WebPath – Uma Ferramenta para Mineração e Visualização de padrões de Navegação do Uso da *Web*

Cristian Tristão, Vantuir Pereira Teixeira, Karin Becker

Faculdade de Informática – Pontifícia Universidade do Rio Grande do Sul (PUCRS)
Av. Ipiranga 6681 – 90619-900 – Porto Alegre – RS – Brazil

ctristao@pop.com.br, vantuir@terra.com.br, kbecker@inf.pucrs.br

Resumo. *A Mineração do Uso da Web (MUW) visa descobrir padrões de navegação a partir da análise de logs de servidores Web. Este trabalho descreve as funcionalidades e a implementação de WebPath, uma ferramenta voltada à descoberta e interpretação de padrões seqüenciais de navegação. As características principais desta ferramenta são a busca dirigida de padrões a partir de critérios especificados por um usuário, bem como os recursos de visualização oferecidos para interpretação dos padrões. Testes preliminares revelam um ótimo desempenho dos algoritmos implementados.*

Abstract. *Web Usage Mining (WUM) aims to discover navigation patterns from the analysis of logs in Web servers. The present work describes the functionalities and implementation of WebPath, a support tool for the discovery of sequential navigation patterns. The striking features are the directed search for patterns based on user-specified criteria, as well as the visualization functionality for pattern interpretation. Preliminary tests disclose a good performance of the implemented algorithms.*

1. Introdução

O fluxo incessante de acessos às páginas *Web* reflete os conteúdos mais diversos, bem como costumes e necessidades pessoais ainda mais distintos, resultando em padrões de utilização ricos e diversificados. A Mineração do Uso da *Web* (MUW) [COO99, SRI00, SPI00] é a área de pesquisa que visa extrair padrões de navegação na *Web* tendo como principal fonte de dados o *log* de um servidor *Web*, que registra todos os acessos às páginas de um *site* feitos pelos usuários. A MUW tem sido aplicada aos mais diversos domínios, abrangendo desde o comércio eletrônico até a educação à distância.

O processo de MUW é dividido em três fases genéricas [COO99]: pré-processamento, descoberta de padrões e análise de padrões. O pré-processamento inclui a limpeza dos dados, identificação de usuários e sessões, complemento de caminho e enriquecimento semântico. A fase de descoberta de padrões restringe-se à aplicação de algoritmos de mineração de dados (e.g. seqüência, associação, classificação), visando a descoberta de padrões. A fase de análise tem por objetivo identificar entre os padrões minerados, aqueles que são interessantes, por meio da interpretação e validação destes no domínio.

A técnica de descoberta de padrões seqüenciais é particularmente interessante para a MUW, por expressar como os usuários interagem com os *sites Web*, isto é, quais páginas são acessadas e em qual ordem. O padrão seqüencial X → Y, por exemplo, mostra que a página

Y foi acessada após (imediatamente ou não) a página X. Uma medida de suporte é associada a este tipo de padrão, neste caso indicando em quantas sessões este padrão de navegação foi observado. Contudo os algoritmos clássicos de seqüência (e.g. [AGR94, MAN95]) possuem limitações quando contextualizados na MUW [SPI00]. Considerando o exemplo acima, mais importante do que conhecer a seqüência X → Y e seu suporte, é saber que apenas 10% dos que acessaram X conseguem convergir para Y. Isto poderia significar, por exemplo, que somente 10% daqueles que colocaram produtos em um carrinho virtual, efetuaram a compra. Este tipo de algoritmo também não captura eventuais desorientações do usuário, já que não distingue entre os vários acessos a uma mesma página no contexto de uma sessão. Finalmente, a análise dos padrões é uma das etapas mais críticas no processo. Tipicamente, os usuários conseguem apenas especificar critérios estatísticos sobre os padrões procurados (e.g. suporte), resultando em um enorme volume de regras a serem interpretadas. A redução do espaço de busca acaba sendo feita na fase de preparação de dados, no qual apenas um subconjunto dos dados é submetido à etapa de mineração.

A ferramenta *WebPath*, descrita neste trabalho, auxilia as fases de descoberta e de análise de padrões. *WebPath* permite dirigir a busca de padrões seqüenciais de navegação, por meio da especificação das propriedades dos padrões que devem ser minerados. As propriedades incluem o conteúdo (e.g. conter a página A), estatística (e.g. possuir pelo menos 5 acessos) ou estrutura (e.g. ser composto de até 5 páginas) dos padrões. Uma interface gráfica permite a formulação da consulta de forma amigável, sem necessidade de aprendizado de uma sintaxe específica. A tarefa de interpretação de padrões é facilitada tanto pelo número mais reduzido de padrões e pela restrição do foco de interesse, quanto pelos recursos oferecidos por *WebPath* para a visualização gráfica dos padrões seqüenciais. Testes de desempenho revelam que os algoritmos que implementam estas funcionalidades apresentam um desempenho muito bom.

O restante deste texto está estruturado como segue. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 apresenta as funcionalidades de *WebPath*, discutindo brevemente sua implementação. Resultados de desempenho são resumidos na Seção 4. Conclusões e trabalho futuros são endereçados na Seção 5.

2. Trabalhos Relacionados

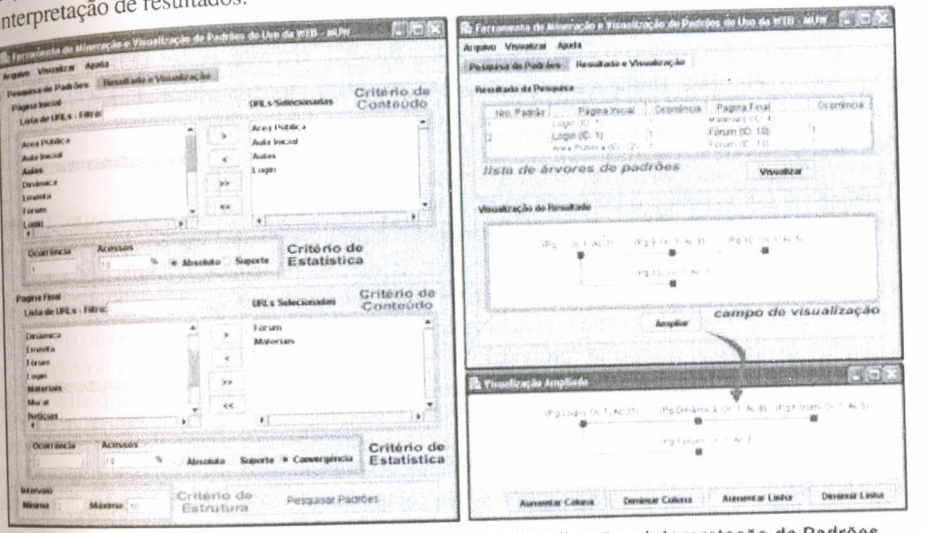
Os objetivos e aplicações da mineração do uso da *Web* são discutidos por Srivastava et al. [SRI00] e Spiliopoulou [SPI00]. Os problemas relacionados à baixa confiabilidade dos *logs* de servidores *Web*, bem como as principais técnicas utilizadas no pré-processamento destes dados são discutidos por Cooley et al. [COO99].

Algoritmos de padrões seqüenciais (e.g. [AGR94, MAN95]) foram propostos no contexto de seqüências de transações. Um problema típico é encontrar seqüências de itens adquiridos por um cliente, analisando suas várias transações distribuídas no tempo.

As limitações deste tipo de algoritmo para o contexto da *Web* são destacadas por Spiliopoulou et al. [SPI98, SPI00], as principais, entre elas, já abordadas na introdução deste trabalho. Estes autores propõem a ferramenta WUM (*Web Usage Miner*), a qual oferece uma linguagem poderosa de consultas (MINT) e recursos para visualização dos resultados. MINT é uma linguagem textual que possibilita dirigir a busca dos padrões seqüenciais usando parâmetros estatísticos, estruturais, e de conteúdo. *WebPath*, apresentada neste trabalho, tem um forte embasamento nos conceitos e mecanismos apresentados na abordagem de Spiliopoulou et al., diferenciando-se por oferecer uma interface gráfica mais intuitiva para formulação de consultas e visualização de resultados. Também, testes de desempenho não foram encontrados na literatura para o sistema WUM.

3. *WebPath*

WebPath é uma ferramenta para a mineração de padrões seqüenciais de uso da *Web*. Focada nas fases de descoberta de padrões e de análise de padrões, ela assume como entrada um *log* pré-processado contendo sessões de navegação. Sobre este *log*, o usuário pode especificar consultas que dirijam a busca dos padrões de navegação e visualizar os resultados. A Figura 1.(a) ilustra a interfaces de consulta, e a Figura 1.(b) mostra a interface para a visualização e interpretação de resultados.



(a) Consultas (b) Visualização e Interpretação de Padrões

Figura 1 – Interfaces de *WebPath*

3.1. Árvore Agregada

A *árvore agregada*, inspirada na abordagem de Spiliopoulou et al. [SPI98], é uma representação interna criada por *WebPath* para representar o *log* de entrada fornecido pelo usuário. Assume-se que o *log Web* tenha passado por todas as tarefas típicas de pré-processamento [COO99], resultando em um conjunto ordenado de registros de acessos a páginas, agrupados por sessão. O objetivo da árvore agregada é representar de forma compacta todas as trilhas de navegação percorridas no *site* pelo conjunto de todos os usuários. As trilhas dos diferentes usuários são unificadas com base nos prefixos em comum. Cada nodo da árvore corresponde a uma ocorrência de página em uma trilha. Cada nodo é representado por uma tripla <E; O; A>, na qual "E" representa a URL da página, "O" a ocorrência da página na trilha (ordem em que a página "E" aparece na trilha) e o "A" o número de acessos àquela ocorrência de página trilha.

A Figura 2 ilustra uma árvore agregada, considerando as 5 sessões contidas no *log* de entrada (topo da figura à esquerda). O nodo "ROOT" é a raiz da árvore e nele é totalizado o número

de sessões contidas no *log* de entrada. Após este nodo, aparecem as trilhas dos usuários unificadas sempre que possuírem prefixos em comum. Assim, as bifurcações representam a parte não comum de trilhas. No exemplo, todas as trilhas iniciam por "X" ou "C". A trilha 5 possui duas visitas à página "C", sendo estas diferenciadas na árvores agregada pela ocorrência 1 e 2. O mesmo para a página "D".

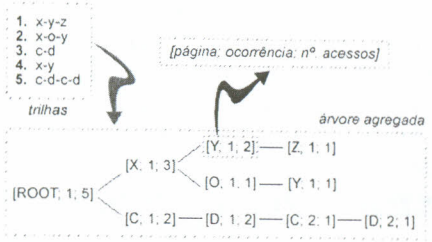


Figura 2. Unificação de trilhas para a geração da árvore agregada

3.2. Recursos de Consulta

Entende-se por consulta a busca de padrões seqüenciais dirigida por critérios especificados pelo usuário. Estes critérios podem ser divididos em 3 categorias: (a) *conteúdo* (URL ou título que identifica a página), (b) *estrutura* (tamanho do padrão) e (c) *estatística* (a ocorrência de uma página na trilha, ou uma medida relativa à freqüência de acessos feitos a uma página). No caso da medida estatística de freqüência de acesso, as opções são o *número absoluto de acessos*, o *suporte* (proporção de acessos sobre o total de sessões no *log*), ou o grau de *convergência* (razão dos acessos entre a página inicial e final de um padrão), este último somente disponível para página final.

A combinação destes diferentes tipos de critérios permite ao usuário realizar três categorias de consulta, as quais são descritas abaixo e ilustradas considerando a árvore agregada exibida no topo da Figura 3.

- *Pesquisa Início*: deseja-se conhecer os diferentes destinos a partir de uma dada página, isto é, para onde se dirigem os usuários. Portanto, as restrições são definidas somente sobre a página inicial, indicando opcionalmente o tamanho do padrão. Na consulta da Figura 3.(a), deseja-se saber todos padrões que iniciem pela primeira ocorrência da página X em uma trilha.
- *Pesquisa Fim*: permite descobrir a origem dos caminhos que chegam a uma determinada página, e portanto são definidos critérios somente para a página final (e opcionalmente, critério de tamanho). A consulta da Figura 3.(b) procura todos caminhos que levaram à re-visita de uma página qualquer (ocorrência igual a 2).
- *Pesquisa Início-Fim*: serve para encontrar os diferentes caminhos percorridos pelos usuários entre dois pontos com propriedades especificadas, sendo os critérios especificados sobre as páginas inicial e final (e opcionalmente, critério de tamanho). A consulta da Figura 3.(c) procura todos os caminhos que iniciaram na página X e terminaram em uma página que possui no mínimo 2 acessos. A Figura 1.(a) apresenta outro exemplo desta categoria de consulta.

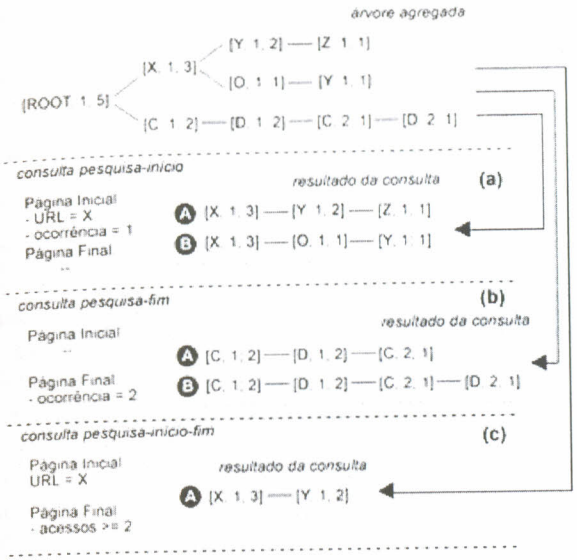


Figura 3. Tipos de Consultas

3.3. Recursos de Visualização

A atividade de interpretação de padrões é de natureza difícil. O volume de padrões a ser interpretado, e a relação destes com o domínio para determinação do grau de interesse estão entre as dificuldades a serem vencidas. Observa-se na Figura 3 que o resultado de cada consulta é um conjunto (possivelmente vazio) de padrões seqüenciais. Cada padrão é uma trilha de navegação encontrada em uma ou mais sessões de usuários. Note-se que há vários padrões que, ainda que representem trilhas diferentes, possuem prefixos comuns, como por exemplo, os padrões da consulta ilustrada na Figura 3.(a). Também pode haver interseção entre os padrões retornados por uma mesma consulta, como no caso da consulta ilustrada na Figura 3.(b), no qual o padrão A está totalmente incluído no padrão B.

Visando reduzir o número de padrões, bem como simplificar a sua visualização e interpretação, os padrões são visualizados na forma de *árvore resultado*, cujos nodos são representados pela mesma tripla <E; O; A> descrita na Seção 3.1. Contudo, a decisão de quais padrões devem ser unificados depende do tipo de consulta feita:

- Pesquisa Início-Fim e Pesquisa Fim: o algoritmo de visualização unifica todos os padrões que possuem a mesma ocorrência de página tanto no nodo inicial do padrão, quanto no nodo final. Assim, tomando-se o conjunto de padrões retornados por uma consulta, para cada combinação de nodo inicial/final de padrão será gerada uma árvore resultado distinta. A Figura 4.(a) representa a árvore resultado gerada para um conjunto de padrões deste caso (i.e. todos os padrões iniciam pela primeira ocorrência de K e terminam pela primeira ocorrência de Y).
- Pesquisa Início: o único critério para unificar os padrões é possuir a mesma ocorrência de página inicial. A Figura 4.(b) representa a árvore resultado para dois padrões retornados por esta categoria de consulta.

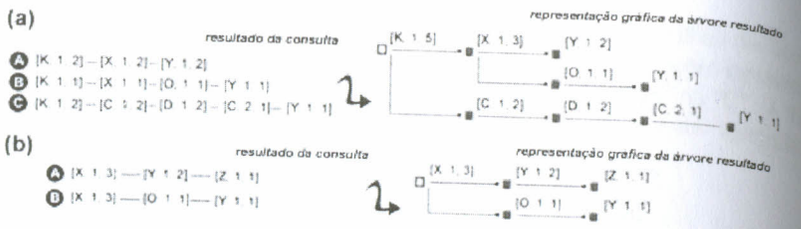


Figura 4. Árvores Resultado

Note-se que em alguns casos os acessos dos nodos unificados são acumulados (e.g. na Figura 4.(a), o nodo unificado K totalizou 5 acessos), e em outros não (e.g. nodo X da Figura 4.(b)). Neste último exemplo, supondo que os padrões sejam aqueles retornados pela consulta 3.(a), trata-se exatamente do mesmo nodo na árvore agregada, portanto a soma implicaria em erro. Tal situação é controlada pelo algoritmo de unificação.

Em termos de interface, o usuário visualiza os padrões de três formas (Figura 1.(b)). Na janela superior (lista de árvores de padrões), são apresentados o nodo inicial e final de todas as árvores resultado geradas pelo algoritmo de unificação de padrões. Selecionando uma árvore resultado, o usuário pode inspecionar sua estrutura (janela ao centro). Sob demanda esta visualização pode ser melhorada (janela inferior), como no exemplo, no qual se coloca o nome das páginas, e não sua identificação interna.

3.4. Projeto da Ferramenta

WebPath foi implementada utilizando a linguagem Java. A Figura 5 apresenta o diagrama de classes UML do projeto, o qual foi desenvolvido de acordo com o padrão arquitetural MVC - *Model-View-Controller*. Os algoritmos não são apresentados por limitações de espaço, podendo ser consultados em Tristão et al. [TRI04]. As classes estão assim divididas:

- Objetos de negócio (*Model*) – *Page*, *PathElement*, *TreeNode* e *StructData* (a qual reúne várias outras classes), utilizadas para representar e manipular a árvore agregada;
- Interface com o usuário ou outro sistema (*View*) – *FullScreen* e *InterfaceScreen* para informar os dados de consulta e visualizar os padrões resultantes;
- Controle do fluxo da aplicação (*Controller*) – *MinerController*, *Loader*, *Miner* e *Unifier* que realizam o controle dos processos existentes na ferramenta.

4. Análise de Desempenho

Para analisar o desempenho da implementação de *WebPath*, foram realizados testes, visando verificar a demanda de memória e de processamento de cada módulo da ferramenta. Foi utilizado um Pentium III - 1Ghz, com 256 MB de memória RAM, e plataforma Windows 2000. Para os testes, utilizou-se um *log* real de um *site* vinculado a PUCRS VIRTUAL, departamento de educação a distância da PUCRS. Ele refere-se a um curso intensivo de extensão com quinze alunos. A relação dos registros foi puramente quantitativa, visando testar um volume típico e expressivo de aplicações de Mineração do Uso da *Web*. Os dados foram pré-processados usando a ferramenta proposta por Marquardt et al. [MAR04].

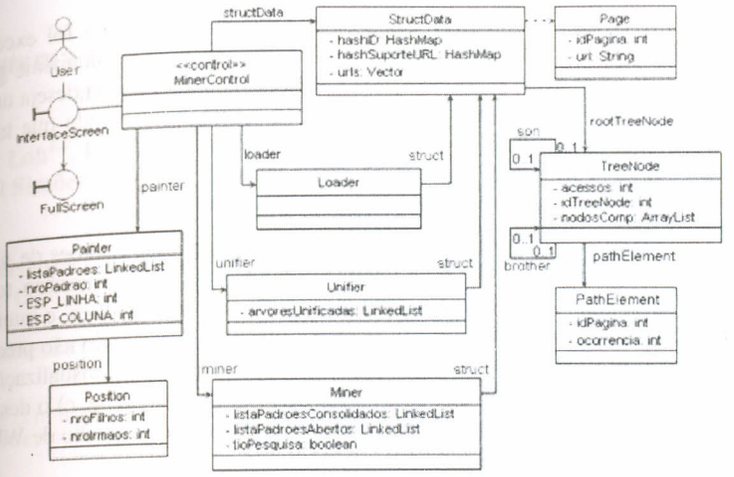


Figura 5. Projeto de *WebPath* (Diagrama de Classes UML)

Testes preliminares mostraram resultados encorajadores. Em termos absolutos, a ferramenta gastou 6 segundos para construir uma árvore agregada para 100.000 registros (42561 nodos), um volume de dados expressivo em MUW. Este tempo é linear. A Figura 6 mostra o tempo tomado para duas dentre as consultas realizadas, considerando árvores com diferentes números de nodos agregados. Estas consultas foram propositadamente bastante genéricas: todas as combinações envolvendo qualquer uma das páginas do *site* (Figura 6.(a)) e qualquer caminho iniciando por uma página com no mínimo 5 acessos (Figura 6.(b)). Apesar de retornarem um volume expressivo de padrões (53060 e 21970, considerando a maior árvore testada para cada consulta), o tempo absoluto para realizar tais pesquisas foi muito bom. O tempo de processamento é em função do número de nodos da árvore agregada e de padrões encontrados, não tendo, ainda, sua complexidade sido totalmente caracterizada. Este bom desempenho deve-se à arquitetura escolhida, a qual armazena todas suas estruturas em memória, evitando assim o acesso a disco, mais custoso. A unificação e visualização de padrões também apresentaram resultados animadores. O conjunto de testes considerando todas as funcionalidades de *WebPath* pode ser consultado em Tristão et al. [TRI04].

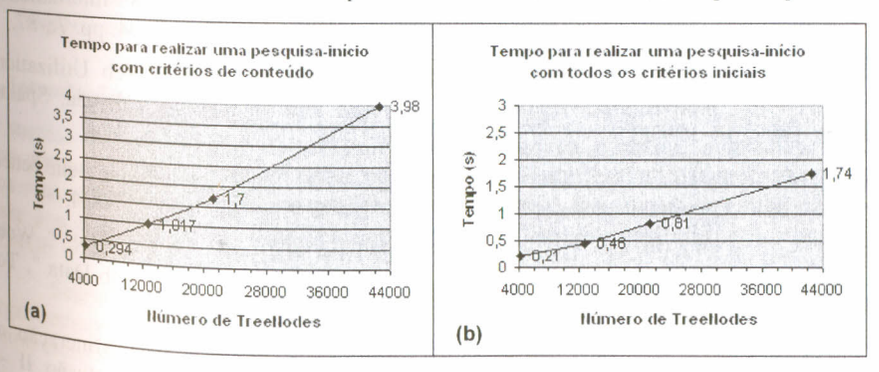


Figura 6 – Pesquisa Início (tempo X nº de nodos)

Tais resultados para uma quantidade expressiva de dados, evidencia a eficácia dos

algoritmos desenvolvidos, salientando que devido o limite de memória foi excedido
somente 2 casos nos testes desenvolvidos, na qual as consultas eram altamente genéric
Contudo, o uso de *WebPath* é justamente recomendado quando o analista deseja uma bus
mais dirigida de padrões, para facilitar o processo de interpretação, sendo que tal tipo
consulta não deve ocorrer na prática.

5. Conclusões

Este trabalho descreveu *WebPath*, uma ferramenta de mineração de padrões de navegaç
seqüenciais na *Web*. Por meio de uma interface gráfica amigável, o usuário tem à s
disposição recursos poderosos para minerar de forma dirigida um *log Web*, e interpretar
resultados. Em comparação com a ferramenta WUM [SPI98]: a) o usuário não precisa sab
nenhuma sintaxe para formular suas consultas; b) os recursos de visualização fora
simplificados sem perda no poder de interpretação, facilitando esta tarefa, c) o desempenh
não pôde ser comparado, pois não foram encontrados testes de desempenho de WUM, ma
os resultados de *WebPath* foram muito bons.

Trabalhos futuros incluem integração de *WebPath* com recursos de pré-processament
[MAR04], exploração de hierarquias de generalização no algoritmo, teste com maior volum
de dados, bem como uma avaliação da complexidade dos algoritmos propostos.

References

[AGR94] AGRAWAL, R.; SRIKANT, R. "Mining Sequential Patterns". Researcl
Report RJ 9910, IBM Almaden Research Center, San Jose, California, Oct. 1994.

[COO99] COOLEY R., MOBASHER B. and SRIVASTAVA J. "Data preparation fo
mining world wide Web browsing patterns". Knowledge and Information Systems
vol. 1, nº 1, Feb. 1999, pp. 5-32.

[MAN95] MANILLA, H., TOIVONEN, H. & VERKANO, A. "Discovering Frequent
Episodes in Sequences". In Proceedings of the First International Conference on
Knowledge Discovery and Data Mining, 1995. pp. 210-215.

[MAR04] MARQUARDT, C.; BECKER, K.; RUIZ, D. A pre-processing tool for web
usage mining in the distance education domain. In: Proc. of the 8th International
Symposium on Database Engineering Applications, Coimbra, July 2004. pp. 78-87.

[SPI98] SPILIOPOULOU, M. and FAULSTICH, L. C. "WUM: a Web Utilization
Miner". In: International Workshop on the Web and Databases, Valencia, Spain,
Mar. 1998, pp. 109-115.

[SPI00] SPILIOPOULOU, M. "Web Usage Mining for Site Evaluation. Making a site better
fit its users". Communications of the ACM, vol.43, nº 8, Aug. 2000, pp. 127-134.

[SRI00] SRIVASTAVA, J.; COOLEY, R.; DESHPANDE, M.; TAN, P.N. " Web
Usage Mining: Discovery and Applications of Usage Patterns from Web Data ", To
appear in SIGKDD Explorations, vol. 1, nº 2, 2000, pp. 12-23.

[TRI04] TRISTÃO, C.; TEIXEIRA, V. P. "Uma Ferramenta para a Mineração e
Visualização de Padrões de Navegação na Web". Trabalho de Conclusão II –
Faculdade de Informática, PUCRS, Porto Alegre, Junho de 2004.