois livros são "High fantasy", os dois livros são escritos pelo mesmo autor "JRR Tolkien ”E os dois livros pertencem à mesma categoria“ romances de fantasia britânicos ”. É importante notar que a interpretação é realmente um passo a posteriori aqui, uma vez que o sistema de recomendação era puramente baseado em medidas estatísticas, ou seja, encontrar os livros mais semelhantes,

11. Discussão

Dada a quantidade de trabalhos de pesquisa discutidos neste documento, é evidente que, especialmente com o advento e o crescimento de Dados Abertos Vinculados, as informações da Web Semântica podem ser usadas de maneira benéfica no processo de mineração de dados e descoberta de conhecimento. Observando os resultados de uma distância maior, no entanto, podemos fazer algumas observações interessantes:

- O DBpedia é usado na grande maioria dos trabalhos de pesquisa discutidos nesta pesquisa, com outras fontes de LOD sendo pouco usadas, e a maioria das centenas de conjuntos de dados de LOD não é usada. Pode haver di ff razões diferentes para isso; variando do modelo de dados relativamente simples da DBpedia e sua ampla cobertura à disponibilidade de ferramentas sofisticadas, como a DBpedia Lookup e o DBpe-dia Spotlight.

Embora isso enfatize a utilidade dessas fontes de conhecimento de uso geral na Web Semântica, também pode ser problemático adaptar e avaliar abordagens apenas para conjuntos de dados únicos, uma vez que limita os insights sobre a aplicabilidade geral das abordagens.

- Muitos documentos usam ontologias e conjuntos de dados personalizados em vez de reutilizar conjuntos de dados abertos da web de dados. Isso é particularmente observado nas ciências da vida e no domínio médico, que, ao mesmo tempo, é um dos maiores domínios mais representados com destaque na nuvem Linked Open Data. Está sujeito a pesquisas futuras descobrir os motivos dessa discrepância, que podem ter ff razões diferentes, como um conhecimento limitado de conjuntos de dados abertos ou uma adequação inferior ao uso desses conjuntos de dados.

- Links entre conjuntos de dados, que são um dos principais ingredientes para *Ligado* Dados abertos, são usados por relativamente poucas abordagens. Isso também pode implicar que muitas das abordagens permaneçam abaixo do que é possível com o Linked Open Data, aproveitando apenas as informações de um conjunto de dados em vez de usar toda a quantidade de conhecimento capturado na Web Semântica. Uma razão pode ser que, mesmo na presença de esquemas interpretáveis por máquina, o desenvolvimento de aplicativos independentes de esquema seja uma tarefa não trivial. Além disso, a construção de abordagens aquele au- siga os links de maneira abrangente e, em última análise, seja capaz de explorar toda a Web de dados vinculados, pois o conhecimento de fundo também levaria a novos desafios de escalabilidade.

- Esquemas / ontologias expressivos e raciocínio sobre esses, que tem sido o principal ponto de venda da Web Semântica há anos, raramente são combinados com mineração de dados e descoberta de conhecimento. Mais uma vez, está sujeito a pesquisas futuras para descobrir se isso se deve a uma disponibilidade limitada de recursos ontológicos adequados. consciência limitada ou adequação imperfeita aos problemas encontrados na prática.

- Na maioria dos casos, o conhecimento da Web Semântica é sobre o domínio dos dados processados, não o domínio de mineração de dados. No entanto, determinados pontos finais, como *myExperiment.org* [45], que fornece muitos fluxos de trabalho científicos (incluindo fluxos de mineração de dados), permitiria soluções que fornecem conselhos aos analistas de dados que criam esses fluxos de trabalho, como o recém-anunciado “Wisdom Recomendações do Operador de Multidões "da Rapid-Miner [46], com base em dados abertos.

Essas observações mostram que, embora exista uma quantidade notável de trabalho na área, a mineração de dados e a descoberta de conhecimento ainda não estão explorando todo o potencial fornecido pela Web Semântica. Os fluxos de trabalho de mineração de dados aproveitam automaticamente as informações de di ff conjuntos de dados diferentes seguindo links além de conjuntos de dados únicos, como o DBpedia, ainda são uma área de pesquisa interessante e promissora.

12. Conclusão e Perspectivas

Neste artigo, fizemos uma pesquisa sobre o uso de dados da Web Semântica, os Dados Abertos Vinculados com mais destaque, para mineração de dados e descoberta de conhecimento. Seguindo o pipeline clássico de fluxo de trabalho de Fayyad, mostramos exemplos para o uso de dados da Web Semântica em todos os estágios do pipeline, bem como abordagens de suporte ao pipeline completo.

Analisando os resultados da pesquisa, a primeira observação é que existem muitos trabalhos de pesquisa na área e existem aplicações em muitos domínios. Um domínio de aplicação frequente é a biomedicina e as ciências da vida, mas as abordagens também são transferidas para vários outros domínios. Além disso, existem alguns aplicativos sofisticados e pilhas de ferramentas que vão além de meros protótipos de pesquisa.

Além disso, vemos que ainda existem alguns territórios desconhecidos no cenário de pesquisa da mineração de dados habilitada para a Web Semântica. Isso mostra que, embora resultados impressionantes já possam ser alcançados hoje, todo o potencial da Web Semântica possibilitou a mineração de dados e O KDD ainda precisa ser desbloqueado.

Reconhecimentos

O trabalho apresentado neste artigo foi parcialmente financiado pela Fundação Alemã de Pesquisa (DFG) sob o número de concessão PA 2373 / 1-1 (Mine @ LOD).

[1] UM Fayyad, G. Piatetsky-Shapiro, P. Smyth, Avanços no conhecimento sobre disco e mineração de dados, American Association para Inteligência Artificial, Menlo Park, CA, EUA, 1996, pp.

45 http: //www.myexperiment.or g

Referências

1–34. URL http://dl.acm.org/citation.cfm?id=25793 8 . 257942

46. https://rapidminer.com/news-posts/rapidminer - faz-snap-move-preditivo-analytics-data-mining - nuvem de aprendizado de máquina /

[2] D. Hand, H. Manni la, P. Smyth, Principles of Data Mining, MIT Press, 2001.

[3] C. Bizer, T. Heath, T. Berners-Lee, Dados vinculados - A história Far., Revista Internacional de Web Semântica e Informação Systems 5 (3) (2009) 1–22.

[4] M. Schmachtenberg, C. Bizer, H. Paulheim, Adoção do Melhores práticas de dados vinculados em Di ff diferentes domínios tópicos, em: Conferência Internacional da Web Semântica, 2014.

[5] G. Stumme, A. Hotho, B. Berendt, mineração de rede semântica - estado da arte e direções futuras, Journal of Web Semantics 4 (2) (2006) 124-143.
URL http://www.kde.cs.uni-kassel.de/hotho/pub/2006 / JWS2006SemanticWebMining.pdf

[6] K. Sridevi, DR UmaRani, Uma pesquisa de soluções semânticas informações sobre mineração na web, International Journal of Emerging Trends e Tecnologia em Ciência da Computação (IJETTS) 1.

[7] QK Quboa, M. Saraee, uma pesquisa de ponta sobre semântica mineração na web, Gerenciamento Inteligente de Informações 5 (2013) 10.

[8] J. Sivakumar, et al., Uma revisão sobre mineração web semântica e suas aplicações.

[9] D. Dou, H. Wang, H. Liu, mineração de dados semântica: uma pesquisa abordagens baseadas em ontologias, em: Computação Semântica (ICSC), Conferência Internacional do IEEE de 2015, IEEE, 2015, pp. 244–251

[10] Tresp, M. Bundschus, A. Rettinger, Y. Huang, Incerteza raciocínio para a web semântica i, Springer-Verlag, Berlin, Hei- Delberg, 2008, cap. Em direção ao aprendizado de máquina no Seminário tic Web, pp. 282-314. doi: 10.1007 / 978-3-540-89765-1_17.
URL http://dx.doi.org/10.1007/978-3-540-89765-1_17

[11] A. Rettinger, U. Lsch, V. Tresp, C. dAmato, N. Fanizzi, Min. web semântica, mineração de dados e descoberta de conhecimento 24 (3) (2012) 613-662. doi: 10.1007 / s10618-012-0253-2)
URL http://dx.doi.org/10.1007/s10618-012-0253-2

[12] TR Gruber, Em direção a princípios para o desenho de ontologias usado para compartilhar conhecimento, International journal of human-computer studies 43 (5) (1995) 907–928.

[13] HO Nigro, SG Cisaro, DH Xodo, Mineração de Dados com Consórcio. : Implementações, Constatações e Estruturas, Infor Science Reference - Impressão de: IGI Publishing, Her Ei, PA, 2007.

[14] COMO. Dadzie, M. Rowe, Abordagens para visualizar dados vinculados: Uma pesquisa, Semantic Web 2 (2) (2011) 89–124.

[15] G. Tummarello, R. Cyganiak, M. Catasta, S. Danielczyk, da 15ª Conferência Internacional do IEEE sobre Informações R. Delbru, S. Decker, Sig. ma: visualizações ao vivo na web de dados, Semântica da Web: Ciência, Serviços e Agentes no Mundo Wide Web 8 (4) (2010) 355–364.

[16] D. Huynh, S. Mazzocchi, D. Karger, Mealheiro: Experiência a web semântica dentro do seu navegador, em: The Semantic Web – ISWC 2005, Springer, 2005, pp. 413–430.

[17] T. Hastrup, R. Cyganiak, U. Bojars, Navegando dados vinculados com fenfire.

[18] O. Pe ̄na. U. Aguilera, D. L´opez-de Ipi ̄na, dados abertos vinculados visualização revisitada: Uma pesquisa, Semantic Web Journal.

[19] B. Mu lu, P. Hoe fler, G. Tschinkel, E. Veas, V. Sabol, F. Stegmaier, M. Granitzer, Sugerindo visualizações para pub dados finais, Proceedings of IVAPP (2014) 267-275.

[20] GA Atemezing, R. Troncy, Para uma visão baseada em dados assistente de sualização, em: Workshop sobre Consumindo Dados Vinculados, 2014.

[21] JM Brunetti, S. Auer, R. García, A visualização de dados vinculados model., in: Conferência Internacional da Web Semântica (Posters & Demos), 2012.

[22] J. Kl´ımek, J. Helmich, M. Neasky, Aplicação do modelo de visualização de dados em dados do mundo real do lod checo cloud, em: 6º Workshop Internacional sobre Dados Vinculados sobre the Web (LDOW14), 2014.

[23] J. Unbehauen, S. Hellmann, S. Auer, C. Stadler, Knowledge extração de fontes estruturadas, em: S. Ceri, M. Brambilla (Eds.), Search Computing, vol. 7538 de Notas de Aula em Ciência, Springer Berlin Heidelberg, 2012, pp. 34–52. doi: 10.1007 / 978-3-642-34213-4_3.
URL http://dx.doi.org/10.1007/978-3-642-34213-4_3

[24] SS Sahoo, W. Halb, S. Hellmann, K. Idehen, TT Jr, S. Auer, J. Sequeda, A. Ezzat, Uma pesquisa de abordagens atuais para mapeamento de bancos de dados relacionais para rdf (01 2009).
URL http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf

[25] D.-E. Spanos, P. Stavrou, N. Mitrou, Trazendo relações bancos de dados na web semântica: uma pesquisa, Semant. teia 3 (2) (2012) 169-209. doi: 10.3233 / SW-2011-0055.
URL http://dx.doi.org/10.3233/SW-2011-0055

[26] C. Bizer, D2rq - tratando bancos de dados não-rdf como rdf virtual gráficos, em: Nos Anais da 3ª Semântica Internacional Conferência na Web, 2004.

[27] V. Mulwad, T. Finin, Z. Syed, A. Joshi, Usando dados vinculados para interpretar tabelas, em: In Proc. 1ª Int. Workshop sobre Consumir Dados vinculados, 2010.

[28] V. Mulwad, T. Finin, Z. Syed, A. Joshi, T2LD: interpretação e representando tabelas como dados vinculados, em: Anais da ISWC Trilha de pôsteres e demonstrações de 2010: resumos coletados, Xangai, China, 9 de novembro de 2010, 2010.
URL http://ceur-ws.org/Vol-658/paper489.pdf

[29] Z. Syed, T. Finin, V. Mulwad, A. Joshi, Explorando uma rede de dados semânticos para interpretação de tabelas, em: In: Anais da Segunda Conferência de Ciência da Web., 2010.

[30] V. Mulwad, DC Proposta: Modelos Gráficos e Probabili- Raciocínio básico para gerar dados vinculados a partir de tabelas, em: Anais da Décima Conferência Internacional da Web Semântica parte II, I. aroyo et al. (eds.) Edição, LCNS, LCNS 7032, Springer-Verlag, 2011, p. 317324, submetido no Doutorado

[31] V. Mulwad, T. Finin, A. Joshi, Mensagem Semântica Gerando dados vinculados a partir de tabelas, em: Anais da 12ª Conferência Internacional da Web Semântica, Springer, 2013.

[32] V. Mulwad, T. Finin, A. Joshi, Interpretação de tabelas médicas como Dados vinculados para gerar relatórios de metanálise, em: Reutilização e Integração, IEEE Computer Society, 2014.

[33] T. Finin, Z. Syed, Criando e explorando uma rede de semântica. data., in: ICAART (1), 2010, pp. 7–18.

[34] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyga- niak, Z. Ives, Dbpedia: núcleo para uma rede de dados abertos, em: Anais da 6ª Internacional A Web Semântica e 2ª Conferência Asiática sobre Conferência Web Semântica Asiática ISWC'07 / ASWC'07, Springer-Verlag, Berlim, Heidelberg Berg, 2007, pp. 722–735.
URL http://dl.acm.org/citation.cfm?id=1785162.1785216

[35] FM Suchanek, G. Kasneci, G. Weikum, Yago: um núcleo de conhecimento semântico, em: Anais da 16ª Conferência Internacional sobre a World Wide Web, WWW '07, ACM, Nova York, NY, EUA, 2007, pp. 697–706. doi: 10.1145 / 1242572.1242667.
URL http://doi.acm.org/10.1145/1242572.1242667

[36] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Free-

base: um banco de dados gráfico criado colaborativamente para estruturar conhecimento humano, em: Proceedings of the ACM SIG 2008 Conferência Internacional MOD sobre Gerenciamento de Dados, SIGMOD '08, ACM, Nova York, NY, EUA, 2008, pp. 1247-1250. doi: 10.1145 / 1376616.1376746.
URL http://doi.acm.org/10.1145/1376616.1376746

[37] GA Miller, Wordnet: um banco de dados lexical para inglês, Commu. publicações do ACM 38 (11) (1995) 39–41.

[38] H. Liu, Rumo à mineração de dados semânticos, em: In Proc. do dia 9 Conferência Internacional da Web Semântica (ISWC2010, 2010.

G. Limaye, S. Sarawagi, S. Chakrabarti, Anotação e Motta, pesquisando tabelas da web usando entidades, tipos e relacionamentos, Proc. VLDB Endow. 3 (1-2) (2010) 1338–1347. doi: 10. 14778 / 1920841.1921005.
URL http://dx.doi.org/10.14778/1920841.1921005

[39] P. Venetis, A. Halevy, J. Madhavan, M. Pas ¸ca, W. Shen, F. Wu, G. Miao, C. Wu, Recuperando a semântica de tabelas na Web, Proc. VLDB Endow. 4 (9) (2011) 528-538. doi: 10.14778 / 2002938.2002939.
URL http://dx.doi.org/10.14778/2002938.2002939

[40] J. Wang, H. Wang, Z. Wang, KQ Zhu, Tabelas de compreensão na web, em: Anais da 31ª Conferência Internacional sobre Modelagem Conceitual, ER'12, Springer-Verlag, Berlim, Heidelberg, 2012, pp. 141–155. doi: 10.1007 / 978-3-642-34002-4_11.
URL http://dx.doi.org/10.1007/978-3-642-34002-4_11

[41] W. Wu, H. Li, H. Wang, KQ Zhu, Probase: um probabilístico taxonomia para a compreensão do texto, em: Anais da edição de 2012 Conferência Internacional ACM SIGMOD sobre Gerenciamento de Data, SIGMOD '12, ACM, Nova York, NY, EUA, 2012, pp. 481-492. doi: 10.1145 / 2213836.2213891.
URL http://doi.acm.org/10.1145/2213836.2213891

[42] Z. Zhang, AL Gentile, I. Augenstein ", vinculou dados Conhecimento pro-ground para extração de informações na web ", SIGWEB Newsl. (Verão) (2014) 5: 1–5: 9. doi: 10.1145 / 2641730.2641735.
URL http://doi.acm.org/10.1145/2641730.2641735

[43] Z. Zhang, comece pequeno, crie completo: E ff eficaz ee m cientifico interpretação de tabelas semânticas usando tableminer, Under transpar revisão: The Semantic Web Journal.

[45] Z. Zhang, aprendendo com dados parciais para tabela semântica entre pretensão, em: K. Janowicz, S. Schlobach, P. Lambrix, E. Hyvnen (Eds.), Engenharia do conhecimento e Homem do conhecimento vol. 8876 de Notas de aula em Ciência da computação, Springer International Publishing, 2014, pp. 607–618. doi: 10.1007 / 978-3-319-13704-9_45.
URL http://dx.doi.org/10.1007/978-3-319-13704-9_45

[46] M. Oita, A. Amarilli, P. Senellart, rede profunda de fertilização cruzada oficina análise e enriquecimento de ontologias, em: M. Brambilla, S. Ceri, T. Furche, G. Gottlob (Eds.), VLDS, CEUR-WS.org , pp. 5-8.

[47] E. Muoz, A. Hogan, A. Mileo, Triplificando as tabelas da wikipedia, nomeado em: LD4IE @ ISWC '13, 2013, pp. –1–1.

[48] E. Mu ˜ noz, A. Hogan, A. Mileo, Usando dados vinculados ao meu o rdf das tabelas da wikipedia, em: Anais da 7ª ACM Conferência Internacional sobre Pesquisa na Web e Mineração de Dados, WSDM '14, ACM, Nova York, NY, EUA, 2014, pp. 533-542. doi: 10.1145 / 2556195.2556266.
URL http://doi.acm.org/10.1145/2556195.2556266

[49] CS Bhagavatula, T. Noraset, D. Downey, Methods for extabelas de exploração e mineração na wikipedia, em: Anais da Workshop ACM SIGKDD sobre Exploração Interativa de Dados e Analytics, IDEA '13, ACM, Nova York, NY, EUA, 2013, pp. 18-26. doi: 10.1145 / 2501511.2501516.
URL http://doi.acm.org/10.1145/2501511.2501516

[50] L. Han, T. Finin, C. Parr, J. Sachs, A. Joshi, Rdf123: From planilhas para rdf, em: Anais da 7ª Internacional Conferência sobre a Web Semântica, ISWC '08, Springer-Verlag, Berlim, Heidelberg, 2008, pp. 451–466. doi: 10. 1007 / 978-3-540-88564-1_29.
URL http://dx.doi.org/10.1007/978-3-540-88564-1_29

[51] A. Langegger, W.W, Xlwrap querying and integrat- planilhas arbitrárias com sparql, em: A. Bernstein, D. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. K. Thirunarayan (Eds.), The Semantic Web - ISWC 2009, vol. 5823 de Notas de aula em Ciência da computação, Springer Berlin Heidelberg, 2009, pp. 359-374. doi: 10.1007 / 978-3-642-04930-9_23.
URL http://dx.doi.org/10.1007/978-3-642-04930-9_23

[52] L. Ding, D. DiFranzo, A. Graves, J. Michaelis, X. Li, DL McGuinness, JA Hendler, Twc corp-data gov: incremengerando dados governamentais vinculados de data.gov ., no: M. Rappa, P. Jones, J. Freire, S. Chakrabarti (Eds.), WWW, ACM, 2010, pp. 1383–1386.
URL http://dblp.uni-trier.de/db/conf/www/www2010.html # DingDGMLMH10

[53] O. Hassanzadeh, SH Yeganeh, RJ Miller, Linking dados semiestruturados na web., em: WebDB, 2011.

[54] PN Mendes, M. Jakob, A. García-Silva, C. Bizer, Dbpedia. holofotes: lançando luz na web de documentos, em: Proresultados da 7ª Conferência Internacional sobre Sistemas Semânticos I-Semantics '11, ACM, Nova York, NY, EUA, 2011, pp. 1–8. doi: 10.1145 / 2063518.2063519.
URL http://doi.acm.org/10.1145/2063518.2063519

[55] J. Daiber, M. Jakob, C. Hokamp, PN Mendes, Melhorando eficiência e precisão na extração de entidades multilíngues, em: da 9ª Conferência Internacional de Sistemas Semânticos. Tempos, ACM, 2013, pp. 121–124.

[56] O. De Clercq, S. Hertling, V. Hoste, SP Ponzetto, H. Paulheim, Identificando tópicos em disputa nas notícias, em: Dados vinculados for Knowledge Discovery (LD4KD), CEUR, 2014, pp. 37–48.

[57] A. Schulz, P. Ristoski, H. Paulheim, vejo um acidente de carro: Real- detecção de incidentes de pequena escala em microblogs, em: Web Semântica: Eventos de Satélite ESWC 2013, Springer, 2013, 22-33.

[58] D. Hienert, D. Wegener, H. Paulheim, Classificação automática e extração de relacionamento para vários idiomas e vários granulares eventos da wikipedia, Detecção, Representação e Exploitação de eventos na Web Semântica (DeRiVE 2012) 902 (2012) 1–10.

[59] M. Schuhmacher, SP Ponzetto, Explorando dbpedia para web agrupamento de resultados de pesquisa, em: Anais do 2013 sobre Construção automatizada de base de conhecimento, ACM, 2013, pp. 91-96.

[60] G. Rizzo, R. Troncy, Nerd: uma estrutura para unificar ferramentas de extração de reconhecimento e desambiguação de entidades, em: Proresultados das manifestações na 13ª Conferência de Capítulo Europeu da Association for Computational Linguistics, Association for Computational Linguistics, 2012, pp. 73-76.

[61] K. Bontcheva, D. Rout, Compreendendo os fluxos de mídia social através da semântica: uma pesquisa, Semantic Web 1 (2012) 1–31.

[62] D. Heckmann, T. Schwartz, B. Brandherm, M. Schmitz, Von Wilamowitz-Moellendor ff, Gumo - o usuário geral model ontology, in: Modelagem de usuários 2005, Springer, 2005, pp. 428-432.

[63] S. Scerri, K. Cortis, I. Rivera, S. Handschuh, Conhecimento dis31

cobertura em atividades distribuídas de compartilhamento na web social, Making Senso de micropostas (# MSM2012) (2012) 26–33.

[64] A. Passant, P. Laublet, Significado de uma etiqueta: um ap colaborativo abordagem para preencher a lacuna entre a marcação e os dados vinculados., em: LDOW, 2008.

[65] G. Solskinnsbakk, JA Gulla, anotação semântica de atributos sociais. dados, em: Anais do Quarto Workshop Internacional sobre Workshop de Dados Sociais na Web, 2011.

[66] L. Qu, C. M¨uller, I. Gurevych, Usando rede semântica de tags para extração de frase-chave em blogs, em: Anais do dia 17 Conferência da ACM sobre gestão da informação e conhecimento, ACM, 2008, pp. 1381–1382.

[67] J. Eisenstein, DH Chau, A. Kittur, EP Xing, et al., Topicviz: Navegação semântica de coleções de documentos, pré-impressão do arXiv arXiv: 1110.6200.

[68] M. Pennacchiot ti, A. -M. Popescu, um ap de aprendizado de máquina - abordagem à classificação de usuário do twitter., ICWSM 11 (2011) 281– 288

[69] F. Abel, Q. Gao, G.-J. Houben, K. Tao, enriquecimento semântico de postagens do twitter para construção de perfil de usuário na web social, in: A Web Semântica: Pesquisa e Aplicações, Springer, 2011, pp. 375–389.

[70] J. Chan, C. Hayes, E. Daly, Fóruns de discussão em decomposição usando funções de usuário comuns.

[71] F. Abel, I. Celik, G. -J. Houben, P. Siehndel, Alavancando as pp. 1–4. semântica de tweets para pesquisa facetada adaptativa no twitter, em: The Semantic Web – ISWC 2011, Springer, 2011, pp. 1–17.

[72] J. Chen, R. Nairn, L. Nelson, M. Bernstein, E. Chi, Short and tweet: experimentos sobre a recomendação de conteúdo de informações fluxos de informação, em: Anais da Conferência SIGCHI sobre Fatores humanos em sistemas de computação, ACM, 2010, pp. 1185–1194

[73] H. Paulheim, F. Probst, interfaces de usuário aprimoradas para ontologia: A Revista Internacional de Web Semântica e Informação Systems (IJSWIS) 6 (2) (2010) 36–59.

[74] X. Wang, HJ Hamilton, Y. Bither, um aplicativo baseado em ontologia abordagem à limpeza de dados (2005).

[75] D. Prez-Rey, A. Anguita, J. Crespo, Ontodataclean: Ontology- integração e pré-processamento baseados em dados distribuídos., em: N. Maglaveras, I. Chouvarda, V. Koutkias, RW Brause (Eds.), ISBMDA, vol. 4345 de Notas de aula em Computador Science, Springer, 2006, pp. 262–272. URL http://dblp.uni-trier.de/db/conf/ismda/ isbmda2006.html # Perez-ReyAC06

[76] J. Phillips, BG Buchanan, conhecimento guiado por ontologia cobertura em bases de dados, em: Anais da 1ª Internacional conferência sobre captura de conhecimento, ACM, 2001, pp. 123–130.

[77] Z. Kedad, E. Métais, Limpeza de dados baseada em ontologia, in: Nat Processamento de idiomas e sistemas de informação, Springer, 2002, pp. 137-149.

[78] D. Mi lano, M. Scannapieco, T. Catarci, Usando ontologias para 2_11 limpeza de dados xml, em: Anais da Confederação OTM de 2005 Conferência Internacional sobre Mudança para o Significado Internet Systems, OTM'05, Springer-Verlag, Berlim, Alemanha Delberg, 2005, pp. 562-571. doi: 10.1007 / 11575863_75. URL http://dx.doi.org/10.1007/11575863_75

[79] S. Br¨uggemann, F. Grunun, Usando o conhecimento do domínio pro fornecidas por ontologias para melhorar o gerenciamento da qualidade dos dados, in: Proceedings of I-Know, 2008, pp. 251–258.

[80] S. Brugugmann, H.-J. Appelrath, operadores de substituição com reconhecimento de contexto para um aprendizado aprimorado a partir de dados de redes semânticas, operações de limpeza de dados, em: Anais do 2011 Simpósio da ACM sobre Computação Aplicada, SAC '11, ACM, Nova York, NY, EUA, 2011, pp. 1700–1704. doi: 10.1145 / 1982185.1982539. URL http://doi.acm.org/10.1145/1982185.1982539

[81] Y. Wang, S. Yang, detecção externa de documentos curtos e maciços usando ontologia de domínio, em: Computação Inteligente e Sistemas Inteligentes (ICIS), Conferência Internacional IEEE 2010 vol. 3, IEEE, 2010, pp. 558–562.

[82] T. Lukaszewski, estendendo o classificador bayesiano com métodos ontológicos

[83] C. Fürber, M. Hepp, Using sparql and spin para qualidade dos dados. gerenciamento na web semântica, em: Informações sobre negócios Systems, Springer, 2010, pp. 35–46.

[84] C. F.urber, M. Hepp, Usando recursos semânticos da web para dados gestão da qualidade, em: Anais da 17ª Internacional Conferência sobre Engenharia e Gestão do Conhecimento Massas, EKAW'10, Springer-Verlag, Berlim, Heidelberg, 2010, pp. 211-225. URL http://dl.acm.org/citation.cfm?id=1948294. 1948316

[85] C. Fürber, M. Hepp, Swiqa-a qualidade da informação da web semântica framework de avaliação., em: ECIS, vol. 15, 2011, p. 19

[86] C. Fürber, M. Hepp, Usando tecnologias semânticas da web para dados gestão da qualidade, in: Handbook of data quality, 2013, pp. 141-161.

[87] L. Moss, D. Corsar, I. Piper, Uma abordagem de dados vinculados a avaliação de dados médicos, em: Sistemas Médicos Baseados em Computador (CBMS), 25º Simpósio Internacional de 2012, IEEE, 2012,

[88] S.-T. Liaw, A. Rahimi, P. Ray, J. Taggart, S. Dennis, S. de Lusignan, B. Jalaludin, AET Yeo, A. Talaei-Khoei, Rumo a uma ontologia para a qualidade dos dados em doenças crônicas integradas facilidade de gerenciamento: Uma revisão realista da literatura., IJ Med- Informática Informática 82 (1) (2013) 10–24. URL http://dblp.uni-trier.de/db/journals/ijmi/ ijmi82.html # LiawRRTDLJYT13

[89] O. Lehmberg, D. Ritze, P. Ristoski, R. Meusel, H. Paulheim, C. Bizer, O mecanismo de junção de pesquisa mannheim, Journal of Web Semântica.

[90] T. Käefer, A. Harth, conjunto de dados do Billion Triples Challenge, Down- carregado de http://km.aifb.kit.edu/projects/btc-2014/ (2014).

[91] R. Meusel, P. Petrovski, C. Bizer, The WebDataCommons Mi- Série de conjuntos de dados crodata, RDFa e Microformat, em: Proc. do 13ª Int. Conferência da Web Semântica (ISWC14), 2014.

[92] H. Paulheim, Explorando dados abertos vinculados como pano de fundo conhecimento em mineração de dados, em: Workshop sobre mineração de dados sobre Dados abertos vinculados, 2013.

[93] T. Pang-Ning, M. Steinbach, V. Kumar, Introdução aos dados. mineração, Pearson, 2006.

[94] S. Kramer, N. Lavra, P. Flach, abordagens de proposicionalização à mineração relacional de dados, em: S. DÅeroski, N. Lavra (Eds.), Mineração de Dados Relacionais, Springer Berlin Heidelberg, 2001, pp. 262-291. doi: 10.1007 / 978-3-662-04599-2_11. URL http://dx.doi.org/10.1007/978-3-662-04599-

[95] H. Paulheim, J. Furnurnran, geração não supervisionada de dados Recursos de mineração de dados abertos vinculados, em: Internacional Conferência sobre Web Intelligence, Mineração e Semântica (WIMS'12), 2012.

[96] VNP Kappara, R. Ichise, O. Vyas, Liddm: Uma mineração de dados. sistema para dados vinculados, em: Workshop sobre Dados Vinculados sobre Web (LDOW2011), 2011.

[97] MA Khan, GA Grimnes, A. Dengel, dois pré-processamento. in: Primeira reunião e conferência da comunidade RapidMiner (RCOMM 2010), 2010.

[98] W. Cheng, G. Kasneci, T. Graepel, D. Stern, R. Herbrich, Geração automatizada de recursos a partir de conhecimento estruturado, em: 20ª Conferência da ACM sobre Gerenciamento de Informações e Conhecimento

(CIKM 2011), 2011.

[99] J. Mynarz, V. Sv´atek, Rumo a um ponto de referência para os descoberta de conhecimento a partir de dados estruturados, em: o segundo workshop internacional sobre descoberta de conhecimento e mineração de dados atende a dados abertos vinculados, CEUR-WS, 2013, 41-48.

[100] T. Kauppinen, GM de Espindola, B. Graler, Sharing e um alyzing dados de observação de sensoriamento remoto para ciência vinculada, em: Anais de pôsteres da Conferência Semântica Estendida na Web, Citeseer, 2012.

[101] T. Kauppinen, GM de Espindola, J. Jones, A. S´anchez, B. Gr¨aler, T. Bartoschek, floresta amazônica brasileira vinculada data, Semantic Web 5 (2) (2014) 151-155.

[102] W. van Hage, M. van Erp, V. Malais, pirataria aberta vinculada: Uma história sobre e-science, dados vinculados e estatísticas, Journal sobre Data Semântica 1 (3) (2012) 187–201. doi: 10.1007 / s13740-012-0009-6.
URL http://dx.doi.org/10.1007/s13740-012-0009-6

[103] H. Paulheim, P. Ristoski, E. Mitichkin, C. Bizer, Data min com conhecimento de base da Web, em: RapidMiner Mundo, 2014.

[104] P. Ristoski, C. Bizer, H. Paulheim, Mineração da rede de dados com quickminer, Journal of Web Semantics.

[105] A. Schulz, C. Guckelsberger, F. Janssen, abstração semântica para generalização da classificação do tweet.

[106] H. Paulheim, Gerando possíveis interpretações para estatística a partir de dados abertos vinculados, em: 9ª Web Semântica Estendida (ESWC), 2012.

[107] H. Paulheim, Ninguém quer morar em uma cidade fria onde não há música foi gravada, em: The Semantic Web: ESWC 2012 Eventos de satélite, Springer, 2012, pp. 387–391.

[108] P. Ristoski, H. Paulheim, Analisando estatísticas com antecedentes conhecimento a partir de dados abertos vinculados, em: Workshop em Semântica Estatísticas, 2013.

[109] H. Paulheim, Identificando links incorretos entre conjuntos de dados por detecção de outlier multidimensional, em: Workshop on Debug ontologias e mapeamentos de ontologias (WoDOOM), 2014, 2014.

[110] P. Ristoski, EL Menc´ıa, H. Paulheim, Uma multi-estratégia híbrida sistema de recomendação usando dados abertos vinculados, em: Semântica Web Evaluation Challenge, Springer, 2014, pp. 150–156.

[111] M. Schmachtenberg, T. Strufe, H. Paulheim, Enhancing a sistema de recomendação baseado em localização, enriquecendo com dados estruturados da web, em: Web Intelligence, Mineração e Semântica, 2014.

[112] P. Ristoski, H. Paulheim, Uma comparação de proposicionalização estratégias de criação para criar recursos a partir de dados abertos vinculados, em: Dados vinculados para descoberta de conhecimento, 2014.

[113] Y. Huang, V. Tresp, M. Nickel, A. Rettinger, H.-P. Kriegel, A abordagem escalável para aprendizado estatístico em gráficos semânticos, Semantic Web 5 (1) (2014) 5–22.

[114] Y. Huang, V. Tresp, M. Bundschus, A. Rettinger, H.-P. Kriegel, Previsão multivariada para aprendizado na web semântica, em: P. Frasconi, F. Lisi (Eds.), Programação Lógica Indutiva, vol. 6489 de Notas de aula em Ciência da computação, Springer Berlin Heidelberg, 2011, pp. 92-104. doi: 10.1007 / 978-3-642-21295-6_13.
URL http://dx.doi.org/10.1007/978-3-642-21295-6_13

[115] Y. Huang, M. Nickel, V. Tresp, H. -P. Kriegel, um kernel escalável 478–484. abordagem à aprendizagem em gráficos semânticos com aplicações para dados vinculados, em: 1º Workshop sobre Mineração do Futuro na Internet, 2010.

[116] N. Fanizzi, C. dAmato, um núcleo declarativo para o conceito de alc. descrições, em: Fundamentos de Sistemas Inteligentes, Springer, 2006, pp. 322-331.

[117] N. Fanizzi, F. Esposito, Aprendizado estatístico para consulta indutiva respondendo às ontologias da coruja, em: Conferência Internacional da Web Semântica (ISWC, 2008.

[118] S. Bloehdorn, Y. Claro, métodos de kernel para instância de mineração dados em ontologias, em: Anais da 6ª Internacional Web Semântica e 2ª Conferência Asiática sobre Ásia Semântica Web Conference, ISWC'07 / ASWC'07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 58–71.
URL http://dl.acm.org/citation.cfm?id=1785162.1785168

[119] V. Bicer, T. Tran, A. Gossen, máquinas relacionais de kernel para aprendendo com dados rdf estruturados em gráficos, em: A Web Semântica: Research and Applications, Springer, 2011, pp. 47–62.

[120] U. Lösch, S. Bloehdorn, A. Rettinger, Kernels de gráfico para rdf. dados, em: The Semantic Web: Research and Applications, Springer, 2012, pp. 134-148.

[121] GKD de Vries, S. de Rooij, um núcleo gráfico rápido e simples for rdf, in: Anais do Segundo Workshop Internacional sobre descoberta de conhecimento e mineração de dados atende a links abertos Data, 2013.

[122] GKD de Vries, Uma rápida aproximação do weisfeiler-lehman graph kernel for rdf data., em: H. Blockeel, K. Kersting, S. Nijssen, F. Zelezn (Eds.), ECML / PKDD (1), vol. 8188 de Notas de aula em ciência da computação, Springer, 2013, pp. 606-621.
URL http://dblp.uni-trier.de/db/conf/pkdd/pkdd2013-1.html # Vries13

[123] P. Bloem, A. Wibisono, G. de Vries, Simplificando dados rdf para aprendizado de máquina baseado em gráficos, em: 11º ESWC 2014 (ESWC2014), 2014.
URL http://data.semanticweb.org/conference/eswc / 2014 / paper / ws / KnowLOD / 8

[124] GKD de Vries, S. de Rooij, gráfico de contagem de subestruturas kernels para aprendizado de máquina a partir de dados rdf, Web Semântica: Ciência, serviços e agentes na Internet.

[125] N. Shervashidze, P. Schweitzer, EJ Van Leeuwen, K. Mehlhorn, KM Borgwardt, Weisfeiler-lehman graph kernels, The Journal of Machine Learning Research 12 (2011) 2539-2561.

[126] GH John, R. Kohavi, K. Pfleger, Características irrelevantes e os problema de seleção de subconjuntos, em: ICML'94, 1994, pp. 121–129.

[127] AL Blum, P. Langley, Seleção de características relevantes e amplos em aprendizado de máquina, INTELIGÊNCIA ARTIFICIAL 97 (1997) 245–271.

[128] P. Ristoski, H. Paulheim, Seleção de características em características hierárquicas espaços de inovação, em: Discovery Science, 2014.

[129] Y. Jeong, S.-H. Myaeng, Seleção de recurso usando uma semântica hierarquia para reconhecimento de eventos e classificação de tipo, em: Conferência Conjunta Internacional sobre Processamento de Linguagem Natural, 2013.

[130] BB Wang, RIB Mckay, HA Abbass, M. Barlow, A estudo parativo para extração guiada de características de ontologia de domínio, in: Conferência Australiana de Ciência da Computação, 2003.

[131] S. Lu, Y. Ye, R. Tsui, H. Su, R. Rexit, S. Wesaratchakit, X. Liu, R. Hwa, redução de recursos baseados em ontologia de domínio para dados mensais de medicamentos e sua aplicação à insuficiência cardíaca em 30 dias previsão de readmissão, em: International Conference Confer- sobre Computação Colaborativa (Collaboratecom), 2013, pp.

[132] J. Euzenat, P. Shvaiko, Ontology Matching, Springer-Verlag Nova York, Inc., Secaucus, NJ, EUA, 2007. [133] FM Suchanek, S. Abiteboul, P. Senellart, PARIS: Probabilistic Alinhamento de relações, instâncias e esquema, PVLDB

5 (3) (2011) 157-168.

[134] A. Bellandi, B. Furletti, V. Grossi, A. Romei, dirigido por ontologia. extração de regras de associação: um estudo de caso, Contexts e Ontolo- Representação e Raciocínio (2007) 10.

[135] C. Antunes, Onto4ar: uma estrutura para associação de mineração , em: Workshop sobre Mineração Baseada em Restrições e Aprendizado ing (CMILE-ECML / PKDD 2007), 2007, pp. 37–48.

[136] C. Antunes, Uma estrutura baseada em ontologia para padrões de mineração de base, em: 1st International Conferência sobre Inteligência Avançada, 2008, pp. 163–168.

[137] ACB Garcia, AS Vivacqua, A ontologia ajuda a sentido de um mundo complexo ou cria uma interpretação tendenciosa ção ?, em: Proc. Oficina de Sensemaking em CHI, vol. 8, 2008. [154] V. Podpe˘can, M. Zemenova, N. Lavra˘c, Orange4ws environ [138] AC Bicharra Garcia, I. Ferraz, A. s. Vivacqua, De dados a mineração de conhecimento, Artif. Intell. Eng. Des. Anal. Manuf. 23 (4) (2009) 427-441. doi: 10.1017 / S089006040900016X. URL http://dx.doi.org/10.1017/ S089006040900016X

[139] M. Zeman, M. Ralbovsk`y, V. Svátek, J. Rauch, Ontologia- preparação de dados direcionada para mineração de associação, Online http: // barril. vse. cz / onto-kdd-draft . pdf.

[140] M. · Z´aková, P. Kremen, F. Zelezny, N. Lavrac, Automação composição do fluxo de trabalho de descoberta de conhecimento através de ontologias planejamento baseado, Engenharia e Ciência da Automação, IEEE Transações em 8 (2) (2010) 253–264.

[141] C. Diamantini, D. Potena, E. Storti, Kddonto: Uma ontologia para descoberta e composição de algoritmos kdd, Third Genera Data Mining: Towards Service-Oriented Dis Dis [158] J.-U. Kietz, F. Serban, A. Bernstein, eproplan: Uma ferramenta para modelar covery (SoKD09) (2009) 13–24. geração automática de fluxos de trabalho de mineração de dados, em:

[142] J. Kietz, F. Serban, A. Bernstein, S. Fischer, Towards cooper planejamento ativo dos fluxos de trabalho de mineração de dados, em: Workshop de Mineração de Dados de Terceira Geração na Eu 2009 [159] J.-U. Kietz, F. Serban, S. Fischer, A. Bernstein, Conferência inrópica semântica sobre aprendizado de máquina (ECML 2009), 2009, pp. 1-12.

[143] M. Hilario, A. Kalousis, P. Nguyen, A. Woznica, A data min ontologia para seleção de algoritmos e meta-mineração, em: Pro- [160] S. D˘zeroski, Towards a General Framework for Data Mining, do Workshop da ECML / PKDD09 sobre terceira geração Data Mining (SoKD-09), 2009, pp. 76–87.

[144] M. Olá Lario, P. Nguyen, H. Do, A. Woznica, A. Kalou sis, Meta-mineração baseada em ontologia do trabalho de descoberta de conhecimento [161] A. Gabriel, H. Paulheim, F. Janssen, Aprendizagem semanticamente coplows, em: Meta-Learning in Computational Intelligence, Springer, 2011, pp. 273–315.

[145] P. Panov, S. Dzeroski, LN Soldatova, Ontodm: An on tecnologia de mineração de dados, em: Data Mining Workshops, 2008. ICDMW'08. Conferência Internacional do IEEE, IEEE, 2008, [163] FM Pinto, MF Santos, Considerando o domínio de aplicação onpp. 752-760. para mineração de dados, transações WSEAS sobre informações

[146] P. Panov, S. D˘zeroski, LN Soldatova, entidades representativas na a ontodm data mining ontology, em: Bancos de dados indutivos e [164] D. Pan, J.-Y. Shen, M.-X. Zhou, Incorporando o conhecimento do domínio Mineração de Dados Baseada em Restrições, Springer, 2010, pp. 27–58.

[147] P. Panov, L. Soldatova, S. DÅeroski, Ontologia dos dados principais entidades de mineração, mineração de dados e descoberta de conhecimentos 28 (5 6) (2014) 1222–1265. doi: 10.1007 / s10618-014-0363-0. [165] D. Pan, Y. Pan, Usando o repositório de ontologia para suportar minURL de dados http://dx.doi.org/10.1007/s10618- ing, in: Controle e Automação Inteligente, 2006. WCICA 0 0

[148] P. Panov, L. Soldatova, S. Dzeroski, Ontodm-kdd: Ontology 5947-5951. por representar o processo de descoberta do conhecimento, em: Discov- [166] H. Ce˘spivová, J. Rauch, V. Svatek, M. Kejkula, M. Tomeckova, Science, Springer, 2013, pp. 126–140.

[149] F. Serban, J. Vanschoren, J.-U. Kietz, A. Bernstein, Uma pesquisa de assistentes inteligentes para análise de dados, o ACM Computing Sur veys (CSUR) 45 (3) (2013) 31.

[150] A. Suyama, N. Negishi, T. Yamaguchi, Camlet: Uma plataforma Semântica para composição automática de sistemas de aprendizagem indutivos ontologias, em: PRICAI98: Topics in Artificial Intelligence, Springer, 1998, pp. 205-215.

[151] A. Bernstein, F. Provost, S. Hill, Rumo à assistência inteligente para um processo de mineração de dados: uma abordagem baseada em ontologia para classificação sensível a custos, engenharia de conhecimento e dados, Transações IEEE em 17 (4) (2005) 503-518.

[152] M. · Záková, P. Kremen, F. Zelezny, N. Lavrac, Planejando aprender ontologia de descoberta de conhecimento, em: SEGUNDO PLAN- NING PARA APRENDER OFICINA (PLANELARN) NO ICML / COLT / UAI 2008, 2008, p. 29

[153] M. Záková, V. Podpecan, F. Zelezn`y, N. Lavrac, Avançando na presença de conhecimento construção de fluxo de trabalho de mineração de dados: uma estrutura e casos usando o kit de ferramentas laranja, Proc. 2nd Intl. Wshop. Terceira Geração Data Mining: Em direção a conhecimento orientado a serviços covery (2009) 39–52.

para mineração de dados orientada a serviços, The Computer Journal (2011) bxr077.

[155] C. Diamantini, D. Potena, anotação e serviços semânticos. para compartilhamento e reutilização de ferramentas do kdd., in: ICDM Workshops, 2008, pp. 761-770.

[156] C. Diamantini, D. Potena, E. Storti, Suporte a usuários no kdd design de processos: uma abordagem de correspondência de similaridade semântica, em: Workshop Planejamento para aprender (PlanLearn10) na ECAI, 2010, pp. 27-34.

[157] J.-U. Kietz, F. Serban, A. Bernstein, S. Fischer, Data min- modelos de fluxo de trabalho para assistência inteligente à descoberta e experimentação automática, mineração de dados de terceira geração: Em direção à descoberta de conhecimento orientada a serviços (SoKD-10) (2010) 1–12.

Vol. 2010, 2010.

lado! mas não vamos informar aos mineradores de dados: Suporte inteligente para mineração de dados, em: A Web Semântica: Tendências e Desafios, Springer, 2014, pp. 706–720.

KDID'06, Springer-Verlag, Berlim, Heidelberg, 2007. URL http://dl.acm.org/citation.cfm?id=1777194. 1777213

regras atuais, em: Interações entre Data Mining e Natural Language Processing, 2014, pp. 49–63.

[162] J. Fürnkranz, Aprendizado de regras para separar e conquistar, Artificial Intelligence Review 13 (1) (1999) 3–54.

Science and Applications 6 (9) (2009) 1478-1492.

borda no processo de mineração de dados: Uma estrutura baseada em ontologia, Jornal da Universidade Wuhan de Ciências Naturais 11 (1) (2006) 165-169.

2006. O Sexto Congresso Mundial, vol. 2, IEEE, 2006, pp.

Papéis da ontologia médica na associação de mineração de crisp-dm cle, in: ECML / PKDD04 Workshop sobre descoberta de conhecimento e Ontologias (KDO04), Pisa, vol. 220, Citeseer, 2004.

[167] P. Ristoski, H. Paulheim, análise visual de dados estatísticos sobre mapas usando dados abertos vinculados, em: The 12th Extended Conferência na Web (ESWC2015). URL http://data.semanticweb.org/conference / eswc / 2015 / paper / demo / 12

[168] M. d'Aquin, N. Jay, Interpretando os resultados da mineração de dados com 3 4

dados vinculados para análise de aprendizado: motivação, estudo de caso e instruções, em: Anais da Terceira Conferência Internacional sobre Análise e Conhecimento de Aprendizagem, LAK '13, ACM, Nova York, NY, EUA, 2013, pp. 155-164. doi: 10.1145 / 2460296.2460327.
URL http://doi.acm.org/10.1145/2460296.2460327

[169] N. Jay, M. dAquin, dados vinculados e classificações on-line para ou padrões de mineração de dados em pacientes, em: AMIA Annual Symposium Proceedings, vol. 2013, American Medical Informatics Association, 2013, p. 681

[170] I. Tiddi, Explicando padrões de dados usando o conhecimento de fundo a partir de dados vinculados (2013).

[171] I. Tiddi, M. d'Aquin, E. Motta, Explicando clusters com programação lógica indutiva e dados vinculados, em: Proceedings da Trilha de Cartazes e Demonstrações da ISWC 2013, Sydney, Austrália, 23 de outubro de 2013, 2013, pp. 257-260.
URL http://ceur-ws.org/Vol-1035/iswc2013_poster_20.pdf

[172] I. Tiddi, M. dAquin, E. Motta, Dedalo: Procurando clusters ex em um labirinto de dados vinculados, em: V. Presutti, C. dAmato, F. Gandon, M. dAquin, S. Staab, A. Tordai (Eds.), The Web Semântica: Tendências e Desafios, vol. 8465 da Palestra Notas em Ciência da Computação, Springer International Publishing, 2014, pp. 333-348. doi: 10.1007 / 978-3-319-07443-6_23.
URL http://dx.doi.org/10.1007/978-3-319-07443-6_23

[173] I. Tiddi, M. d'Aquin, E. Motta, Andando dados vinculados: um gráfico abordagem transversal para explicar os agrupamentos, em: 5º Workshop Internacional sobre Consumo de Dados Vinculados (COLD 2014) co-localizado com o 13º International Semantic Conferência na Web (ISWC 2014), Riva del Garda, Itália, outubro 20, 2014., 2014.
URL http://ceur-ws.org/Vol-1264/cold2014_TiddiDM.pdf

[174] I. Tiddi, M. dAquin, E. Motta, Usando redes neurais para agregar - 18–36. regras de dados vinculados a porta, em: K. Janowicz, S. Schlobach, P. Lambrix, E. Hyvnen (Eds.), Engenharia do conhecimento e conhecimento edge Management, vol. 8876 de Notas de aula em Computador Science, Springer International Publishing, 2014, pp. 547–562. doi: 10.1007 / 978-3-319-13704-9_41.
URL http://dx.doi.org/10.1007/978-3-319-13704-9_41

[175] W. Klsgen, descoberta de conhecimento em bancos de dados e dados mineração, em: Z. Rath, M. Michalewicz (Eds.), Foundations of Sistemas Inteligentes, vol. 1079 de Notas de aula em computador Science, Springer Berlin Heidelberg, 1996, pp. 623–632. doi: 10.1007 / 3-540-61286-6_186.
URL http://dx.doi.org/10.1007/3-540-61286-6_186

[176] I. Trajkovski, N. Lavra6, J. Tolar, Segs: pesquisa de conjuntos de genes em dados de microarrays, Journal of biomedical informatics 41 (4) (2008) 588–601.

[177] V. Podpecan, N. Lavrac, I. Mozetic, PK Novak, I. Trajkovski, L. Langohr, K. Kulovesi, H. Toivonen, M. Petek, H. Motaln, K. Gruden, fluxos de trabalho Segmine para dados semânticos de microarranjos análise em orange4ws, BMC Bioinformatics (2011) 416–416.

[178] L. Eronen, H. Toivonen, Biomine: predição de ligações entre 2014, pp. 129-143. entidades biológicas usando modelos de rede de métodos heterogêneos bancos de dados, BMC bioinformtics 13 (1) (2012) 119.

[179] PK Novak, A. Vavpeti6, I. Trajkovski, N. Lavra6, N .: Topara mineração de dados semânticos com g-segs, em: In: Proceedings da 11ª Sociedade Internacional de Informação Multiconferência (IS, 2009.

[180] N. Lavra6, A. Vavpeti6, L. Soldatova, I. Trajkovski, PK Novak, Utilizando ontologias na mineração de dados semânticos com segs e g-segs, em: Anais da 14ª Conferência Internacional em Discovery Science, DS'11, Springer-Verlag, Berlim, Heidelberg, 2011, pp. 165–178.
URL http://dl.acm.org/citation.cfm?id=2050236.2050251

[181] A. Vavpeti6, N. Lavra6, sistemas de descoberta de subgrupos semânticos e fluxos de trabalho no sdm-toolkit, Comput. J. (2013) 304-320.

[182] A. Vavpeti6, PK Novak, M. Gr6ar, I. Mozeti, N. Lavra6, Semineração de dados mantic de artigos de notícias financeiras, em: J. Frnkranz, E. Hllermeier, T. Higuchi (Eds.), Discovery Science, vol. 8140 Notas de Aula em Ciência da Computação, Springer Berlin Heidelberg, 2013, pp. 294-307. doi: 10.1007 / 978-3-642-40897-7_20.
URL http://dx.doi.org/10.1007/978-3-642-40897-7_20

[183] N. Lavra6, PK Novak, Mineração de dados relacionais e semânticos para pesquisa biomédica (2012).

[184] A. Vavpeti6, V. Podpe˘can, N. Lavra6, subgrupo semântico plantações, Journal of Intelligent Information Systems 42 (2) (2014) 233–254. doi: 10.1007 / s10844-013-0292-1.
URL http://dx.doi.org/10.1007/s10844-013-0292-1 1

[185] B. Schäfer, P. Ristoski, H. Paulheim, O que é especial sobre Belém, Pensilvânia? identificação de fatos inesperados sobre entidades dbpedia, em: Conferências Internacionais Semânticas da Web, Cartazes e demonstrações, 2015.

[186] H. Paulheim, R. Meusel, A Decomposição do Declínio Outlier Problema de proteção em um conjunto de problemas de aprendizado supervisionado, Aprendizado de máquina (2-3) (2015) 509-531.

[187] R. Srikant, R. Agrawal, Regras gerais de associação de mineração, em: VLDB, vol. 95, 1995, pp. 407-419.

[188] X. Zhou, J. Geller, Raising, para aprimorar a mineração de regras na web marketing com o uso de uma ontologia, Data Mining with Ontologias: Implementações, Resultados e Estruturas (2007)

[189] MA Domingues, SO Rezende, Usando taxonomias para facilitar facilitar a análise das regras de associação, a pré-impressão do arXiv arXiv: 1112.1734.

[190] C. Marinica, F. Guillet, pós-administração interativa baseada no conhecimento regras de associação usando ontologias, conhecimento e dados Engineering, IEEE Transactions em 22 (6) (2010) 784–797.

[191] Z. Huang, H. Chen, T. Yu, H. Sheng, Z. Luo, Y. Mao, Semântica. mineração de texto com dados vinculados, em: INC, IMS e IDC, 2009. NCM '09. Quinta Conferência Conjunta Internacional de 2009, pp. 338-343. doi: 10.1109 / NCM.2009.131.

[192] V. Svátek, J. Rauch, M. Ralbovská, Associações aprimoradas em ontologia. mineração de ciação, em: M. Ackermann, B. Berendt, M. Grobelnik, A. Hotho, D. Mladeni6, G. Semeraro, M. Spiliopoulou, G. Stumme, V. Svátek, M. van Someren (Eds.), Semântica, Web e Mineração, vol. 4289 de Notas de aula em Computer Sci - Springer Berlin Heidelberg, 2006, pp. 163-179. doi: 10.1007 / 11908678_11.
URL http://dx.doi.org/10.1007/11908678_11

[193] T. Di Noia, I. Cantador, VC Ostuni, habilitado para dados abertos vinculados sistemas de recomendação: desafio da Eswc 2014 nas recomendações de livros recomendação, em: Semantic Web Evaluation Challenge, Springer, 2014, pp. 129-143.

[194] T. Di Noia, R. Mirizzi, VC Ostuni, D. Romito, M. Zanker, Dados abertos vinculados para oferecer suporte a sistemas de recomendação baseados em conteúdo em: Anais da 8ª Conferência Internacional sobre Semantic Systems, I-SEMANTICS '12, ACM, Nova York, NY, EUA, 2012, pp. 1-8. doi: 10.1145 / 2362499.2362501.
URL http://doi.acm.org/10.1145/2362499.2362501