Regras de Associação
Extracção de Conhecimento de Dados II
Mestrado em Inteligência Artificial e Sistemas Inteligentes

António Jesus Monteiro de Castro
Aluno Nº 040594004

Faculdade de Engenharia da Universidade do Porto
frisky.antonio@gmail.com

17 de Abril de 2005

Resumo

Neste trabalho analisei um conjunto de dados correspondentes à utilização pelos tripulantes da TAP Air Portugal do Portal DOV, um portal *Web* [1], que se encaixa na categoria de portais B2E *Business to Employee*. Através da análise dos *Web Logs* correspondentes aos servidores de produção e utilizando o programa *Caren* [2], uma implementação do algoritmo *APRIORI* [8], procurei conhecer os hábitos de navegação dos tripulantes, extraindo regras de associação das páginas consultadas pelos tripulantes. As regras foram analisadas, tentando conhecer as que são mais interessantes para os fins em vista, tendo sido utilizado o *PEAR* [7] e outras ferramentas para fazer esse trabalho. Os resultados obtidos permitirão alterar o funcionamento do Portal de forma a facilitar a navegação, colocando de uma forma dinâmica as páginas mais consultadas de acordo com as regras obtidas.

Palavras-chave: Regras Associação, APRIORI, Portal DOV, Caren, PEAR.

1 Introdução

Para dar cumprimento ao enunciado deste trabalho, vou analisar um conjunto de dados correspondente à utilização do site Portal DOV [1], um site de acesso exclusivo pelos tripulantes da TAP Air Portugal. O objectivo é analisar e seleccionar regras de associação e conjuntos frequentes, para que possamos saber como é que os tripulantes navegam no site, quais os temas que mais os interessam e a forma como os relacionam. Com esta informação pretende-se, posteriormente, implementar mecanismos no site que permitam sugerir para cada página ou conjuntos de páginas *links* interessantes e implementar um sistema de menus personalizados que serão adaptados dinamicamente aos hábitos de navegação de cada utilizador.
Para atingir este objectivo vou efectuar as seguintes etapas:

- Seleccionar os dados.
- Limpar os dados.
- Identificar os utilizadores.
- Identificar as sessões.
- Identificar as transacções.
- Gerar as regras de associação com o Caren [2].
- Seleccionar e explorar as regras tendo em conta os objectivos.

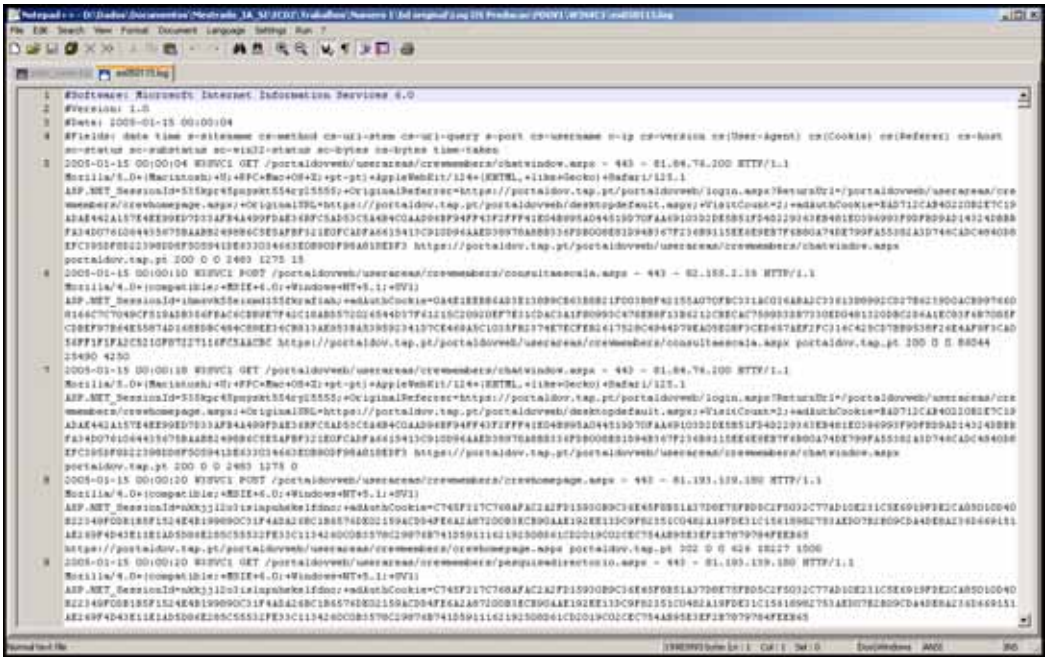
A secção 2 explica de onde vieram os dados e como os obtive. A secção 3 descreve o trabalho que realizei para limpar os dados. A secção 4 descreve a tentativa de identificar os utilizadores e de os relacionar com os dados existentes nos *Logs* dos servidores. A secção 5 explica o processo de identificar todas as sessões existentes nos registos. A secção 6 mostra como dos dados obtidos até esta fase passei para a identificação das transacções de forma a ter os registos no formato *Basket*. Na secção 7 descrevo o processo de geração das regras de associação, utilizando o Caren [2], quais os parâmetros utilizados e porquê. Na secção 8 procuro seleccionar as regras mais interessantes de acordo com os objectivos definidos, falando um pouco das medidas de interesse utilizadas e utilizando folhas de cálculo e o programa PEAR [7] como auxiliares para realizar esta tarefa. Na secção 9 falo um pouco da minha experiência com uma versão Beta do SQL Server 2005 [3].
Finalmente, este documento também contém uma secção dedicada às conclusões e melhoramentos futuros.

2 Selecção dos Dados

Pelo simples facto dos tripulantes navegarem no site são criados dois tipos de registos:

- Os *Logs* dos servidores *Web*.
- Os registos de entrada quando um tripulante faz *Login*.

Um exemplo dos *Web Logs* no seu formato original é o seguinte:



Os registos dos *Logins* têm o seguinte aspecto:

TRANS_DATE	TRANS_SEQ	HOST_NAME	PROCESS_ID	HOST_USR_ID	APPL_ID	USR_ID	TRANS_CD	TRANS_DETL	REM	ROWID
01-02-2005 8:10	810712	cra	0	CRA	CRA	CRA	LOGIN	ICRA	12956.9 CCB J.C VASCONCELOS	AAAA23AAKAABP0tAAD
01-02-2005 8:14	810731	cra	0	CRA	CRA	CRA	LOGIN	ICRA	13084.9 CCB R.M SILVEIRA	AAAA23AAKAABP0tAAW
01-02-2005 8:15	810749	cra	0	CRA	CRA	CRA	LOGIN	ICRA	15676.0 OPT A.E ENCARNACAO	AAAA23AAKAABP0tAAo
01-02-2005 8:01	810638	cra	0	CRA	CRA	CRA	LOGIN	ICRA	13184.7 CPT F.D RODRIGUES	AAAA23AAKAABP0uAAO
01-02-2005 8:02	810640	cra	0	CRA	CRA	CRA	LOGIN	ICRA	09611.5 SCB V.M SANTOS	AAAA23AAKAABP0uAAI
01-02-2005 7:57	810593	cra	0	CRA	CRA	CRA	LOGIN	ICRA	21210.0 CAB M.T LIMA	AAAA23AAKAABP0vAAs
01-02-2005 7:50	810527	cra	0	CRA	CRA	CRA	LOGIN	ICRA	12507.0 CPT A.C FERREIRA	AAAA23AAKAABP0wAA+
01-02-2005 7:28	810360	cra	0	CRA	CRA	CRA	LOGIN	ICRA	18155.2 CAB F.C FIGUEIROA	AAAA23AAKAABP0yAA0
01-02-2005 6:57	810117	cra	0	CRA	CRA	CRA	LOGIN	ICRA	09647.0 SCB A.M CARNEIRO	AAAA23AAKAABP01AAy
01-02-2005 6:49	810045	cra	0	CRA	CRA	CRA	LOGIN	ICRA	12040.2 CCB F.M VIANNA	AAAA23AAKAABP02AAt
01-02-2005 6:37	809880	cra	0	CRA	CRA	CRA	LOGIN	ICRA	13929.5 CAB H.R VENTURA	AAAA23AAKAABP04AAa
01-02-2005 6:19	809713	cra	0	CRA	CRA	CRA	LOGIN	ICRA	23162.1 CAB P.A JERONIMO	AAAA23AAKAABP06AAJ
01-02-2005 5:52	809652	cra	0	CRA	CRA	CRA	LOGIN	ICRA	26017.4 CAB P.S MILHEIRAS	AAAA23AAKAABP07AAT
01-02-2005 3:45	809540	cra	0	CRA	CRA	CRA	LOGIN	ICRA	17961.4 CAB L.F OLIVEIRA	AAAA23AAKAABP09AAq
01-02-2005 3:53	809552	cra	0	CRA	CRA	CRA	LOGIN	ICRA	15399.9 CCB A.D MONIZ	AAAA23AAKAABP0BAA2
01-02-2005 3:53	809553	cra	0	CRA	CRA	CRA	LOGIN	ICRA	15399.9 CCB A.D MONIZ	AAAA23AAKAABP0BAA3
01-02-2005 2:40	809429	cra	0	CRA	CRA	CRA	LOGIN	ICRA	11936.2 OPT J.L FERREIRA	AAAA23AAKAABP0+AAB
01-02-2005 2:40	809430	cra	0	CRA	CRA	CRA	LOGIN	ICRA	10737.7 CAB M.M SARAIVA	AAAA23AAKAABP0+AAC
01-02-2005 2:40	809431	cra	0	CRA	CRA	CRA	LOGIN	ICRA	11936.2 OPT J.L FERREIRA	AAAA23AAKAABP0+AAD
01-02-2005 2:40	809432	cra	0	CRA	CRA	CRA	LOGIN	ICRA	11936.2 OPT J.L FERREIRA	AAAA23AAKAABP0+AAE
01-02-2005 2:41	809433	cra	0	CRA	CRA	CRA	LOGIN	ICRA	25854.1 OPT F.M CARVALHO	AAAA23AAKAABP0+AAF
01-02-2005 2:49	809436	cra	0	CRA	CRA	CRA	LOGIN	ICRA	25986.1 CAB R.A ROCHA	AAAA23AAKAABP0+AAI

A informação seleccionada diz respeito ao primeiro trimestre de 2005 e contém 3,899,823 registos antes de se fazer qualquer preparação dos dados.
Como se pode verificar o trimestre possui demasiados registos e muitos deles são irrelevantes para o tipo de trabalho que vou fazer. Para facilitar a tarefa de limpeza dos registos que não interessam, resolvi carregar os dados para uma base de dados em MySQL [4], tendo sido necessário criar uma tabela com a estrutura correspondente aos dados dos vários *logs* cuja estrutura é a constante da seguinte tabela:

Nome	*Tipo*	*Descrição*
Date	Varchar	Data do registo
Time	Varchar	Hora do registo
s-sitename	Varchar	Código do site a que o utilizador acedeu. Neste caso W3SVC1
Cs-method	Varchar	Tipo de acesso GET/POST
Cs-uri-stem	Varchar	Página, imagem, script, outros, disponibilizados pelo servidor
Cs-query	Varchar	Parâmetros passados na consulta das páginas
s-port	Numeric	Porto através do qual se efectuou o acesso
Cs-username	Varchar	Identificação do utilizador (se Windows Authentication)
c-ip	Varchar	Endereço IP do computador que acedeu ao servidor
Cs-version	Varchar	Versão do HTTP
Cs-useragent	Varchar	User agent correspondente ao browser utilizado
Cs-cookie	Varchar	Id do cookie de sessão/autorização
Cs-referer	Varchar	Referer da consulta efectuada (se houver)
Cs-host	Varchar	Host acedido, neste caso, portaldov.tap.pt
Sc-status	Numeric	Código que identifica o sucesso ou não do acesso. 200=sim
Sc-subsystem	Numeric	
Sc-win32-status	Numeric	
Sc-bytes	Numeric	
Time-taken	Numeric	Tempo que demorou a passar a informação

Depois de ter a base de dados criada é necessário carregar os ficheiros que estão em vários ficheiros de texto para MySQL. Como o site está a funcionar numa *Web Farm* composta por dois servidores de produção, existe um ficheiro de *Web Logs* para cada servidor. Assim, carreguei a informação de ambos os servidores através do seguinte comando SQL:

LOAD DATA INFILE nome ficheiro INTO TABLE weblog FIELDS TERMINATED BY ' ' LINES TERMINATED BY '\\n';

Esta instrução foi repetida para cada dia do trimestre e para os *Logs* de cada servidor de produção.

3 Limpeza dos Dados

Tendo a informação em MySQL é mais fácil apagar os registos que não interessam. Neste caso, foram apagados os registos que correspondem às condições indicadas na tabela seguinte:

Condição	*Descrição*
Cs-cookie = null	Só interessam os acessos pós ter havido uma autenticação.
Cs-host = localhost	Só interessam os acessos ao site do Portal.
Sc-status <> 200	Só interessa analisar os acessos efectuados com sucesso. Os que originaram erro não interessam.
Cs-method <> GET	Só interessa a informação obtida pelo utilizador.
Cs-uri-stem LIKE '/portaldovweb/IgScripts'	Os acessos correspondentes à execução dos scripts não interessam.
Cs-uri-stem LIKE '/portaldovweb/css/'	Idem para as CSS
Cs-uri-stem LIKE '/portaldovweb/IgImages'	Idem para as Imagens dos controlos
Cs-uri-stem LIKE '/portaldovweb/Images'	Idem para as Imagens do site.
Cs-uri-stem LIKE '/portaldovweb/login.aspx'	O acesso à página para fazer o login não interessa incluir.
Cs-uri-stem LIKE '/portaldovweb/errorpages'	As páginas de erro padrão também não interessam.

Após esta etapa a base de dados ficou com 490,960 registos. Bastante menos que os perto de 4 milhões iniciais.
Agora é necessário eliminar os atributos desnecessários. Como o que interessa é ter a informação final num formato *Basket* entendi que os atributos necessários para a realização das próximas fases são os constantes da tabela seguinte, tendo sido eliminados todos os outros.

Atributo	*Descrição*
Date	Data do registo
Time	Hora do registo
Cs-cookie	Identificador da sessão e do cookie de autorização
Cs-uri-stem	Página consultada pelo utilizador
c-ip	Endereço IP do computador que acedeu ao site

Com esta estrutura e somente com os registos necessários já é possível passar para a fase seguinte.

4 Identificação dos Utilizadores

Em termos teóricos é possível e desejável efectuar a associação entre o tripulante que fez o *Login* e a sua ficha pessoal na empresa, obtendo, desta forma, outro tipo de informações muito interessantes, tais como a idade, a categoria profissional, a frota a que pertence, entre outros.
Infelizmente, ao olhar para o formato de ambos os registos, verifico que não é possível acrescentar mais informação sobre o utilizador, pois não existe maneira eficaz de relacionar os *Web Log* com o registo do *Login* dos tripulantes. Apesar de ter no *Web Log* o endereço IP da máquina do utilizador e, também, o *Cookie* da sessão e o da autorização, o ficheiro com os Logins não regista nenhuma dessa informação. É uma pena porque a informação sobre a idade, categoria profissional e a frota, era essencial para se conhecer melhor os hábitos de navegação dos utilizadores.
Uma maneira possível de resolver este assunto seria aplicar qualquer um dos métodos referidos no capítulo *5.2 User Identification* do artigo de Cooley [5] ou, então, usar um método semelhante mas que em vez de utilizar os endereços IP e características do *Browser* conforme indicado em [5], usasse a data e hora do *login* e, através desta informação, tentasse obter os registos da sessão que está registada nos *Web Logs*.
No entanto, entendi que para efeito deste trabalho, não valeria a pena estar a aumentar o grau de complexidade e, consequentemente de tempo, para a realização do mesmo.
De qualquer das formas, esta análise serviu para proceder à alteração do registo do *Login* do tripulante no Portal, de forma a incluir o identificador de sessão e, assim, poder no futuro fazer uma análise mais detalhada.

5 Identificação das Sessões

A identificação das sessões no conjunto de dados que analisei ficou muito mais facilitada do que pensei porque os *Web Logs* registam o identificador da sessão na coluna *Cs-cookie*, embora esse identificador esteja misturado com outros dados, como se poderá ver através deste exemplo:

ASP.NETSessionId=hkiryhasg4agpl45dx4a5w45;
+adAuthCookie=42CF7B26EA7517AA56A7FE1D63703633E03FF9

De notar que o acesso autenticado ao Portal sem ter os *Cookies* activos no *Browser* do utilizador não é permitido, pelo que é razoável supor que os acessos registados se referem a todos os acessos realmente efectuados ao Portal.
Queria realçar que **não tive em consideração os problemas relacionados com *Proxies Servers, Firewalls, Caches*** conforme indicado no artigo do Cooley [5].
Para poder isolar a informação referente ao identificador de sessão, criei uma nova coluna na tabela a que chamei *SessionID* e, através de comandos SQL, copiei para lá somente a informação correspondente ao identificador de sessão de cada registo. Para ter a certeza que estava correcto, ordenei os registos por Sessão, Data e Hora.

6 Identificação das Transacções

Antes de poder utilizar o software CAREN para analisar os dados é necessário ter os registos num formato *Basket*, ou seja, uma coluna com a identificação da transacção e a outra coluna com o artigo que foi objecto da aquisição. No caso que estou a analisar, a transacção corresponde ao identificador da sessão e as páginas consultadas correspondem aos artigos adquiridos.
Para poder realizar esta operação criei uma tabela com as colunas conforme indicado na seguinte tabela:

Atributo	*Descrição*
SessionID	Identificador da Sessão
URL	Página Consultada

Para terminar esta fase é necessário garantir que não existem registos repetidos dentro da mesma transacção. Para isso utilizei o comando adequado do MySQL e eliminei os registos em duplicado. Um exemplo dos dados nesta nova tabela é o seguinte:

SessionID	URL
SessionId=015iqx45yuscht55ybz5zxne	crewhomepage.aspx
SessionId=015iqx45yuscht55ybz5zxne	consultatripulacoes.aspx
SessionId=015iqx45yuscht55ybz5zxne	mostratripulacao.aspx
SessionId=01fwhlnaf3uqoouks3ktk555	crewhomepage.aspx
SessionId=01fwhlnaf3uqoouks3ktk555	pesquisadirectorio.aspx
SessionId=01fwhlnaf3uqoouks3ktk555	consultatrocas.aspx
SessionId=01fwhlnaf3uqoouks3ktk555	crewhomepage.aspx
SessionId=01fwhlnaf3uqoouks3ktk555	consultarotacao.aspx
SessionId=01fwhlnaf3uqoouks3ktk555	mostratripulacao.aspx
SessionId=03pvl2ympvnqbn45v4gbia55	crewhomepage.aspx
SessionId=03pvl2ympvnqbn45v4gbia55	consultatrocas.aspx
SessionId=03pvl2ympvnqbn45v4gbia55	mostraescala.aspx
SessionId=03pvl2ympvnqbn45v4gbia55	consultarotacao.aspx
SessionId=03pvl2ympvnqbn45v4gbia55	mostraescala.aspx
SessionId=03pvl2ympvnqbn45v4gbia55	mostraescala.aspx
SessionId=03pvl2ympvnqbn45v4gbia55	mostraescala.aspx
SessionId=03pvl2ympvnqbn45v4gbia55	mostraescala.aspx
SessionId=03pvl2ympvnqbn45v4gbia55	mostraescala.aspx
SessionId=04ipi4ykjo5yvp45dguf2j45	crewhomepage.aspx

Para facilitar a leitura das regras retirei do URL de cada registo as palavras que eram comuns e que em nada afectam a análise que se pretende fazer. Por exemplo, /portaldovweb/ ou /portaldovweb/userareas, deixando ficar somente o nome da página consultada.
Com esta fase a base de dados ficou pronta para ser exportada para CSV e, assim, ser analisada usando o CAREN.

Alguns dados finais:

- Número total de registos na tabela: 451.768
- Número total de registos não repetidos (Sessão + URL): 193.598
- Número total de transacções (sessões): 59.214
- Número total de artigos (páginas): 32

7 Geração das Regras de Associação

Para gerar as regras de associação escolhi o programa Caren [2] que implementa o algoritmo Apriori [8]. A razão para a sua escolha tem a ver com o facto de trabalhar com dados no formato *Basket* e, também, ser bastante flexível permitindo a utilização de vários parâmetros e dando várias opções de *output* dos dados que serão bastante úteis para a secção seguinte.
O algoritmo Apriori gera todas as regras que satisfazem o suporte e a confiança que forem indicadas. As definições de suporte e confiança são as seguintes:

- **Suporte** duma regra, por exemplo, sup(A -> B) é a percentagem das transacções que contêm todos os elementos de A U B.
- **Confiança** duma regra, por exemplo, conf(A -> B) é a percentagem das transacções que contendo A também contêm B, relativamente a todas as transacções que contêm A.

Para utilizar o Caren é necessário indicar o valor do suporte e da confiança, para além de outros parâmetros. Logicamente que a definição deste parâmetros depende muito do tipo de análise que pretendemos fazer embora se saiba que quanto mais baixo ambos os valores maior número de regras serão geradas. Baseando-me num trabalho prévio [6] resolvi usar os valores de 0.02 para o suporte e de 0.1 para a confiança. A tabela seguinte indica todos os parâmetros que usei:

Parâmetro	*Valor*	*Observações*
DataSetName	pdovLogFinalCarenBasket.txt	Ficheiro com os dados
minSup	0.02	Suporte mínimo a considerar
minConf	0.1	Confiança mínima a considerar
formato DataSet	-Bas	Indica que os dados estão no formato Basket
-o[opt]	-opmPdovResCarenS002C04pmml	Guardar as regras no ficheiro no formato PMML

Todos os restantes parâmetros ficaram com os valores definidos por defeito.
Convém realçar que o Caren permite muitos outros parâmetros (daí a sua flexibilidade) inclusive a utilização de outros filtros para além do suporte e da confiança.
Na sebenta da cadeira de extracção de conhecimento de dados II [9] é possível encontrar com bastante detalhe instruções sobre a forma de instalar e invocar o Caren. Por essa razão abstive-me de repetir neste trabalho essas instruções.
Após invocar o Caren com os parâmetros acima, obtive 17 itens frequentes e 320 regras considerando um total de 59.214 transacções. Este valores parecem-me razoáveis tendo obtido um número de regras que não é excessivo.
Segue-se uma pequena amostra do conteúdo do ficheiro gerado pelo Caren com os itens:

NumFreqItems = 17 in the dataset pdovLogFinalCarenBasket.txt with 59214 transactions
using minsup = 0.02
Itemsets calculation Time spent = 0 hrs 0 mts 15 secs

contactarempresa.aspx s()=1215
contactarempresa.aspx crewhomepage.aspx s()=1199
pedferias.aspx s()=1568
pedferias.aspx crewhomepage.aspx s()=1530
mostraescala.aspx s()=1847
mostraescala.aspx consultatrocas.aspx s()=1846
mostraescala.aspx consultatrocas.aspx consultarotacao.aspx s()=1433
mostraescala.aspx consultatrocas.aspx consultarotacao.aspx mostratripulacao.aspx s()=1248
mostraescala.aspx consultatrocas.aspx consultarotacao.aspx mostratripulacao.aspx crewhomepage.aspx s()=1234

Uma amostra de algumas regras geradas pelo Caren, já importadas para Excel, é a que consta da seguinte figura.

suporte	confiança	lift	conviction	consequente	antecedente
0,0352	0,5821	1,6840	1,5657	consultaescala,aspx	consultahoras,aspx & consultarotacao,aspx
0,0295	0,5796	1,6766	1,5561	consultaescala,aspx	consultahoras,aspx & consultarotacao,aspx & mostratripulacao,aspx
0,0345	0,5779	1,6718	1,5500	consultaescala,aspx	consultahoras,aspx & consultarotacao,aspx & crewhomepage,aspx
0,0289	0,5748	1,6631	1,5391	consultaescala,aspx	consultahoras,aspx & consultarotacao,aspx & mostratripulacao,aspx & crewhomepage,aspx
0,0324	0,5653	1,6354	1,5052	consultaescala,aspx	consultahoras,aspx & mostratripulacao,aspx
0,0488	0,5612	1,6237	1,4913	consultaescala,aspx	consultatripulacoes,aspx & consultarotacao,aspx & mostratripulacao,aspx
0,0317	0,5604	1,6213	1,4885	consultaescala,aspx	consultahoras,aspx & mostratripulacao,aspx & crewhomepage,aspx
0,0512	0,5602	1,6209	1,4880	consultaescala,aspx	consultatripulacoes,aspx & consultarotacao,aspx
0,0469	0,5520	1,5970	1,4605	consultaescala,aspx	consultatripulacoes,aspx & consultarotacao,aspx & mostratripulacao,aspx & crewhomepage,aspx
0,0493	0,5513	1,5960	1,4584	consultaescala,aspx	consultatripulacoes,aspx & consultarotacao,aspx & crewhomepage,aspx
0,0201	0,5399	1,5621	1,4223	consultaescala,aspx	consultatrocas,aspx & mostratripulacao,aspx
0,0201	0,5238	1,5155	1,3742	consultaescala,aspx	chatwindow,aspx & chat,aspx & consultarotacao,aspx
0,0201	0,5238	1,5155	1,3742	consultaescala,aspx	chatwindow,aspx & consultarotacao,aspx
0,0202	0,5216	1,5090	1,3678	consultaescala,aspx	chat,aspx & consultarotacao,aspx
0,0251	0,5066	1,4658	1,3263	consultaescala,aspx	consultaferias,aspx
0,0245	0,5035	1,4566	1,3179	consultaescala,aspx	consultaferias,aspx & crewhomepage,aspx
0,0497	0,5015	1,4510	1,3127	consultaescala,aspx	consultahoras,aspx
0,0483	0,4951	1,4323	1,2959	consultaescala,aspx	consultahoras,aspx & crewhomepage,aspx
0,0257	0,4932	1,4269	1,2912	consultaescala,aspx	consultatrocas,aspx
0,0229	0,4699	1,3595	1,2344	consultaescala,aspx	consultatrocas,aspx & crewhomepage,aspx
0,0291	0,4645	1,3439	1,2220	consultaescala,aspx	chatwindow,aspx & chat,aspx
0,0291	0,4645	1,3439	1,2220	consultaescala,aspx	chatwindow,aspx
0,0293	0,4619	1,3364	1,2161	consultaescala,aspx	chat,aspx

O Caren também gerou um ficheiro em formato PMML com as regras que, por ser extenso, não o coloquei aqui. No entanto, esse ficheiro é enviado juntamente com este relatório, tal como está indicado nos anexos.

8 Selecção e Exploração das Regras

Agora que as regras estão geradas é necessário encontrar aquelas que são mais interessantes para o objectivo da minha análise. Tal como referi na introdução os objectivos são:

- Implementar mecanismos no site que permitam sugerir para cada página ou conjuntos de páginas *links* interessantes.
- Implementar um sistema de menus personalizados que serão adaptados dinamicamente aos hábitos de navegação de cada utilizador.

Para além das duas medidas de interesse indicadas na secção anterior é necessário definir outras duas que vão ser usadas a partir deste momento: *Lift* e *Conviction*.
As suas definições são:

- *Lift* duma regra, por exemplo, lift(A -> B) mede a informatividade de A relativamente a B. Dito de outro modo, mede quão distantes estão A e B da independência. Se Lift=1 significa que A e B são independentes. Esta medida de interesse permite eliminar regras com confiança elevada mas com pouco interesse. Nesta perspectiva interessa ter valores de *Lift* superiores a 1.
- *Conviction* duma regra, por exemplo, conv(A -> B) mede a independência de A e negação de B. Quando A e negação de B é verdade é um contra-exemplo da regra pois é o único caso em que a implicação lógica está errada. Assim a *Conviction* mede a validade da direcção da implicação de A para B. Se Conviction=1 significa que A e B são independentes.

Existem outras medidas de interesse definidas na literatura (ver, por exemplo [10]). No entanto e como o Caren exporta esta informação durante a geração das regras, vou limitar-me às quatro indicadas: (1) Suporte, (2) Confiança, (3) Lift e (4) Conviction.

Para atingir o objectivo 1 será necessário obter regras que tenham interesse e, para isso, vou aplicar as medidas de interesse referidas anteriormente para seleccionar algumas regras.
Vou utilizar uma folha de cálculo, obtida a partir do ficheiro CSV gerado pelo Caren, que contém as regras exportadas e filtrar somente as que têm Suporte superior a 0.05, Confiança

superior a 0.3 e Lift superior a 1.2. Utilizando uma linguagem mais corrente, vou seleccionar as regras que sejam suportadas por mais de 5% das transacções e que mais de 30% das sessões que incluam os antecedentes também incluam o consequente e que exista um grau de dependência razoável entre o antecedente e o consequente (neste caso, 1.2).
As regras obtidas aparecem na seguinte figura:

suporte	confiança	lift	conviction	consequente	antecedente
0,0512	0,5602	1,6209	1,4880	consultaescala,aspx	consultatripulacoes,aspx & consultarotacao,aspx
0,0626	1,0000	15,7652	+oo	chat,aspx	chatwindow,aspx
0,0620	1,0000	15,7652	+oo	chat,aspx	chatwindow,aspx & crewhomepage,aspx
0,0850	0,9509	1,7690	9,4114	mostratripulacao,aspx	consultatripulacoes,aspx & consultarotacao,aspx & crewhomepage,aspx
0,0869	0,9509	1,7689	9,4095	mostratripulacao,aspx	consultatripulacoes,aspx & consultarotacao,aspx
0,0699	0,9063	1,6860	4,9338	mostratripulacao,aspx	consultatripulacoes,aspx & consultaescala,aspx & crewhomepage,aspx
0,0732	0,9017	1,6774	4,7030	mostratripulacao,aspx	consultatripulacoes,aspx & consultaescala,aspx
0,1869	0,8760	1,6296	3,7284	mostratripulacao,aspx	consultatripulacoes,aspx & crewhomepage,aspx
0,1914	0,8725	1,6232	3,6278	mostratripulacao,aspx	consultatripulacoes,aspx
0,0603	0,8421	1,5665	2,9279	mostratripulacao,aspx	consultahoras,aspx & consultarotacao,aspx & crewhomepage,aspx
0,0609	0,8412	1,5649	2,9114	mostratripulacao,aspx	consultahoras,aspx & consultarotacao,aspx
0,1695	0,8331	1,5500	2,7716	mostratripulacao,aspx	consultaescala,aspx & consultarotacao,aspx & crewhomepage,aspx
0,1768	0,8328	1,5493	2,7661	mostratripulacao,aspx	consultaescala,aspx & consultarotacao,aspx
0,4250	0,8235	1,5320	2,6197	mostratripulacao,aspx	consultarotacao,aspx & crewhomepage,aspx
0,4325	0,8224	1,5299	2,6036	mostratripulacao,aspx	consultarotacao,aspx
0,0626	0,9870	15,7652	71,8543	chatwindow,aspx	chat,aspx
0,0620	0,9868	15,7631	71,1274	chatwindow,aspx	chat,aspx & crewhomepage,aspx
0,0732	0,3637	1,6580	1,2269	consultatripulacoes,aspx	consultaescala,aspx & mostratripulacao,aspx
0,0699	0,3631	1,6551	1,2256	consultatripulacoes,aspx	consultaescala,aspx & mostratripulacao,aspx & crewhomepage,aspx
0,1914	0,3561	1,6232	1,2123	consultatripulacoes,aspx	mostratripulacao,aspx
0,1869	0,3546	1,6163	1,2095	consultatripulacoes,aspx	mostratripulacao,aspx & crewhomepage,aspx
0,0503	0,8888	1,6901	4,2622	consultarotacao,aspx	consultahoras,aspx & mostratripulacao,aspx & crewhomepage,aspx
0,0509	0,8880	1,6887	4,2336	consultarotacao,aspx	consultahoras,aspx & mostratripulacao,aspx
0,1695	0,8806	1,6746	3,9705	consultarotacao,aspx	consultaescala,aspx & mostratripulacao,aspx & crewhomepage,aspx
0,1768	0,8781	1,6698	3,8884	consultarotacao,aspx	consultaescala,aspx & mostratripulacao,aspx
0,4250	0,8063	1,5334	2,4483	consultarotacao,aspx	mostratripulacao,aspx & crewhomepage,aspx
0,4325	0,8045	1,5299	2,4255	consultarotacao,aspx	mostratripulacao,aspx
0,0522	0,7437	1,4142	1,8498	consultarotacao,aspx	pedcurtoprazovoos,aspx & crewhomepage,aspx
0,0536	0,7347	1,3972	1,7875	consultarotacao,aspx	pedcurtoprazovoos,aspx

Um exemplo da aplicação destas regras no Portal DOV poderia ser sugerir ao tripulante a consulta das páginas CONSULTATRIPULACOES, CONSULTAROTACAO e CREWHOMEPAGE quando este se encontra a consultar a página MOSTRATRIPULACAO. Como se pode ver pela figura abaixo, corresponde à regra que tem melhores indicadores, ou seja, Sup=0.0850, Conf=0.9509 e Lift=1.7690. De realçar que para obter esta informação, para além de filtrar as regras pelas medidas de interesse já indicadas, também as filtrei pelo consequente de maneira a obter somente as regras que se aplicam à consulta da página MOSTRATRIPULACAO.

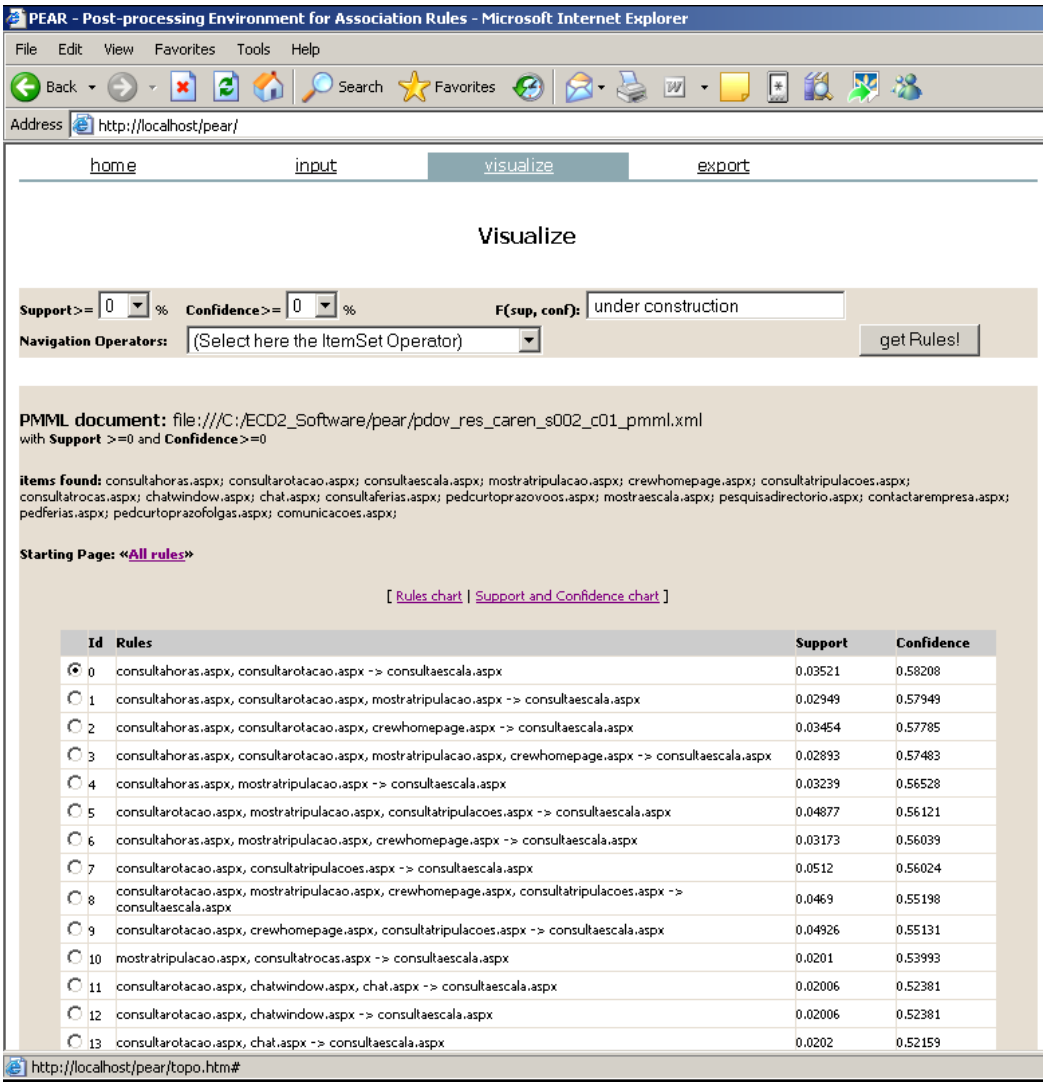
suporte	confiança	lift	conviction	consequente	antecedente
0,0850	0,9509	1,7690	9,4114	mostratripulacao,aspx	consultatripulacoes,aspx & consultarotacao,aspx & crewhomepage,aspx
0,0869	0,9509	1,7689	9,4095	mostratripulacao,aspx	consultatripulacoes,aspx & consultarotacao,aspx
0,0699	0,9063	1,6860	4,9338	mostratripulacao,aspx	consultatripulacoes,aspx & consultaescala,aspx & crewhomepage,aspx
0,0732	0,9017	1,6774	4,7030	mostratripulacao,aspx	consultatripulacoes,aspx & consultaescala,aspx
0,1869	0,8760	1,6296	3,7284	mostratripulacao,aspx	consultatripulacoes,aspx & crewhomepage,aspx
0,1914	0,8725	1,6232	3,6278	mostratripulacao,aspx	consultatripulacoes,aspx
0,0603	0,8421	1,5665	2,9279	mostratripulacao,aspx	consultahoras,aspx & consultarotacao,aspx & crewhomepage,aspx
0,0609	0,8412	1,5649	2,9114	mostratripulacao,aspx	consultahoras,aspx & consultarotacao,aspx
0,1695	0,8331	1,5500	2,7716	mostratripulacao,aspx	consultaescala,aspx & consultarotacao,aspx & crewhomepage,aspx
0,1768	0,8328	1,5493	2,7661	mostratripulacao,aspx	consultaescala,aspx & consultarotacao,aspx
0,4250	0,8235	1,5320	2,6197	mostratripulacao,aspx	consultarotacao,aspx & crewhomepage,aspx
0,4325	0,8224	1,5299	2,6036	mostratripulacao,aspx	consultarotacao,aspx

Uma outra possível aplicação seria tentar aumentar as consultas de uma determinada página que, por algum motivo, não é muito consultada. Assim, em vez de nos concentrarmos somente nas regras que têm melhores medidas de interesse, poderíamos, nos casos em que as regras com medidas de interesse inferiores contêm no antecedente uma página para a qual queremos aumentar os acessos, escolher essa regra e indicar como *links* as páginas correspondentes aos antecedentes.

Uma outra ferramenta interessante para analisar e explorar regras de associação é o PEAR [7]. Esta ferramenta aceita como *Input* um ficheiro no formato PMML que o Caren também permite gerar. O PEAR permite navegar nas regras geradas, definindo restrições e aplicando operadores. As restrições são especificadas através de valores mínimos de suporte e de confiança e os operadores através da escolha numa *Listbox* dos vários valores possíveis. Alguns exemplos de operadores são:

- **Antecedent Generalization (AntG)** que produz regras semelhantes à regra escolhida mas com um antecedente sintacticamente mais simples.
- **Consequent Generalization (ConsG)** que produz regras em que o novo consequente é obtido eliminando átomos no consequente.
- **Antecedent Specialization (AntS)** que produz regras com suporte mais baixo mas confiança mais alta do que a regra escolhida.
- **Consequent Specialization (ConsS)**
- **Focus on Antecedent (FAnt)** que obtém todas as regras que tenham exactamente o mesmo antecedente da regra escolhida.
- **Focus on Consequent (FCons)** que obtém todas as regras que tenham exactamente o mesmo consequente da regra escolhida.

Para fazer a mesma análise que fiz anteriormente, usando o PEAR, terei de, depois de carregar as regras para o PEAR através da página *Input* e de fornecer o nome e caminho do ficheiro em formato PMML, definir como restrições um valor de 0.05 para o suporte e um valor de 0.3 para a confiança. Nesta versão do PEAR não é possível definir um valor de Lift como restrição. As regras obtidas são as seguintes:

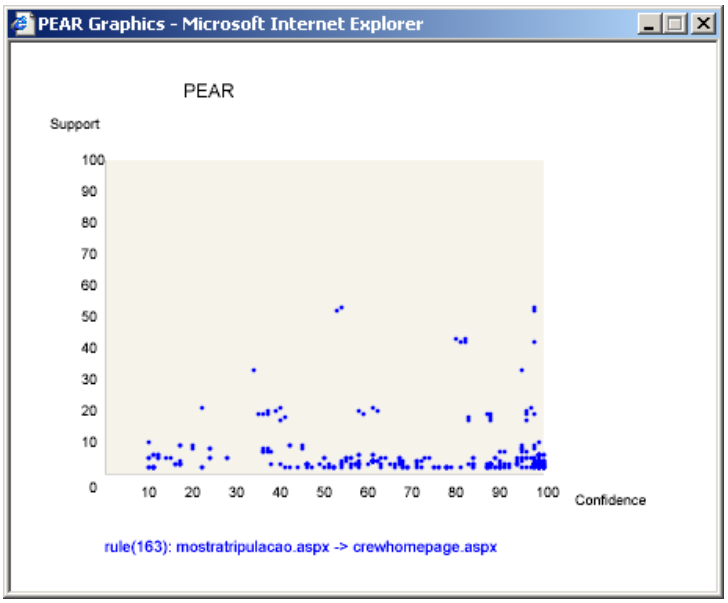


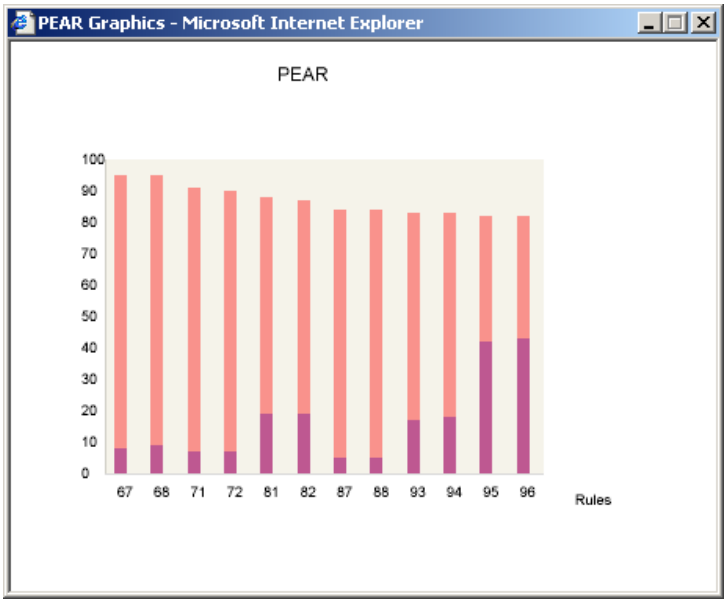
Como se pode verificar, por não se poder restringir também pelo *Lift*, aparecem muito mais regras do que quando utilizei a folha de cálculo. No entanto o PEAR tem uma visualização

gráfica que resume o conjunto de regras da página. É possível obter dois tipos de visualização gráfica:

- Gráfico de Confiança X Suporte.
- Histograma de Confiança / Suporte.

Na duas figuras seguintes é possível ver estes dois tipos de visualização resultantes da regras seleccionadas anteriormente.





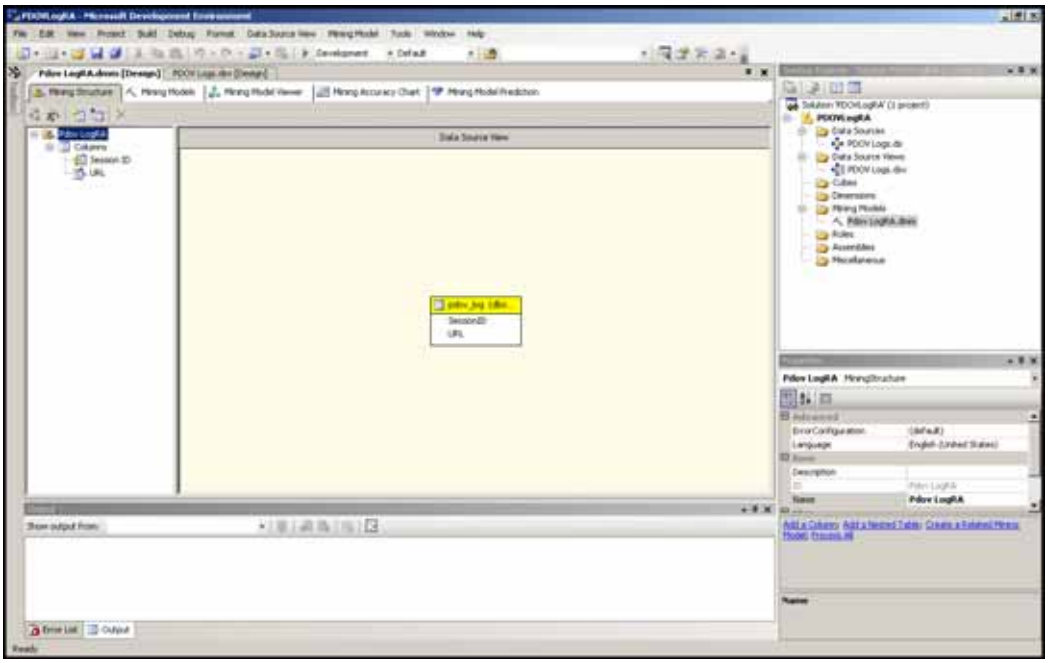
Quer num caso quer noutro os gráfico mostram qual a regra associada ao ponto ou à barra, quando se passa o ponteiro do rato por cima. No caso do primeiro gráfico, aparece indicado a regra que à partida será a mais interessante uma vez que é a que tem o suporte e a confiança mais elevados.

Para atingir o objectivo 2 é necessário ter mais dados do que aqueles que utilizei neste trabalho. Como o que se pretende é um menu personalizado, ou seja, cuja estrutura reflicta

os hábitos de navegação do utilizador (por exemplo, as páginas mais consultadas separadas das menos consultadas), será necessário ligar os dados da sessão com a identificação do utilizador. Como já referi na secção 3, não foi possível fazê-lo pelas razões apresentadas.
Gostaria de realçar que nesta secção apresentei apenas alguns exemplos do potencial desta tecnologia e destas ferramentas, quando aplicadas ao Portal DOV. Logicamente que muito mais haveria para explorar.

9 Análise com o SQL Server 2005 Beta 2

O enunciado do trabalho incentiva a utilização de outras ferramentas que não somente aquelas que foram apresentadas nas aulas. Por essa razão, lembrei-me de experimentar as ferramentas de *Business Intelligence* que estão disponíveis com o Microsoft SQL Server 2005 [3] embora ainda em versão Beta.
A passagem dos dados para uma base de dados SQL Server 2005 ocorreu sem problema, tendo a importação sido feita a partir do ficheiro CSV. A partir deste momento foi possível começar a construir um projecto de *Data Mining* utilizando o *Business Intelligence Development Studio*. Poderão ver um *Screenshot* desta aplicação na imagem seguinte:



Esta ferramenta permite construir modelos a partir de dados numa base de dados operacional (relacional) ou num *Data Wahrehouse*.
Inclui também as ferramentas para se fazer pré-processamento dos dados sem ser necessário recorrer a qualquer outro tipo de software.
Os modelos possíveis de construir são:

- Regras de Associação.
- Agrupamento (*Clustering*).
- Árvores de Decisão.
- *Naive Bayes*.
- Redes Neuronais.

Do pouco que pude verificar na literatura disponível (relembro que ainda é uma versão Beta) os algoritmos implementados apesar de se basearem em algumas implementações conhecidas, são específicos da *Microsoft*.

Para além de poder gerar o modelo de *Data Mining* pretendido tem ainda as seguintes ferramentas:

- **Mining Model Viewer** permite navegar pelos *Itemsets* através do valor mínimo de suporte e/ou do tamanho do *itemset*, permitindo também filtrar por parte do conteúdo do *itemset*. Da mesma forma que se pode navegar nos *itemsets* também é possível fazê-lo nas regras (à semelhança do PEAR). Finalmente, também incluí a visualização da Rede de Dependências sendo possível escolher um nó da rede e, a partir daí, ver as dependências em mais detalhe.
- **Mining Accuracy Chart** permite visualizar um gráfico dos valores *Lift* e, também, obter a Matriz Confusão.
- **Mining Model Prediction** permite obter previsões para novos dados usando um modelo previamente construído.

Como se pode verificar é uma ferramenta bastante completa e que, para utilizar num ambiente empresarial, tem todas as condições para se fazer *Data Mining* a partir dos dados da empresa sem ser necessário utilizar outro ambiente de análise.

Apesar de ter conseguido utilizar os dados fornecidos juntamente com o Tutorial com sucesso, o mesmo não aconteceu com os dados objecto desta análise. Não sei se o problema tem a ver com algo que ainda não esteja devidamente implementado (dado ser uma versão Beta) ou se está relacionado com o formato dos dados que utilizei. De qualquer das formas a impressão geral foi bastante positiva e será, concerteza, uma ferramenta que utilizarei no futuro.

10 Conclusões

Em primeiro lugar o objectivo deste trabalho, do meu ponto de vista, foi plenamente atingido: familiarizar-me com todo o processo necessário à realização de uma análise com regras de associação e, também, com a utilização de ferramentas adequadas para fazer essa análise. Foi importante ter-me apercebido de que a fase de pré-processamento é das mais importantes e, também, das mais demoradas. Também muito importante foi obter conhecimentos relacionados com a fase de pós-processamento. Todo o processo de escolha das regras mais interessantes e das medidas de interesse que existem para ajudar nesse processo (apesar de eu não ter utilizado todas as que existem) é muito importante. Deu para notar que o interesse das regras está muito relacionado com o objectivo da análise. Aquelas regras que possam parecer menos importantes num determinado contexto podem ser mais importantes noutro.

Em segundo lugar e como utilizei um conjunto de dados reais, permitiu-me implementar algumas alterações na forma como o registo dos *Logins* dos tripulantes é efectuado, de forma a permitir no futuro uma análise mais completa e que abranja hábitos de navegação específicos de cada utilizador.

Recomendações de Trabalho Futuro

Gostaria de realçar dois aspectos importantes que deverão ser realizados num futuro próximo:

- A alteração do registo dos *Logins* dos tripulantes de forma a incluir a identificação da sessão e, desta forma, ser possível ligar os dados da sessão ao utilizador.
- A implementação no Portal da recomendação de páginas a consultar a partir da página em que o utilizador se encontra. Corresponde ao objectivo 1 relatado na secção 8.
- A implementação no Portal do menu personalizado de acordo com os hábitos de navegação do utilizador. Corresponde ao objectivo 2 da secção 8.
- A implementação das páginas recomendadas e do menu personalizado relatado nos itens anteriores, ser feita dinamicamente e não com base num modelo previamente classificado e que é actualizado de tempos a tempos. O objectivo é que, por cada navegação efectuada pelo utilizador esse modelo possa ser actualizado e esse conhecimento possa ser reflectido no menu e na informação das páginas recomendadas.

Agradecimentos

Gostaria de agradecer à TAP Air Portugal o acesso aos dados que permitiram realizar este trabalho.

Referências

[1] TAP Air Portugal. Portal da Direcção de Operações de Voo da Tap Air Portugal. http://portaldov.tap.pt, 2005.

[2] Paulo J. Azevedo, CAREN - A Java based Apriori Implementation for Classification Purposes, 2003.

[3] Microsoft, SQL Server 2005 - Beta 2, http://www.microsoft.com/sql/2005, 2005.

[4] MySQL AB, The World's Most Popular Open Source Database. http://www.mysql.com.

[5] Robert Cooley, Bamshad Mobasher, Jaideep Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns, 1998

[6] A. Jorge, M. Alves, P. Azevedo, Recommendation With Association Rules: A Web Mining Application, 2002.

[7] A. Jorge, J. Poças, P. Azevedo, A Post-processing Environment for Browsing Large Sets of Association Rules.

[8] Rakesh Agrawal, Ramakrishnan Srikant, Fast Algorithms for Mining Association Rules.

[9] A. Jorge, Sebenta da Cadeira de Extracção de Conhecimento de Dados II, 2005.

[10] A. Silberschatz, A. Tuzhilin, On Subjective Measures of Interestingness in Knowledge Discovery, 1995.

Ficheiros Anexos ao Relatório

Juntamente com este relatório foram enviados os seguintes ficheiros:

1. **Web_Logs_Originais_Dos_Servidores_Producao.rar** Contém todos os ficheiros dos *Web Logs* retirados dos dois servidores de produção do Portal, no seu formato original.

2. **Web_Logs_Originais_Formato_MYSQL.rar** Contém a mesma informação que o anterior mas já numa tabela do MySQL.

3. **Logins_Portal_Dov.XLS** Contém os registos dos *Logins* dos tripulantes num folha de cálculo com a informação registada pelo Portal.

4. **Web_Logs_Apos_Preparacao_Formato_MYSQL.rar** Contém os registos em formato MySQL após a preparação dos dados.

5. **Pdov_Log_Final_Caren_Basket.txt** Registo já no formato final e que serviram de *Input* ao programa Caren para gerar as regras de associação.

6. **Pdov_Log_Final_Basket.csv** O mesmo que o anterior mas em formato CSV.

7. **Items_S002_C01.txt** Ficheiro com os itens gerados pelo Caren.

8. **Pdov_Res_Caren_S002_C01.xls** Ficheiro com as regras geradas pelo Caren, já importadas para uma folha de cálculo, e que serviu de base a algumas das análises da secção 8.

9. **Pdov_Res_Caren_S002_C01_Pmml.xml** Ficheiro com as regras geradas pelo Caren em formato PMML. Foi utilizado para efectuar análises utilizando o PEAR.

10. **Pdov_Res_Caren_S002_C01.csv** Ficheiro com as regras geradas pelo Caren em formato CSV.