Knowledge-based Association Rules Mining: Survey

Marcelo A. R. d’Almeida

Instituto de Computação – Universidade Federal Fluminense (UFF)  
Av. Gal. Milton Tavares de Souza, s/nº, 24210-346 – Niterói – RJ – Brazil

md@id.uff.br

**Abstract.** This survey aims to provide an overview of what has been done regarding the application of knowledge in the association rules extraction process, which has the basic purpose of discovering existing patterns and generate new knowledge.

**Resumo.** Este survey tem por objetivo prover um panorama geral do que se tem feito com relação à aplicação de conhecimento no processo de extração de regras de associação, o qual possui o intuito básico de descobrir padrões existentes e gerar novos conhecimentos.

# 1. Introdução

A extração de Regras de Associação é um assunto por si só muito estudado e é uma das técnicas principais na área de Mineração de Dados. O uso de algum conhecimento já adquirido surge como uma extensão da definição do problema original para facilitar (otimizar) o processo de extração, reduzir a quantidade de regras ditas óbvias ou desinteressantes, abordar regras que antes não eram consideradas ou aumentar a qualidade das regras já adquiridas.

Existem estudos que são motivados em melhorar a representação da informação. Neste survey, focaremos na extração das regras e a Representação do Conhecimento apenas como uma ferramenta utilizada, sem adentrar nas vantagens e desvantagens de cada uma.

A seção 2 apresenta alguns conceitos importantes para a compreensão do estudo de revisão. A seção 3 apresenta alguns dos termos mais comuns encontrados ao se pesquisar sobre o assunto. A seção 4 trata dos tipos de abordagens disponíveis na área. A seção 5 relata as primeiras abordagens a serem consideradas por aqueles que contribuíram para formular o problema. A seção 6 tem por objetivo ilustrar tipos de abordagem diferentes. A seção 7 conclui o survey.

# 2. Conhecimento prévio

Para melhor compreender a aplicação de conhecimento do domínio na extração de novos conhecimentos, é necessário revisar alguns conceitos importantes. Revisaremos a ideia básica inerente às Regras de Associação, uma das técnicas principais no processo de Mineração de Dados, bem como a sua forma mais específica voltada ao uso de Taxonomias. Para então apresentar, de uma forma geral, o conceito de Representação de Conhecimento, o qual será necessário para expandir o uso de Regras de Associação para uma utilização mais abrangente do que a associada somente com taxonomias.

# 2.1. Regras de Associação

Uma Regra de Associação representa um padrão de relacionamento entre itens de dados de um domínio de aplicação que ocorre com uma determinada frequência na base de dados.

{candidíase} ⇒ {pneumonia}

{café, leite} ⇒ {pão, manteiga, queijo}

A primeira regra indica, com um determinado grau de certeza, que se o paciente contraiu candidíase, então também teve pneumonia. A segunda, que quem compra café e leite também compra pão, manteiga e queijo.

Seja I = {i1, i2, ..., in} o conjunto de itens do domínio da aplicação. Seja D = {t1, t2, ..., tm} uma base de dados transacional, onde cada transação ti, 1≤i≤m, é um subconjunto de I. Uma regra de associação R definida sobre D é uma implicação da forma: X ⇒ Y onde X ⊂ I, Y ⊂ I, X ≠ ∅, Y ≠ ∅ e X ∩ Y = ∅. X é dito o antecedente e Y o consequente da regra.

O suporte de uma regra X ⇒ Y é definido pela porcentagem de transações que incluem todos os itens do conjunto X U Y. Representa a fração das transações que satisfazem tanto o antecedente quanto o consequente da regra. O suporte de uma regra tenta indicar sua relevância. Seja R a regra X ⇒ Y. Seja T o número de transações da base. Seja TXUY o número de transações que incluem os elementos de X U Y.

**Suporte(R) = TXUY / T**

Seja X um conjunto de itens. Seja TX o número de transações que incluem os elementos de X.

**Suporte(X) = TX / T**

A confiança de uma regra X ⇒ Y é definido pela porcentagem de transações que incluem os itens X e Y em relação a todas que incluem os itens de X. Representa o grau de satisfatibilidade do consequente, em relação às transações que incluem o antecedente. A confiança tenta indicar a validade da regra. Seja R a regra X ⇒ Y. Seja TX o número de transações que incluem os elementos de X. Seja TXUY o número de transações que incluem os elementos de X U Y.

**Confiança(R) = TXUY / TX**

Entrada: Base de dados de transações; Suporte mínimo; Confiança mínima. Saída: Todas as regras de associação que possuem suporte e confiança maiores ou iguais ao suporte e à confiança mínimos.

Seção 2.1 [9]

**2.2. Regras de Associação com Taxonomia**

Suponha que os itens de dados do domínio da aplicação estejam organizados em taxonomias que os classificam. Taxonomia: grafo direcionado acíclico. Uma aresta (x,y) indica que y é um elemento do tipo x (ou da classe x). Um caminho entre x e z indica, por transitividade, que z é do tipo x.

É possível que as duas regras a seguir não tenham o suporte desejado.

{calça-social} ⇒ {sandália}

{calça-social} ⇒ {sapato}

Porém, a regra envolvendo a generalização calçado-social pode ser relevante.

{calça-social} ⇒ {calçado-social}

Uma regra entre taxonomias pode relacionar itens de diferentes níveis. Em muitas aplicações que envolvem taxonomias, as folhas são os produtos com suas marcas.

Seja I = {i1, i2, ..., in} o conjunto de itens do domínio da aplicação. Seja C = {c1, c2, ..., cm} um conjunto de classes. Seja G um grafo direcionado acíclico sobre I∪C, representando um conjunto de taxonomias. Uma regra de associação em taxonomias R definida sobre I e G é uma implicação da forma  **X ⇒ Y** onde X ⊂ I∪C, Y ⊂ I∪C, X ≠ ∅, Y ≠ ∅, X ∩ Y = ∅ e nenhum item em Y é ancestral de algum item em X. Esta última restrição evita regras do tipo {calça-social} ⇒ {calça}. Os conceitos de suporte e confiança se aplicam como nas regras convencionais.

Entrada: Base de dados de transações; Um conjunto de taxonomias; Suporte mínimo; Confiança mínima. Saída: Todas as regras de associação em taxonomias que possuem suporte e confiança maiores ou iguais ao suporte e à confiança mínimos.

Seção 2.2 [9]

# 2.3. Representação do Conhecimento

O objetivo é estabelecer um relacionamento entre o conhecimento humano e sua representação, por meio de formalismos. O computador poderá usar uma estratégia de raciocínio.

Representar conhecimento e em domínios complexos, por exemplo: agente que compra na Internet, agente que dirige um carro e como os agentes se relacionam com demais agentes e com o ambiente.

Taxonomias são um bom exemplo de representações simples, as quais já foram apresentadas na seção anterior. Redes Semânticas e Ontologias são formas de representação mais completas e mais complexas do que a taxonomia e ilustram bem o poder de expressividade que possa vir a ser útil a ser aplicado.

Redes Semânticas representam objetos individuais, categorias de objetos e relacionamentos entre objetos.

Ontologias podem descrever: especificação formal dos conceitos em um domínio; relacionamento entre esses conceitos; axiomas em uma linguagem formal; vocabulário de termos, seus significados, como se relacionam.

Servem para compartilhar conhecimento comum (entre pessoas ou software), possibilitar reuso de conhecimento sobre um domínio e para explicitar suposições de um domínio.

Podem ser construídas (‘populadas’) de diversas formas: extração de informação a partir de textos; através da inclusão de conhecimento de especialistas, por meio de uma equipe de ontologistas/lógicos; importando, de forma automática ou semiautomática, informações presentes em bancos de dados; ou até mesmo de forma colaborativa, através de pessoas diversas, incluindo, por exemplo, conhecimento de senso comum.

Componentes de uma Ontologia :

Conceito: classes de entidades em um domínio;

Relações: interações entre os conceitos, por exemplo, taxonomias;

Papéis e propriedades;

Instâncias: representantes de um conceito;

Axiomas: restringem valores de classes e definem regras gerais.

Seção 2.3 [10]

# 3. Diversas nomenclaturas e variações

Quando se trata de extração de Regras de Associação e aplicação de conhecimento existem duas formas principais de se nomeá-las, Regras de Associação Generalizadas (*Generalized Association Rules*) e Regras de Associação Multi-Níveis (*Multi-Level Association Rules*). Ambas, de maneira prática, expressam o uso de termos mais gerais, presente em camadas com maior abstração, em substituição de termos mais específicos, ou especializados. Como por exemplo, o uso do termo ‘roupas leves’ como abstração para termos mais específicos como ‘camisa-regata’, ‘short’ e ‘camiseta’.

Existem também diversas variações envolvendo essas duas nomenclaturas. Para citar algumas: *Preknowledge-based Generalized Association Rules*; *Cross-Level Association Rules* e; *Constraint-based Multi-level Association Rules*. No geral, estas variações apenas são outras formas de se denominar e não diferem muito das duas principais.

# 4. Tipos de abordagens disponíveis

Existem diversas formas de se aplicar conhecimento no processo de extração. Nesse estudo, as abordagens foram divididas em dois grupos principais: um referente à etapa do processo de mineração no qual é aplicado e outro referente ao tipo de representação do conhecimento utilizado.

No primeiro grupo, as etapas são: Pré-processamento, Processamento, Pós-processamento e Auxiliar. Eles se diferenciam em que parte do processo, o conhecimento é aplicado.

No caso do Pré-processamento, o que é modificado, geralmente, são os dados de entrada, o que por muitas vezes não exige ou exige pouca alteração do algoritmo de mineração de Regras de Associação usado. Aumentar a base transacional para incluir as generalizações é um exemplo de abordagem de Pré-processamento.

As abordagens de Processamento visam a modificação do algoritmo de mineração propriamente dito para atender as novas necessidades. Geralmente, se tratam de algoritmos baseados no apriori.

Na etapa de Pós-processamento, de forma tão intuitiva quanto as demais, o que é modificado são os dados de saída do algoritmo de mineração. Geralmente são utilizadas como generalização das regras encontradas ou como mecanismo de poda para filtrar o resultado. Não necessitam de modificações no algoritmo de extração de regras.

Enquanto as demais estratégias estão intimamente relacionadas com o processo de mineração em si, a etapa ‘Auxiliar’ visa uma estratégia diferente, como por exemplo, melhorar a qualidade da taxonomia.

No segundo grupo, estão incluídos, mas não limitados: Taxonomias, Taxonomias Fuzzy, Ontologias, Grafos de Conhecimento e DataCubes. Eles se diferenciam pelo nível de detalhe necessário ou disponível e pela forma como descrevem e relacionam o conhecimento.

O uso de Taxonomias na extração de regras de associação, o qual motivou a realização deste estudo exploratório, é uma das formas principais de utilização. Isso se dá pela sua forma simples de representação, uma hierarquia ‘é-um’, a qual pode representar noção de generalização/especialização. Pode ser representado através de uma árvore ou mais especificamente com um Grafo Direcionado Acíclico, no caso do uso de múltiplas taxonomias.

Taxonomias Fuzzy é um caso específico de Taxonomia (geralmente de taxonomias múltiplas, sendo então representado por um Grafo Direcionado Acíclico) onde se aplica a Lógica Fuzzy. Sob esta lógica, um elemento pode pertencer aos conjuntos diferentes (ou ser especialização de outros elementos, no caso de uma taxonomia) com um certo grau de inclusão para cada um deles. Por exemplo, A pode pertencer 0.6 a B e 0.4 a C.

Ontologias e Grafos de Conhecimento foram introduzidos na seção de Representação de Conhecimento. DataCubes serão apresentados mais adiante.

# 5. Primeiras abordagens

As primeiras propostas de solução que surgiram juntamente com a formulação do problema foram os algoritmos ‘*Basic*’, *Cumulate, Stratify, Estimate* e *EstMerge*.

# 5.1. *‘Basic’ Algorithm*

O algoritmo ‘Basic’ se baseia em uma ideia intuitiva. Ele adiciona todos os ancestrais para cada item da transição na transição, e então roda qualquer algoritmo de mineração de Regras de Associação nessas ‘transações estendidas’. Ele resolve o problema de forma simples, porém não é muito rápido.

# 5.2. *Cumulate, Stratify, Estimate* *and* *EstMerge Algorithms*

Os algoritmos Cumulate e o EstMerge são propostos e surgem a partir de otimizações feitas, de modo que ele é de 2 a 5 vezes mais rápido do que o algoritmo ‘Basic’, e até mais de 100 vezes mais rápido se considerar datasets da vida real.

O algoritmo Cumulate se baseia em 3 estratégias para otimizar o processo: Filtrar os ancestrais adicionados na transação; Pré-computar os ancestrais; Podar itemsets contendo um item e seu ancestral.

O algoritmo Stratify utiliza a ideia por de trás da estratificação. Ele conta o suporte mínimo dos itemsets mais generalizados. Caso o mínimo não seja atingido, não há necessidade de calcular o suporte de itemsets ‘derivados’, isto é, o mesmo itemset, com algum elemento substituído por alguma de suas especializações. As otimizações do Cumulate continuam valendo para o Stratify.

O algoritmo Estimate utiliza a noção de amostragem para tentar prever se o elemento tem suporte mínimo ou não. Em cada passada, ele só calcula o suporte de fato caso o elemento tenha o suporte esperado maior que o suporte mínimo ou caso seja descendente direto de alguém que o tenha. Ele averigua qual é o suporte uma vez que é a única maneira de ter certeza se ele possui ou não o suporte mínimo. Necessita de uma segunda passagem para contar os descendentes do que ele havia estimado abaixo do mínimo.

O algoritmo EstMerge se comporta de maneira similar ao Estimate, porém ele desconsidera a segunda passagem e aceita os descendentes sem checar se o suporte mínimo é atingido. Possui a vantagem de economizar a segunda passagem em toda a base, porém ele gera mais candidatos.

Seção 5 [1]

# 6. Abordagens adicionais

Outras soluções, abordagens e aplicações surgiram posteriormente. Para citar alguns que exemplifiquem os diferentes tipos de abordagens disponíveis: GART Algorithm e GARPA, Ontology-Driven, G−L³ Classifier, FDiff\_ET e FDiff\_ET\* Algorithms, Using genetic algorithms e COGARE Algorithm

# 6.1. *Generalization of Association Rules using Taxonomies – (GART Algorithm)* e *Generalized Association Rule Post-Processing Approach – (GARPA)*

Diferencial:

É realizado na etapa de pós-processamento

A proposta começa com uma etapa de pós-processamento. Após as regras serem geradas pelo algoritmo apriori ou qualquer outro algoritmo para extrair Regras de Associação ‘comuns’, isto é, sem Taxonomia, O usuário então cria uma taxonomia e então usa o algoritmo proposto para gerar Regras de Associação Generalizadas. O Algoritmo GART generaliza apenas um lado da regra, a ser definido pelo usuário.

Em outro momento, o trabalho evoluiu na abordagem denominada GARPA, o qual possui as diferenças básicas: possibilita generalizar de apenas um dos lados, ou dos dois (de acordo com a vontade do usuário); a generalização passou a ocorrer entre os itens das regras e não somente entre as regras; não é necessário que tenha uma regra especializada para cada item contido na taxonomia; entre outras diferenças.

Seção 6.1 [5 e 6]

# 6.2. *Ontology-Driven Association Rule Extraction*

Diferencial:

Explora a expressividade da Ontologia

Este estudo substitui o uso da Taxonomia por uma Ontologia. Com isso, restrições podem ser definidas a partir das propriedades especificas presentes. Possui a vantagem de ser mais flexível, pois ao se adicionar uma informação, não se precisa adaptar o modelo de representação, como seria necessário ao usar uma Taxonomia.

Utiliza o apriori em conjunto com uma linguagem para consultar a Ontologia. O algoritmo apriori age a partir do resultado dos itens que obedecem as restrições presentes na consulta e consequentemente, no conhecimento.

Seção 6.2 [7]

# 6.3. *Improving Classification Models with Taxonomy Information (G−L³ Classifier)*

Diferencial:

Integra Taxonomia em modelos de classificação.

O classificador G−L³ estende o classificador associativo L³ ao utilizar conhecimento provido por uma taxonomia durante o treinamento.

A fase de treinamento se baseia em um processo de duas etapas: mineração de regras de classificação generalizada; e seleção de regras através de uma poda ‘preguiçosa’. A primeira etapa é conseguida com uma extensão do algoritmo de mineração FP-growth, o qual permite a descoberta de regras de classificação generalizadas.

Seção 6.3 [3]

# 6.4. *Updating Generalized Association Rules with Evolving Fuzzy Taxonomies (FDiff\_ET and FDiff\_ET\* Algorithms)*

Diferencial:

Atualiza as regras a partir do uso de Taxonomias Fuzzy que podem se modificar

O algoritmo começa similar ao apriori. Primeiro gere os itemsets candidatos dos frequentes descobertos do passo anterior. Diferencie os itemsets afetados dos não afetados. Incorpore os itemsets generalizados frequentes, os candidatos afetados e os candidatos não afetados para determinar se um itemset é frequente ou não na base de dados estendida resultante. Por fim, escaneie o banco de dados com a taxonomia atualizada para contar o suporte dos itemsets que não foram determinados no passo anterior.

Seção 6.4 [2]

# 6.5. *Generalized Association Rule Mining Using Genetic Algorithms*

Diferencial:

Utiliza uma combinação de algoritmo apriori com algoritmos genéticos

O processo começa com o processo apriori. Após encontrar os itemsets cujo suporte é maior do que o mínimo definido pelo usuário e extrair as regras, as Regras de Associação Generalizadas são escolhidas através da proposta de utilizar algoritmos genéticos para podar a taxonomia dos itemsets frequentes gerados. Algumas medidas são usadas utilizadas para extrair regras interessantes.

Seção 6.5 [8]

# 6.6. *COmplexity Guided Association Rule Extraction (COGARE Algorithm)*

Diferenciais:

Utiliza DataCubes como Representação de Conhecimento que incorporam, inclusive, Lógica Fuzzy;

Projetado para representar e lidar com imprecisão em seu processo;

Reduz a complexidade do resultado final, sem reduzir a qualidade do conjunto de regras.

DataCubes são uma forma de representar dados de múltiplas dimensões (high dimensional data) e podem conter hierarquias em sua estrutura.

Quando se trata de Regras de Associação em DataCubes, alguns tipos de associação são definidos: Intra-Dimensional Association, quando a associação é entre elementos dentro da dimensão; Inter-Dimension Association quando a associação é entre dimensões; e Hybrid Association, a junção dos dois tipos anteriores (analisando os conjuntos frequentes em uma mesma dimensão e depois entre as dimensões). O algoritmo COGARE extrai Regras de Associação Inter-Dimensionais usa conceitos Fuzzy definidos nas dimensões e hierarquias para reduzir a complexidade das regras obtidas.

Ele se baseia em dois passos principais. Geração da Regra e Processo de Generalização. Na Geração da Regra, se o item não for frequente, o método generaliza o item. O processo é repetido até o item se tornar frequente ou não puder mais ser generalizado. No Processo de Generalização, o resultado do passo anterior é generalizado, de modo a reduzir a complexidade do resultado final, ao utilizar elementos mais abstratos e reduzir a cardinalidade. Nesse passo, a perda de qualidade é controlada.

Seção 6.6 [4]

# 7. Conclusão

O estudo em torno da extração de Regras de Associação com uso de conhecimento como parte do processo de mineração se expandiu além do uso de Taxonomias, como nos inúmeros exemplos de representação relacionados acima, e além da extração das regras relacionando seus elementos como aplicação somente, como no caso do classificador citado.

Este survey visa prover uma introdução ao assunto e apresentar algumas das abordagens disponíveis. Existem outras abordagens que não foram exploradas. Este survey também possui o objetivo de prover um panorama geral, amplo e ao mesmo tempo enxuto. E por esse motivo, detalhes mais aprofundados de como funciona cada proposta podem ser conferidos diretamente nos artigos, listados nas referências

# Referências

[1] Ramakrishnan Srikant, Rakesh Agrawal, “Mining generalized association rules”, Future Generation Computer Systems 13 (1997) 161-180

[2] Wen-Yang Lin • Ja-Hwung Su • Ming-Cheng Tseng, “Updating generalized association rules with evolving fuzzy taxonomies”, Soft Comput (2012) 16:1109–1118

[3] Luca Cagliero, Paolo Garza, “Improving classification models with taxonomy information”, Data & Knowledge Engineering 86 (2013) 85–101

[4] Nicolas Marin, Carlos Molina, José M. Serrano, and M. Amparo Vila, “A Complexity Guided Algorithm for Association Rule Extraction on Fuzzy DataCubes”, IEEE Transactions on Fuzzy Systems, vol. 16, no. 3, june (2008)

[5] Marcos Aurélio Domingues and Solange Oliveira Rezende, “Using Taxonomies to Facilitate the Analysis of the Association Rules”, Proc. Second Int’l Workshop Knowledge Discovery and Ontologies, held with ECML/ PKDD, pp. 59-66, (2005).

[6] Veronica Oliveira de Carvalho, Solange Oliveira Rezende, Mário de Castro, “Obtaining and Evaluating Generalized Association Rules”, Proceedings of the 9th International Conference on Enterprise Information Systems (ICEIS), vol. 2 - Artificial Intelligence and Decision Support Systems, 310–315. (2007)

[7] Andrea Bellandi, Barbara Furletti, Valerio Grossi, and Andrea Romei, “Ontology-Driven Association Rule Extraction: A Case Study”, Proc. Workshop Context and Ontologies: Representation and Reasoning, pp. 1-10, (2007).

[8] Peter P. Wakabi-Waiswa, Venansius Baryamureeba and K. Sarukesi, “Generalized Association Rule Mining Using Genetic Algorithms”, International Journal of Computing and ICT Research, Vol. 2 No. 1, pp:59-69, (2008)

[9] Prof. Alexandre Plastino, Aula 07 – Mineração de Dados, “Regras de Associação” (2015)

[10] Prof. Aline Paes, Aula 13 – Inteligência Artificial, “Representação de Conhecimento” (2015)