

# Mineração de regras de associação em servidores Web com RapidMiner\*

Fabício J. Barth  
fabricio.barth@gmail.com

## Resumo

Este texto apresenta o uso da ferramenta RAPIDMINER na mineração de regras de associação em servidores Web através de um estudo de caso onde o objetivo é identificar padrões de navegação em uma livraria virtual. Ao longo do texto serão apresentados conceitos sobre mineração de padrões de comportamento na Web e sobre algoritmos para geração de regras de associação. O funcionamento e características da ferramenta RAPIDMINER são ilustrados durante a apresentação do estudo de caso.

## 1 Introdução

Com o contínuo aumento e proliferação de sites de *e-commerce*, *web services* e portais, o volume de *clickstream* e dados sobre os usuários coletados por organizações do mercado Web, nas suas operações do dia-a-dia, tem alcançado proporções astronômicas. Analisar este tipo de dado coletado pode ajudar estas organizações, por exemplo: na definição de estratégias de marketing; na avaliação da efetividade de campanhas promocionais; na otimização das funcionalidades de aplicações Web; no fornecimento de conteúdo personalizado aos usuários, e; na definição de uma estrutura de navegação mais adequada ao usuário.

Este tipo de análise envolve a descoberta automática de padrões e relações úteis existentes numa grande coleção de dados semi-estruturados, geralmente armazenados em servidores de *logs*, e estruturados, armazenados em banco de dados. O objetivo da *Mineração de padrões na Web*<sup>1</sup> é capturar, modelar e analisar os padrões de comportamento e perfis de usuários que interagem com um web site [4].

---

\*Versão preliminar.

<sup>1</sup>do inglês, *Web Usage Mining*

O objetivo deste texto é apresentar o uso da ferramenta RAPIDMINER [6] na mineração de regras de associação em servidores Web. Na seção 2 será apresentada uma descrição genérica do processo para a mineração de padrões na Web; na seção 3 serão apresentados as etapas e algoritmos necessários para a geração de regras de associação que descrevem o comportamento dos usuários de um determinado web site; na seção 4 é apresentada como a mineração de regras de associação em servidores Web pode ser implementada utilizando a ferramenta RAPIDMINER<sup>2</sup> [6], e; na seção 5 são apresentadas algumas considerações finais sobre o tema examinado neste texto.

## 2 Processo para mineração de padrões na Web

Seguindo o processo padrão de mineração de dados [3], o processo de Mineração de Padrões na Web é composto por três independentes estágios: coleta e pré-processamento de dados, descoberta de padrões e análise de padrões.

Na fase de coleta e pré-processamento dos dados, o *clickstream* do web site é coletado, filtrado e particionado em conjuntos de transações de usuários que representam as atividades de cada usuário durante diferentes visitas ao web site. O *clickstream* de um web site costuma ser armazenado em arquivos de *log* do servidor Web, como por exemplo: *Apache HTTP Server*<sup>3</sup>. O formato tradicional destes dados é apresentado na figura 4.

Em um arquivo tradicional de *log* são armazenadas informações, tais como: o IP do usuário ("1.2.3.4"), o recurso que o usuário acessou ("/classes/cs589/papers.html"), em que servidor ("maya.cs.depaul.edu"), o tipo e versão do browser do usuário, o sistema operacional do usuário e a data de acesso do usuário ("2006-02-01 00:08:43").

Em alguns casos, onde o web site possui uma infra-estrutura mais robusta, são implementados componentes específicos para a coleta de dados de *clickstream*, inclusive de forma mais detalhada e focada nas necessidades do site. Informações sobre o identificador de um usuário cadastrado no site, o *cookie*<sup>4</sup> de um usuário não cadastrado no site ou uma ação de impressão de informações sobre um determinado item em um site de e-commerce podem ser facilmente capturadas.

Dependendo dos objetivos da análise, os dados brutos de *clickstream* são transformados e agregados em diferentes níveis de abstração. Na Mineração de padrões na Web, o nível mais básico de abstração dos dados é o *pageview*, ou

---

<sup>2</sup><http://rapid-i.com>

<sup>3</sup><http://httpd.apache.org/>

<sup>4</sup>*Cookie* é um grupo de dados trocados entre o navegador e o servidor de páginas, colocado num arquivo de texto criado no computador do usuário. A sua função principal é a de manter a persistência de sessões HTTP. Mas também pode ser utilizado para rastrear um usuário.

1	2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://dataminingresources.blogspot.com/
2	2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://maya.cs.depaul.edu/~classes/cs589/papers.html
3	2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200 318814 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1) http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey
4	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/
5	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html
6	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/header.gif - 200 6027 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html

Figura 1: Exemplo típico de *log* [4].

seja, uma ação única de um usuário, tal como: leitura de um artigo, visualização de um produto ou adição de um produto no carrinho de compras. O próximo nível de abstração é o de seção<sup>5</sup>, que é uma sequência de *pageviews* realizados por um único usuário numa única visita ao site.

Para a identificação de seções existem algumas heurísticas, entre elas, heurísticas baseadas em tempo e heurísticas baseadas na mudanças do padrão de navegação. As heurísticas baseadas em tempo são as mais utilizadas. Basicamente, estas heurísticas especificam que todos os *pageviews* que ocorreram dentro de uma faixa de tempo fazem parte da mesma seção. Um exemplo de identificação de seção utilizando uma heurística orientada por tempo é apresentado na figura 2.

<sup>5</sup>Muitas vezes o termo utilizado é *session*, mesmo em textos em português.

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

User 1

Session 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C

Session 2

1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Figura 2: Exemplo de seção utilizando uma heurística orientada por tempo.

Este trabalho de limpeza dos dados, identificação de *pageview* e identificação de seções faz parte da etapa de coleta e pré-processamento dos dados. Na figura 3 é possível identificar em que parte do processo de mineração de padrões na Web a etapa de pré-processamento se encontra.

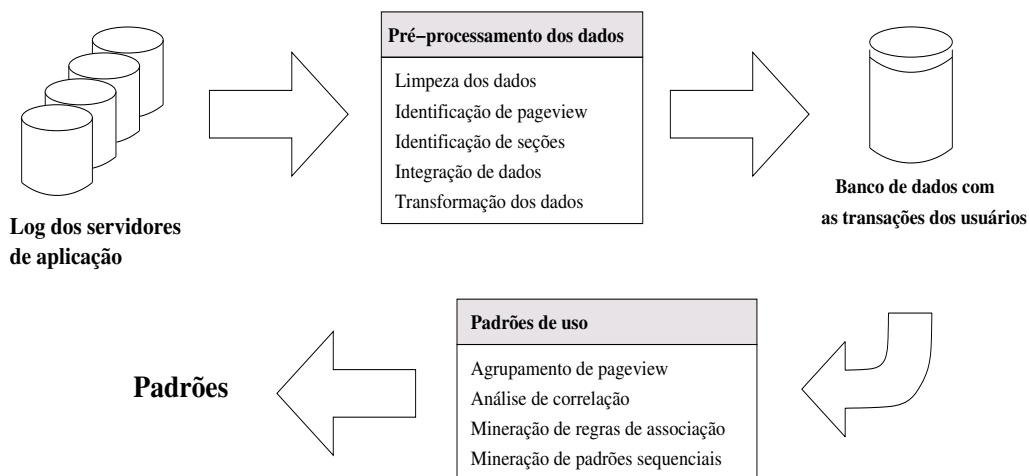


Figura 3: Processo de mineração de padrões na Web.

Após realizada a etapa de pré-processamento, as transações dos usuários podem ser armazenados em um único repositório. Desta forma, os algoritmos utilizados na identificação dos padrões de uso acessam uma única fonte de dados.

Existem diversos padrões que podem caracterizar a forma como um Web site é utilizado, entre eles: agrupamentos, correlações e regras de associação. O objetivo deste texto é explorar regras de associação. Para conhecer outros algoritmos sugere-se a leitura dos textos [4] e [7].

### 3 Minerando regras de associação em servidores Web

Descoberta de regras de associação podem auxiliar na identificação de grupos de páginas ou itens que são normalmente acessados ou comprados em conjunto. Uma Regra de Associação caracteriza o quanto a presença de um conjunto de itens no registro de uma base de dados implica na presença de algum outro conjunto distinto de itens nos mesmos registros. Uma de suas típicas aplicações é a análise de transações de compra. A partir de uma base de dados que armazena produtos comprados por clientes de, por exemplo, uma livraria virtual, uma estratégia para a mineração de regras de associação poderia gerar o seguinte exemplo:  $\{arquitetura\} \rightarrow \{artes\}[0.20, 0.85]$ . Esta regra é utilizada para indicar que os clientes que comprem livros da categoria *arquitetura*, tendem (em 85% dos casos) a comprar também livros da categoria *artes*.

As abordagens mais comuns para a descoberta de regras de associação são originárias do algoritmo *Apriori* [1, 5]. Este algoritmo encontra grupos de itens que ocorrem frequentemente juntos em muitas transações. A quantidade de transações mínimas é definida pelo parâmetro de suporte mínimo (em inglês, *minimum support*) do algoritmo *Apriori*. Este grupo de itens é chamado de conjunto de itens freqüentes (em inglês, *frequent itemsets*). As regras de associação são geradas a partir do conjunto de itens freqüentes.

Uma regra de associação é uma expressão na forma  $X \rightarrow Y[sup, conf]$ , onde  $X$  e  $Y$  são conjunto de itens,  $sup$  é o suporte do conjunto de itens  $X \cup Y$ , representando a probabilidade de  $X$  e  $Y$  ocorrerem juntos em uma transação, e  $conf$  é a confiança da regra, definida por  $sup(X \cup Y)/sup(X)$ , representando a probabilidade condicional de  $Y$  ocorrer em uma transação dado que  $X$  ocorreu nesta transação.

### 4 Minerando regras de associação em servidores Web com RapidMiner

RAPIDMINER é uma ferramenta implementada em Java, *open-source*, que fornece a implementação de algoritmos utilizados em problemas de aprendizagem de máquina e uma interface gráfica para o desenvolvimento rápido de projetos para a criação de modelos preditivos [6]. Ao utilizar esta ferramenta é possível definir um processo de tratamento dos dados, inserindo operadores responsáveis por: I/O (entrada e saída); algoritmos de aprendizagem (supervisionados ou não); funções de *on-line analytical processing*; pré-processamento; validação, e; visualização. A atual versão do RAPIDMINER é a 5.1.11. Todos os exemplos

utilizados neste texto são realizados utilizando esta versão do software.

Trata-se de uma ferramenta muito útil no desenvolvimento e prototipação de processos para análise de dados, sejam eles dados estruturados, dados não estruturados ou um misto de dados provenientes da web.

A ferramenta RAPIDMINER fornece uma interface gráfica onde o processo de tratamento dos dados pode ser definido inserindo os operadores dentro de um fluxo (parte central da figura 4). Os operadores que podem ser inseridos no processo estão disponíveis do lado esquerdo da interface e toda vez que um operador é selecionado, os detalhes de configuração do respectivo operador são apresentados do lado direito da interface.

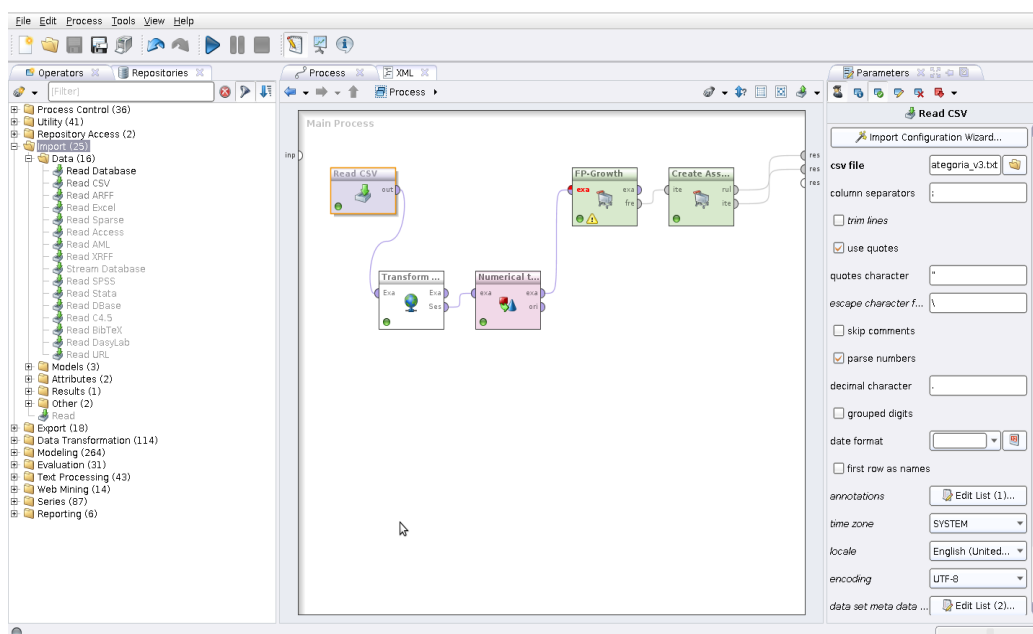


Figura 4: Visão geral da ferramenta RAPIDMINER.

Os principais grupos de operadores disponíveis do lado esquerdo da interface na última versão da ferramenta são:

- *Import*: possui os operadores utilizados na leitura de dados, modelos e resultados. Existem diversos operadores para diversos formatos, por exemplo: arquivos CSV, acesso à banco de dados via JDBC, Access e outros formatos utilizados por outras ferramentas similares ao RAPIDMINER;
- *Export*: possuir os operadores utilizados na escrita de dados, modelos e resultados. Estes operadores manipulam basicamente os mesmos formatos que os operadores utilizados para leitura;

- *Data Transformation*: em inúmeras situações são necessárias algumas operações para conversão de tipos, ordenação de valores, filtragem e limpeza dos dados antes da aplicação de algum algoritmo de aprendizagem. É no grupo *data transformation* que são encontrados os operadores responsáveis por esta tarefa;
- *Modeling*: este grupo possui os operadores que implementam os algoritmos de aprendizagem de máquina, tais como: os algoritmos utilizados na geração de regras de produção, algoritmos para problemas de classificação (*Naive Bayes*, *SVM*) e algoritmos para problemas de agrupamento (*Knn*, *EM*, *UPGMA*);
- *Evaluation*: antes de colocar qualquer modelo preditivo em produção, é necessário avaliar a sua acurácia. Para medir a acurácia de um modelo preditivo são utilizadas algumas métricas. Os componentes responsáveis por medir todas estas métricas estão presentes no pacote *evaluation* do RAPIDMINER;
- *Text Processing*: este grupo possui operadores especiais para processos que necessitam manipular textos (informação não estruturada). Na maioria dos casos, são operadores para pré-processamento do texto - etapa que tem como objetivo transformar o dado não estruturado em dado estruturado, e;
- *Web Mining*: este grupo possui operadores que implementam algoritmos relevantes para projetos de *Web Mining*, como por exemplo: algoritmos para identificação de seções (figura 2).

Na próxima seção é apresentado um estudo de caso onde o objetivo é identificar padrões de navegação em uma livreria virtual, mais especificamente, identificar por quais categorias de livros (i.e., artes, tecnologia, administração) o mesmo usuário navega.

## 4.1 Estudo de caso

Para realizar esta análise é necessário seguir o processo ilustrado na figura 3. A primeira atividade que deve ser realizada é acessar o *log* dos servidores de aplicação e limpar os dados. Neste caso, a informação que deve permanecer é o identificador do usuário e a categoria do livro. Todas as outras informações (i.e., *timestamp*, *user-agent*) podem ser excluídas. Além disso, na maioria dos casos, a informação que está no *log* não é a categoria do livro, mas sim o identificador do livro. Portanto, além de excluir as informações que não são relevantes, também

será necessário substituir o identificador do livro pelo identificador da categoria do livro.

Feito o tratamento inicial, tem-se um arquivo, neste caso no formato CSV, com duas colunas: *idusuario* e *categoriaid*. Para ler este conteúdo é utilizado o operador *Read CSV* (grupo: *Import - Data*) (figura 4). Este operador lê um arquivo CSV e gera um conjunto de treinamento, uma tabela com duas colunas e  $n$  linhas que representam as transações capturadas a partir do servidor de *log* analisado.

A tabela criada a partir do operador *Read CSV* pode ser acessada por qualquer outro operador através da linha *out* do operador *Read CSV*. Neste estudo de caso, o operador que deve ser usado em seguida é o operador que transforma a tabela gerada pelo operador *Read CSV* por uma tabela que sumariza as transações do usuário (por exemplo, a tabela 1). Na ferramenta RAPIDMINER este operador é o *Transform log to session* (grupo: *Web Mining - Server Log Processing*).

Tabela 1: Exemplo de tabela com as transações dos usuários

usuário	<i>categoria<sub>1</sub></i>	<i>categoria<sub>2</sub></i>	<i>categoria<sub>3</sub></i>	...	<i>categoria<sub>m</sub></i>
<i>user<sub>1</sub></i>	0	2	0	...	1
<i>user<sub>2</sub></i>	1	1	0	...	0
<i>user<sub>3</sub></i>	2	0	1	...	0
<i>user<sub>4</sub></i>	0	1	0	...	0
...	...	...	...	...	...
<i>user<sub>n</sub></i>	1	1	0	...	1

Ao usar o operador *Transform log to session* deve-se conectar a saída do operador *Read CSV* na entrada do operador *Transform log to session* e configurar os dois parâmetros: *session attribute* e *resource attribute*. No nosso caso, o *session attribute* recebe o nome da coluna *idusuario* e o *resource attribute* recebe o nome da coluna *categoriaid*. Detalhes sobre esta configuração podem ser vistos na figura 5.



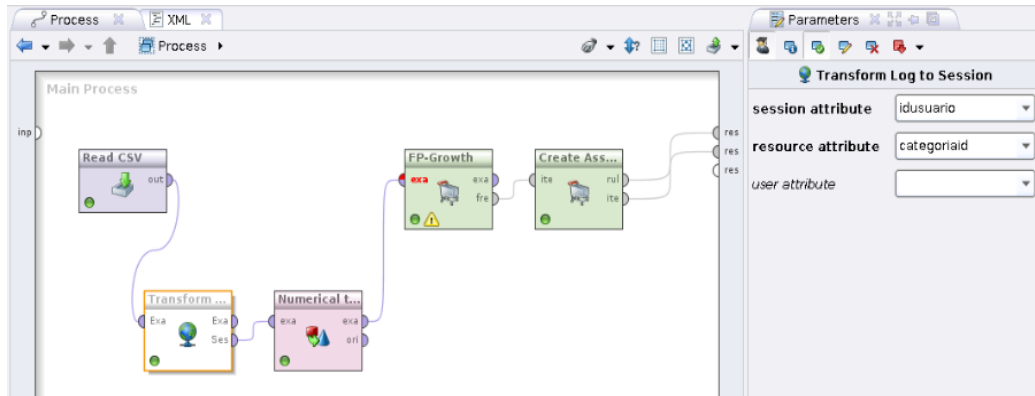


Figura 5: Detalhes sobre a configuração do operador *Transform log to session* - operador marcado em amarelo no diagrama.

O operador *Transform log to session* possui duas portas de saída, uma com o resultado do processamento e outra com os dados originais (figura 5). Neste exercício nós só vamos utilizar a porta que fornece o resultado do processamento deste operador, que é uma estrutura parecida com a apresentada na tabela 1.

No entanto, não podemos utilizar diretamente a saída do operador *Transform log to session* na entrada do operador *FP-Growth* (grupo: *Modeling - Association and Item Set Mining*) - operador responsável por identificar os conjuntos de itens frequentes - visto que o operador *FP-Growth* só manipula dados binários (0 e 1). Desta forma, é necessário transformar todos os dados numéricos da tabela 1 em dados binários usando a função:

$$f(x) = \begin{cases} 0 & \text{se } x = 0 \\ 1 & \text{se } x > 0 \end{cases} \quad (1)$$

O operador que implementa a função acima é o *Numerical to Binomial*. Este operador está dentro das pastas *Data Transformation* e *Type Conversion* (figura 6).

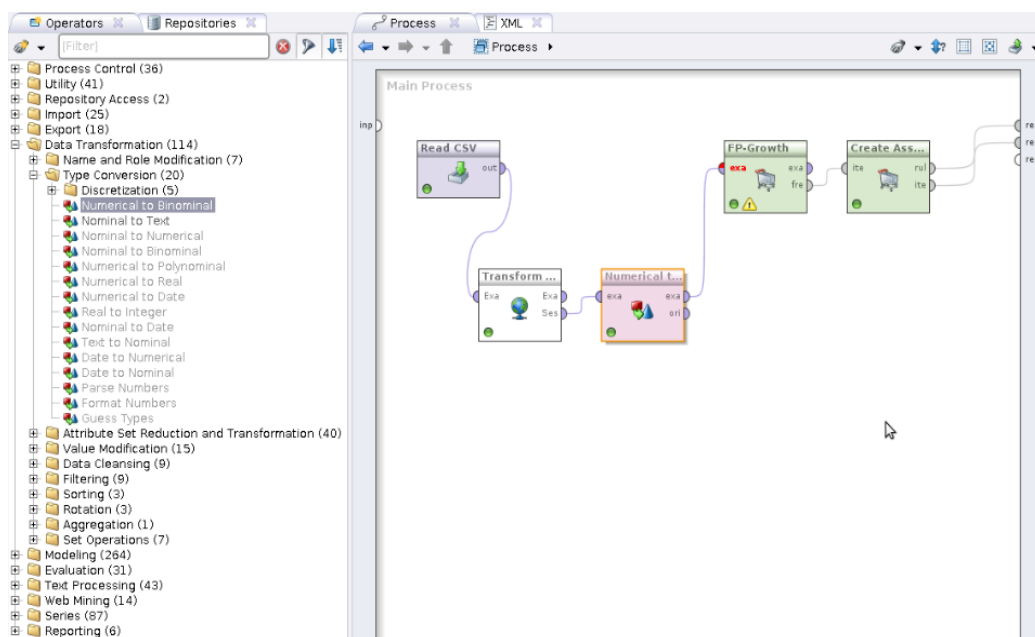


Figura 6: Exemplo do uso do operador *Numerical to Binomial*.

Depois de utilizar o operador *Numerical to Binomial* então é possível fazer uso do operador *FP-Growth* para a identificação dos itens freqüentes (figura 7).

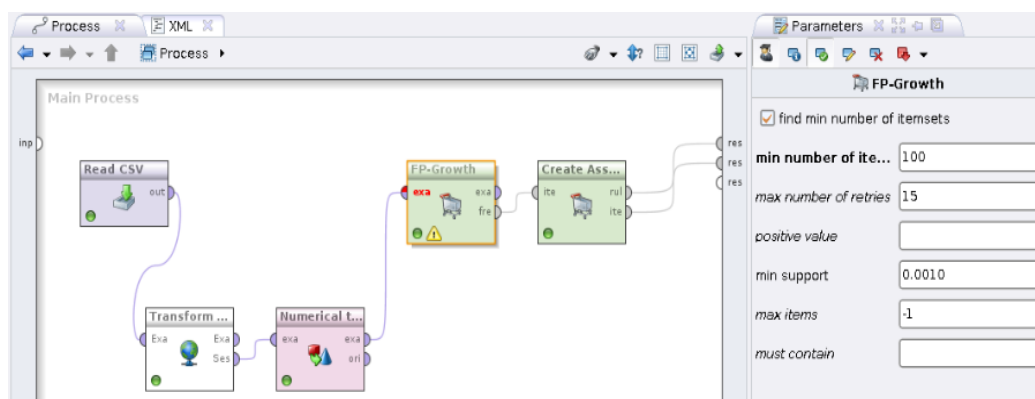
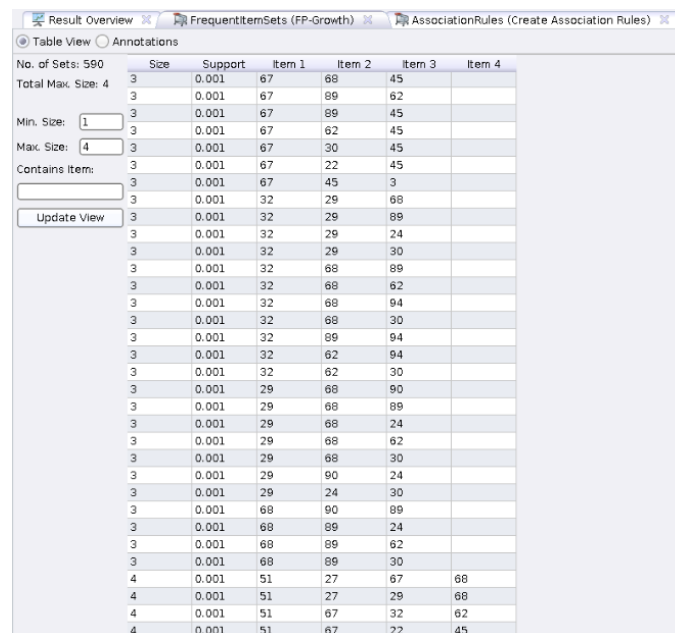


Figura 7: Exemplo de utilização do operador *FP Growth*.

Na figura 7 é possível visualizar os parâmetros necessários na configuração do operador. Existem duas maneiras de configuração: uma onde o operador deve encontrar no mínimo  $n$  conjuntos de itens freqüentes - que é o caso da configuração apresentada na figura 7. E outra onde é informado o valor do suporte mínimo. Quando a opção *find min number of itemsets* está habilitada

então o operador *FPGrowth* deve retornar um número mínimo de conjuntos de itens frequentes que também é configurável. Se a opção *find min number of itemsets* não for habilitada então o operador *FPGrowth* irá retornar apenas os conjuntos de itens frequentes que tiverem um suporte mínimo  $x$ . No caso da figura 7 este parâmetro está configurado com o valor 0.0010.

Ao executar este *script* é possível obter o resultado que é apresentado na figura 8. Nesta figura são apresentados conjuntos de itens frequentes com três e quatro identificadores de categorias de livros. Todos os conjuntos possuem um suporte de 0.001. Isto significa que estes itens foram acessados conjuntamente em 0.01% das transações. Este valor pode parecer baixo, no entanto, é um valor bom para análise de *log* de aplicações Web.



The screenshot shows a software interface with three tabs: 'Result Overview', 'FrequentItemSets (FP-Growth)', and 'AssociationRules (Create Association Rules)'. The 'FrequentItemSets (FP-Growth)' tab is active, displaying a table of frequent itemsets. On the left side of the table, there are controls for 'Min. Size' (set to 1) and 'Max. Size' (set to 4), and a 'Contains Item:' field. An 'Update View' button is located below these controls. The table itself has columns for 'Size', 'Support', and four item categories labeled 'Item 1', 'Item 2', 'Item 3', and 'Item 4'. The data shows various combinations of items with a consistent support of 0.001.

	Size	Support	Item 1	Item 2	Item 3	Item 4
	3	0.001	67	68	45	
	3	0.001	67	89	62	
	3	0.001	67	89	45	
	3	0.001	67	62	45	
	3	0.001	67	30	45	
	3	0.001	67	22	45	
	3	0.001	67	45	3	
	3	0.001	32	29	68	
	3	0.001	32	29	89	
	3	0.001	32	29	24	
	3	0.001	32	29	30	
	3	0.001	32	68	89	
	3	0.001	32	68	62	
	3	0.001	32	68	94	
	3	0.001	32	68	30	
	3	0.001	32	89	94	
	3	0.001	32	62	94	
	3	0.001	32	62	30	
	3	0.001	29	68	90	
	3	0.001	29	68	89	
	3	0.001	29	68	24	
	3	0.001	29	68	62	
	3	0.001	29	68	30	
	3	0.001	29	90	24	
	3	0.001	29	24	30	
	3	0.001	68	90	89	
	3	0.001	68	89	24	
	3	0.001	68	89	62	
	3	0.001	68	89	30	
	4	0.001	51	27	67	68
	4	0.001	51	27	29	68
	4	0.001	51	67	32	62
	4	0.001	51	67	22	45

Figura 8: Extrato do resultado do processamento do operador *FPGrowth*.

Os conjuntos de itens frequentes, como os apresentados na figura 8, são resultado do processamento do operador *FPGrowth* e podem ser obtidos pela porta de saída, chamada *freq*, do próprio operador (figura 9).

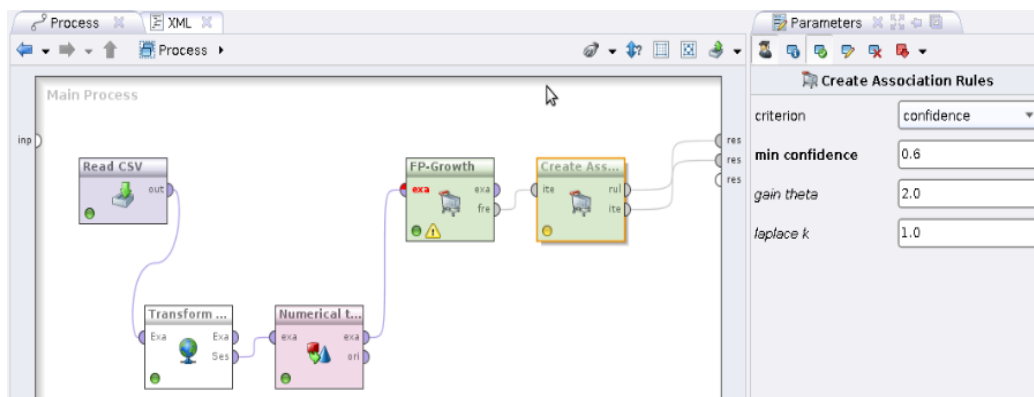


Figura 9: Exemplo de utilização do operador *Create Association Rules*.

Estes mesmos dados são a entrada para o operador *Create Association Rules* (*Modeling - Association and Item Set Mining*). Com base nos itens frequentes e no valor, parametrizável, de confiança mínimo (*min confidence*), o operador *Create Association Rules* gera as regras de associação. Um exemplo de como as regras de associação são visualizadas no RAPIDMINER é apresentado na figura 10

Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
51, 22, 45	67	0.001	0.947	1.000	-0.001	0.001	13.881	17.703
45, 3	67	0.001	0.895	1.000	-0.001	0.001	13.110	8.852
30, 45	67	0.001	0.825	1.000	-0.001	0.001	12.082	5.311
22, 45	67	0.001	0.814	1.000	-0.002	0.001	11.926	5.008
62, 45	67	0.001	0.788	1.000	-0.002	0.001	11.544	4.393
51, 67, 22	45	0.001	0.771	1.000	-0.002	0.001	53.939	4.312
67, 22, 45	51	0.001	0.771	1.000	-0.002	0.001	6.720	3.873
67, 32, 62	51	0.001	0.758	1.000	-0.002	0.001	6.603	3.659
89, 45	67	0.001	0.757	1.000	-0.002	0.001	11.088	3.831
51, 32, 62	67	0.001	0.746	1.000	-0.002	0.001	10.931	3.669
27, 67, 68	51	0.001	0.742	1.000	-0.002	0.001	6.467	3.437
32, 45	67	0.001	0.718	1.000	-0.002	0.001	10.520	3.303
68, 45	67	0.001	0.718	1.000	-0.002	0.001	10.520	3.303
68, 45	51	0.001	0.705	1.000	-0.002	0.001	6.142	3.002
27, 45	67	0.001	0.704	0.999	-0.002	0.001	10.311	3.145
51, 45	67	0.002	0.696	0.999	-0.003	0.002	10.198	3.065
51, 3	67	0.001	0.680	0.999	-0.002	0.001	9.964	2.912
22, 45	51	0.001	0.663	0.999	-0.002	0.001	5.774	2.625
51, 67, 32	62	0.001	0.653	0.999	-0.002	0.001	16.818	2.768
51, 67, 62	32	0.001	0.653	0.999	-0.002	0.001	9.976	2.692
51, 67, 68	27	0.001	0.636	0.999	-0.002	0.001	7.559	2.518
89, 45	51	0.001	0.635	0.999	-0.002	0.001	5.533	2.426
22, 45	51, 67	0.001	0.628	0.999	-0.003	0.001	86.901	2.668
62, 35	51	0.001	0.627	0.999	-0.002	0.001	5.459	2.371
89, 30	51	0.001	0.625	0.999	-0.003	0.001	5.444	2.361
51, 67, 45	22	0.001	0.621	0.999	-0.003	0.001	29.594	2.581
27, 45	51	0.001	0.617	0.999	-0.002	0.001	5.377	2.313
51, 22	67	0.001	0.614	0.999	-0.003	0.001	8.997	2.414
32, 45	51	0.001	0.603	0.999	-0.002	0.001	5.249	2.227
27, 29, 68	51	0.001	0.603	0.999	-0.002	0.001	5.249	2.227

Figura 10: Exemplo de regras de associação gerados pelo RAPIDMINER.

Cada linha na tabela apresentada na figura 10 é uma regra de associação. Uma regra de associação possui, no mínimo, uma lista de premissas, uma lista

de conclusões, o valor do suporte da regra e o valor da confiança da regra. Por exemplo, em algumas linhas da tabela é possível visualizar os valores apresentados na tabela 2.

Tabela 2: Extrato das regras de associação.

Premissas	Conclusões	Suporte	Confiança
51, 22, 45	67	0.001	0.947
45, 3	67	0.001	0.895
51, 67, 22	45	0.001	0.771

A partir dos valores apresentados na primeira linha da tabela 2 é possível gerar uma descrição de mais alto nível que explica o padrão encontrado: *em 94.7% dos casos quando um usuário acessa livros das categorias 51 e 22 e 45 ele também irá acessar livros da categoria 67*. Assim como para a segunda linha da tabela 2 tem-se a seguinte descrição: *em 89.5% dos casos quando um usuário acessa livros das categorias 45 e 3 ele também irá acessar livros da categoria 67*.

Além das informações sobre as premissas, conclusões, suporte e confiança de cada regra, o operador *Create Association Rules* da ferramenta RAPIDMINER gera os valores de *LaPlace*, *Gain*, *p-s*, *Lift* e *Conviction*. Todos estes valores fornecem dados para auxiliar na análise das regras geradas. No entanto, este texto não irá abordar a interpretação destes dados. Para obter maiores informações sobre a interpretação destes atributos é sugerido a leitura do artigo [2].

## 5 Considerações finais

Este texto apresentou o uso da ferramenta RAPIDMINER na mineração de regras de associação em servidores Web através de um estudo de caso onde o objetivo era identificar padrões de navegação em uma livraria virtual, mais especificamente, identificar por quais categorias de livros (i.e., artes, tecnologia, administração) o mesmo usuário navega.

Regras de associação é um formalismo útil para a identificação de itens que ocorrem em conjunto. No entanto, às vezes pode ser um tanto quanto difícil fazer a análise delas, principalmente, quando o número de regras geradas é muito alto. A ferramenta RAPIDMINER é uma ferramenta muito útil para a prototipação destes processos pois facilita a configuração de parâmetros, tais como: suporte mínimo e confiança. Também permite a fácil leitura dos dados brutos, eventuais transformações e visualização das regras geradas.

Além da execução dos *scripts* via interface gráfica, a ferramenta RAPIDMINER permite a execução dos scripts em modo *batch* e a integração dos seus

componentes em outras aplicações.

A ferramenta RAPIDMINER possui inúmeras funcionalidades e, conseqüentemente, pode ser aplicada em inúmeras situações. O objetivo deste texto não era explorar todas estas possibilidades, mas sim, fornecer apenas um exemplo de aplicação junto com conceitos de uma área nova de atuação e pesquisa: a área de mineração de padrões na web.

## Referências

- [1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *20th International Conference on Very Large Data Bases*, pages 478–499. Morgan Kaufmann, Los Altos, CA, 1994.
- [2] Eduardo C.Gonçalves. Regras de associação e suas medidas de interesse objetivas e subjetivas. *INFOCOMP Journal of Computer Science*, 4(1):26–35, 2005.
- [3] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, November 1996.
- [4] Bing Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer, 1st ed. 2007. corr. 2nd printing edition, January 2009.
- [5] Bing Liu, Wynne Hsu, and Yiming Ma. Integrating classification and association rule mining. In *Fourth International Conference on Knowledge Discovery and Data Mining*, pages 80–86. AAAI Press, 1998.
- [6] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. Yale: Rapid prototyping for complex data mining tasks. In Lyle Ungar, Mark Craven, Dimitrios Gunopulos, and Tina Eliassi-Rad, editors, *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940, New York, NY, USA, August 2006. ACM.
- [7] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, second edition, 2005.