Um estudo sobre a interação entre Mineração de Dados e Ontologias

Cássio Oliveira Camilo *João Carlos da Silva*

Technical Report - RT-INF_002-09 - Relatório Técnico
August - 2009 - Agosto

Instituto de Informática
Universidade Federal de Goiás
www.inf.ufg.br

Um estudo sobre a interação entre Mineração de Dados e Ontologias

Cássio Oliveira Camilo [*] cassio@inf.ufg.br

João Carlos da Silva [†] jcs@inf.ufg.br

Abstract. *The first work on the use of ontologies in data mining is dated from the beginning of 2000. Since then, several studies have been done on how we can solve some problems related to these technologies, such as improving the mined association rules, how to better define the extent of similarity between groups, how to embed the knowledge acquired during the mining process, how to automating the choice of better algorithms, among others. In this article we examine how this interaction is being made recently.*

Keywords: Ontology, Data Mining.

Resumo. *Os primeiros trabalhos relacionados com o uso de ontologias na mineração de dados é datada do início do ano de 2000. Desde então, diversos estudos têm sido feitos sobre como podemos resolver certos problemas ligados a essas tecnologias, tais como melhorar as regras de associação mineradas, definir melhor a medida de similaridade entre agrupamentos, inserir o conhecimento adquirido nas fases da mineração, automatizar a escolha dos melhores algoritmos, entre outros. Neste artigo, analisaremos como essa interação tem sido feita recentemente.*

Palavras-Chave: Ontologia, Mineração de Dados.

1 Introdução

Duas tecnologias têm se destacado nos últimos anos: a Mineração de Dados e as Ontologias. A Mineração de Dados visa identificar padrões em grandes volumes de dados e está ligada, entre outras, as técnicas estatísticas e de banco de dados. As Ontologias são uma forma de representação do conhecimento e tem uma ligação íntima com a Web Semântica.

Apesar de não existir uma dependência explícita entre estas duas tecnologias, alguns estudos tem demonstrado que uma interação entre elas pode trazer bons resultados. Iremos apresentar, através de estudos recentes, como a interação pode ocorrer e quais resultados têm sido obtidos.

O restante do artigo esta organizado da seguinte forma: na Sessão 2 relembramos o conceito de Mineração de Dados, na Sessão 3 o conceito de Ontologia e na Sessão 4 apresentamos alguns estudos de como a Mineração de Dados e as Ontologias estão sendo usadas, classificando-as quanto aos problemas que visam resolver.

[*] Mestrando em Ciência da Computação - INF/UFG

[†] Orientador - INF/UFG

2 Mineração de Dados

Considerada como uma área interdisciplinar a Mineração de Dados, do inglês *Data Mining*, teve grande influência de três outras áreas: Estatística, Banco de Dados e Aprendizado de Máquina. Destas áreas tem-se as principais definições para o termo:

- Em Hand et al. [25], a definição é dada de uma perspectiva estatística: "Mineração de Dados é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensível ao dono dos dados".
- Em Cabena et al. [6], a definição é dada de uma perspectiva de banco de dados: "Mineração de Dados é um campo interdisciplinar que junta técnicas de máquinas de conhecimentos, reconhecimento de padrões, estatísticas, banco de dados e visualização, para conseguir extrair informações de grandes bases de dados".
- Em Fayyad et al. [16], a definição é dada da perspectiva do aprendizado de máquina: "Mineração de Dados é um passo no processo de Descoberta de Conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados."

2.1 Tarefas

Larose [35] classifica a Mineração de Dados de acordo com as tarefas que podem ser realizadas. As tarefas mais comuns são:

Descrição (*Description*) É a tarefa utilizada para descrever os padrões e tendências reveladas pelos dados. A descrição geralmente oferece uma possível interpretação para os resultados obtidos. A tarefa de descrição é muito utilizada em conjunto com as técnicas de análise exploratória de dados, para comprovar a influência de certas variáveis no resultado obtido.

Classificação (*Classification*) Uma das tarefas mais comuns, a Classificação visa identificar a qual classe um determinado registro pertence. Nesta tarefa, o modelo analisa o conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertence, afim de 'aprender' como classificar um novo registro (aprendizado supervisionado). Por exemplo, categorizamos cada registro de um conjunto de dados contendo as informações sobre os colaboradores de uma empresa: Perfil Técnico, Perfil Negocial e Perfil Gerencial. O modelo analisa os registros e então é capaz de dizer em qual categoria um novo colaborador se encaixa.

Estimação (*Estimation*) ou Regressão (*Regression*) A estimação é similar à classificação, porém é usada quando o registro é identificado por um valor numérico e não um categórico. Assim, pode-se estimar o valor de uma determinada variável analisando-se os valores das demais. Por exemplo, um conjunto de registros contendo os valores mensais gastos por diversos tipos de consumidores e de acordo com os hábitos de cada um. Após ter analisado os dados, o modelo é capaz de dizer qual será o valor gasto por um novo consumidor.

Predição (*Prediction*) A tarefa de predição é similar às tarefas de classificação e estimação, porém ela visa descobrir o valor futuro de um determinado atributo. Alguns métodos de classificação e regressão podem ser usados para predição, com as devidas considerações.

Agrupamento (*Clustering*) A tarefa de agrupamento visa identificar e aproximar os registros similares. Um agrupamento (ou *cluster*) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Esta tarefa difere da classificação pois não necessita que os registros sejam previamente categorizados (aprendizado não-supervisionado). Além disso, ela não tem a pretensão de classificar, estimar ou predizer o valor de uma variável. Ela apenas identifica os grupos de dados similares.

Associação (*Association*) A tarefa de associação consiste em identificar quais atributos estão relacionados. Apresentam a forma: *SE* atributo X *ENTÃO* atributo Y. É uma das tarefas mais conhecidas devido aos bons resultados obtidos, principalmente nas análises da "Cestas de Compras"(*Market Basket*), onde identificamos quais produtos são levados juntos pelos consumidores.

2.2 Ferramentas

Atualmente, diversas ferramentas podem ser encontradas para auxiliar no processo de Mineração de Dados, algumas são livres (*open source*) e outras são proprietárias. Na grande maioria dos trabalhos apresentados nesse artigo, é utilizado o software WEKA [59].

O WEKA (*Waikato Environment for Knowledge Analysis*) implementa uma série de algoritmos para as tarefas de mineração. Os algoritmos podem ser aplicados diretamente na ferramenta, ou utilizados por programas Java através da sua API. Fornece as funcionalidades para pré-processamento, classificação, regressão, agrupamento, regras de associação e visualização. Em [61] a ferramenta é apresentada em detalhes.

3 Ontologia

O termo ontologia tem origem na Filosofia e é relativo à existência do ser [19]. Em [42], outros significados ainda no campo da filosofia podem ser encontrados. Entretanto, no campo da computação o termo possui outro significado. Ainda não há consenso na definição formal do termo ontologia [22]. Iremos adotar aqui uma das definições mais referenciadas. Definido por Gruber[19] [20], inicialmente no campo da Inteligência Artificial, como "a especificação explícita de um conceito"e mais recentemente no contexto da Ciência da Computação como [21]:

> "... No contexto da Ciência da Computação e da Informação, uma ontologia define um conjunto de representações primitivas com o qual se modela um domínio de conhecimento ou discurso..."

Apesar de não haver um consenso sobre o termo, os diversos autores concordam que as ontologias são um meio de permitir o compartilhamento e a reutilização do conhecimento. Em [23], Guarino apresenta uma definição formal para o conceito de Ontologia.

Diversas são as áreas de aplicação das ontologias (Recuperação de Informações, Gestão do Conhecimento, Educação, Processamento de Linguagem Natural, Mineração de Dados, entre outras). Dentre elas, a Web Semântica tem obtido grandes avanços. Atualmente, as ontologias são uma recomendação do W3C [60] para a Web Semântica [54] como padrão de

vocabulário comum para a troca de dados, tornar o conhecimento reutilizável e facilitar a comunicação de sistemas heterogêneos [21].

3.1 Tipos de Ontologias

Segundo [17] e [22] podemos classificar as ontologias quanto a sua função em:

Ontologias Genéricas São ontologias que descrevem conceitos mais amplos. Não dependem de um problema específico (domínio). Geralmente representam conceitos da natureza, relativos ao espaço e tempo.

Ontologias de Domínio Representam conceitos de um domínio (área) específico, tais como: medicina, genética, computação. São os tipos de ontologias mais comuns.

Ontologias de Aplicação Descrevem conceitos que estão relacionados com um domínio específico (área) e com uma tarefa específica.

Ontologias de Representação Descrevem os conceitos que são usados para a representação do conhecimento.

Ontologias de Tarefa Descrevem conceitos que são usados por processos (tarefas e atividades) de uma maneira geral, sem dependência com um domínio específico. Por exemplo, processo de compra e venda.

Em [1], é feito um resumo sobre as diversas classificações encontradas na literatura.

3.2 Linguagens

As Ontologias precisam ser descritas em alguma linguagem, para que possam então ser processada pelas máquinas. Existem diversas linguagens para a representação das Ontologias. Iremos destacar nesse trabalho a linguagem OWL, devido à sua grande utilização nos trabalhos pesquisados, além de ser atualmente a recomendação da W3C, como a linguagem padrão para a representação das Ontologias.

As ontologias OWL podem ser classificadas em três espécies, de acordo com a sub-linguagem utilizada: OWL-Lite, OWL-DL e OWL-Full. A característica principal de cada sub-linguagem é a sua expressividade: a OWL-Lite é a menos expressiva, a OWL-Full é a mais expressiva e a expressividade da OWL-DL está entre a OWL-Lite e a OWL-Full.

OWL-Lite A OWL-Lite é a sub-linguagem sintaticamente mais simples. Destina-se a situações em que apenas são necessárias restrições e uma hierarquia de classe simples. Por exemplo, o OWL-Lite pode fornecer uma forma de migração para tesauros existentes, bem como de outras hierarquias simples.

OWL-DL A OWL-DL é mais expressiva que a OWL-Lite e baseia-se em lógica descritiva, um fragmento de Lógica de Primeira Ordem, passível portanto de raciocínio automático. É possível assim computar automaticamente a hierarquia de classes e verificar inconsistências na ontologia. Este tutorial utiliza a OWL-DL.

OWL-Full A OWL-Full é a sub-linguagem mais expressiva. Destina-se a situações onde alta expressividade é mais importante do garantir a decidibilidade ou completeza da linguagem. Não é possível efetuar inferências em ontologias OWL-Full.

3.3 Ferramentas

Dentre as diversas ferramentas disponíveis para se trabalhar com as Ontologias, destacamos o *Protégé* e o Jena, devido à sua grande utilização nos trabalhos apresentados.

O Protégé [32] é uma ferramenta livre que permite a criação de Ontologias, incluindo o formato OWL. Possui um Mecanismo de Inferência (*reasoner*) baseado em lógica descritiva para verificar a consistência da Ontologia e para computar a hierarquia das classes. Um tutorial da ferramenta pode ser encontrado em [27].

O Jena [34] é um *framework* Java, voltado para a construção de aplicações da *Web* Semântica. Ele possui um ambiente de programação para os ambientes RDF, RDFS, OWL, SPARQL e inclui uma *engine* para realizar inferências.

4 Mineração de Dados e Ontologias

As Ontologias, introduzidas na Mineração de Dados pela primeira vez no começo de 2000, podem ser usadas de diferentes formas [44]: Ontologias de Domínio e de Conhecimento, Ontologias para o processo de Mineração ou Ontologias de Metadados. A primeira, organiza os conhecimentos sobre um domínio e desempenha um papel importante no processo de mineração. A segunda, representa a descrição do processo de mineração e auxilia na escolha da melhor tarefa para o problema. A terceira, armazena o conhecimento sobre os itens e os relacionamentos que compõem uma ontologia. A figura 1, representa um *framework* geral para esse processo. O processo começa com a extração dos dados a serem minerados, podendo para isso, usar um *data warehouse* ou outras fontes de dados. Esses dados servem de subsídio para a ontologia de metadados. Com os dados selecionados, uma ontologia de domínio pode ser usada para preparar os dados. Em seguida, os algoritmos de mineração são aplicados. Uma ontologia para a mineração de dados pode ser usada. O resultado do algoritmo pode ser utilizado para a visualização ou para a tomada de decisão.

Neaga [43], apresenta um estudo que visa contribuir para a unificação das metodologias existentes nas áreas de mineração e da engenharia de ontologias, de forma que uma possa se beneficiar da outra e uma interação bidirecional possa existir.

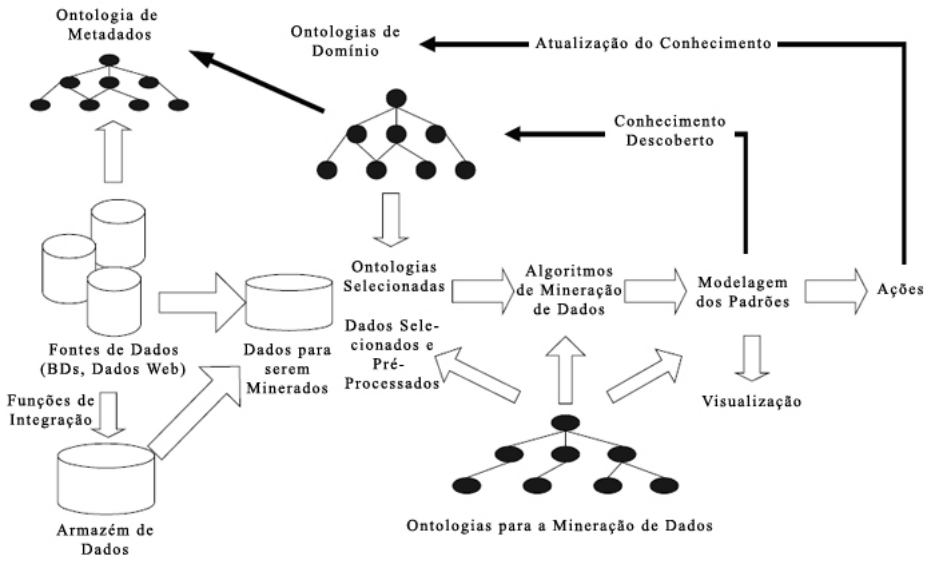

Figura 1: *Framework* geral para a integração entre Mineração de Dados e Ontologia [44]

4.1 Geração de Ontologias

Criar, manter e evoluir Ontologias de forma automatizada são alguns dos principais campos de pesquisa da atualidade. A ideia da geração de todas as ontologias de forma manual não é muito bem aceita para uma aplicação real, onde o domínio sobre um assunto é muito vasto. Além disto, essas ontologias precisam ser combinadas. Assim, cria-se um cenário em que a Mineração de Dados pode auxiliar neste processo, como apresentado nos estudos seguintes.

Um abordagem para a geração de uma ontologia de conceito, através da mineração de textos é apresentada em [12]. O método proposto inicialmente utiliza-se do conceito de *Rought Set* e por fim aplica as técnicas de agrupamento para a definição da hierarquia de conceitos.

Em Kiu et al. [31], é proposto um *framework* chamado OntoDNA. Concebido inicialmente para o uso em sistemas de aprendizado, pode ser estendido para outros domínios. Em suma, o OntoDNA utiliza a técnica de *Formal Concept Analysis - FCA* para capturar os atributos da ontologia, o *Self-Organizing Map(SOM)* e o *k-means* para minerar e medir a similaridade semântica entre os conceitos da ontologia. O resultado é uma ontologia compartilhada. Experimentos demostraram que o OntoDNA conseguiu uma precisão de 94.55%, comparados com a montagem da ontologia de forma manual.

Liu et al [39], propõem uma maneira de auxiliar na resolução do problema da evolução automática das ontologias. A proposta é utilizar as técnicas de mineração para coletar informações de *thesaurus*, dicionários e textos.

Elsayed et al. [14], propõem um sistema para a geração de ontologias baseando-se na árvore de decisão obtida pelo algoritmo C4.5. A ontologia é construída na linguagem OWL [58] e é voltada para os domínios específicos: doenças da soja e doenças animais.

Uma abordagem usando a geração de ontologias através da base de dados e a incorporação de metadados durante o processo de mineração é proposto em [36]. Inicialmente, o processo usa o *plugin* da ferramenta Protégé [32], chamado DataGenie, para analisar a estrutura do banco de dados e gerar a ontologia de forma automática. Em seguida, através da linguagem OWL, são realizadas inferências sobre esta ontologia. O resultado é o conhecimento para o processo de mineração.

Em [7], é apresentado um método chamado RTAXON, para geração de ontologias a partir de bases de dados relacionais. O processo consiste basicamente de três passos: normalização da base de dados, aprendizado das classes e das propriedades e a população da ontologia. A abordagem proposta combina as técnicas clássicas de análise dos metadados e a identificação de padrões dos dados através de uma mineração hierárquica. O método proposto foi verificado em diversas bases e obteve resultados satisfatórios.

4.2 Regras de Associação

Um dos grandes desafios para as regras de associação, geradas por alguns algoritmos de mineração, é produzir resultados que sejam úteis para a tomada de decisão. A tarefa de mineração comumente costuma produzir um número elevado de regras, o que algumas vezes inviabiliza uma análise eficiente. Os trabalhos seguintes propõem algumas soluções para este problema.

Em [28], um algoritmo que gera regras de associação de auto-nível, através da mineração de regras de associação de baixo-nível usando ontologias é apresentado. O algoritmo é aplicado sobre um repositório de uma companhia de energia, para auxiliar na detecção de falhas. O uso da ontologia na geração de regras de auto-nível é importante pois regras de baixo-nível são mais difíceis de serem interpretadas e podem produzir regras desnecessárias, além disso, o volume de regras produzidas também será menor.

Farzanyar et al. [15], propõem um novo algoritmo que combina Lógica Nebulosa e Ontologias. O algoritmo proposto, funciona basicamente em três passos. Inicialmente, o algoritmo analisa os registros e os classifica de acordo com os conceitos da ontologia. A relação semântica entre esses conceitos presente na ontologia, é chamada de *Meta Rules*, e é ela que define a relação semântica entre as classes. O próximo passo consiste em recuperar os itens mais frequentes. Porém, ao invés da maneira tradicional, apenas os itens que possuem relações semânticas definidas na ontologia são considerados. Por fim, as regras de associação são recuperadas do conjunto de itens frequentes obtido. Segundo os autores, os resultados obtidos são mais interessantes e compreensivos.

Pan e Pan [49], propõem um repositório central de ontologias de diversos domínios, chamado Ontobase, para a integração, o compartilhamento e o reuso de forma genérica. O *framework* utiliza o protocolo XMI para comunicação e pode ser usado para a mineração de regras de associação, mineração sequencial e classificação. O estudo ressalva que a ferramenta oferece uma alternativa ao serviço de ontologias da FIPA, e que permite aos sistemas multiagentes usarem ontologias, com a vantagem de não depender do protocolo OKBC.

Xuping et al. [64], propõem uma variação no algoritmo Apriori para a mineração de regras de associação baseada em Ontologias, chamado de ARMO (*Association Rules Mining based on Ontology*), sobre base de dados de informações de *e-commerce*. A variação envolve inicialmente o uso das ontologias para geração de um número menor de regras e depois a aplicação da mineração em múltiplos níveis.

Hong et al. [26] propõem um sistema que minera as regras de associação e as representa através da linguagem OWL (*Web Ontology Language*) [58]. O sistema é composto de 5 subsistemas: conversor de consultas, sistema de inferência de regras, sistema de gerenciamento de ontologias, sistema de geração de conhecimento e o sistema de gerenciamento de conhecimento. A ideia é permitir que as regras de associação mineradas possam ser disponibilizadas na internet através da linguagem OWL [58]. O conversor de consultas, é a entrada do sistema. A consulta do usuário é convertida para o formato OWL-QL (*OWL Query Language*).O subsistema de inferência de regras tenta extrair as respostas certas para a consulta desejada. O subsistema de gerenciamento de ontologias, é responsável por construir e gerenciar as ontologias de domínios que estão relacionadas com os dados armazenados em uma base transacional orientada a objetos. O subsistema de geração de conhecimento é o responsável por gerar as regras de associação da base objeto relacional. Por fim, o subsistema de gerenciamento de conhecimento é responsável por controlar as regras de associação e convertê-las para o formato OWL [58].

Marinica et al. [40], apresentam uma abordagem focada na fase de pós-processamento. O conhecimento do usuário é modelado usando uma ontologia, que também estará relacionada com os dados. O artigo propõe a criação de um Esquema de Regras (*Rule Schema*) que contém as expectativas do usuário, e alguns operadores são propostos para auxiliar a tomada de decisão do usuário. Este interage com as regras geradas utilizando a ontologia e os operadores criados.

Em [57], é proposta uma medida de Distância Semântica (DS), que faz o uso de ontologias, para realizar a filtragem dos itens extraídos nas tarefas de mineração de regras de associação.

4.3 Expansão de Consultas

Um dos grandes problemas que temos atualmente na busca por informações é relativo à maneira como os usuários escrevem o que desejam. Certos conceitos podem ter diferentes significados, ou sinônimos, e podem estar presentes em certos documentos e em outros não.

Assim, o espaço de busca do usuário, fica comprometido. Para auxiliar a resolução desses problemas, as Ontologias têm sido usadas como fontes de expansão de consultas, para que os dados possam ser então, minerados.

Em sua tese de doutorado, Wives [62] propõe a utilização de conceitos para a realização dos agrupamentos (*clustering*). Os conceitos extraídos dos documentos são representados através de Ontologias.

Em Barth et al. [2], a consulta a uma base de dados criminal contendo dados estruturados e não estruturados é expandida por meio de uma ontologia de domínio. Os documentos recuperados são submetidos a algoritmos de agrupamento. O trabalho conclui que a busca de informações contextualizada, com o uso da ontologia, possui um desempenho superior se comparada com o processo de recuperação tradicional.

Em [3], é feita a proposta do uso de Ontologias para a expansão de consultas de uma maneira geral.

4.4 Ontologias de Mineração

O resultado do processo da Mineração depende da aplicação do algoritmo correto. A escolha de tais algoritmos ainda é uma tarefa complexa e muitas vezes necessita de um especialista, o que torna o uso da mineração restrito. As Ontologias tem sido usadas para representar o conhecimento sobre o processo de mineração, de forma que as escolhas possam ser automatizadas, como demonstra os estudos a seguir.

Um dos trabalhos pioneiros na criação de sistemas de auxilio ao processo de mineração é proposto em [4]. O artigo apresenta o IDA (*Intelligent Discovery Assistant*) como forma de auxiliar na tarefa de escolha das melhores técnicas de mineração. O IDA enumera e sugere as várias técnicas possíveis utilizando uma ontologia de mineração como referência. Uma outra proposta de um assistente inteligente é feita em [8]. Inicialmente, as informações básicas de um processo de mineração são capturadas (o trabalho utiliza o modelo CRISP-DM). Depois, o sistema gera as recomendações para os usuários não especialistas nas atividades de mineração utilizando-se da técnica de CBR (*Case-Based Reasoning*). A proposta apresentada, no entanto, refere-se somente às tarefas de classificação.

Lin et al. [38], propõem a criação de uma ontologia, chamada *mining ontology*, que contém o conhecimento sobre as técnicas de mineração e é base para a escolha de qual algoritmo usar. Para a validação da proposta foi utilizado o software WEKA [59] e várias bases de dados. Uma outra abordagem, proposta por Liang et al. [37], utiliza o software Protégé [32] para criar a ontologia, o arcabouço Jena [34] para trabalhar com a API OWL e a *engine* RACER [56] para realizar as inferências.

Zagoruikot et al. [66], é proposta a primeira versão de uma ontologia de mineração de dados voltada para o "domínio de assuntos". O domínio de assuntos está relacionado com a descoberta de regularidade em vetores de dados estatísticos ou experimentais e o uso dessa regularidade para solucionar problemas de reconhecimento de padrões, predição, agrupamento e outros que exigem uma analise profunda. A ontologia possui classes que representam o conhecimento tanto da fase de pré-processamento, quanto de pós-processamento.

A proposta de Panov et al. [50] é construir uma ontologia de mineração de dados chamada OntoDM. Diferente de outras ontologias, os autores definem a OntoDM como uma ontologia profunda e pesada (com um alto grau de formalidade) e é construída seguindo as recomendações da Engenharia de Ontologias. A proposta apresentada é base para o projeto de construção de um *framework* geral de mineração de dados. A OntoDM foi desenvolvida para um propósito geral. A linguagem atual usada é a OWL-DL.

Em [48], é proposta uma ontologia para a Mineração de Textos, chamada MT-Ontology. Essa ontologia estende os conceitos de uma ontologia de Mineração de Dados, chamada DAMON, e incorpora elementos específicos da Mineração de Textos.

4.5 Recuperação de Informações

O tratamento da informação é uma atividade primordial para a grande maioria das organizações. Usar de forma adequada a informação obtida, alinhando-a com a necessidade do usuário, é cada vez mais necessário. A Mineração e as Ontologias têm sido aplicadas nesse sentido, como vemos nos trabalhos relacionados a seguir.

Medina et al. [41] propõem o uso de ontologias leves (com pouco formalismo) para auxiliar na recuperação de informações de múltiplas fontes. A comunicação com as diversas fontes é feita através do protocolo OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*). Sobre os documentos recuperados, o algoritmo FIHC (*Frequent Itemset Hierarchical Clustering*) é aplicado, gerando uma árvore de grupos. A partir dos grupos gerados, uma Ontologia é construída e armazenada em formato XML. É com base nesta ontologia que os documentos mais relevantes são recuperados.

Robal e Kalja [52] propõem a aplicação de técnicas de mineração da internet e o uso de ontologias para sugerir melhorias em um determinado sítio, otimizando a navegabilidade dos usuários, baseando-se nas experiências já adquiridas de outros usuários. As recomendações são feitas aos usuários em forma de outras sessões que ele possa visitar, ou então, destacando certo texto na página em que se encontra. Inicialmente, é realizada a mineração nos arquivos de registros de acesso (*log*), e com base no comportamento do usuário o padrão do usuário é extraído. Esse padrão é combinado com a ontologia da internet para dar sentido aos dados minerados.

Uma outra abordagem, ainda na linha da mineração da internet, foi proposta em [63]. Nela, os autores criaram um *framework* de aprendizado não supervisionado para criar a ontologia de um sítio específico baseando-se em várias páginas deste sítio. Além do uso de várias páginas para a construção da ontologia, os autores destacam que a abordagem proposta utiliza-se de uma inferência gramatical estocástica para aprender sobre a regularidade dos conceitos presentes nos formatos dos *layouts* e com isso gerar a estrutura hierárquica da ontologia.

Grande parte dos sistemas de serviço de informações desenvolvidos falham na personalização do serviço e na falta de semântica das informações, o que causa um serviço de má qualidade. Em [18], é proposto um sistema de serviço de informação que usa as ontologias e a mineração de regras de associação para propor um sistema cooperativo de recomendação que seja orientado ao conteúdo, trabalhando assim, com a semântica das informações. O sistema usa uma ontologia de domínio e o comportamento do usuário para gerar uma ontologia de interesse do usuário, utilizando posteriormente a técnica de mineração de regras de associação ponderada.

Em seu trabalho [29], Hui et al. propõem um modelo para descoberta de *Web Services* baseado em ontologias. A ideia central é descobrir de uma forma mais precisa quais dos serviços disponíveis na internet podem de fato auxiliar um usuário em suas necessidades. O trabalho propõe uma ontologia baseada na qualidade do serviço (QoS - *Quality of Service*) para garantir que a demanda do usuário seja atendida de forma correta. Em [11], é proposto o *WebS Composer* para realizar a localização e a composição de serviços que melhor atendam as necessidades dos usuários.

4.6 Outras aplicações

Diversas outras áreas que já faziam uso das técnicas de Mineração de Dados para otimizar seus resultados passaram também a combiná-las com as ontologias. A seguir, algumas dessas iniciativas são descritas.

Zhou, Jiang e Wang [67], utilizam a mineração de dados e as ontologias para melhorar o compartilhamento das informações dos sistema de PDM (*Product Data Management)*) de diferentes proprietários. A proposta utiliza a ferramenta Protégé [32] para criação de um modelo de ontologias de domínio dos sistemas PDMs. Após a geração das ontologias, é realizada a mineração desses dados.

Em [10], uma ontologia para vigilância de doenças contagiosas, usando várias línguas, é proposta. O estudo faz parte de um sistema de vigilância na região da Ásia e do Pacífico, chamado BioCaster. O sistema utiliza a mineração de textos para monitorar automaticamente notícias da internet e de outras fontes. O trabalho apresenta a ontologia chamada de BCO (*BioCaster Ontology*). Com ela é possível uma melhor análise e entendimento dos dados minerados além de possibilitar o sistema de fazer melhores inferências.

Kuo et al. em [33], propõem o uso de uma ontologia de domínio voltada para a mineração de regras de associação de uma base de dados contendo informações sobre o tratamento de pacientes que sofreram de doenças crônicas dos rins. O estudo propõe então que seja usada uma ontologia para entender e selecionar as variáveis adequadas, além de poder interpretar melhor as regras geradas. Para a realização do experimento, o trabalho utiliza a ontologia UMLS (*Unified Medical Language System*).

O trabalho de Youn et al. [65], utilizam as ontologias para auxiliar os filtros de *emails* a identificar mensagens que podem ser classificadas como *spam*. Inicialmente é utilizado o algoritmo J48, que realiza a classificação mediante a árvore de decisão criada pelo algoritmo C4.5, sobre uma base de treinamento. Após feita a classificação, uma ontologia representando-a é criada. Essa ontologia serve como base para o servidor de *email* filtrar as mensagens. O trabalho realizou os experimentos utilizando uma base obtida no *UCI Machine Learning Lab*, o software WEKA [59] para executar os algoritmos e o Jena [34] para trabalhar com as ontologias.

Em [5], é proposto o uso de ontologias para otimizar a mineração em base de dados multi-relacionais. A ideia central consiste em representar o modelo de dados por meio de uma ontologia e esta servir como base para os algoritmos de mineração.

No campo da mineração multimídia, um trabalho recente [9], mostra os resultados que estão sendo obtidos com a aplicação de uma ontologia (ainda em desenvolvimento). Em [51] é apresentado o *framework* MI-MERCURY, que utiliza agentes móveis para mineração de informações multimídia na internet, e as ontologias para a conceituação semântica das informações.

Um estudo sobre o estado da arte do uso de ontologias na mineração de dados de bases biomédicas é apresentado em [53]. O trabalho mostra as duas dificuldades encontradas: geração da ontologia e adaptação dos algoritmos de mineração para usarem a ontologia criada. O artigo discute as ontologias PO (*Protein Ontology*) e a GO (*Gene Ontology*) e como alguns algoritmos de mineração podem ser adaptados para utilizarem as ontologias.

A mineração de dados combinada com a ontologia tem sido recentemente usada em diversas áreas não convencionais. Em [24], é proposta um solução para a detecção de falhas em turbinas de vento. Em [47], elas são utilizadas para o gerenciamento de doenças e prescrições de medicamentos. Já em [45] e [46], o uso dessas técnicas se dá no campo do petróleo. Em [55], elas são usadas para o processamento de dados de sensores. Em [13], as técnicas são usadas em arquivos com dados de satélites de observação da terra. Em [30], o uso da mineração é voltada para a construção, manutenção e evolução, de forma automática, de uma ontologia de domínio para produtos de aviação.

5 Conclusão

Através dos diversos artigos citados, observa-se que a interação entre Mineração de Dados e Ontologias é fundamental para obtenção de melhores resultados. Agregar conhecimento e semântica ao processo de Mineração de Dados possibilita obter resultados mais claros e voltados para as necessidades específicas. Da mesma forma, utilizar a Mineração de Dados para automatizar a construção, manutenção e evolução das Ontologias torna-se cada vez mais necessário para o seu uso em aplicações reais.

6 Agradecimentos

Ao Prof. Dr. Cedric Luiz de Carvalho, pela avaliação do presente texto e pelas sugestões feitas, as quais muito contribuíram para a melhoria do texto original.

Referências

[1] ALMEIDA, M. B; BAX, M. P. **Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção**. Revista Ciência da Informação, 32(3):7–20, set./dez. 2003.

[2] BARTH, F. J; BELDERRAIN, M. C; QUADROS, N. L. P; FERREIRA, L. L; TIMOSZCZUK, A. P. **Recuperação e mineração de informações para área criminal**. VI Encontro Nacional de Inteligência Artificial - ENIA - XXVII SBC, 2007.

[3] BARTH, F. J; TIMOSZCZUK, A. P. **Expansão Automática de Consultas utilizando Ontologias**. SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL, 2008.

[4] BERNSTEIN, A; PROVOST, F; ; HILL, S. **Toward Intelligent Assistance for a Data Mining Process: An Ontology-Based Approach for Cost-Sensitive Classification**. TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2005.

[5] BORTOLETO, S; EBECKEN, N. F. F. **Ontology model for multi-relational data mining application**. International Conference on Intelligent Systems Design and Applications, 2008.

[6] CABENA, P; HADJINIAN, P; STADLER, R; JAAPVERHEES; ZANASI, A. **Discovering Data Mining: From Concept to Implementation**. Prentice Hall, 1998.

[7] CERBAH, F. **Mining the Content of Relational Databases to Learn Ontologies with Deeper Taxonomies**. International Conference on Web Intelligence and Intelligent Agent Technology, 2008.

[8] CHAREST, M; DELISLE, S; CERVANTES, O; SHEN, Y. **Intelligent Data Mining Assistance via CBR and Ontologies**. International Conference on Database and Expert Systems Applications, 2006.

[9] COLANTONIO, S; SALVETTI, O; GUREVICH, I. B; TRUSOVA, Y. **An ontological framework for media analysis and mining**. In: PATTERN RECOGNITION AND IMAGE ANALYSIS, volume 19, p. 221Ű230. Pleiades Publishing, 2009.

[10] COLLIER, N; KAWAZOE, A; JIN, L; SHIGEMATSU, M; DIEN, D; BARRERO, R. A; TAKEUCHI, K; KAWTRAKUL, A. **A multilingual ontology for infectious disease surveillance: rationale, design and challenges**. In: LANG RESOURCES & EVALUATION, p. 405Ű413. Springer Science+Business, 2006.

[11] DE ANDRADE, F. G. **Webs composer: Uma ferramenta baseada em ontologias para a descoberta e composição de serviços na web**. Master's thesis, UNIVERSIDADE FEDERAL DE CAMPINA GRANDE - UFCG, 2006.

[12] DEY, L; RASTOGI, A. C; KUMAR, S. **Generating Concept Ontologies Through Text Mining**. International Conference on Web Intelligence, 2006.

[13] DURBHA, S. S; KING, R. L. **Knowledge Mining in Earth Observation Data Archives: A Domain Ontology Perspective**. International Geoscience and Remote Sensing Symposium, 2004.

[14] ELSAYED, A.-E; EL-BELTAGY, S. R; RAFEA, M; HEGAZY, O. **Applying data mining for ontology building**. Conference On Statistics, Computer Science, and Operations Research, 2007.

[15] FARZANYAR, Z; KANGAVARI, M; HASHEMI, S. **A New Algorithm for Mining Fuzzy Association Rules in the Large Databases Based on Ontology**. International Conference on Data Mining -Workshops, 2006.

[16] FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. American Association for Artificial Intelligence, 1996.

[17] G. VAN HEIJS, A. T. S; WIELINGA, B. J. **Using Explicit Ontologies in KBS Development**. International Journal of Human and Computer Studies, 1996.

[18] GE, J; QIU, Y; CHEN, Z. **Cooperative Recommendation Based on Ontology Construction**. International Conference on Computer Science and Software Engineering, 2008.

[19] GRUBER, T. R. **What is an Ontology?** `http://www-ksl.stanford.edu/kst/what-is-an-ontology.html`, acessado em Maio de 2009, 1992.

[20] GRUBER, T. R. **Toward Principles for the Design of Ontologies Used for Knowledge Sharing**. International Journal Human-Computer Studies, p. 907–928, 1993.

[21] GRUBER, T. R. **Ontology**. In: Liu, L; Özsu, M. T, editors, ENCYCLOPEDIA OF DATABASE SYSTEMS. Springer-Verlag, 2008.

[22] GUARINO, N. **Understanding, Building, And Using Ontologies**. `http://ksi.cpsc.ucalgary.ca/KAW/KAW96/guarino/guarino.html`, acessado em Maio de 2009, 1996.

[23] GUARINO, N. **Formal Ontology and Information Systems**. Formal Ontology in Information Systems, 1998. In.

[24] GUO, Q; ZHANG, M. **A novel Approach for fault diagnosis of steam turbine based on neural network and genetic algorithm**. International Joint Conference on Neural Networks, 2008.

[25] HAND, D; MANNILA, H; SMYTH, P. **Principles of Data Mining**. MIT Press, 2001.

[26] HONG, T.-P; DONG, J.-S; LIN, W.-Y. **An Integrated OWL Data Mining and Query System**. IEEE International Conference on Systems, Man and Cybernetics - SMC, 2008.

[27] HORRIDGE, M; KNUBLAUCH, H; RECTOR, A; STEVENS, R; WROE, C. **A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tool**. The University Of Manchester, 2004.

[28] HOU, X; GU, J; SHEN, X; YAN, W. **Application of Data Mining in Fault Diagnosis Based on Ontology**. International Conference on Information Technology and Applications, 2005.

[29] HUI, Z; WEIYING, G. **A Research on QoS-based Ontology Model for Web Services Discovery**. International Workshop on Knowledge Discovery and Data Mining, 2009.

[30] JING, M; QINGQING, S; SIFENG, L. **Research on OWL-based Construction, Merging, Mapping and Evolution of Ontology in Aviation Product Domain**. International Conference on Grey Systems and Intelligent Services, 2007.

[31] KIU, C.-C; LEE, C.-S. **A Data Mining Approach for Managing Shared Ontological Knowledge**. International Conference on Advanced Learning Technologies, 2006.

[32] KNAUBLOCK, H. **Protégé-OWL**. `http://protege.stanford.edu/overview/protege-owl.html`, acessado em Junho de 2009, 2003.

[33] KUO, Y.-T; LONIE, A; SONENBERG, L; PAIZIS, K. **Domain Ontology Driven Data Mining - A Medical Case Study**. Workshop on Domain Driven Data Mining, 2007.

[34] LABS, H. **Jena: A semantic web framework**. `http://jena.sourceforge.net/`, acessado em Junho 2009.

[35] LAROSE, D. T. **Discovering Knowledge in Data: An Introduction to Data Mining**. John Wiley and Sons, Inc, 2005.

[36] LI, G; SHENG, H; FAN, X. **Incorporating Metadata into Data Mining with Ontology**. Institute of Electronics, Information and Communication Engineers, 2007.

[37] LIANG, Z; XUEMING, L. **An Ontology Reasoning Architecture for Data Mining Knowledge Management**. Wuhan University Journal of Natural Sciences, 2008.

[38] LIN, M.-S; ZHANG, H; YU, Z.-G. **An Ontology for Supporting Data Mining Process**. IMACS Multiconference on Computational Engineering in Systems Applications (CESA), 2006.

[39] LIU, Y; CHEN, X; SUI, Z. **Intelligent Information Systems and Data Mining Study on Evolution of Domain Ontology**. Innovative Computing, Information and Control, 2007.

[40] MARINICA, C; GUILLET, F; BRIAND, H. **Post-Processing of Discovered Association Rules Using Ontologies**. International Conference on Data Mining Workshops, 2008.

[41] MEDINA, M. A; SŤANCHEZ, J. A; PAZ, J. A. **Document Retrieval from Multiple Collections by using Lightweight Ontologies**. International Conference on Computing, 2006.

[42] MORAIS, E. A. M; AMBRÓSIO, A. P. L. **Ontologias: conceitos, usos, tipos, metodologias, ferramentas e linguagens**. Technical report, Universidade Federal de Goiás, 2007.

[43] NEAGA, E. I. **Semantics enhancing knowledge discovery and ontology engineering using mining techniques: A crossover review**. In: KNOWLEDGE DISCOVERY AND DATA MINING: CHALLENGES AND REALITIES, p. 163Ű188. Information science reference, 2007.

[44] Nigro, H. O; Císaro, S. E. G; Xodo, D. H, editors. **Data Mining with Ontologies: Implementations, Findings, and Frameworks**. Information Science Reference, 2007.

[45] NIMMAGADDA, S. L; DREHER, H. **Ontology Based Data Warehouse Modelling Ű a Methodology for Managing Petroleum Field Ecosystems**. International Conference on Digital Ecosystems and Technologies, 2008.

[46] NIMMAGADDA, S. L; DREHER, H. **Petroleum Ontology: an effective data integration and mining methodology aiding exploration of commercial petroleum plays**. International Conference on Industrial Informatics, 2008.

[47] NIMMAGADDA, S. L; NIMMAGADDA, S. K; DREHER, H. **Ontology based data warehouse modeling and managing ecology of human body for disease and drug prescription management**. International Conference on Digital Ecosystems and Technologies, 2008.

[48] OLIVEIRA, D; BAIÃO, F; MATTOSO, M. **MF-Ontology, uma ontologia para o processo de mineração de textos**. SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL, 2008.

[49] PAN, D; PAN, Y. **Using Ontology Repository to Support Data Mining**. World Congress on Intelligent Control and Automation, 2006.

[50] PANOV, P; DZEROSKI, S; SOLDATOVA, L. N. **OntoDM: An Ontology of Data Mining**. International Conference on Data Mining Workshops, 2008.

[51] PAPADAKIS, N; DOULAMIS, A; LITKE, A; DOULAMIS, N; SKOUTAS, D; VARVARIGOU, T. **Mi-mercury: A mobile agent architecture for ubiquitous retrieval and delivery of multimedia information**. In: MULTIMED TOOLS APPL, p. 147Ű184. Springer Science+Business Media, 2008.

[52] ROBAL, T; KALJA, A. **Applying User Profile Ontology for Mining Web Site Adaptation Recommendations**. ADBIS, 2007.

[53] SIDHU, A. S; KENNEDY, P. J; SIMOFF, S; DILLON, T. S; CHANG, E. **Knowledge discovery in biomedical data facilitated by domain ontologies**. In: KNOWLEDGE DISCOVERY AND DATA MINING: CHALLENGES AND REALITIES, p. 189Ű202. Information science reference, 2007.

[54] TIM BERNERS-LEE, J. H; LASSILA, O. **The semantic web**. Scientific American Magazine, 2001.

[55] TRIFAN, M; IONESCU, B; IONESCU, D; PROSTEAN, O; PROSTEAN, G. **An Ontology based Approach to Intelligent Data Mining for Environmental Virtual Warehouses of Sensor Data**. Virtual Environments, Human-Computer Interfaces, and Measurement Systems, 2008.

[56] V, H; R, M. **RACER System Description**. In: LECTURE NOTES IN ARTIFICIAL INTELLIGENCE., 2001.

[57] VIVACQUA, A. S; GARCIA, A. C. B. **MINERAÇÃO DE DADOS BASEADA EM ONTOLOGIA**. SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL, 2008.

[58] W3C. **OWL**. `http://www.w3.org/TR/owl-features/`, acessado em Junho de 2009.

[59] WAIKATO, U. O. **WEKA**. `http://www.cs.waikato.ac.nz/ml/weka/`, acessado em Junho de 2009.

[60] WEB, W. S. **W3C Semantic Web Activity**. `http://www.w3.org/2001/sw/`, acessado em Maio de 2009, 2003.

[61] WITTEN, I. H; FRANK, E. **Data Mining - Practical Machine Learning Tools and Techniques**. Elsevier, 2005.

[62] WIVES, L. K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos**. Dr.scient. afhandling, Universidade Federal do Rio Grande do Sul, 2004.

[63] WONG, T.-L; CHOW, K.-O; WANG, F. L. **AN UNSUPERVISED LEARNING FRAMEWORK FOR DISCOVERING THE SITE-SPECIFIC ONTOLOGY FROM MULTIPLE WEB PAGES**. International Conference on Machine Learning and Cybernetics, 2008.

[64] XUPING, W; ZIJIAN, N; HAIYAN, C. **Research on Association Rules Mining Based-on Ontology in E-commerce**. Wireless Communications, Networking and Mobile Computing, 2007.

[65] YOUN, S; MCLEOD, D. **Efficient Spam Email Filtering using Adaptive Ontology**. International Conference on Information Technology, 2007.

[66] ZAGORUIKO, N. G; GULYAEVSKII, S. E; KOVALERCHUK, B. Y. **Ontology of the data mining subject domain**. In: PATTERN RECOGNITION AND IMAGE ANALYSIS, volume 17, p. 349Ű356. Pleiades Publishing LTD, 2007.

[67] ZHOU, C; JIANG, B; WANG, Q. **The ontology construction and data mining research of PDM system**. Computer-Aided Industrial Design and Conceptual Design, 2006.