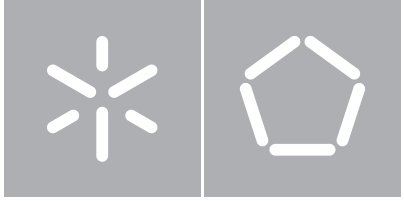
Universidade do Minho
Escola de Engenharia

Paulo Alexandre Ricardo Monsanto

Geração de Perfis Web Baseada em Assinaturas

Janeiro de 2013



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Paulo Alexandre Ricardo Monsanto

Geração de Perfis Web Baseada em Assinaturas

Dissertação de Mestrado
Mestrado em Engenharia Informática

Trabalho realizado sob orientação de

Professor Doutor Orlando Manuel de Oliveira Belo

Janeiro de 2013

À minha mulher Filomena.

Pelo seu amor, carinho e apoio.

Aos meus filhos, Alexandre e Ricardo.

Que nunca cessem na procura de conhecimento e o apliquem para o bem comum.

Agradecimentos

Queria expressar os meus agradecimentos a todos aqueles que, por diversas formas, permitiriam a realização deste trabalho e em particular agradeço:

- Ao meu orientador o Professor Doutor Orlando Manuel de Oliveira Belo, por todo o acompanhamento e sabedoria que pôs à minha disposição e que me permitiu realizar este trabalho.
- À DTSI (Direção de Tecnologias e Sistemas de Informação) da Universidade do Minho e em particular ao seu Diretor, Mestre José Manuel Machado Fernandes, pelo apoio e disponibilização dos dados que permitiram a elaboração desta dissertação.
- Aos meus pais, que, pela educação que me deram, me permitiram chegar à realização deste projeto e de ter a perseverança para o concluir.
- Aos meus irmãos, pelo incentivo e pelo acreditar.
- Aos meus amigos, pelo alento que deram. Em particular ao Paulo Serafim que partilhou comigo muitas e longas jornadas de trabalho e que pela sua presença e amizade ajudou à realização deste trabalho.
- Em último agradecer à minha mulher e filhos a quem dediquei esta dissertação, pelo apoio, paciência e alegria que me deram e continuam a dar. Pelas muitas horas e longos dias de ausência, que necessitei para a execução deste projeto.

Resumo

Geração de Perfis Web Baseada em Assinaturas

No mundo atual, com o crescimento da Internet e o consequente aumento de informação e serviços que são oferecidos pelas empresas e organizações no seu ambiente torna-se premente desenvolver técnicas que facilitem a navegação dos utilizadores por este enorme espaço virtual. A forma como interagimos com os diversos sítios presentes na Internet define um determinado comportamento, os nossos hábitos, os nossos costumes. De facto, no nosso dia-a-dia, e depois de frequentar durante bastante tempo um mesmo estabelecimento, apreciamos o cuidado com que, por vezes, sem nada dizer, o que mais apreciamos é posto à frente e à nossa disposição sem que sejamos consultados. Simplesmente conhecem-nos. Os sítios na Internet cada vez mais tentam ter esse mesmo cuidado com os seus utilizadores. Todavia, a comunidade cibernauta é, como sabemos, muito vasta e heterógena e, como tal, saber os hábitos e costumes de tantos indivíduos é uma tarefa complicada.

O uso de perfis é uma ação normal na caracterização de utilizadores, seja por questões de segurança ou funcionais. Um determinado utilizador pode sempre enquadrar-se num ou noutro perfil, em que cada um deles determina o acesso a este ou àquele tipo de informação ou funcionalidade usualmente oferecida por um sítio presente na Internet. Este tipo caracterização pode permitir o agrupamento de utilizadores por diversos perfis, facilitando a gestão de informação e serviços, aproximando-os às necessidades reais dos utilizadores. Contudo uma das questões relacionadas com este tipo de caracterização de perfis é o facto de ela ser estática ao longo do tempo. Os nossos comportamentos e hábitos, como é conhecido, podem não o ser. O conhecimento de "Quem nós somos" num sítio pode sofrer alterações ao longo do tempo. As

nossas características de consumo e as nossas preferências podem mudar, o que nos define perante ele, a nossa assinatura, pode ter variações.

As características de utilização que os diversos sítios presentes Internet valorizam mais varia de sítio para sítio. Questões de consumo, por exemplo, serão provavelmente mais valorizadas em sítios que ofereçam produtos e serviços, em detrimento de outras questões, dependendo, obviamente dos objetivos de negócio do sítio em questão. Com certeza que, a definição de quais as características mais importantes num determinado contexto irá definir os atributos da assinatura dos utilizadores para esse sítio, tendo cada utilizador um valor diferente de acordo com esses atributos. Cada utilizador terá a sua assinatura. O valor da assinatura dos utilizadores ao longo do tempo tem que ser determinada por processos de cálculo específicos e ajustados ao contexto em questão e à informação disponível. O uso de técnicas de mineração de dados e de extração de conhecimento é, assim, essencial para este processo.

A definição e o cálculo da variação de assinaturas de utilizadores para um dado sítio permitem a realização de várias análises. Uma delas é a análise do perfil de um utilizador ao longo do tempo. A variação da assinatura de um utilizador poderá indiciar uma alteração no seu perfil de comportamento perante o sítio que frequenta. O sítio, sabendo dessa alteração, poderá reagir de forma dinâmica e de diversas formas. Por exemplo, alterando o conteúdo ou a estrutura do próprio sítio. Neste trabalho de dissertação foram exploradas estas temáticas. Mais especificamente pretendeu-se aprofundar a definição de assinatura de um utilizador Web e a sua associação aos diversos padrões de utilização de um sítio. No âmbito deste trabalho, esses padrões foram extraídos recorrendo a técnicas de mineração de dados a partir de diversas fontes de informação disponibilizadas pelos servidores que alojam os sítios. As técnicas utilizadas na extração desse conhecimento são também abordadas ao longo desta dissertação, com o objetivo de fornecer uma perspetiva global tanto do processo de mineração de dados em si, como da posterior associação do conhecimento extraído às assinaturas definidas para os utilizadores de um sítio específico que escolhemos como alvo para o nosso estudo.

Palavras-Chave: Assinaturas, Assinaturas Web, Perfis, Web, Perfis Web, Personalização, Personalização Web, Mineração de Dados, Mineração de Dados Web, Mineração de Dados de Utilização, *Data Warehouse*, *Data Warehousing*, Padrões, Descoberta de Padrões, Mineração de Padrões.

Abstract

Web Profiling Based on Signatures

In today's world, with the continuous growth of the Internet and the consequent increase of information and services that are offered by companies and organizations, it is urgent to develop techniques that ease users' navigation throughout this virtual space. The way we interact with the various sites on the Internet defines a particular behavior, our habits and our customs. In fact, in our daily life and after attending for a long time the same establishment, we appreciate that sometimes, without saying a word, the things that we like the most are put at our disposal without we being consulted. They just know us. The websites increasingly try to have that same care with their users. However and as we know, the cybernetic community is very large and heterogeneous and as so, knowing the habits and customs of so many individual users is a complicated task.

The use of profiles is a normal procedure in user characterization, either for security or functional reasons. A given user can always be fitted in one profile and each profile will give access to various types of information or functionalities, usually given by a website. This characterization may allow the grouping of users by different profiles, easing the information and services management, bringing them closer to the real needs of users. However one of the issues with this type of profile characterization is that it is static over time. Our behaviors and routines, as it is known, may not be. The knowledge of "Who we are" in a website may change over time. Our consumption habits and our preferences may change, that which defines us to that website, our signature, may have variations.

The users' characteristics that websites value the most, change from website to website. Consumption issues, for example, will probably be more valued to websites that sell products and services over others, depending, obviously of the business goals of the website in question. It is certain that the definition of which are most important characteristics in a given context, will define the attributes of the signature for that website, having each user a different value according to those attributes. Each user will have its signature. The value of the users' signatures over a period of time must be determined by specific calculation processes and must be adjusted to the context in question and to the information available. The use of data mining techniques and knowledge extraction is thus, crucial to this process.

The definition and calculation of users' signature variation for a given website enables several analyses. One of those is the chance to analyze a users profile over a period of time. Signature variation may indicate a change in its behavior profile to the website that he attends. The website, knowing of this change, may react to that change dynamically and in several ways. It can, for example, change its contents or structure. In this dissertation these issues were explored. More specifically it was intended to deepen the definition of a web user's signature and its association with the usage patterns of a website. As part of this work, these patterns were extracted by data mining techniques from various sources of information, particularly those provided by the web servers that host these websites. The techniques used in the extraction of this knowledge are also addressed in this dissertation with the purpose of giving a global perspective of both the data mining process itself, and of the subsequent association of the extracted knowledge to the user´s signatures that were defined by a specific website that we selected as target for our study.

Keywords: Signatures, Web Signatures, Profiles, Web, Web Profiling, Personalization, Web Personalization, Data Mining, Web Mining, Web Usage Mining, Data Warehouse, Data Warehousing, Patterns, Pattern Discovery, Pattern Mining.

Índice

Índice de Figuras

Índice de Tabelas

Lista de Siglas e Acrónimos

CRISP-DM	*CRoss-Industry Standard Process for Data Mining*
CERN	*Conseil Europeen pour le Recherche Nucleaire*
E-Learning	*Eletronic Learning*
ETL	*Extraction, Transformation and Loading*
FAIS	*The Financial Crimes Enforcement Network AI System*
FINRA	Financial Industry Regulatory Authority
HHTP	*HyperText Transfer Protocol*
IBM	*International Business Machines*
IP	*Internet Protocol*
KDD	*Knowledge Discovery in Database*
KDP	*Knowledge Discovery Process*
NASD	*National Association of Securities Dealers*
NASDAQ	*National Association of Securities Dealers Automated Quotations*
NIC	*Network Interface Card*
NSCA	*National Center for Supercomputing Applications*
OLAP	*Online Analytical Processing*
ROC	*Receiver Operating Characteristic*
SONAR	*The NASD Securities Observation, News Analysis & Regulation System*
SQL	*Structured Query Language*
T-SQL	*Transact-SQL*
URI	*Uniform Resource Identifier*
URL	*Uniform Resource Locator*
WWW	*World Wide Web*

Capítulo 1

Introdução

1.1 Identificação de perfis

O crescimento da Internet tem sido exponencial nos últimos anos, estimando-se que tenha crescido, em termos de utilizadores, nos últimos doze anos cerca de 566,4%, com um total de utilizadores estimados em Junho de 2012 de 2.405.518.376 utilizadores (Internet World Stats, 2012).

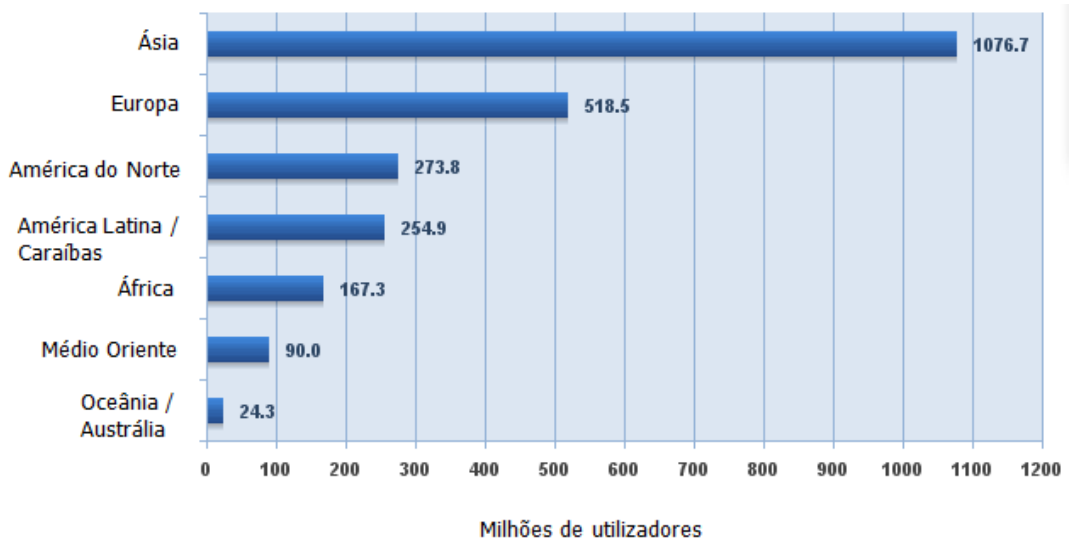


Figura 1.1: Utilizadores da Internet no mundo – figura adaptada de (Internet World Stats, 2012)

Com este crescimento do número de utilizadores e com a proliferação de sítios e de serviços disponibilizados, a procura e gestão de conhecimento relacionada com a informação disponível na Internet tornou-se uma tarefa cada vez mais desafiante.

A presença das empresas e organizações na Internet é hoje em dia quase obrigatória, pois a procura de produtos, o acesso a serviços ou a simples procura de um esclarecimento sobre uma qualquer questão por parte dos utilizadores passa por abrir um navegador e efetuar uma pesquisa na Internet. Se há alguns anos atrás a presença neste espaço dava a ideia de uma imagem vanguardista, hoje a não presença passa uma imagem retrógrada e pouco esclarecida. Todavia, a disponibilização de conteúdos na Internet pelas empresas e pelas organizações torna-se cada vez mais um processo criterioso e estudado. Já não basta disponibilizar simplesmente a informação. Faz-se a sua disponibilização de uma forma mais personalizada, tendo em conta frequentemente o perfil do utilizador que acede aos nossos sítios. É aqui que, hoje, se requer a aplicação de processos seletivos de informação no sentido de "caminhar" na direção das necessidades e preferências dos utilizadores.

Para personalizar a informação que disponibilizamos aos utilizadores que visitam o nosso sítio é necessário primeiro recolher alguns dados sobre eles, conhecer os seus perfis, os seus hábitos, para que, assim, possamos, dar a cada um deles aquilo que julgamos ser o mais adequado ao seu perfil. Neste contexto a informação que é recolhida pelos servidores que alojam e suportam a operacionalidade dos sítios é cada vez mais uma fonte de informação bastante preciosa e útil, podendo ser trabalhada e conjugada com outras fontes de informação de modo a dar aos responsáveis pela gestão dos sítios uma visão melhorada sobre os padrões de utilização dos utilizadores no seu portal.

A mineração de conteúdos da Web é cada vez mais comum (e essencial) para conseguir alcançar muitos dos objetivos que as empresas e organizações traçaram para a sua exposição na Internet, seja ele o aumento das suas vendas, a satisfação dos seus clientes ou a otimização dos seus próprios recursos. Existem três grandes áreas a nível da mineração de conteúdos da Web: a mineração de conteúdo, a mineração de estrutura e a mineração de utilização. Será sobre esta última área que assentarão, essencialmente, os trabalhos desta dissertação. A mineração de utilização visa compreender a atividade de navegação dos utilizadores em cada um dos sítios que visitam, de modo a conhecer as suas preferências e as suas rotinas. Com esse conhecimento é

possível traçar os diversos perfis dos utilizadores e fazer, quando possível, as adaptações necessárias sobre a estrutura e os serviços do sítio aos seus utilizadores. Esta adaptação pode ser realizada a diversos níveis, como por exemplo em adaptações de estrutura ou na apresentação dos seus conteúdos.

A identificação de perfis de utilizadores pode ser feita recorrendo a várias fontes de informação, desde a informação recolhida em servidores de alojamento até à recolhida diretamente na rede. Depois de identificadas as fontes de informação existe a aplicação dos processos de manipulação desses dados, que podem envolver desde a seleção de dados relevantes, sua limpeza e transformação, até à fase da mineração de dados, onde posteriormente é feita a interpretação e consolidação dos dados para extração do conhecimento. Depois de interpretados e consolidados os dados, e de ter sido feita a extração do conhecimento, é realizada a adequação do conhecimento extraído aos propósitos que pretendemos alcançar com ele. Tais propósitos podem ser vários e abranger várias vertentes de atuação. Desde logo, a vertente comercial, utilizada pelos portais de comércio eletrónico, sendo uma das mais utilizadas pela sua visibilidade e retorno de investimento para as empresas e que permite, também, oferecer aos clientes produtos e serviços personalizados, tendo em conta o perfil de compra ou os hábitos de navegação de cada utilizador. Mas esta não é a única vertente de utilização. Sabendo quais os perfis dos utilizadores que utilizam os nossos produtos e serviços bem como os seus hábitos e preferências, podemos oferecer-lhes algo que não seria possível sem ter tal conhecimento. O aumento de mecanismos de segurança como a deteção atempada de acessos indevidos ou fraudulentos e a disponibilização de conteúdos reservados e personalizados são apenas dois exemplos de serviços que se poderiam melhorar a partir do conhecimento dos perfis dos utilizadores.

1.2 Assinaturas

Uma assinatura, tal como o próprio nome indica, é algo que identifica um determinado individuo, num determinado contexto e num determinado instante no tempo. Esta identificação pode ir de uma simples assinatura escrita a um conjunto de características e comportamentos que indicam a uma determinada entidade que quem se apresenta perante ela é de facto o individuo que afirma ser.

O conceito de assinatura tem emergido nos sistemas de informação para complementar a informação relativamente estática que normalmente é recolhida de um utilizador, num determinado sistema, e possibilitar a obtenção de outros parâmetros de análise que não seriam possível sem a implementação deste conceito. A extração e atribuição de uma assinatura a um utilizador, que pode ser feita de forma completamente oculta ao utilizador, não tendo ele conhecimento dessa atribuição, irá caracterizá-lo perante o sistema que a gerou. A assinatura é composta por um conjunto de características que podem ser recolhidas de diversas fontes, que pode integrar informação recolhida por registos de atividade nos servidores ou informação preenchida pelo próprio utilizador. Tudo dependerá do sistema aplicacional e da análise que se pretende efetuar.

Um determinado utilizador pode ter várias assinaturas perante os diversos sistemas aplicacionais que utiliza, dependendo da utilização que faz de cada um deles. Os atributos de uma assinatura num determinado sistema irão variar, podendo um utilizador ter várias assinaturas num mesmo sistema, inclusive, para diferentes parâmetros de análise. Atributos considerados genéricos, como tempos de sessão ou horas de conexão, ou atributos de análise mais específicos, como por exemplo montantes transacionados, poderão ser introduzidos numa assinatura, caso se trate, por exemplo, de um sistema de acesso a banca *online*. Uma das utilizações mais frequentes para este conceito de assinatura ao nível de mineração de dados é o uso em sistemas de deteção de fraude. A variação abrupta na assinatura de um utilizador, num determinado intervalo de tempo, poderá indiciar uma situação anómala, uma potencial situação de fraude.

Nesta dissertação abordar-se-á a questão da assinatura de um utilizador em termos da mineração de conteúdos Web, fazendo a sua caraterização para os utilizadores de um sítio e verificando a sua variação para determinar, também, possíveis variações nos perfis de utilização. O enquadramento de um utilizador num perfil de utilização correto irá permitir, posteriormente, uma reação do sistema à entrada de um utilizador com uma determinada assinatura, podendo este adaptar-se a esse utilizador de diferentes formas, seja a nível estrutural, seja a nível de conteúdos, ou mesmo dos serviços que disponibiliza ao utilizador que naquele instante o visita.

1.3 Motivação e objetivos

Com a expansão da Internet, a procura de informação no seu espaço virtual torna-se uma tarefa cada vez mais complexa. Os sítios fornecem cada vez mais produtos e serviços, que são utilizados por pessoas nas suas tarefas diárias, seja no lazer seja em contexto laboral. A filtragem de conteúdos e a personalização de serviços tornam-se tarefas cada vez mais prementes por forma a facilitar a vida diária dos utilizadores. O uso do conceito de assinatura, a sua aplicação a nível de mineração de dados Web ou a sua utilização para contextualizar utilizadores em perfis de utilização, pode ser um dos meios para conseguir essa personalização de conteúdos e serviços.

O objetivo principal deste trabalho de dissertação consistiu na aplicação das diversas fases envolvidas num processo de mineração de comportamentos de utilização para um dado sítio, no caso presente foi um sítio de *e-learning*, para definição de uma assinatura e extração dos respetivos dados que a compõem. Posteriormente, fazer uma análise detalhada à variação dessas assinaturas ao longo do tempo e relacionar as variações com alterações ao nível dos perfis de utilização definidos para o sistema de *e-learning* e para os seus diversos utilizadores em análise. Este conhecimento permitirá, no futuro, a adaptação do sítio aos diversos perfis de utilização, de modo a que os seus utilizadores possam ter uma experiência de navegação mais facilitada e personalizada. Assim, os objetivos associados à realização deste trabalho de dissertação foram, sumariamente:

- Analisar as fontes de informação e respetivos dados para o caso de estudo selecionado, o sítio de *e-learning* da Universidade do Minho.
- Implementar um processo de extração, transformação e carga para as fontes e dados identificados na primeira etapa deste trabalho.
- Definir a estrutura de uma assinatura a utilizar no âmbito deste trabalho.
- Implementar um processo de extração, transformação e carga para os dados necessários à geração de assinaturas dos utilizadores.
- Implementar um processo de cálculo para a variação das assinaturas dos utilizadores.
- Analisar os resultados obtidos no processo de cálculo de variação das assinaturas, em particular com a confrontação dos perfis definidos para os utilizadores no caso de estudo.

1.4 A Estrutura da dissertação

Para além deste primeiro capítulo, esta dissertação incorpora outros quatro capítulos, que estão organizados da seguinte forma:

- **Capítulo 2** - Perfis Web. Inicialmente, neste capítulo, apresentam-se os conceitos essenciais à compreensão da área de trabalho desta dissertação. De seguida é feita uma abordagem aos processos de mineração Web, a um caso específico dos processos de mineração e com particular relevância para este trabalho. Dentro do âmbito da mineração Web deu-se especial atenção à mineração de dados de utilização, apresentando-se, complementarmente, algumas ferramentas utilizadas no domínio da extração de padrões de utilização. Ainda neste capítulo, aborda-se a questão da personalização, bem como a forma como esta pode ser implementada e algumas referências à sua aplicação.
- **Capítulo 3** – Assinaturas. Este capítulo começa por apresentar a definição daquilo que consideramos ser uma assinatura, bem como as suas características específicas. De seguida, apresenta-se uma possível arquitetura computacional para a sua implementação em sistemas que as possam utilizar. A questão da deteção de anomalias, explicando o seu conceito, tipos e formas de o fazer é exposta e discutida logo a seguir, terminando-se este capítulo explorando-se algumas formas de atualização de bases de assinaturas bem como de alguns dos seus possíveis domínios de aplicação.
- **Capítulo 4** – O Caso de Estudo. Aqui é apresentado o caso de estudo que serviu de base a todo o trabalho efetuado nesta dissertação. Depois, abordamos o processo de recolha de dados e do processo de extração, transformação e carga dos dados provenientes das fontes de informação, e descrevemos a estrutura e características da assinatura que foi definida no âmbito deste trabalho. Na parte final deste capítulo descreve-se o processo de cálculo de variação das assinaturas e a análise dos resultados obtidos.
- **Capítulo 5** – Conclusões e Trabalho Futuro. Por fim, neste capítulo, é efetuada uma reflexão acerca do trabalho realizado. Após uma pequena síntese apresenta-se um balanço geral de todo o trabalho, confrontando-o com os objetivos propostos inicialmente propostos e os objetivos realmente alcançados. Apresenta-se ainda algumas das limitações sentidas ao longo de todo este processo e pontos passíveis de melhoria. O capítulo termina elencando algumas tarefas possíveis para implementação futura de forma a ser possível dar algum tipo de seguimento ao trabalho efetuado nesta dissertação.

Capítulo 2

Perfis Web

2.1 Definições

Apesar de hoje serem conceitos e termos sobejamente conhecidos por todos aqueles que utilizam de uma forma regular a Web, achamos ser necessário, nem que seja para um mero enquadramento deste trabalho de dissertação, iniciar este capítulo com uma revisão sumária daquilo que consideramos importante para a compreensão do projeto que desenvolvemos. Assim, tendo em conta o estabelecimento de um qualquer perfil Web é necessário relembrar alguns conceitos e definições (Tabela 2.1). Num contexto Web pode-se considerar como utilizador um individuo ou uma aplicação automatizada que acede direta ou indiretamente aos seus recursos. Um recurso pode ser visto como algo que tenha uma identidade própria, sendo que um URI (*Uniform Resource Identifier*) é uma sequência compacta de caracteres que identifica um recurso físico ou abstrato (Universal Resource Identifiers, 2013), como por exemplo um ficheiro HTML, uma imagem ou um serviço web. Um URL (*Uniform Resource Locator*) é um URI que além de identificar um recurso fornece os meios primários de acesso a esse mesmo recurso, ou seja a sua localização na rede. Um recurso Web é um recurso que pode ser acessível através de uma qualquer versão do protocolo HTTP. Um servidor Web por seu turno fornece acesso a recursos Web. Um *link* ou *hyperlink* é uma conexão entre dois recursos Web (Links, 2013), será usado o termo hiperligação ao longo desta dissertação. Por sua vez, um navegador Web (*browser)*, ou um

cliente Web, é um programa informático que envia pedidos de recursos Web, trata as respostas e mostra ao utilizador os URIs respetivos. Por fim, uma página Web. Esta pode ser vista como um conjunto de dados que constituem um ou mais recursos Web acessíveis através de um URI.

A visualização de uma página (*page view*) ocorre quando um navegador mostra uma página Web que é solicitada quando um utilizador clica numa hiperligação. Os recursos dessa página são, então, mostrados ao utilizador. O pedido desses recursos pode ser explícito ou implícito. O pedido explícito ocorre quando o utilizador manualmente inicializa o pedido (por exemplo com a referida ação de clicar ou quando se coloca um endereço URL no navegador). Os pedidos implícitos são gerados pelo próprio navegador quando esta precisa de recursos que estejam embebidos na própria página para a mostrar ao utilizador (imagens, ficheiros multimédia, etc.). Uma sequência de cliques (*clickstream*) é uma lista ordenada de visualizações de páginas. Durante um certo período, essa sequência, que pertencente a um utilizador, será mais tarde a base daquilo que reconhecemos ser a sua sessão (*user session*). Dentro da sua sessão de utilizador, um episódio (*episode*) é caracterizado como um conjunto de cliques que entre si estão sequencialmente ou semanticamente relacionados. Por exemplo, uma visita ao sítio do Google pode originar vários episódios distintos, uma procura de imagens de um modelo de carro ou uma verificação de cotações de ações em bolsa geraria dois episódios diferentes.

Termo	**Descrição**
Utilizador	Um individuo ou uma aplicação automatizada que acede a recursos de um servidor através da WWW.
Servidor Web	Papel adotado por um computador que fornece recursos Web.
Recurso web	Algo com identidade própria, acessível através de um qualquer versão do protocolo HTTP.
Navegador	Programa informático que envia pedidos de recursos web, trata as respostas e mostra-as ao utilizador.
Visualização página	Página dada a visualizar ao utilizador num determinado navegador ou aplicação cliente.
Sequência de cliques	Lista ordenada de visualizações de páginas.
Sessão de utilizador	Conjunto de cliques de um determinado utilizador, num determinado servidor e durante um período específico de tempo.
Episódio	Conjunto de cliques que entre si estão sequencialmente ou semanticamente relacionados.
Perfil	Conjunto de informação caracterizadora de um determinado utilizador.

Tabela 2.1: Terminologia e conceitos da Web

O perfil de um utilizador contém informação sobre esse mesmo utilizador, podendo incluir informação demográfica, tal como o seu nome, idade, país, estado civil, etc., ou mesmo informação sobre os seus interesses e preferências. Toda essa informação pode ser recolhida por processos de registo ou questionários, mas também pode ser inferida a partir dos registos de *log*, tal como Srivastava et al. (2000) apresenta num dos seus trabalhos.

```
#Version: 1.0 #Date: 12-Jan-1996
00:00:00 #Fields: time cs-method
cs-uri 00:34:23 GET /foo/bar.html
12:21:16     GET     /foo/bar.html
12:45:52     GET     /foo/bar.html
12:57:34 GET /foo/bar.html
```

Figura 2.1: Excerto de um ficheiro *log* em formato ECLF

Um ficheiro Web *log* é um registo cronológico dos pedidos feitos pelos utilizadores a um determinado sítio e é armazenado pelo servidor Web que alberga esse mesmo sítio. Os formatos mais populares, desenvolvidos pelo CERN e NCSA, são o *Common Log Format* (CLF) (Logging Control In W3C, 1995) e uma versão estendida deste formato *Extended Common Log Format* (ECLF) (Extended Log File Format, 1995). Na Figura 2.1 apresenta-se um exemplo com um excerto de um ficheiro ECLF.

Campo	**Descrição**
Remote host	Nome ou endereço IP de domínio de origem do pedido.
Rfc931	Identificação do cliente que está a efetuar o pedido.
AuthUser	Identificação do utilizador/código, se servidor solicitar autenticação.
Date	Hora e data do pedido.
Offset	Diferença horária local relativamente a Greenwhich.
Method	Método do pedido (Get, Post, Head, etc).
URI	Endereço completo da página ou pedido, tal como veio do cliente.
Protocol	Protocolo de comunicação HTTP usado pelo cliente.
Status	O estado HTTP do pedido retornado para o cliente.
Bytes	Tamanho do conteúdo transferido para o utilizador, em *bytes*.
Referrer	URI de origem do pedido.
Agent	Identificação do navegador e sistema operativo usado.

Tabela 2.2: Descrição dos campos de um ficheiro de Web *log*

A informação guardada num ficheiro *log* varia consoante o formato e a informação que cada servidor consegue registar. A Tabela 2.2 apresenta uma descrição dos campos que podem ser encontrados nos diferentes formatos.

2.2 Processos de Mineração

O termo "descoberta de conhecimento" foi inicialmente apresentado em 1989 durante o congresso inaugural sobre descoberta de conhecimentos em bases de dados (Piatetsky, 1991) e referia-se ao processo genérico de procura de conhecimento em dados. O fator principal seria que o conhecimento é o produto final de um processo de descoberta baseado em dados. Este foi definido como um processo não trivial de identificação de padrões em dados que fossem válidos, novos, úteis e, em última análise, compreensíveis. O processo de descoberta de conhecimento tem vários passos que são, no fundo, um conjunto de procedimentos que terão que ser executados ao longo do processo de descoberta. Essencialmente, esta estruturação do trabalho de descoberta é usada para planear e reduzir o custo e esforço de qualquer projeto (figura 2.2) (Cios et al., 2007).

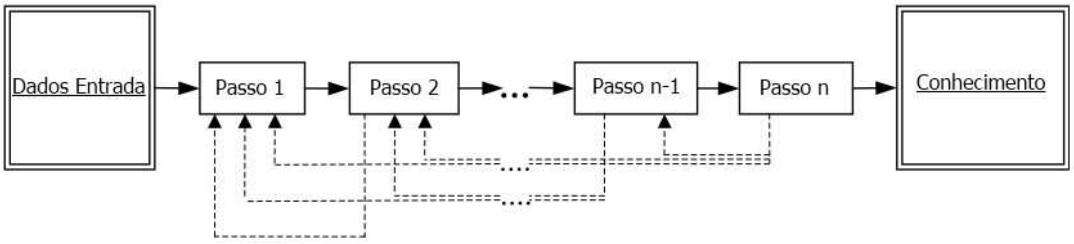


Figura 2.2: Estrutura sequencial do processo KDP (Cios et al., 2007)

Os esforços iniciais para o estabelecimento de um modelo de *Knowledge Discovery Process* (KDP) foram desenvolvidos em termos académicos e moldados de modo a que os utilizadores dos processos de mineração tivessem uma ajuda no desenvolvimento desse processo de descoberta de conhecimento. O modelo inicial tido como referência foi o desenvolvido por Fayyad, que é definido como sendo *"o processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em bases de dados "* (Fayyad et al., 1996).

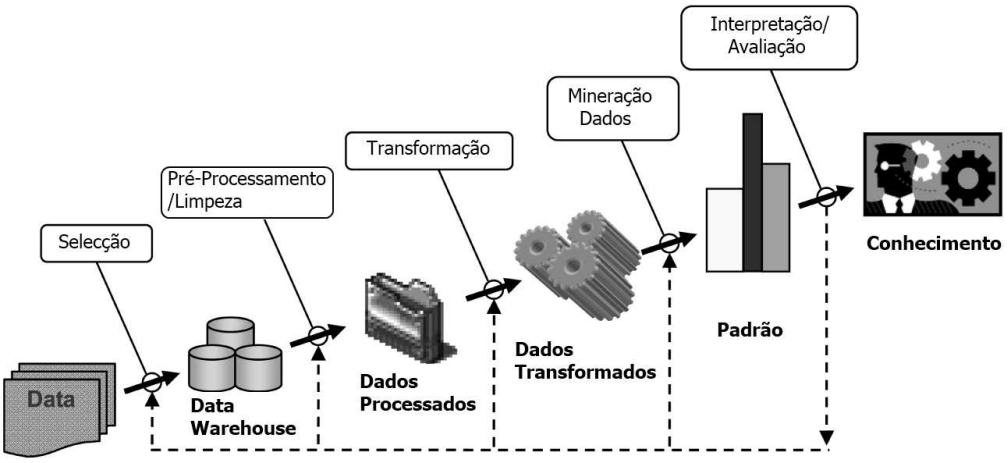


Figura 2.3: O processo de KDD (Fayyad et al., 1996)

O processo descrito por Fayyad et al. (1996) tem essencialmente cinco passos (Figura 2.3) que consistem essencialmente em:

1) *Seleção*. Este passo consiste em criar um conjunto de dados sobre o qual o processo se vai centrar, selecionando o subconjunto de variáveis ou amostras de dados para os restantes passos do processo.
2) *Pré-Processamento/Limpeza*. É a etapa na qual se trata de obter dados consistentes a partir dos dados selecionados para a resolução do problema.
3) *Transformação*. Aqui os dados obtidos são transformados usando métodos que podem passar pela redução dimensional dos dados ou processos de transformação.
4) *Mineração*. A etapa em que se faz a procura de padrões de interesse, através de uma análise particular que depende do objetivo do processo KDD. Em geral é um processo iterativo, em que os analistas podem refinar a procura que pretendem efetuar. Várias técnicas, algoritmos e ferramentas de mineração são utilizados neste passo.
5) *Interpretação/Avaliação*. Este passo consiste na análise e interpretação dos padrões descobertos na etapa anterior, traduzindo-os para uma linguagem mais corrente para que o conhecimento extraído possa ser utilizado e documentado.

Rapidamente, alguns modelos industriais se seguiram aos modelos académicos e várias abordagens, individuais e de empresas, foram também surgindo. Uma das iniciativas que se tornou referência no mundo industrial foi a metodologia CRISP-DM (*CRoss-Industry Standard Process for*

Data Mining). Este modelo foi criado no fim dos anos noventa por quatro empresas: a Integral Solutions Ltd, uma empresa que comercializa produtos de mineração; a NCR, fornecedor de produtos de bases de dados; a Daimler Chrysler, um produtor de automóveis e a OHRA, uma empresa na área dos seguros. Além do apoio empresarial contou também com o apoio da comissão europeia pelo programa ESPRIT (Esprit, 1999).

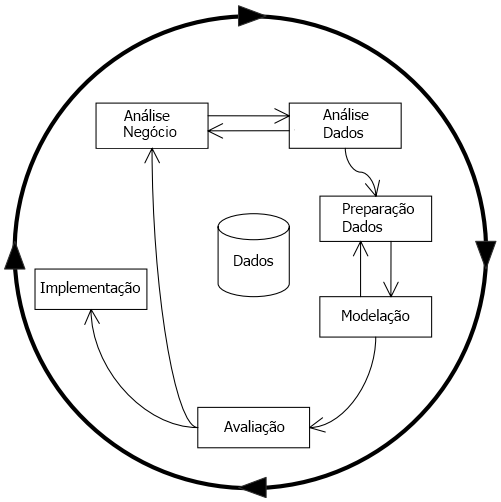


Figura 2.4: O modelo processual da CRISP-DM

O modelo assenta em seis passos (Figura 2.4), que podem ser sumarizados da seguinte maneira (Chapman et al. 2000):

1) *Análise negócio*. Este passo centra-se na compreensão dos objetivos e requisitos do ponto de vista da componente de negócio. Será necessário converte-los também em objetivos de mineração, fazendo para isso um plano de projeto preliminar.
2) *Análise dados*. Neste passo é feita a familiarização com os dados, procedendo-se a uma recolha inicial. Os dados são então analisados, descritos e avaliados em termos de qualidade. Subconjuntos de dados de interesse são também detetados neste passo.
3) *Preparação dados*. Este passo inclui todas as atividades necessárias para a construção do conjunto de dados finais que será a fonte dos passos subsequentes. Inclui a seleção de tabelas, atributos, limpeza dados, construção de novos atributos e transformação de dados.

4) *Modelação*. Neste passo são aplicadas técnicas de modelação para que, dada a natureza do problema, se obtenham os melhores valores possíveis. Será necessário fazer também a identificação das técnicas a aplicar, das ferramentas a utilizar, e da criação de modelos de teste e confirmação da aplicação dos modelos gerados.
5) *Avaliação*. Depois de se ter um ou mais modelos gerados, estes serão avaliados do ponto de vista de negócio quando confrontados com os dados de análise. Os objetivos iniciais são analisados de modo a verificar se algum deles não está a ser cumprido. No final terá que haver uma decisão sobre se os resultados da mineração vão ser ou não utilizados. Consoante essa decisão o próximo passo será especificado.
6) *Implementação*. O conhecimento então extraído terá agora que ser organizado e apresentado de uma forma que possa ser utilizado. Pode ser algo simples como a geração de um relatório ou algo mais complexo que requeira um plano de implementação, monitorização e manutenção.

O desenvolvimento de modelos académicos e modelos industriais levou ao desenvolvimento de modelos híbridos. Um desses modelos é o de o modelo de seis passos de (Cios e Kurgan, 2005).

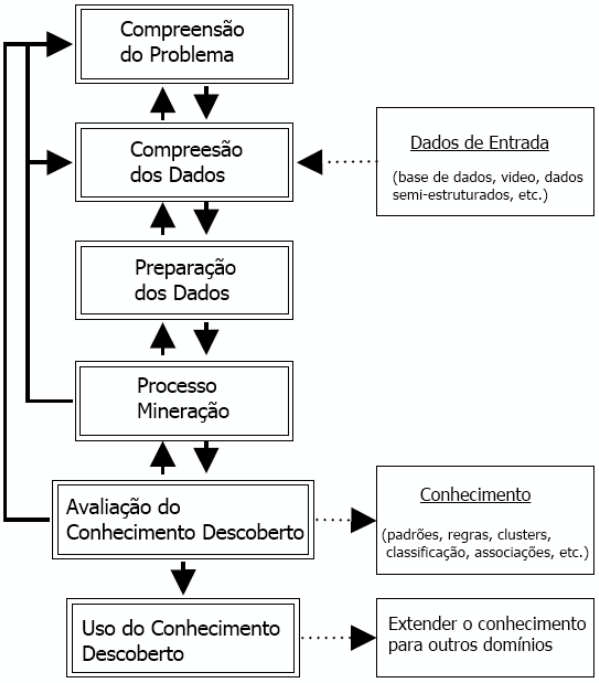


Figura 2.5: Os passos do processo KDP (Cios e Kurgan, 2005)

Os seis passos estão representados na figura 2.5 e podem ser descritos como (Cios e Kurgan, 2005):

1) *Compreensão do problema*. Este passo envolve o trabalhar em conjunto com as pessoas que têm o domínio do problema, peritos com os quais se possa aprender e estudar possíveis soluções para o problema. Envolve também a compreensão de terminologia específica que será usada ao longo do processo. Uma descrição do problema, incluindo as suas restrições é preparada. Finalmente os objetivos do projeto são transformados em objetivos do processo de mineração e é feita a seleção das ferramentas a ser usadas ao longo de todo este processo.
2) *Compreensão dos dados*. Este passo inclui a recolha de amostras de dados, verificação de tipos de dados e formatos, para decisão de quais serão os necessários para os restantes passos. Os dados são verificados para averiguar se são completos, redundantes, quais os atributos que possuem, entre outras características. Desta análise serão selecionados os dados que mais se enquadram para o restante processo de mineração.
3) *Preparação dos dados*. Neste passo é feita a preparação dos dados que serão a fonte para os próximos passos do processo. Inclui a execução de alguns testes como os de correlação e significância, bem como a limpeza dos próprios dados. Valores em falta, ruído que os dados possam ter também são aqui tratados. Algum tratamento adicional pode ainda ser realizado, como a derivação de novos atributos, seleção, execução de algoritmos de extração ou mesmo sumarização, para que os dados melhor se enquadrem nos objetivos a atingir e para que também melhor se adaptem às ferramentas que serão usadas no passo seguinte.
4) *Processo de mineração*. Aqui é realizado o próprio processo de extração do conhecimento. Neste passo são usados os vários métodos disponíveis para mineração, para extrair conhecimento dos dados pré-processados que foram obtidos no passo anterior.
5) *Avaliação do conhecimento*. Após termos o conhecimento obtido no passo anterior é necessário avaliar a sua qualidade. É necessário conhecer os resultados obtidos, verificar se correspondem a conhecimento novo e interessante e, obviamente, interpretá-los. Este passo tem a participação de peritos no domínio de conhecimento em questão de modo a melhor conhecer o impacto da informação obtida. Somente conhecimento validado é considerado. Caso seja necessário poderá ser feita uma reavaliação de todo o processo

realizado de modo a verificar o que pode ser melhorado para que se obtenha conhecimento que garanta melhores resultados.

6) *Utilização do conhecimento descoberto*. Por fim, é feito o planeamento do que se irá fazer com o conhecimento obtido. A área de aplicação do conhecimento pode ser estendida para outros domínios, diferentes do domínio inicial, caso se verifique essa possibilidade. Além do planeamento do que será feito, será também elaborado um plano para monitorizar a implementação do conhecimento recolhido. Todo o processo deverá ser documentado e, então, o conhecimento obtido poderá ser utilizado de acordo com os objetivos inicialmente traçados.

2.3 Mineração Web

A proliferação de informação disponível na Web é constante e aumenta de uma forma tremenda. Com mais de dezassete biliões de páginas indexadas (World Wide Web Size, 2013), a Web é um repositório de dados de enormes dimensões, com um potencial imenso para extração de conhecimento. Contudo tem algumas características diferentes das tradicionais fontes de informação como as comuns bases de dados. Diferenças a nível de configuração, volume e na própria coerência da informação. Tal, criou a necessidade de tratar a mineração Web de uma forma específica. Além disso a Web gera informação em vários formatos e suportes. Por exemplo os registos guardados pelos servidores que nela estão instalados permitem obter informação sobre padrões de acesso de utilizadores, o que pode ser usado para personalização ou melhoria na estrutura do próprio sítio.

O termo mineração Web foi introduzido por Etzioni (1996) para definir o uso de técnicas de mineração com os objetivos de descobrir documentos e serviços, extrair informação dos recursos que a WWW oferece e encontrar padrões dessa mesma informação. De modo a clarificar o que caracteriza este processo de mineração Kosala e Blockeel (2000) sugerem a decomposição do processo nas seguintes tarefas:

1) Procura de recursos, que se define como a tarefa de procurar e identificar os recursos Web pretendidos e que serão utilizados nas restantes tarefas do processo.
2) Seleção da informação e pré-processamento, na qual se seleciona e pré-processa a informação recolhida sobre os recursos escolhidos.

3) Generalização, em que de uma forma automática se descobre os padrões gerais em sítios Web individuais ou mesmo entre diferentes sítios Web.
4) Análise, na qual se faz a validação e a interpretação dos padrões encontrados.

Existem três categorias de mineração de dados aplicada a sítios da Internet, que se diferenciam pela informação sobre a qual a mineração vai incidir (Kosala e Blockeel, 2000). Estas categorias são a mineração de conteúdo (*Web Content Mining*), a mineração de estrutura (*Web Structure Mining*) e a mineração de utilização (*Web Usage Mining*). Existem duas outras variantes em que estas categorias são reduzidas de três para duas categorias apenas. Numa a mineração de estrutura é tratada como uma parte da mineração de conteúdo (Han e Kamber, 2001), na outra a mineração de utilização é tratada como parte da mineração de estrutura (Wang et al., 1999).

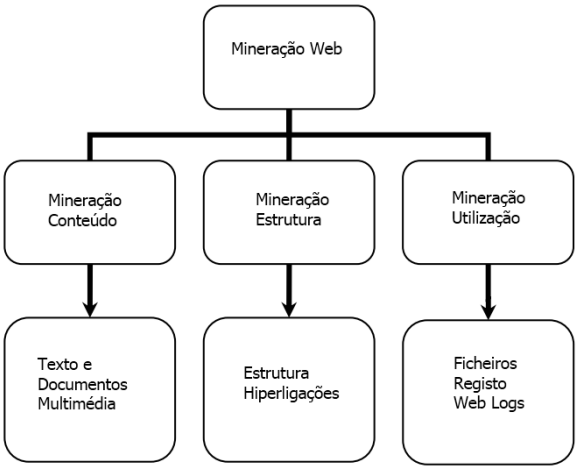


Figura 2.6: Categorias da mineração Web e seus objetos (Kosala e Blockeel, 2000)

Nos tópicos seguintes será feita uma breve explicação sobre o que está subjacente a cada uma das categorias de mineração Web, tendo especial atenção para a mineração de utilização.

2.3.1 Mineração de Conteúdo e de Estrutura

A mineração de conteúdo (Chakrabarti, 2000) centra-se na procura de informação útil para o utilizador que está disponível *online*, ou seja, dada uma determinada pergunta (*query*) encontrar objetos relevantes que satisfaçam essa mesma pergunta. Algumas das suas tarefas são a organização e o agrupamento de objetos que, posteriormente, irão permitir aos motores de busca

pesquisar esses mesmos objetos por palavras-chave, por categorias ou por conteúdos. Estes objetos são tradicionalmente coleções de documentos de texto mas podem ser também documentos multimédia como por exemplo imagens, vídeos e ficheiros de áudio, que estão referenciados, ou mesmo embebidos, como recursos nas páginas Web.

A mineração de estrutura (Hou e Zhang, 2003) foca-se na procura das páginas mais relevantes sobre a estrutura de hiperligações dos sítios. O objetivo deste tipo de mineração é o de encontrar as páginas que são referenciadas mais vezes para um determinado conteúdo, através das referências presentes nas várias páginas. Se estas são referenciadas mais vezes, então é provável que o conteúdo presente nessa página seja útil e relevante para a pesquisa que se faça sobre um determinado conteúdo. O uso de mineração de estrutura minimiza dois grandes problemas que se colocam hoje na Web devido à quantidade de informação disponível e que são os resultados de pesquisa irrelevantes e a incapacidade de indexar toda a informação disponível. Por isso necessário encontrar formas de somente indexar páginas cujo conteúdo seja considerado de facto útil.

2.3.2 Mineração de Utilização

A mineração de utilização analisa as atividades dos utilizadores durante a sua navegação na Web. Fazendo-se essa análise é possível compreender as suas preferências e assim poder melhorar os sítios que visitam em várias vertentes. Essas melhorias podem ser feitas a nível de personalização do sítio (Eirinaki e Vazirgiannis, 2003) ou na otimização da sua estrutura e performance (Pei et al., 2000). Alguns exemplos de sítios nos quais se pode aplicar este tipo de mineração são os sítios de comércio eletrónico, sítios que utilizem publicidade personalizada dirigida ao utilizador, sítios que recorram a deteção de fraude ou de fornecimento de informação para sítios adaptativos. Como já foi referido existem três tipos de arquivos que podem ser utilizados para este tipo processo, nomeadamente os arquivos que são obtidos do lado do servidor, do lado do cliente ou do lado dos servidores *proxy* que fornecem o acesso. O facto de esta informação estar distribuída complica a tarefa de mineração. No entanto, e dado ser mais difícil recolher a informação obtida do lado do cliente, a maioria dos algoritmos que recorrem a este tipo de análise usa somente a informação do lado do servidor, sendo, assim, esta a principal fonte de informação para a mineração de utilização.

Tanto a mineração de conteúdo como a mineração de utilização, mas em particular a mineração de utilização, apresentam alguns problemas à análise dos dados. Questões como a filtragem e

integração de várias fontes (registos de acesso, perfis, entre outros), dificuldades de identificação de utilizadores, identificação de sessões ou transações a partir dos registos de acesso, refletem alguns desses problemas. Estas questões têm sido debatidas de forma assídua. Kosala e Blockeel (2000), bem como Wang (2000) sugeriram algumas abordagens.

A mineração de utilização é essencialmente a aplicação de técnicas de mineração para a descoberta de padrões de utilização, para assim melhor compreender e satisfazer as necessidades dos utilizadores (Srivastava et al., 2000). Podemos identificar três passos principais neste processo (Srivastava et al., 2000): pré-processamento, descoberta de padrões e análise de padrões.

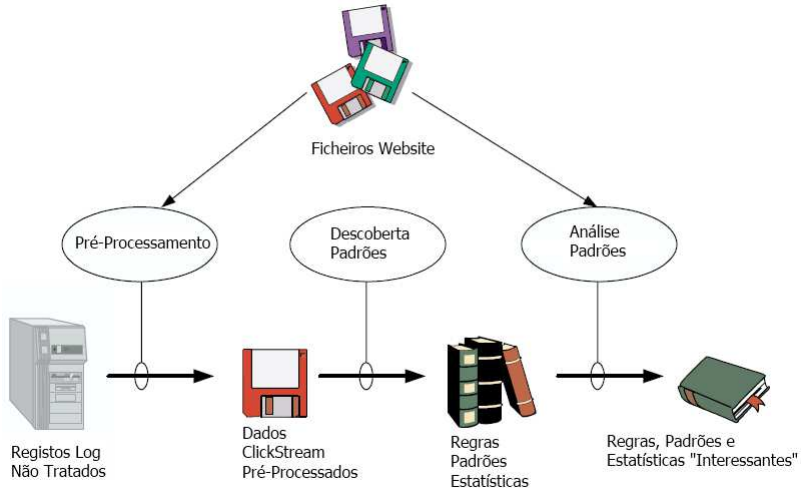


Figura 2.7: Esquema do processo de mineração de utilização segundo Srivastava et al., (2000)

Existem, porém, algumas questões a serem levadas em consideração em cada um desses passos. Assim, na fase de pré-processamento deveremos ter em conta tarefas como a:

1) *Limpeza de dados*, tarefa esta que consiste na remoção de acessos que não devem ser tomados em consideração para a análise, como o acesso a itens irrelevantes (imagens de botões, por exemplo), acessos feitos por utilizadores não humanos (*web crawlers*) ou pedidos falhados.
2) *Identificação de utilizadores*, na qual identificação do endereço IP pode não ser o melhor método para identificar univocamente um utilizador (Rosenstein, 2003). Muitos utilizadores utilizam o mesmo IP e o mesmo utilizador pode ter diferentes endereços IP numa mesma

sessão, considerando o possível uso de *proxies* e *gateways*. Além disto, o mesmo computador pode servir de acesso a vários utilizadores. Assim, e apesar de se puder aplicar algumas técnicas para prever estas situações, nunca será garantida a completa distinção dos diversos utilizadores. Uma possível solução poderia ser utilizar neste processo de identificação alguma informação do lado do cliente, neste caso a informação armazenada em *cookies* para uma melhor identificação, dado que se estima que cerca de 90% dos utilizadores têm esta opção ativa nos seus navegadores (Baldi et al., 2003). O uso de identificação de utilizador e palavra-chave é também uma excelente forma de identificação. Contudo o seu uso se aplica a sítios Web que requeiram autenticação para a sua navegação.

3) *Identificação de sessões e complemento de informação*, na qual se faz a identificação das sessões realizadas. Uma sessão pode ser definida como o conjunto de páginas visitadas pelo mesmo utilizador durante a duração de uma visita particular a um sítio (Pierrakos et al., 2003). Esta identificação pode-se revelar árdua, porque alguns servidores poderão não guardar toda a informação necessária nos seus registos de *log*. A restante terá que ser inferida ou reconstruída usando por exemplo heurísticas temporais.
4) *Identificação de transações*, relativamente a esta tarefa alguns autores propõem agrupar as sessões de modo a obter um conteúdo mais significativo. As páginas visitadas durante a sessão de um utilizador podem ser categorizadas como auxiliares ou de conteúdo. As páginas auxiliares são páginas em que o utilizador não tem interesse nelas. São usadas somente para navegar para páginas que ele realmente pretende, as usualmente designadas por páginas de conteúdo. A identificação de transações pretende agrupar as chamadas páginas de conteúdo, separando-as das páginas auxiliares para assim obter informação mais relevante.
5) *Formatação*, em que se faz a formatação da informação de acordo com o tipo de mineração que se pretende realizar.

Depois da fase de pré-processamento temos a fase de descoberta de padrões, na qual os métodos e os vários algoritmos de mineração selecionados podem ser aplicados aos dados já tratados na fase anterior. Estes métodos por vezes passam por uma adaptação para melhor se enquadrarem ao objetivo a alcançar. Podemos, contudo, desde logo, identificar alguns tipos de análise que podem ser realizados: simples análises estatísticas (*statistical analysis*), descoberta de regras de associação (*association rules*), descoberta de padrões sequências (*sequential patterns*),

agrupamento (*clustering*), classificação (*classification*) e modelação de dependências (*dependency modeling*). Descrevamos então, um pouco, cada uma delas:

1) *A análise estatística*. Esta é uma ciência matemática que diz respeito à recolha, análise, interpretação ou explicação e finalmente à apresentação de dados (Hill e Lewicki, 2006). A palavra em si é o plural de estatística e refere-se à aplicação de um algoritmo estatístico a um conjunto de dados. Os métodos estatísticos podem ser usados para sumarizar ou descrever uma coleção de dados ou mesmo para modelar padrões existentes nesses mesmos dados, de modo a contar com a sua aleatoriedade e incerteza assim como inferir informação a partir deles. Existem muitos métodos estatísticos que podem ser utilizados, desde os mais básicos como a média, confiança ou intervalos, até aos mais avançados como as correlações e regressões. É o método mais simples para extrair informação dos utilizadores que acedem a um sítio, produzindo dados estatísticos que se podem revelar muito úteis nomeadamente para o trabalho dos administradores desses mesmos sítios.
2) *Regras de associação*, estas suportam técnicas para descoberta de padrões frequentes, associações e correlações entre conjuntos de itens. São usadas para descobrir relacionamentos entre as páginas que são visualizadas durante uma sessão, mesmo que aparentemente nenhuma relação possa ser extraída numa primeira instância. Esse relacionamento pode revelar associações entre grupos de utilizadores com interesses específicos. É um dos métodos de mineração mais estudados (Ceglar e Roddick, 2006), sendo utilizadas em vários domínios de problemas, como tarefas de reestruturação de sítios, fazendo uma melhor distribuição de páginas onde possam ser encontradas relações (Mobasher et al., 1999) e, também, em estratégias de marketing, nomeadamente em sítios de comércio eletrónico, podendo por exemplo descobrir associações entre transações executadas pelo mesmo utilizador numa determinada sessão (Jian et al., 2004).
3) *Padrões sequenciais*, que são descobertos por algoritmos específicos (Han et al., 2005) revelam padrões que ocorrem entre sessões. A descoberta de padrões sequenciais é uma extensão da aplicação das regras de associação. Contudo introduz uma noção de tempo e de relações entre as sessões. O objetivo será tentar descobrir conjuntos de itens, páginas, entre sessões que sigam uma sequência temporal, conjuntos de páginas que sejam consultadas após outras. O uso destes tipos de algoritmos foi inicialmente introduzido para o estudo de sequências de compras de clientes (Agrawal e Srikant, 1995) e o uso deste tipo de informação pode ser usado para prever futuras visitas de utilizadores para

colocação, por exemplo, de anúncios ou recomendações. Pode também ser utilizado em estratégias de *prefeteching* de páginas, isto é, colocar as páginas preparadas para visualização antes mesmo de estas serem invocadas. Dois tipos de métodos têm sido usados para descobrir padrões sequenciais: métodos determinísticos que gravam o comportamento de navegação do utilizador; e métodos estocásticos que usam a sequência de páginas visitadas para prever as páginas seguintes que previsivelmente serão requisitadas. Um exemplo de um método determinístico foi apresentado por Spiliopoulou et al. (1999) para a procura sequencial de padrões. O exemplo típico de métodos estocásticos são os modelos de Markov, que são os mais utilizados para previsão de hiperligações. As cadeias de Markov, grafos acíclicos com um conjunto de estados associados e um conjunto de transições entre estados, aplicados a um exemplo de um cesto de compras podem ser caracterizados da seguinte forma: cada estado corresponde a um elemento do cesto de compras e no caso de navegação pela Internet, cada estado é uma página. Borges e Levene (2007) usaram os modelos de Markov para analisar o comportamento de utilizadores que navegam na Web.

4) *Agrupamento*, uma técnica que é usada em grande parte dos métodos na descoberta de padrões na Web. O agrupamento é utilizado para descobrir grupos de itens com características semelhantes. Em (Han e Kamber, 2001) foi proposta uma taxonomia para os métodos de agrupamento:
    a. Os métodos de partição, que criam *k* grupos de um dado conjunto de dados, em que cada grupo representa um agrupamento;
    b. Os métodos hierárquicos que decompõem um dado conjunto de dados para criar uma estrutura hierárquica de grupos;
    c. Os métodos baseados em modelos, que procuram o melhor modelo matemático para um dado conjunto de dados.

    Ao nível mais específico da Web podemos desde logo distinguir dois tipos de agrupamento, agrupamento de utilizadores e agrupamento de páginas (Srivastava et al., 2000). O agrupamento de utilizadores agrupa utilizadores com comportamentos de navegação semelhantes. O agrupamento de páginas agrupa páginas com características semelhantes. Algoritmos de agrupamento como os algoritmos de otimização por colónia de formigas (*ant-based Optimization*) ou otimização por grupo de partículas (*Particle Swarm Optimization*) podem ser usados nesta fase (Handl e Meyer, 2007). O conhecimento

extraído pode ser utilizado em sítios de comércio eletrónico para segmentação de mercado ou para tarefas genéricas de personalização.

5) *Classificação*, que é um processo que mapeia um item numa ou várias classes pré-definidas. Um classificador é o mapeamento de um espaço X (discreto ou continuo) para um conjunto discreto de valores Y (Duda et al., 2000). Geralmente um domínio Web é utilizado para mapear um utilizador num conjunto de categorias em que cada categoria tem um conjunto de atributos definidos e o utilizador é inserido nessas categorias, consoante as suas próprias características. Neste processo são utilizados algoritmos de classificação em que alguns dos mais comuns são árvores de decisão, classificadores *naïve bayes* ou redes neuronais.
6) *Modelação de dependências*, que é utilizada para determinar se existe alguma dependência significante entre variáveis num domínio Web. Este tipo de técnica é útil para determinar futuros consumos de recursos de um sítio ou no desenvolvimento de estratégias de negócio.

No passo final temos a análise dos padrões que envolve a validação e análise dos padrões descobertos. A validação implica eliminar regras ou padrões tidos como irrelevantes e escolher de facto a informação que suscita interesse para a análise que se pretende efetuar. Interpretar os resultados obtidos não é uma tarefa que possa ser efetuada por interpretação humana direta. Geralmente essa análise é feita por mecanismos de perguntas diretas à base de dados (*querys*), em que os resultados estão guardados, ou carregando os dados dos resultados em cubos de dados e, então, efetuar operações OLAP sobre esses cubos, usando também técnicas de visualização para facilitar a interpretação dos resultados da análise efetuada. Kimball e Merz (2000) descreveram uma visão integrada de mineração de utilização aplicada à Web a que chama *"Web data warehousing"* e que inclui ferramentas OLAP e desenvolvimento de esquemas para o armazenamento de dados Web.

2.3.3 Ferramentas e Aplicações para a Descoberta de Padrões

Genericamente, o processo de mineração pode ser definido como sendo a descoberta de padrões interessante, que não são óbvios em grandes coleções de dados (Klosgen e Zytkow, 2002). A nível da Web a mineração é uma poderosa ferramenta para empresas e organizações que queiram retirar mais informação dos seus sítios, seja para fins comerciais seja para oferecer melhores serviços aos seus clientes. A nível de aplicações comerciais, a utilização de processos de mineração

em sítios de comércio eletrónico oferece vantagens mais evidentes. A aplicação de mineração de utilização para reunir informação sobre a atividade dos seus clientes *online* ajuda as empresas a ganhar um bom conhecimento do negócio, na forma de comportamentos e padrões que descrevem quer a navegação dos seus utilizadores quer o seu comportamento consumista (Buchner et al., 1999). Este conhecimento de padrões, de perfis de consumo e mesmo da segmentação do seu mercado dá às empresas algumas vantagens competitivas, que podem fazer a diferença face aos seus concorrentes. A utilidade dos padrões não se restringe à sua utilização comercial. Este tipo de conhecimento permite, também, que sejam efetuadas melhorias quer a nível de sistemas quer a nível do próprio sítio de modo a torná-lo mais amigável.

A análise de registos de acessos, os ficheiros de *log*, é o primeiro passo na mineração de utilização. Hoje, praticamente todas as ferramentas comerciais a suportam. A maioria das ferramentas públicas e pagas são simples analisadores de *logs* ou de tráfego, direcionadas para pequenas e médias empresas que pretendem tão-somente uma análise do tráfego ou dos acessos ao seu próprio sítio, sendo que as funcionalidades disponibilizadas ficam limitadas ao fornecimento de relatórios estatísticos. Nessa categoria temos aplicações como o Analog[1], AWStats[2] ou o WUM[3]. Contudo nem todas as aplicações têm funcionalidades limitadas à apresentação de estatísticas. O WUM, por exemplo, é uma aplicação mais avançada, permitindo mineração de sequências utilizando uma linguagem de *querys* específica, a MINT, que ajuda no processo de descoberta de padrões (Spiliopoulou, 1999).

Existem bastantes ferramentas no mercado que oferecem funcionalidades mais avançadas, que são disponibilizadas ou em produtos específicos de mineração ou em soluções integradas, em conjunto com os serviços *data warehousing.* Algumas delas disponibilizam também algoritmos de mineração, relatórios e visualização por gráficos e diagramas, em que se incluem, por exemplo, os produtos disponibilizados pela Microsoft ou pela Oracle. Alguns destes produtos integram com o CRM das empresas permitindo combinar os dados obtidos dos ficheiros de *log* com a informação disponível de clientes, e outros dados operacionais, fornecendo às empresas uma análise mais abrangente sobre os seus processos de negócio. Exemplos de algumas aplicações comerciais que oferecem estas funcionalidades, e que estão disponíveis atualmente no mercado, são por exemplo

[1] http://www.analog.cx
[2] http://awstats.sourceforge.net
[3] http://ka.rsten-winkler.de/hypknowsys/wum/

os produtos da SAS[4], os produtos da WebTrends[5], o software DBMiner[6] ou o produto da IBM SPSS Clementine[7].

Existem ainda algumas ferramentas de domínio público disponíveis, como sendo o WEKA[8], o KEEL[9] ou o RapidMiner[10]. O software WEKA (Witten e Frank, 2005), é um produto *open source* que disponibiliza uma série de algoritmos para tarefas de mineração de informação bem como para a sua visualização. Estas funcionalidades estão também presentes no KEEL (Alcalá et al., 2004), outra ferramenta *open source* desenvolvida para construir e usar diferentes modelos de mineração, como algoritmos de pré-processamento, árvores de decisão, métodos estatísticos, redes neuronais, entre outros. Dentro destas ferramentas de domínio público há ainda a mencionar o RapidMiner, uma ferramenta de mineração bastante utilizada, com diferentes versões que podem incluir algumas funcionalidades adicionais, mas que na sua versão comunitária possui já um vasto leque de características que lhe permitem ser considerado uma das melhores ferramentas *open source* disponíveis. Um estudo comparativo de alguns destes softwares pode ser consultado em (Magdalena et al., 2009).

Os resultados da descoberta de padrões podem ser aplicados em várias áreas, das quais de destacam as seguintes a nível da Web: 1) recolha de informação para aplicação em melhorias no sistema; 2) melhorias de design do sítio; 3) emissão de recomendações ou inclusão de tópicos adicionais para o utilizador; 4) personalização Web; e 5) compreensão do comportamento dos utilizadores/clientes, que pode ser inserido num contexto de conhecimento de negócio. Vejamos, com um pouco mais de detalhe cada uma destas áreas:

1) A recolha de informação para melhorias no sistema pode ser feita a nível de desempenho ou de segurança do próprio sistema. A nível de desempenho visa essencialmente a melhoria da satisfação de navegação dos utilizadores, nomeadamente com a introdução de políticas de *web caching*, *load balancing* e melhorias na transmissão na rede ou

[4] http://www.sas.com
[5] http://www.webtrends.com
[6] http://www.dbminer.com
[7] http://www.spss.com/clementine
[8] http://www.cs.waikato.ac.nz/ml/weka/
[9] http://www.keel.es/
[10] http://rapid-i.com/content/view/181/196/lang,en/

distribuição de dados. A nível de segurança pode ser aplicado em deteções de intrusão, deteção de fraude e tentativas de entradas não autorizadas no sistema.

2) A melhoria do sítio pode ser feita a nível de reorganização de páginas de uma forma manual (Melody e Marti, 2002) ou automática. Os sítios adaptativos (Perkowitz e Etzioni 2000) melhoram automaticamente a sua organização e apresentação, aprendendo com os padrões dos seus visitantes. A mineração é feita sobre os logs dos servidores para permitir uma navegação mais amigável. Técnicas de *clustering mining* e *conceptual clustering* são aplicadas para sintetizar os índices das páginas que são tidas como centrais de forma a possibilitar a organização do sítio.
3) Os sítios de comércio eletrónico usam bastante as técnicas de mineração para lhes dar sugestões de compras aos seus clientes ou informação que os auxilie na sua navegação. Estas recomendações ou informações são baseadas nas compras que o cliente realizou no passado ou no seu padrão de navegação. Várias tecnologias têm sido propostas para sistemas de recomendação (Sarwar et al., 2000), tendo muitos sítios de comércio eletrónico incorporado já estes sistemas (Schafer et al., 2000).
4) A personalização a nível da Web permite reorganizar as páginas de um sítio para que este se possa adaptar melhor às necessidades dos utilizadores (Spiliopoulou 2000). É uma área vasta que engloba os sítios adaptativos e sistemas de recomendação como casos especiais. Uma visão geral sobre abordagens para incorporar conhecimento semântico e personalização é dada no seguinte artigo (Dai e Mobasher, 2003).
5) A compreensão de comportamentos de utilizadores na Web tem também uma aplicação clara em sítios de comércio eletrónico. O sistema usa técnicas de mineração para descobrir regras que descrevam o comportamento dos utilizadores. Em (Fu et al., 2000) foi proposto um algoritmo para agrupar utilizadores baseado nos seus padrões de navegação, organizados por sessões de exploração e que representam iterações entre utilizador e sítio. Usando indução baseada em atributos, as sessões são então generalizadas de acordo com a hierarquia de páginas. Nesta hierarquia as páginas estão organizadas por generalização. As sessões generalizadas são finalmente agrupadas usando um método de agrupamento hierárquico.

A área de deteção de fraude tem também sido uma importante aplicação da descoberta de padrões, usando algoritmos e técnicas de mineração para descobrir padrões de utilização invulgares, que, ao serem posteriormente analisados, podem revelar comportamentos

fraudulentos. Uma das aplicações específicas é a nível de deteção de movimentos bancários fraudulentos, ou seja, detetar dentro nos registos de transações bancárias dos diversos clientes dos bancos, padrões de transações anormais que podem revelar falhas de segurança ou comportamentos fraudulentos. A rápida análise dos dados e a deteção destes movimentos pode permitir o aparecimento de métodos pró-ativos para a deteção de fraude, avançando para sistemas em tempo real que possam atuar antes mesmo que este tipo de transações possa ser efetuada (Edge e Sampaio, 2009). Uma outra aplicação na área de deteção de fraude pode ser a nível de telecomunicações (Ferreira et al. 2007). Nesse artigo foi utilizado uma técnica de agrupamento dinâmico e um modelo baseado em assinaturas para a deteção de situações consideradas anómalas num ambiente de telecomunicações móveis.

2.4 Personalização

O foco central da personalização é fazer com que os sistemas de informação tenham a capacidade de se adaptar às necessidades dos seus utilizadores. Esta preocupação está a ser cada vez mais importante devido à enorme quantidade de informação que atualmente existe *online*, assim as organizações tentam adicionar valor aos seus sítios criando sistemas centrados nas necessidades dos seus clientes para que estes ofereçam conteúdos personalizados a esses mesmos clientes em tempo real e sem que estes tenham qualquer tipo de intervenção no processo. Este será sempre o principal objetivo: dar aos utilizadores o que eles querem ou precisam, antecipando o seu pedido explícito (Mulvenna et al., 2000).

Os passos para essa personalização incluem a recolha de informação, a modelação e a sua categorização, a análise dos dados recolhidos e a determinação das ações que deverão ser tomadas. Os meios empregues na análise da informação recolhida incluem:

1) *A filtragem de conteúdo*, que se baseia somente na preferência dos utilizadores; o sistema verifica as preferências passadas do utilizador e recomenda itens similares aos que ele escolheu no passado; existem também combinações entre filtragem colaborativa e filtragem de conteúdo (Melville et al., 2002).
2) *A filtragem colaborativa*, na qual o sistema convida os utilizadores a classificar objetos ou a assinalar as suas preferências e interesses para assim retornar informação que

previsivelmente será interessante para eles; em (Adomavicius e Tuzhilin, 2005) podemos encontrar algumas aplicações do uso de filtragem colaborativa.

3) *A filtragem baseada em regras*, em que os utilizadores são convidados a responder a uma série de questões; estas perguntas são derivadas de uma árvore de decisão e ao responder a cada uma delas o sistema irá apresentar ao utilizador o resultado da travessia dessa árvore de decisão, que pode ser, por exemplo, um conjunto de produtos que deverão corresponder às necessidades do utilizador.
4) *A utilização de mineração dos dados de utilização*, que, como já vimos, corresponde a uma análise de padrões para a descrição do comportamento de navegação dos utilizadores, para que assim, depois da análise desses mesmos padrões se possam estabelecer as medidas de personalização mais adequadas.

Conceptualmente poderemos ver um sistema de personalização Web como um conjunto de módulos interligados, que visam a alteração concertada e dinâmica de um sítio baseada em informação recolhida de diversas fontes de informação.

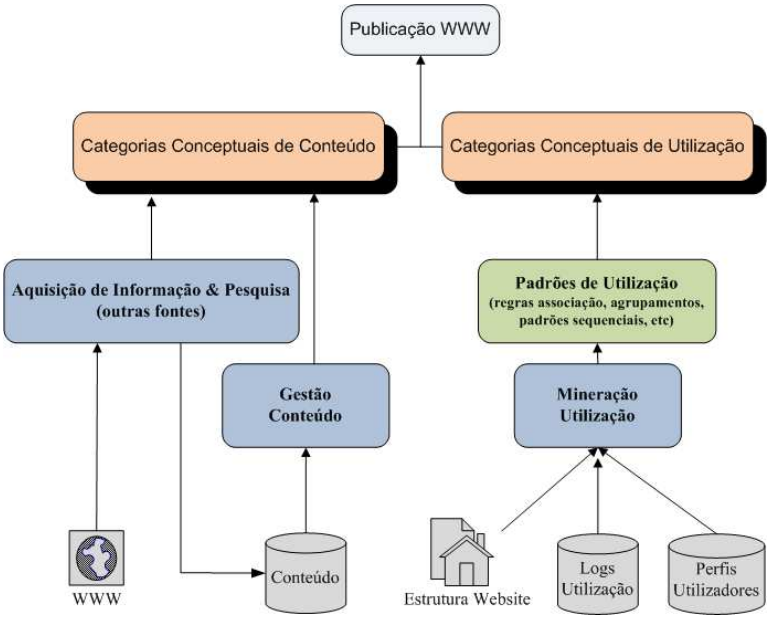


Figura 2.8: Módulos de um sistema de personalização Web

Analisando um pouco a figura 2.8 e os diversos módulos que nela estão representados, podemos ver que o módulo de gestão de conteúdo processa o conteúdo de um sítio, classificando-o

posteriormente em categorias conceptuais de conteúdo. Ainda relacionado com a classificação conceptual de conteúdo podemos melhorar essa mesma classificação com informação adicional adquirida de outras fontes, usando técnicas avançadas de pesquisa. Analisando agora as categorias conceptuais de utilização, podemos identificar desde logo três fontes distintas de informação. A informação obtida com base na estrutura do sítio e informação resultante de mineração de utilização. Por sua vez, a informação resultante de mineração de utilização deriva da análise de padrões de utilização, perfis, análise de comportamentos, dados de sessões entre outros dados considerados relevantes. A informação que é recolhida do processo de mineração pode ser posteriormente adicionada ao próprio perfil de utilizador, para que, deste modo, da próxima vez que a informação de perfil de utilizador seja utilizada, já contenha este enriquecimento de informação. A informação sobre conteúdos, hiperligações e comportamentos típicos é posteriormente conceptualmente abstraída e classificada em categorias semânticas. Qualquer informação extraída das diversas relações entre conhecimento adquirido com processos de mineração e os conteúdos do sítio em si farão a base para possíveis alternativas para a reestruturação desse mesmo sítio. Um mecanismo de publicação fará então a modificação do sítio em si, garantindo que cada utilizador tem a estrutura ideal para a sua navegação. As opções disponíveis de conteúdo serão organizadas de acordo com as preferências do utilizador.

2.5 Sítios Adaptativos

Os utilizadores não revelam uma utilização de sítios homogénea. Como tal, os sistemas têm que se adaptar às necessidades dos utilizadores, caso pretendam apresentar uma melhor qualidade de serviço. A adaptabilidade é uma funcionalidade particular de um sistema, que pode atenuar algumas dificuldades de navegação, fazendo a distinção entre interações de diferentes utilizadores dentro do seu espaço de informação (Brusilovsky e Nejd, 2004). Uma das alternativas mais viáveis para a promoção desta adaptabilidade aos diferentes perfis de utilizadores são os sítios adaptativos (Brusilovsky, 2004). De acordo com Koch (2004) os sistemas baseados em sítios adaptativos potenciam a abordagem centrada no utilizador, na qual o sistema adapta os aspetos visíveis de acordo com o perfil do utilizador em causa, baseado numa recolha de dados do próprio utilizador, gerando um interface com informação apropriada bem como uma disposição gráfica adequada a cada utilizador. Os sítios adaptativos têm a potencialidade de oferecer aplicações e serviços personalizados, promovendo automaticamente a sua organização e apresentação de acordo com os padrões de acesso dos utilizadores. Segundo Perkowitz e Etzioni (2000) as páginas tornam-se

mais acessíveis, sendo possível destacar hiperligações interessantes, conectar páginas relacionadas ou promover o agrupamento de documentos semelhantes.

Um modelo possível para este tipo de sítios pode ser consultado na figura 2.9. Todo o processo de adaptação do sítio pode ser dividido em duas componentes distintas, a componente *offline* e a componente *online*. Na componente *offline* podemos incluir todo o pré-processamento dos dados, limpeza, e tarefas de preparação e modificação de dados, que resultam na informação final pronta com os registos necessários para análise. Estes registos alimentam o módulo de descoberta de padrões, que é responsável pela descoberta de padrões de utilização por via de técnicas de mineração. O resultado da descoberta de padrões alimenta já a segunda componente do sistema, que funciona em modo *online* que, e que, tendo em consideração os dados resultantes da descoberta de padrões, procede a uma adaptação do sítio ao perfil do utilizador em questão. O próprio sítio em si, pelos registos de utilização que gera, seja através de ficheiros de *logs* ou outros sistemas de armazenamento, vai ser a fonte do módulo de recolha de dados, que alimentará novamente a componente *offline*, completando desta forma o círculo de personalização.

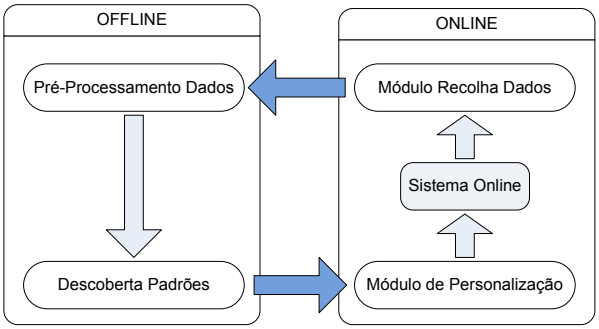


Figura 2.9: Um possível modo de funcionamento de um sítio adaptativo

O nível de adaptação e personalização de um sítio relativamente ao utilizador que o acede é variável e depende de vários fatores e, obviamente, dos objetivos pretendidos, sendo que, genericamente, essa adaptação é feita de uma forma semiautomática, melhorando a sua organização e apresentação, e aprendendo (se possível) com os padrões de acesso dos utilizadores do sistema. Podemos sumarizar em alguns pontos a que níveis os processos de adaptação podem ocorrer, como resultado das tarefas executadas no módulo de personalização representado na figura 2.9. De seguida apresentaremos aqueles que achamos mais relevantes no âmbito deste trabalho.

Conteúdo

Ao mudar o conteúdo de um sítio esse mesmo sítio proporcionará (potencialmente) uma maior satisfação aos requisitos de cada utilizador que o acede. O conteúdo pode ser adicionado, removido ou mesmo adaptado (Kilfoil et al., 2003). As alterações podem envolver várias áreas de atuação:

- *Explicações adicionais ou detalhes* que podem ou não surgir ao utilizador consoante o conhecimento ou interesse do utilizador no assunto em questão (Kobsa et al., 2001).
- *Detalhes adicionais* sobre um determinado tópico que podem ser ou não incluídos na página mediante a perceção de interesse do utilizador (Kobsa et al., 2001).
- *Personalizações recomendadas*, principalmente a nível de sítios de comércio eletrónico nos quais são oferecidos produtos que podem interessar ao utilizador ou mesmo hiperligações para outras subsecções, nas quais o utilizador poderá procurar algo que seja do seu interesse (Kobsa et al., 2001).
- *Dicas de interesse*, que são oportunamente apresentadas em determinados contextos de utilização e mediante os interesses demonstrados do utilizador, em que ele poderá encontrar informação adicional (Kobsa et al., 2001).
- *Substituição de conteúdo*, dependendo das capacidades do navegador que o utilizador está a utilizar para acesso ao sistema ou dos seus interesses, determinados tipos de conteúdo podem ser substituídos por outros, com requisitos funcionais menores ou maiores. Como exemplo, podemos apontar a substituição de uma imagem de um mapa por uma descrição textual ou mesmo um conteúdo multimédia, como um vídeo, que pode ser substituído por uma imagem fixa e uma hiperligação para o vídeo, caso a conexão detetada do utilizador indique uma largura de banda baixa (Nielsen, 1999).

Apresentação

Além de se mudar o conteúdo que é apresentado ao utilizador, altera-se também a forma como a informação lhe é apresentada. Algumas das modificações possíveis a nível de apresentação podem ser:

- *Variantes de página*, diferentes versões de todas as possíveis variantes adaptadas para os diferentes tipos de utilizador são guardadas no sistema e, em tempo real, uma determinada página é mostrada ao utilizador. Um exemplo comum desta técnica são os

sítios multilingue nos quais existem diferentes versões da mesma página no sistema, mas em que cada uma delas é traduzida no seu respetivo idioma. Consoante a seleção de idioma do utilizador uma determinada versão da página é então mostrada (Kobsa et al., 2001).

- *Fragmentos variantes*, um pouco similar às variantes de página, esta técnica guarda pequenos pedaços de fragmentos de conteúdo (átomos) e, também em tempo real, seleciona os fragmentos que são adequados para construir uma página de informação para apresentação ao utilizador (Kobsa et al., 2001). Esta técnica pode ser aplicada em sítios que tenham fragmentos de informação que podem ser agregados como os sítios de notícias (Ardissono et al., 2001).
- *Coloração de fragmentos*, que consiste em fazer sobressair ou diminuir a intensidade de apresentação de determinados fragmentos de informação, através do uso de técnicas de cor. Neste caso o conteúdo das páginas em si é o mesmo para todos os utilizadores, variando somente a forma como cada um dos utilizadores visualiza a informação (Kobsa et al., 2001).
- *Geração de texto em linguagem natural adaptada*, também semelhante às variantes de página, em que as páginas são guardadas e recuperadas quando necessário, esta técnica alterna descrições de texto para diferentes utilizadores (Kobsa et al. 2001). Uma aplicação desta técnica pode ser vista em tradutores de páginas *online* como o Babel Fish do Altavista[11].

Hiperligações

A adaptação da navegação também pode ser conseguida pela adaptação das várias hiperligações presentes num sítio (Kobsa et al., 2001). Esta adaptação visa aumentar a rapidez de procura de uma página em particular e impede que os utilizadores se percam em navegações desnecessárias. Algumas técnicas possíveis para adaptação de hiperligações são:

- *Orientação direta*. Nesta técnica uma hiperligação dinâmica é dada ao utilizador, como por exemplo um botão de "seguinte", que o guiará dinamicamente para uma página que o sistema entende que será a melhor para o utilizador em questão. É uma das técnicas mais simples que pode ser usada neste contexto (Brusilovsky, 1997).

[11] http://babelfish.altavista.com

- *Classificação*. Consiste em selecionar as páginas mais relevantes, baseado nos interesses do utilizador ou dos seus objetivos de pesquisa e posterior apresentação dessas páginas através de uma lista ordenada de hiperligações. As hiperligações mais relevantes são sempre apresentadas em primeiro lugar. Contudo o utilizador poderá consultar as restantes hiperligações, caso pretenda (Brusilovsky 1997). Alguns dos problemas no uso desta técnica é a dificuldade de lidar com os índices necessários para a respetiva indexação de páginas e ter em consideração que a variação na ordenação das hiperligações poderá causar alguma confusão ao utilizador.
- *Ocultação, inabilitação e remoção*. São três técnicas bastante semelhantes. Ocultar hiperligações irrelevantes para o utilizador torna a sua navegação menos confusa e permite aumentar a velocidade de acesso à informação que ele considera relevante (Kilfoil et al., 2003). Ao ocultar uma hiperligação esta fica igual ao texto circundante, passando despercebida ao utilizador, causando menos confusão. Isto não implica alteração na estrutura da informação em si dado que a hiperligação existe, contudo o utilizador não a consegue encontrar. A diferença para a inabilitação é que na inabilitação a hiperligação é apresentada e o utilizador consegue vê-la. Porém, este não consegue seguir essa mesma hiperligação, uma vez que está desabilitada (Kobsa et al., 2001). A operação de remoção remove completamente a hiperligação ao utilizador que fica indisponível e por isso o utilizador não a consegue seguir, deixando somente o texto âncora ou imagem de contexto (De Bra e Calvi, 1998).
- *Anotação de hiperligações*. Esta técnica usa diferentes símbolos, como ícones, cores, tamanho de letra, tipo de letra ou outras, para indicar hiperligações potencialmente interessantes ao utilizador (Brusilovsky, 1997). Esta técnica pode ser uma alternativa bastante relevante à técnica de esconder ou desabilitar uma hiperligação que pode provocar a sensação de confusão de hiperligações ao utilizador (Brusilovsky, 1996), estimulando o uso de fontes, tipos de letras especiais e cores para tornar a navegação do utilizador mais intuitiva (Virvou et al., 2001).

Estrutura

Aparte de modificações temporárias à estrutura de um sistema, feitas a pedido, é possível que a adaptação de um determinado sítio seja feita ao nível da sua estrutura permanente. De um modo geral esta decisão não deverá ser tomada por um qualquer processo automático, tendo que ter alguma intervenção humana dado que altera a estrutura permanente do sítio. Estas alterações

podem ocorrer a nível das páginas dos sítios ou dos átomos de informação aí presentes. Todavia, podem também ser alterações que necessariamente precisem da intervenção de um administrador como o mapeamento de imagens (Brusilovsky, 1996). Apesar de ter que ter alguma intervenção humana, o sistema poderá prestar algum auxílio para que esta decisão possa ser feita de uma forma mais eficaz, fornecendo alguns indicadores e sugestões, como por exemplo:

- *Novas indexações de página*. Baseando-se em padrões de navegação dos utilizadores, o sistema poderá sugerir a criação de novos índices que suportem de um ponto central as hiperligações mais utilizadas (Perkowitz e Etzioni, 1998).
- *Estatísticas sobre o uso de páginas*. O sistema ao gerar estatísticas relativamente ao uso de páginas e conjuntos de páginas mais utilizadas, o que vai permitir ao administrador do sistema, ou a quem tenha o poder de decisão de alteração de estrutura do sítio, verificar se a espectativa de visualizações de páginas está a ser cumprida. Isto poderá permitir a reclassificação de páginas e hiperligações ou mesmo a sua total remoção.
- *Sugestões permanentes*. Poderá ser sugerido pelo sistema que determinadas hiperligações sendo temporárias ou dinâmicas possam ser classificadas como permanentes, dada a afluência de acessos pelos utilizadores do sistema.

2.6 Sistemas de Recomendação

Os sistemas de recomendação têm sido alvo de uma pesquisa bastante ativa durante a última década, muito por influência dos sistemas que utilizam técnicas de recomendação como forma de apresentação de conteúdo, produtos ou serviços. Os sítios de comércio eletrónico são um exemplo típico do uso destas técnicas. Mas não só. Podemos ver também a sua aplicação noutras áreas, como nas redes sociais, em plataformas como o FaceBook[12].

Uma das bases dos sistemas de recomendação é o conhecimento das preferências dos seus utilizadores. São essas preferências que servirão para a tarefa de fornecer os itens que teoricamente serão mais do interesse do utilizador do sistema. A recolha dessas preferências pode ser feita recorrendo a várias técnicas de recolha de informação (secção 2.4). Fundamentalmente, esse processo de recolha pode ser feito de uma forma explícita, em que o utilizador de alguma

[12] http://www.facebook.com

forma classifica as suas preferências, ou de uma forma implícita, em que essa recolha é feita analisando dados recolhidos e armazenado no sistema.

A figura 2.10 retrata genericamente o funcionamento de um sistema de recomendação. As preferências do utilizador são, numa primeira instância, recolhidas e classificadas segundo a sua importância, para posterior seleção. A importância de um determinado item pode também ser dada de uma forma explícita pelo utilizador, ou de forma implícita pelo sistema. Tendo as preferências do utilizador já selecionadas, o próximo passo será prever quais os itens que o sistema irá apresentar como recomendações. Esta tarefa é feita por um módulo que na figura 2.10 está representado como motor de previsão que através da utilização de técnicas de previsão classifica e apresenta ao utilizador os melhores itens desta lista. Mais tarde estes serão apresentados ao utilizador como recomendações. Existem algumas abordagens que os sistemas de recomendação utilizam para a previsão dos itens que serão recomendados ao utilizador. De seguida discutiremos algumas delas.

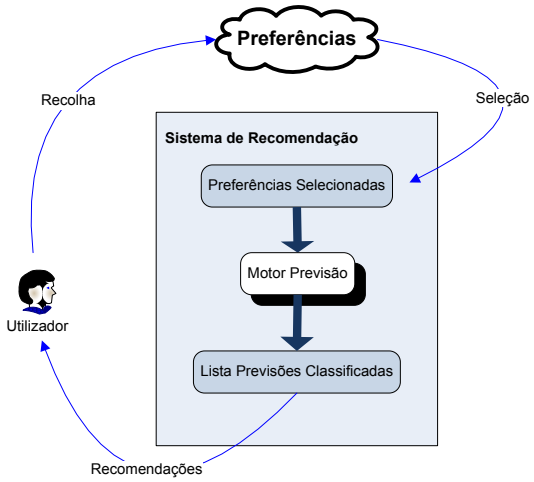


Figura 2.10: Funcionamento genérico de um sistema de recomendação

Abordagens clássicas.

Algumas técnicas, já abordadas anteriormente na secção 2.4, podem ser agrupadas essencialmente em três categorias. A primeira é a filtragem de conteúdo, também conhecida como filtragem cognitiva, que se baseia na classificação dos itens feita pelo próprio utilizador para recomendar itens semelhantes presentes no sistema, com conteúdos ou características

semelhantes (Pazzani e Billsus, 2007). A filtragem de conteúdo é das técnicas clássicas mais usadas, fundamentalmente para itens do tipo texto ou documentos (Adomavicius e Tuzhilin, 2005). A segunda destas técnicas, a filtragem colaborativa, é talvez a mais popular a nível de técnicas de recomendação. É também conhecida como filtragem social, porque se baseia em preferências coletivas de outros utilizadores. Essencialmente, funciona com base na consulta e comparação das opiniões ou preferências de utilizadores com os mesmos gostos do utilizador alvo (Sarwar et al., 2000). Muitos sítios de comércio eletrónico, como o Amazon[13], fazem uso desta técnica. Existem ainda algumas outras técnicas mistas ou híbridas que tentam balancear as vantagens e desvantagens das técnicas clássicas já descritas, partindo da premissa que nenhuma técnica isolada é a melhor solução para todos os utilizadores em todas as situações (Adomavicius e Tuzhilin, 2005). Um sistema híbrido de recomendação tem que ter pelo menos duas ou mais técnicas conjuntas para produzir as suas recomendações. Em (Burke, 2002) podemos encontrar mais informação sobre estes sistemas, nomeadamente algumas outras classificações de sistemas híbridos feitas pelo autor.

Sistemas baseados em classificação

A abordagem baseada em classificação tem uma estratégia um pouco diferente: encontrar itens de interesse para o utilizador, recorrendo a técnicas de classificação que se baseiam em informação recolhida pela categorização de objetos. O objetivo de recomendação é, todavia, o mesmo, dado um conjunto de itens o sistema tentar encontrar um subconjunto que seja do interesse do utilizador. No entanto, por vezes, as abordagens tradicionais recorrem a técnicas de correspondência de itens para incluir nas recomendações ao utilizador. Isto pode ser feito, por exemplo, recorrendo à criação de um "conjunto de palavras" que por correspondência semântica dará um conjunto de itens de potencial interesse. Para efetuar esta correspondência ou semelhança entre itens existem vários algoritmos disponíveis (Baeza-Yates e Ribeiro-Neto, 2011). Existem, contudo, situações em que esta correspondência não pode ser feita desta maneira mais simplista, uma vez que pode conduzir a resultados de recomendação pobres (Gabrilovich e Markovitch, 2005). Isto acontece nos casos em que a relação semântica entre objetos não pode ser estabelecida simplesmente pela comparação de termos diretamente entre si. Uma maneira de lidar com este problema é enriquecer a informação do objeto, categorizando-o através de outras fontes de informação (Ribeiro-Neto et al., 2005). Isto significa que os itens que serão fornecidos

[13] http://www.amazon.com

como recomendação não são somente os itens relacionados mas aqueles que são obtidos através de uma categorização dos mesmos. Isto faz com que um utilizador que esteja interessado por exemplo em "flores" seja aconselhado também com tópicos como "rosa" ou "jardim". Existem vários trabalhos realizados nesta área para enriquecimento da informação dos objetos, bem como para a sua estruturação e aplicação em sistemas de recomendação, nomeadamente os trabalhos de (Ziegler et al., 2004) e (Weng et al., 2008).

Sistemas baseados em etiquetagem

Também conhecida como *Folksonomia*, esta técnica faz uso do conceito de etiquetagem de objetos. Este termo foi inicialmente criado por Vander Wal (2004) e pode ser definido como "… *o resultado da etiquetagem, pessoal e livre, de informação e de objetos*…" (Vander Wal, 2007). Ou seja, um determinado utilizador coloca a sua própria informação num determinado objeto, por exemplo um produto, um bem ou um serviço. Esta etiqueta fornece uma breve descrição do conteúdo do objeto. O próprio ato de etiquetar um objeto pode ser feito de uma forma abrangente ou restrita. Na forma mais abrangente a etiquetagem é feita por um conjunto alargado de pessoas que colocam as etiquetas nos objetos, podendo cada uma dessas pessoas colocar as suas próprias etiquetas. Na forma mais restrita, essa etiquetagem é feita por uma pessoa ou por um grupo restrito de pessoas que usam as suas etiquetas para basicamente poderem rastrear novamente os objetos que etiquetaram. Desde 2006 que tem sido uma forma de enriquecer a informação presente em objetos e, tal como a técnica anterior, este enriquecimento de informação irá servir para aplicar técnicas de recomendação com base na informação dos objetos e na informação presente na etiqueta. Ao nível das recomendações, a pesquisa foca-se, essencialmente, em quais as etiquetas a recomendar aos utilizadores e na forma de como o fazer (Sen et al., 2009). A este nível têm aparecido diversas abordagens, como a coocorrência de etiquetas (Li et al., 2008), regras de associação (Heymann et al., 2008) ou as redes de ligação (Au Yeung, 2009).

Capítulo 3

Assinaturas

3.1 Definição e Características

A assinatura de um utilizador WWW pode ser caracterizada por um conjunto de atributos ou características que o permitem identificar perante um dado sítio. Pode ser algo tão único como o seu *login* de entrada ou como algo tão genérico como informações de contexto que de alguma forma a ele estejam ligadas. Informação de contexto é tudo aquilo que, direta ou indiretamente, se poderá atribuir ou associar a um utilizador e que deriva da sua atividade num dado sítio. Exemplos típicos dessa atividade de utilizador poderá ser a hora em que um utilizador normalmente estabelece ligação a um sítio, o tempo máximo, mínimo e médio das sessões que realiza, o tipo de conteúdos que costuma consultar ou as pesquisas efetuadas (Maty´aˇs e Cvrˇcek, 2004).

As assinaturas têm sido utilizadas como um conceito chave para responder a novas necessidades de análise de informação que são feitas sobre os sítios. Juntamente com alguma informação estatística, mais resumida, do perfil ou comportamento de um utilizador, as assinaturas fornecem-nos elementos bastante úteis para que possamos de uma forma mais eficaz tentar prever o seu comportamento futuro.

O que deveremos considerar como característico de uma assinatura tem também muito a ver com a aplicação que queremos dar à mineração de dados em si e aos padrões que pretendemos analisar. A nível de deteção de intrusão os perfis padrão de utilização podem ser os comportamentos normais de utilização de um sítio e qualquer desvio desse padrão pode indiciar uma intrusão ao sistema (Yeung et al., 2002). Ao nível da deteção de casos de fraude bancária podemos considerar um padrão que incorpore informação como a hora de acesso, a data e o tipo de transação. Qualquer possível desvio significativo nos valores desse mesmo padrão de utilização pode ser um indício de fraude (Cahill et al., 2000). A assinatura de um ataque informático do tipo *denial-of-service* poderá ser definida como um número elevado e quase consecutivo de conexões TCP, de um ou vários endereços IP, sem o correspondente reconhecimento do servidor. Numa análise de deteção de fraude aplicada a clientes de telecomunicações, a assinatura pode corresponder a um "vetor" de variáveis, cujos valores são obtidos num determinado período de tempo. A escolha das variáveis da assinatura poderá ser influenciada por diversos fatores como, por exemplo, a sua complexidade, os dados disponíveis ou o grau de computação necessário para efetuar o seu cálculo (Ferreira et al. 2006).

A definição que damos à assinatura é relevante para a comparação que será feita a nível dos padrões encontrados como resultado da mineração efetuado, para, se aplicável, fazer um enquadramento do utilizador e selecionar a personalização ou tratamento que a ele será aplicado.

Recentemente têm também surgido outras definições e tipos de assinaturas, utilizando esquemas criptográficos que usando os atributos de um utilizador permitem outras funcionalidades adicionais para serem utilizadas ao nível da segurança. As assinaturas baseadas em atributos estendem a assinatura de um utilizador baseada na sua identidade (Shamir, 1984). Um utilizador passa a ser identificado não somente por um conjunto de caracteres que compõem a sua assinatura mas sim por um conjunto de atributos. Esses atributos terão que ser certificados por uma entidade competente, assegurando assim que esse utilizador possui determinado tipo de características que poderão ser utilizadas num qualquer sistema de verificação ou autenticação. Este tipo de assinatura tem diversas utilizações. Uma delas é a possibilidade de autenticação anónima, um utilizador não tem que se revelar perante um qualquer sistema de autenticação. O sistema terá somente de verificar se o utilizador tem as características necessárias para a ele aceder. Outra possibilidade de uso de uma assinatura é ao nível da utilização restrita de serviços, quaisquer que

eles sejam, em que pode ser necessário restringir o acesso a utilizadores que tenham somente um determinado tipo de características próprias.

Uma variante das assinaturas baseadas em atributos são os esquemas de assinaturas de grupo baseadas em atributos (Khader, 2007). Um esquema de assinatura em grupo é um tipo de assinatura em que a certificação da identidade de um membro de um grupo pode ser invocada por algo ou alguém, tendo a assinatura desse membro que responder perante um determinado conjunto de características (Delerablée e Pointcheval 2006). Contudo, a identidade desse membro é sempre mantida em segredo e não é revelada perante o grupo. Aliás, o próprio grupo tem também uma assinatura própria que pode ser usado para determinado membro ver a sua filiação, que é revogada se as características presentes na sua assinatura não estiverem em conformidade com a norma do grupo.

Num trabalho recente (Phua et al., 2010), que recaiu sobre casos de estudos efetuados a nível de deteção de fraude, foi efetuada uma análise relativamente ao tamanho dos conjuntos de dados usados para comparação de uma assinatura, bem como no número de atributos que compõem a essa mesma assinatura. Genericamente, os atributos podem ser classificados como binários, numéricos (escalares ou intervalo de valores), categóricos (nominais ou escalas ordinais) ou uma sua combinação.

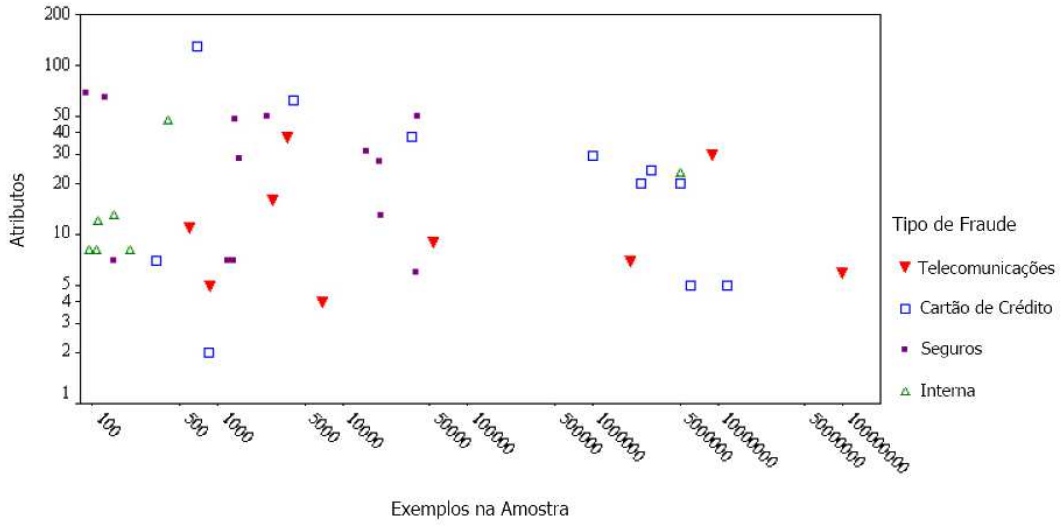


Figura 3.1: Gráfico de um estudo de fraude – figura adaptada de (Phua et al., 2010)

Neste estudo, e em termos de atributos, de uma forma resumida, temos uma situação em que 16 conjuntos de dados têm menos de 10 atributos, 18 têm entre 10 e 49 atributos, 5 entre 50 e 99 atributos e somente 1 desses conjuntos de dados tem mais do que 100 atributos. Adicionalmente, e relativamente aos tamanhos das amostras, vemos que a maior parte das amostras de dados utilizadas têm menos que 50 000 exemplos, sendo que os exemplos para deteções de fraudes internas às próprias empresas têm os tamanhos inferiores. Entre 500 000 e 10 000 000 de exemplos, tamanhos já consideráveis, temos quase exclusivamente as deteções de fraudes relacionadas com cartões de crédito e com telecomunicações. De referir que, este último tipo de fraude teve o maior caso de estudo, uma amostra de dados com 100 milhões de exemplos.

Neste estudo temos ainda um dado interessante, que está apresentado no gráfico da figura 3.2.

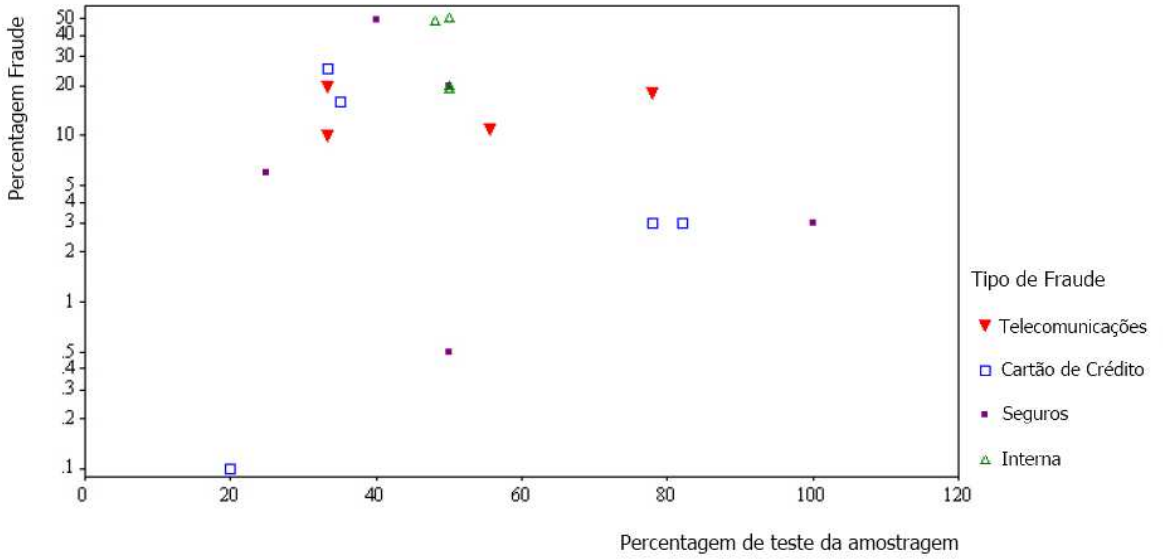


Figura 3.2: Gráfico de um estudo de fraude-percentagem deteção fraude (Phua et al., 2010)

Esse caso revela a percentagem de deteção de fraude, relativamente à percentagem dos exemplos testados nas amostras. Vemos aqui que, em média, foram somente testados 50% dos exemplos das amostras, sendo que, maioritariamente, a percentagem de fraude detetada foi de 30%, para os exemplos testados. O que este estudo nos permite obter é uma perspetiva prática das variantes a considerar para a definição daquilo que poderá ser considerada a nossa assinatura, tendo em consideração o funcionamento do nosso sistema e o equilíbrio entre a sua eficácia e *performance* de funcionamento.

3.2 Arquitetura e Modelos de Implementação

Tendo nesta altura a caracterização de uma assinatura, passamos de seguida a definir uma possível arquitetura computacional para acolhimento da sua implementação (figura 3.3). Nessa figura podemos identificar um módulo de processamento de assinaturas. Este módulo recebe pedidos do sistema central no qual está integrado – que pode ser por exemplo um sistema para deteção de fraude ou no caso deste trabalho um sistema para deteção de alterações/criação de perfis. O módulo pode ser representado como uma caixa negra na qual entram os pedidos do sistema central que, genericamente, são solicitações de execução de procedimentos de verificação ou cálculos sobre uma determinada assinatura presente no sistema de processamento de assinaturas. Como resultado dessa atuação o sistema central recebe a correspondente resposta. Depois, existem algumas tarefas que são executadas dentro do próprio módulo. Primeiro o carregamento das assinaturas presentes no sistema, pois será sobre elas que serão executados os procedimentos pedidos. De seguida, é realizado o cálculo de uma nova assinatura para o instante em que o pedido foi feito, que será confrontada com as assinaturas já existentes. O resultado desta confrontação será a resposta que o sistema dará. Finalmente proceder-se-á à salvaguarda da nova assinatura, fazendo-se a atualização da base de dados de assinaturas.

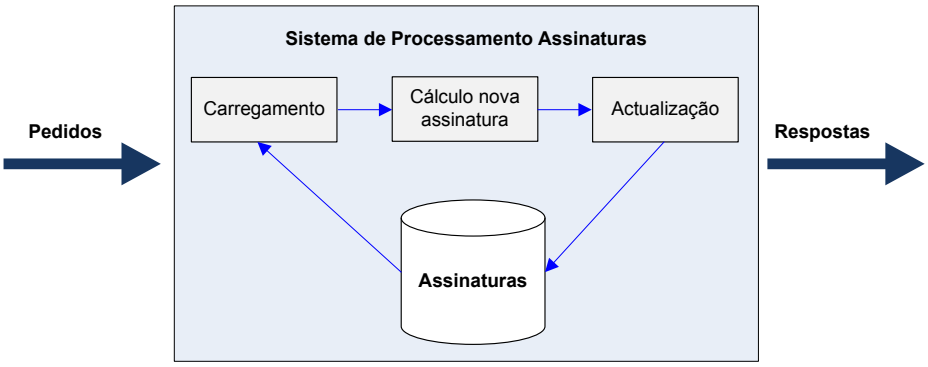


Figura 3.3: Uma arquitetura genérica para um sistema baseado em assinaturas

A forma como as assinaturas são usadas estabelece a forma de implementação do sistema de processamento de assinaturas. Cortes e Pregibon (2001), a nível de sistemas de deteção de fraudes, descreveram duas possíveis formas distintas de implementação que podem ser transpostas para outras aplicações, nomeadamente:

- *A deteção baseada em perfis.* O que é guardado na base de dados de assinaturas são os comportamentos e padrões anormais ou fraudulentos detetados ao longo do tempo. As assinaturas são calculadas quando são recebidos novos dados e existe uma confrontação com as assinaturas guardadas. Se existir uma correspondência então estaremos perante um padrão anormal ou fraudulento.
- *A deteção baseada em anomalias***.** As assinaturas ao longo do tempo são guardadas e constituem a base para a comparação ou deteção que o sistema irá efetuar. Quando os novos dados são recebidos uma nova assinatura é calculada. Se essa assinatura se desviar de uma forma significativa das assinaturas guardadas no sistema, então o módulo de processamento de assinaturas irá alertar para esse mesmo desvio, tomando o sistema as medidas adequadas.

A deteção baseada em perfis pressupõe a presença de assinaturas de perfis conhecidas, ou seja, terá que existir uma correta caracterização das assinaturas logo *a priori* para que o sistema tenha o comportamento desejado. Isto pode ser visto, em analogia, como o conjunto de treino de métodos supervisionados. Em sistemas para deteção de fraudes (Edge e Sampaio, 2009), esta implementação pode ser vantajosa em cenários em que a assinatura de casos fraudulentos seja bem conhecida e possa ser bem caracterizada, mantendo uma base de dados de casos que possa ser usada para os sistemas de deteção efetuarem as devidas comparações. Contudo em cenários em que novas formas de fraude estejam constantemente a surgir, a (necessária) adaptação do sistema poderá não ser feita em tempo adequado e, nestes casos, a deteção baseada em anomalias poderá fazer mais sentido, pois logo à partida não existe uma descriminação de dados.

Ainda relativamente aos modelos de implementação, os sistemas baseados em assinaturas podem ser categorizados em dois modelos de processamento distintos, consoante a sua forma e altura de atuação. De referir (Edge e Sampaio, 2009):

- *Os sistemas orientados ao tempo*. Estes são sistemas que guardam informação ao longo de vários períodos de tempo, que podem ser horas, dias ou meses, e de seguida fazem o processamento dessa informação através do seu módulo de assinaturas. Este processamento é feito *a posteriori*, ou seja o sistema reage de uma forma reativa à informação que chega ao sistema, com uma latência correspondente ao intervalo definido como o período de tempo a considerar.

- *Os sistemas orientados ao evento***.** Estes sistemas reagem imediatamente à chegada de nova informação. Os dados são examinados e comparados com a informação armazenada em tempo real, o que permite, por exemplo, abortar uma transação antes que ela seja completada. Estes sistemas são por isso proactivos, dado que podem atuar antes que uma determinada ação seja executa no sistema.

O uso de sistemas orientados a eventos seria sempre o cenário mais desejável, uma vez que permite uma resposta mais imediata para a deteção e para a análise de assinaturas. Porém esse nem sempre é o cenário possível de implementação. Nos trabalhos realizados por Cortes e Pregibon (2001) e também por Ferreira et al. (2006) é demonstrado que existem condicionantes externas ao próprio sistema de processamento de assinaturas que terão que ser levadas em consideração, como o tempo de carregamento e processamento das assinaturas por parte do sistema. Isto poderá levar a que a abordagem mais equilibrada seja a orientação por intervalos temporárias de processamento. Assim será evitada uma carga excessiva no sistema e tempos de respostas não compatíveis com os próprios requisitos da solução.

3.3 Anomalias

Uma anomalia pode ser definida como um desvio de uma regra comum, tipo ou forma. As anomalias são padrões que se encontram em dados que não são conformes com uma definida noção de comportamento normal (Chandola et al., 2007). Por exemplo, relativamente à análise de tráfego de uma rede, uma anomalia é um desvio significativo do tráfego normal de rede esperado (Kopka et al., 2010). A figura 3.4 apresenta um caso simples de uma anomalia.

Na figura 3.4 vemos a representação de duas áreas, os conjuntos de dados $S1$ e $S2$ que contêm dados considerados normais, isto porque grande parte dos dados se localiza nestas áreas. Os pontos $X1$ e $X2$ representam anomalias dado que se afastam significativamente destas duas áreas. O conjunto $X3$ também representa uma anomalia apesar de conter mais que uma instância de dados. Os dados são sempre parâmetro de entrada em todos os sistemas de deteção de anomalias. As anomalias podem ser categorizadas em três tipos, nomeadamente:

- *Anomalias pontuais*, que acontecem quando uma determinada instância de dados é considerada anómala relativamente aos restantes dados. Se isso acontecer então essa

instância é considerada uma anomalia pontual. Na figura 3.4 vemos que os pontos $X1$ e $X2$ representam anomalias pontuais. Se pensarmos num exemplo no domínio da deteção de fraude bancária, e numa análise simplista feita ao valor transacionado, se houver uma transação em que o valor é muito alto relativamente ao intervalo das restantes transações, essa será uma anomalia pontual.

- *Anomalias contextuais*, que ocorrem se uma determinada instância de dados é anómala mas somente num determinado contexto. São também conhecidas como anomalias condicionais (Song et al., 2007). A noção de contexto é dada pela estrutura dos dados em si, pois existe nessa estrutura dois conjuntos de atributos: os atributos de contexto e os atributos de comportamento. Os atributos de contexto são usados para determinar o contexto de cada registo, como por exemplo atributos de georreferenciação em que a latitude e a longitude dão um contexto de localização. Os atributos de comportamento definem as características não contextuais dos registos, por exemplo, em dados que contenham informação sobre a pluviosidade - a quantidade de chuva que cai num determinado local é um atributo comportamental. Um exemplo de uma anomalia contextual no caso de fraude bancária pode ser o período de compra. Durante o ano um consumidor pode ter um padrão de compra regular, mas no período de Natal esse padrão muda abruptamente. Mas no contexto do período de Natal é considerado um padrão normal, fora do período de Natal é que já não o será.
- *Anomalias coletivas*, que é uma anomalia que ocorre quando um determinado conjunto de registos de dados estiver presente, verificando uma determinada condição. Essa situação será considerada anómala, mas os registos, por si só, isoladamente não o serão. Poderemos também considerar uma sequência de registos cardíacos. Uma determinada leitura isolada de uma leitura cardíaca baixa pode não ser anómala, mas muitos registos de leituras cardíacas baixas seguidas será de certeza uma situação anómala. Outro exemplo poderá ser a deteção de intrusão. Mais uma vez, uma determinada sequência de operações por parte de um utilizador pode caraterizar uma tentativa de intrusão, sendo que essas operações analisadas separadamente não serão consideradas operações fora do comum.

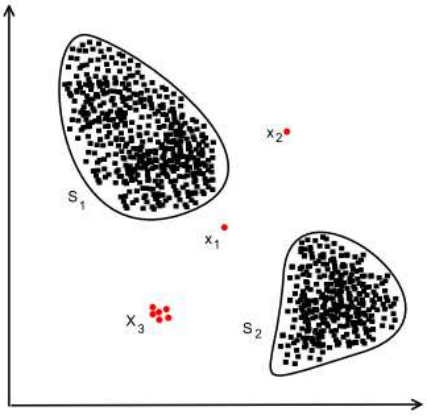


Figura 3.4: Anomalia: um exemplo simples – figura adaptada de Kopka et al. (2010)

Ao fazer a deteção de anomalias fará com que haja uma categorização dos dados, classificando-os como sendo normais ou anómalos. Esta categorização, também chamada de etiquetagem de dados fará com que haja uma etiqueta associada a cada instância de dados, precisamente para fazer a essa distinção a nível de dados. As técnicas de deteção de anomalias podem ser divididas em três tipos:

- *Deteção supervisionada.* As técnicas supervisionadas assumem que existe um conjunto de treino que tenha dados já categorizados e etiquetados, ou seja dados que sejam considerados normais e dados anómalos. Uma abordagem clássica é a construção de um modelo preditivo para as classes normal e anomalia. Novas instâncias de dados que cheguem ao sistema são comparadas com o modelo para determinar a que classes pertencem esses novos dados. Existem contudo duas questões relativamente à deteção supervisionada de anomalias. A primeira é que as instâncias de dados anómalas serão previsivelmente muito inferiores às classes de dados normais, o que fará com que não haja um balanceamento a nível de classes de dados. A segunda questão está relacionada com a correta e precisa etiquetagem de dados normais/anómalos que se pode revelar complicada, especialmente para a classe anomalia. Algumas técnicas propõem a injeção de anomalias artificiais em dados normais de modo a obter conjuntos de dados etiquetados (Steinwart et al., 2005).
- *Deteção semi-supervisionada*. As técnicas que operam em modo semi-supervisionado assumem que somente existe um conjunto de dados de treino para a classe normal. Como não precisam da etiquetagem de dados para as anomalias, têm um espectro de atuação

maior que as técnicas supervisionadas. Nestas técnicas a abordagem clássica é a construção de um modelo para a classe normal e a utilização do modelo para identificar anomalias nos dados a analisar. O uso da classe anomalia em deteção semi-supervisionada não é muito normal dado ser mais difícil a obtenção de um conjunto de dados de treino que contemple as anomalias.

- *Deteção não supervisionada*. Na deteção não supervisionada não existem dados de treino, o que torna a sua área de aplicação ainda mais alargada. Estas técnicas partem do pressuposto que os registos de dados normais são em maior número que os registos anómalos, se isto não estiver correto então o número de falsos positivos será elevado. Por vezes existe uma adaptação das técnicas semi-supervisionadas às técnicas não supervisionadas, usando um conjunto de dados de treino que não é etiquetado. Esta adaptação pressupõe que o número de anomalias nesse conjunto de dados de treino é baixo.

O resultado da aplicação das técnicas de deteção de anomalias aos dados que queremos analisar pode ser geralmente dividido em dois tipos, a atribuição de uma pontuação ou de uma etiqueta. Quando existe a atribuição de uma pontuação esta atribuição é feita a cada instância de dados e descreve o grau em que essa mesma instância é considerada uma anomalia. Dessa forma, o resultado final será uma lista ordenada de anomalias em que o analista poderá inclusive escolher analisar somente as anomalias que estejam no topo dessa lista. Quando o resultado é a atribuição de uma etiqueta ao registo que queremos analisar, então essa etiqueta irá classificá-lo como anómalo ou normal, atribuindo-lhe portanto um valor binário simples, não dando muita versatilidade de análise ao analista que fará a verificação de dados.

3.4 Indicadores de Desempenho

O resultado da confrontação do cálculo de variação das assinaturas irá ser sujeito a análise, seja indicando uma anomalia, como visto anteriormente, seja um outro qualquer resultado de análise produzido pelo sistema de processamento de assinaturas. O resultado dessa análise irá verificar a veracidade do sinal produzido pelo sistema, isto é, se esse sinal foi ou não bem produzido.

Existem custos associados à produção de um falso sinal pelo sistema de processamento de assinaturas, que poderá ser um custo implícito ou explícito. No caso das deteções de fraude pode

ser feita uma análise dos custos monetários reais envolvidos nas deteções efetuadas, através de uma análise explícita dos custos entrando em consideração com os custos de não deteções ou de alarmes falsos (Phua et al., 2004), ou recorrendo-se a uma análise dos benefícios obtidos com o modelo implementado (Wang et al., 2003), como por exemplo a eficiência ou a facilidade de utilização.

Os estudos relacionados com a deteção de fraude que usam algoritmos supervisionados, desde 2001, têm abandonado gradualmente os indicadores de desempenho como os verdadeiros positivos (divisão das fraudes detetadas pela totalidade das fraudes) ou a precisão com medida definida (número de instâncias previstas corretamente divididas pelo total de instâncias). Na deteção da fraude os custos de uma má classificação (falsos positivos e falsos negativos) podem ser desiguais e incertos diferindo de caso para caso podendo ainda variar ao longo do tempo, sendo que um falso negativo tem geralmente custos superiores a um falso positivo. Existem, contudo, estudos de fraude a nível de transações com cartões de créditos (Chen et al., 2004) e em telecomunicações (Kim et al. 2003) que têm ainda como objetivo a melhoria da precisão. Todavia, existem outros que fazem também uso da análise ROC (*Receiver Operating Characteristic*) para analisar a taxa de verdadeiros positivos com a taxa de falsos positivos. Para outras abordagens, como as semi-supervisionadas, que existem por exemplo na deteção de anomalias (Lee e Xiang, 2001) são propostos indicadores como a entropia, entropia condicional, entropia condicional relativa ou ganho de informação. Para algoritmos não supervisionados para deteção de valores extremos em companhias de seguros (Yamanishi et al., 2004) foram usadas as distâncias de Hellinger e análises logarítmicas. A distância Hellinger foi também utilizada para deteção de diferenças de perfis de comportamento de curto e longo prazo em dados de telecomunicações móveis (Burge e Shawe-Taylor, 2001]). Em (Bolton e Hand, 2001), para análise de fraude em contas de crédito, é recomendado o uso da estatística t como um valor para calcular a distância normalizada do valor da conta em análise para um determinado centróide de grupo de pares, bem como para deteção de grandes variações de gastos em contas.

Existem contudo outros indicadores que deverão ser levados em consideração. Para deteção de fraude a rapidez com que uma fraude é detetada e o respetivo alarme é emitido são também fatores bastante importantes. Outo fator importante pode também ser a quantidade de tipos de fraude que se consegue abarcar com o algoritmo de deteção, sendo que um maior espectro poderá implicar um maior nível de deteção. A forma como a deteção é feita é um outro fator, se

essa deteção é feita em tempo real ou em modo *offline* e ainda se é um sistema orientado ao tempo ou um sistema orientado ao evento (Ghosh e Reilly, 1994). Todas as problemáticas e envolvências específicas a cada área de aplicação e a cada caso de estudo deverão ser levadas em consideração na escolha e aplicação de indicadores de desempenho do sistema, pois essas características particulares de cada problema poderão influenciar os resultados que são obtidos.

3.5 Formas de atualização

A atualização das assinaturas presentes no sistema de processamento de assinaturas é também um passo essencial para o correto funcionamento do sistema. É com a incorporação de nova informação acerca das assinaturas que o sistema se tornará mais robusto e preciso na avaliação de assinaturas que precisa de efetuar. Existe contudo a necessidade de saber quando efetuar novas atualizações ao seu historial de assinaturas, tornando os registos mais antigos obsoletos e substituindo-os por nova informação. No trabalho de Ferreira et al. (2006), num sistema de assinaturas ligado às telecomunicações, é demonstrado que isto pode ser conseguido recorrendo a uma fórmula que determina a taxa com os dados são envelhecidos em cada reavaliação do valor da assinatura. Essa fórmula é a seguinte:

$$S_{t+1} = \beta . S_t + (1 - \beta) . P_c$$

O valor de β é uma constante regulada pelo analista no sistema de assinaturas S para um novo conjunto de assinaturas processadas P_c . Este valor ajusta o efeito de novas atualizações sobre o valor atual da assinatura e a devida adaptação da atualização da assinatura com base no modelo de processamento adotado. Em sistemas orientados ao tempo o peso deste fator é geralmente uma constante que determina a taxa com que os dados mais antigos deixam de se tornar relevantes para o sistema. Em (Cortes e Pregibon, 2001) é demonstrando que um valor de 0,85 para β pode ser usado para eliminar dados anteriores a 30 dias e um valor de β de 0,5 elimina dados anteriores a 7 dias. Em sistemas orientados ao evento, β é normalmente uma função da periodicidade de chegada de novos registos. Isto porque apesar de ser orientado ao evento, o valor da assinatura pode-se manter durante um determinado período de tempo que pode ser por exemplo uma semana ou um mês. Geralmente esta não atualização minimiza a falsa deteção de anomalias e assegura que a atualização da assinatura é proporcional à chegada de novos dados.

Em sistemas mais complexos, como por exemplo os sistemas para deteção de fraude, existe mais uma variável a ter em consideração, que é o facto de lidar com transações ambíguas. O sistema terá que decidir se atualiza o valor da assinatura em todas as transações ou se só atualiza o valor da assinatura em transações legitimas, que são sejam etiquetadas como fraudulentas. Nos trabalhos de Cahill et al. (2000) e de Cortes e Pregibon (2001) é defendido que o valor da assinatura só deverá ser revisto com base em dados legítimos, remetendo as fraudes para análise por quem de direito. Em contraste, no trabalho de Ferreira et al. (2006) é referido que todos os dados recebidos, incluindo possíveis fraudes, deverão ser usados para atualizar o valor da assinatura, isto porque poderão estar a se postos de parte registos que são falsos positivos e assim não existe perda de informação que se poderá revelar importante.

A forma como atualizamos a nossa base de dados de assinaturas e a periodicidade com que o fazemos, depende ainda de outros fatores importantes, como a capacidade de processamento do nosso sistema, os objetivos a que nos propomos e a janela temporal que temos para fazer essa atualização. Sistemas com uma capacidade de processamento mais baixa não poderão efetuar uma atualização muito intensiva dos valores das suas assinaturas, sob pena do sistema pura e simplesmente não conseguir responder às exigências. Os objetivos postos ao sistema de processamento de assinaturas também influenciarão, pois se estes objetivos exigirem que os valores das assinaturas não podem estar muito desatualizados então a taxa de atualização terá que ser mais elevada. A janela temporal varia um pouco com os objetivos que temos e com a nossa capacidade de processamento, tendo que contrabalançar um pouco estes dois fatores fazendo com que os objetivos sejam cumpridos sem influenciar um desempenho adequado do sistema de processamento de assinaturas em si.

3.6 Alguns Domínios de aplicação

Um dos domínios de aplicação mais usuais da mineração de dados e no uso das assinaturas é a deteção de fraude. A nível da própria deteção de fraude existem ainda diferentes áreas de aplicação como a fraude financeira (Edge e Sampaio, 2009) ou a fraude em sistemas de telecomunicação (Ferreira et al., 2007). Contudo estas aplicações têm também relação com outros domínios de aplicação, menos conhecidos, onde estas técnicas poderão ser aplicadas. De seguida apresentamos algumas dessas áreas.

Terrorismo

Sistemas de vigilância para deteção de ações de terrorismo, bioterrorismo ou terrorismo químico, frequentemente dependem de dados com características geográficas e temporais. Aqui, costumam aplicar-se técnicas não supervisionadas que também são aplicadas a nível de deteção de fraude. Por exemplo, em (Neill e Moore, 2004) foram aplicadas estatísticas temporais e outras técnicas para a deteção de aglomerações de pacientes em departamentos de emergência, com base em dados de venda de medicamentos da tosse e constipação. Por seu lado, a deteção a nível de bioterrorismo procura encontrar irregularidades em dados temporais, em processos semelhantes e existentes, por exemplo, na deteção de fraude, em que os dados têm que ser parcialmente simulados, injetando dados artificiais de epidemias. Depois é feita uma avaliação de performance a nível de tempo de deteção de falsos positivos. No trabalho realizado por (Wong et al., 2003) foram aplicadas redes bayesianas para descobrir ataques simulados de antraz em dados de departamentos de emergência.

Crime financeiro

Os crimes financeiros, aparte das mais conhecidas atividades de deteção de fraude bancária, incluem lavagem de dinheiro, negócios bolsistas ilícitos ou a captação indevida de informação privilegiada. Neste domínio podemos citar nos Estados Unidos o trabalho de Senator et al., (1999) sobre o sistema FAIS (*The Financial Crimes Enforcement Network AI System*) que opera sobre um sistema bayesiano de inferência e que dá como resultado uma classificação de atos suspeitos e respetivas ligações para análise e visualização dos respetivos suspeitos e contas associadas. No trabalho de Kirkland et al. (1999), sobre uma organização bolsista, a NASD (*National Association of Securities Dealers*), que em 2007 se fundiu com a comissão de regulação do mercado bolsista norte americano para formar a FINRA (*Financial Industry Regulatory Authority*) foi apresentado um sistema que tenta encontrar semelhanças de padrões de regras com comportamentos suspeitos pré-definidos, enquanto que um outro motor tenta encontrar sequências de eventos e comportamentos semelhantes ao longo do tempo em dados do mercado bolsista que possam revelar padrões anómalos. Regras de associação e árvores de decisão são usadas para descobrir novos padrões ou novas regras para adaptação do sistema às mudanças comportamentais do próprio mercado. O sistema SONAR (*Securities Observation, News Analysis & Regulation System*), desenvolvido também pela NASD (Goldberg et al., 2003), usa mineração de texto, regressão estatística, inferência de regras e outras técnicas para encontrar relações implícitas e explicitas entre as entidades e eventos. Isto para formar cenários ou episódios com identificadores

específicos para análise. O SONAR começou a operar em 2001 no índice norte-americano NASDAQ (*National Association of Securities Dealers Automated Quotations*), analisando mais de 10 000 notícias diárias e a evolução de mais de 25 000 cotada, gerando cerca de 50 a 60 alertas diários para análise, tendo sido reportados casos de sucessos na deteção de uso de informação privilegiada no mercado bolsita e falsificação de notícias. A redução do tempo de deteção com o uso, por exemplo, de novas fontes relevantes de informação, é outra área de estudo complementar aos estudos existentes a nível de deteção de fraude. Em (Donoho, 2004) foi explorado o uso de um algoritmo de árvores de decisão, redes neuronais, algoritmos de agrupamento para instâncias positivas entre outras técnicas para procura de indícios de negociações bolsistas, com base em informação privilegiadas antes mesmo das noticias serem divulgadas.

Sistemas de deteção de intrusão e *spam*

Usualmente, uma intrusão numa rede de computadores é um ataque ou atividade maliciosa que possa comprometer a estabilidade e segurança do próprio ambiente de rede. Existem várias fontes de informação que podem ser usadas nestes sistemas. Em particular, em (Otey et al, 2003) é defendida a deteção da ações de intrusão a nível dos próprios NIC (*Network Interface Card*) de rede. Por sua vez, Leckie e Yasinsac (2004) fazem uso de informação presente nos metadados de utilizadores para a deteção de anomalias em redes seguras. Em comparação com a deteção de fraude existem muitos estudos que fazem uso de grupos de dados reais para análise, sendo que um dos mais utilizados são os dados presentes no dados de intrusão do *benchmark KDD cup 1999*. Em complemento existe também o uso de dados semirreais em que as intrusões são simuladas recorrendo por vezes a dados de outros utilizadores.

Não existem, pois, garantias que métodos aplicados numa determinada área possam ser aplicados diretamente num outro domínio completamente diferente com sucesso. Contudo estas aplicações em áreas tão distintas como as de contra terrorismo ou dos sistemas de deteção de intrusão, podem de facto contribuir para avanços na utilização de técnicas diferentes. O mesmo pensamento poder-se-á aplicar ao uso de técnicas baseadas em assinaturas em diferentes domínios de conhecimento.

Capítulo 4

O Caso de Estudo

4.1 Apresentação Geral

Após a análise das técnicas existentes para recolha, armazenamento e tratamento de toda a informação necessária para a implementação de análise de perfis de utilizadores baseada em assinaturas passamos à parte de aplicar essas mesmas técnicas num caso de estudo concreto, sobre um sítio alvo. Antes de avançar para a aplicação de qualquer tipo de técnica, é necessário estudar profundamente o funcionamento do sítio alvo e de todo o tipo de informação envolvida de modo a fazer o correto enquadramento e definição do plano de trabalho da mineração de dados, pois, como sabemos, cada caso de estudo tem as suas particularidades. Para este trabalho de dissertação, a escolha do caso de estudo recaiu num sistema de *e-learning*. Os sistemas de *e-learning*, sistemas eletrónicos de aprendizagem, podem ser definidos como meios de instrução, instrução essa que é fornecida sobre meios digitais, tais como os computadores ou dispositivos móveis (Clark e Mayer, 2007).

No caso específico da Universidade do Minho, o sítio de *e-learning* sofreu, em Julho de 2012, uma atualização que permitiu o seu acesso através de dispositivos móveis, além das plataformas de acesso que até então já eram suportadas. Este sistema é utilizado por docentes e estudantes para partilha de informação relacionada com as diversas unidades curriculares que são lecionadas na

universidade. O caso de estudo selecionado para este trabalho de análise de perfis baseados em assinaturas é, assim, o sítio de *e-learning* da Universidade do Minho[14]. Este sítio assenta na plataforma de *e-learning* BlackBoard™[15] e pretende ser o mecanismo preferencial para a partilha de toda a informação relacionada com as unidades curriculares referentes aos cursos que são lecionados na Universidade do Minho[16].

Figura 4.1: Página de entrada do sítio de *e-learning* da Universidade do Minho

O acesso a este sítio está restringido a utilizadores devidamente credenciados. O registo de utilizadores é um processo controlado pela própria universidade, abrangendo os docentes e alunos dos diversos ciclos de estudo da Universidade do Minho. Como já foi referido, este sítio tem por base o sistema de e-*learning* BlackBoard, por isso inclui todas as funcionalidades inerentes a esta plataforma. Porém, este sistema sofreu diversas personalizações, como a introdução de módulos que foram desenvolvidos pela própria Universidade de modo a introduzir ferramentas adicionais de auxílio orientadas para os processos de aprendizagem específicos dos seus diversos projetos de ensino. No que diz respeito às funcionalidades gerais disponibilizadas pela plataforma, podemos mencionar as zonas de avisos, o calendário, as tarefas, as notas ou o envio de correio eletrónico. Contudo, uma das suas componentes principais prende-se com toda a informação disponibilizada a nível das diversas unidades curriculares inseridas na plataforma. Os utilizadores são registados nas unidades curriculares com dois perfis principais distintos: estudantes ou docentes. Cada um destes perfis tem permissões e funcionalidades específicas a nível de cada unidade curricular.

[14] http://elearning.uminho.pt
[15] http://www.blackboard.com
[16] http://www.uminho.pt

O perfil de um utilizador numa unidade curricular determina as zonas de informação a que este pode aceder e quais as operações que pode efetuar. Algumas dessas zonas e operações incluem o acesso a conteúdos da própria unidade curricular, informações, grupos de discussão ou o dossier pedagógico da unidade curricular. O dossier pedagógico contém, por exemplo, trabalhos e mecanismos de avaliação específicos da cada unidade curricular. Sucintamente e no que diz respeito a permissões de acesso, podemos desde logo definir dois níveis distintos que são atribuídos aos utilizadores do sistema, os perfis base que lhe darão acesso a um conjunto de funcionalidades e operações a nível da plataforma em si e o perfil que lhe é atribuído a nível de uma unidade curricular que determina o seu nível de acesso dentro dessa mesma unidade curricular.

4.2 O Processo de Recolha de Dados

A plataforma de *e-learning* BlackBoard faz uso de vários mecanismos de armazenamento de informação no que diz respeito a acessos e operações efetuados pelos utilizadores da aplicação. Existem dois grandes grupos de *logs* usados pela plataforma:

- *Ficheiros físicos guardados no sistema operativo*. Estes ficheiros de log incluem ficheiros guardados pelo próprio servidor Web, bem como ficheiros específicos da própria aplicação. Nestes ficheiros estão salvaguardados os dados relativos a pedidos feitos pelo cliente ao servidor Web, eventos de erro, exceções, atualizações feitas ao sistema e outras operações específicas feitas a nível da própria plataforma BlackBoard.
- *Armazenamento em bases de dados*. A plataforma BlackBoard utiliza duas bases de dados para armazenar informação que permitem o rastreamento da atividade efetuada na plataforma. Uma delas é uma base de dados de produção que guarda diariamente dados de estatística sumarizada e detalhada com as respetivas tabelas de apoio para referência. Diariamente, parte desta informação de produção é copiada para uma outra base de dados de histórico que irá assim acumular todos os dados relativos à atividade na plataforma ao longo do tempo.

Para o trabalho desta dissertação, e tendo em conta a definição dos atributos da assinatura a que se chegou e que será apresentada posteriormente, os dados da base de histórico foram suficientes pelo que foi esta a única fonte de informação utilizada.

A nível da base de dados de histórico, usada neste trabalho como fonte de informação, uma tabela é primordial para a recolha dos dados pretendidos: a tabela ACTIVITY_ACCUMULATOR. Esta tabela, tal como o próprio nome indica é a tabela principal de armazenamento de toda a atividade da plataforma armazenada em base de dados. O processo de recolha de informação desde a fonte até à base de dados final de armazenamento dos dados relativos à assinatura definida foi executado em duas fases.

- *Primeira fase*. Consistiu na construção de um *data warehouse* para armazenamento dos dados transformados e filtrados da base de dados de histórico do sistema de *e-learning*. Com a criação deste *data warehouse* foi criado também um processo ETL para fazer o povoamento deste *data warehouse*. O objetivo foi a construção de uma fonte de informação intermédia para que posteriormente a extração dos dados necessários para a assinatura fosse executado de uma forma mais eficiente.
- *Segunda fase*. Nesta fase foi realizada a criação de uma segunda estrutura multidimensional num *data warehouse*, que serviu para proceder ao armazenamento da informação relativa aos dados das assinaturas dos utilizadores ao longo do tempo. A tabela de assinaturas tinha como chave a informação do ano, mês, semana, dia e a identificação do utilizador. Os atributos armazenados são os atributos definidos para a assinatura neste caso de estudo, com os respetivos valores para o dia em questão e para os diversos utilizadores. Foi nesta segunda estrutura que foi efetuada a análise da variação temporal das assinaturas. Um segundo processo de ETL foi também criado para extrair e transformar os dados já armazenados no *data warehouse* com a informação de histórico proveniente da fonte.

Para efeitos de amostragem e dado o volume de dados existente na fonte, optou-se por proceder à recolha de dados de apenas um mês de histórico - Janeiro de 2011. Esta amostra envolveu um conjunto de amostra com cerca de 15 milhões de registos na tabela de atividade, para um universo geral na amostragem de mais de 40 mil utilizadores e quase 5 mil unidades curriculares. As bases de dados para armazenamento dos *data warehouse* de informação da fonte e da base de dados de assinatura foram criadas utilizando o Microsoft SQL Server 2008 R2 ®.

4.3 Processo ETL da fonte de dados

O *data warehouse* desenvolvido para armazenamento dos dados integra uma tabela de factos ("TFActivity") e quatro tabelas de dimensão, correspondentes ao tipo de acesso ("AccessType"), operação ("Operation"), dimensão tempo ("DimTime") e dimensão tipo de utilizador ("DimUserType"). Duas delas são degeneradas, o tipo de acesso e operação – figura 4.2.

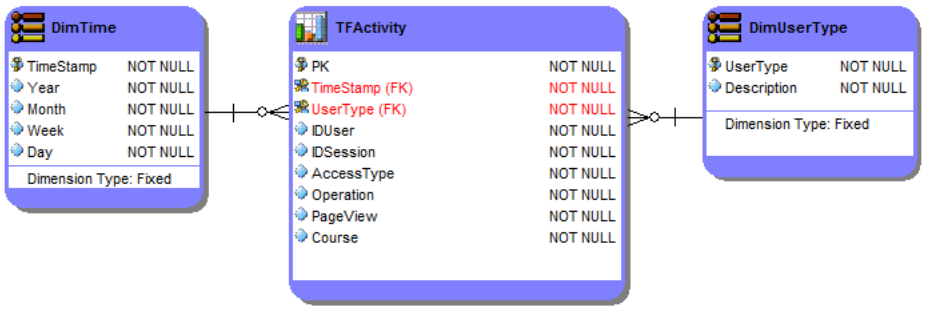


Figura 4.2: Modelo dimensional do *data warehouse* com os registos provenientes da fonte

De seguida, apresenta-se uma breve descrição de todos os atributos dessas tabelas. Refira-se que, a dimensão "tempo" foi organizada em quatro atributos: "Year", "Month", "Week" e "Day".

Atributo	**Descrição**	**Exemplo**
TimeStamp	Data do registo	2011-01-01 00:00:19.270
Year	Ano do registo	2011
Month	Mês do registo	Janeiro
Week	Semana do registo	4
Day	Dia do registo	1

Tabela 4.1: Descrição da tabela de dimensão "Dim Time"

A dimensão relativa ao tipo de utilizador ("DimUserType") contém informação sobre o tipo de utilizador contido na amostra. Os tipos de utilizador alvos de análise serão primordialmente alunos, tipo de utilizador S (*Student*), e docentes, tipo de utilizador P (*Professor*). Existem registos que não contêm nenhum tipo de utilizador, registos nulos na fonte, a que irão corresponder ao tipo de utilizador U (*Unknown)*, utilizadores desconhecidos. Todos os outros registos com outros tipos de utilizador, por não terem relevância para a análise a efetuar, foram agrupados num tipo de utilizador O *(Outros)*.

Atributo	Descrição	Exemplo
UserType	Tipo de utilizador	S/P/U/O
Description	Descrição do tipo de utilizador	Aluno (*Student*) / Docente (*Professor*)/ Desconhecido (*Unknown*) / Outros (*Other*)

Tabela 4.2: Descrição da tabela de dimensão "DimUserType"

A tabela de factos ("TFActivity") desta base de dados irá conter os dados transformados do registo de atividades da tabela de histórico na fonte. Será a base para a construção do registo de assinaturas. Para efeitos de anonimato, a identificação dos utilizadores foi alterada recorrendo a uma chave de substituição.

Atributo	Descrição	Exemplo
PK	Identificador único de registo	129705980
TimeStamp	Data do registo	2011-01-01 00:00:19.270
UserType	Tipo de utilizador	U
IDUser	Chave de substituição do utilizador	47051
IDSession	Identificador da sessão	15253311
AccessType	Tipo de acesso feito pelo utilizador. Dimensão degenerada.	Acesso a Página
Operation	Operação efetuada pelo utilizador. Dimensão degenerada.	Operação sobre conteúdos
PageView	Indica se a operação efetuada constitui uma visualização de página.	Sim/Não
Course	Descrição da unidade curricular acedida	[10-11] Projeto

Tabela 4.3: Descrição da tabela de factos "TFActivity"

Além dos atributos que guardam a data de registo ("TimeStamp"), o tipo de utilizador ("UserType") e a identificação do utilizador ("IDUser"), temos ainda nesta tabela o identificador de sessão ("IDSession") que será usado para por exemplo para determinar tempos de sessão. O atributo "AcessType" está tipificado pela plataforma, correspondendo ao tipo de acesso a que aquele registo diz respeito. Pode corresponder, por exemplo, a um acesso a um módulo, a um acesso a uma página, a um acesso a uma unidade curricular, a um acesso a conteúdo ou a início e fim de sessão. O atributo *"Operation"* irá armazenar informação sobre a operação realizada pelo utilizador, correspondente ao evento de navegação que ele despoletou. Porém, nem todos os registos na tabela de atividade correspondem a um evento de navegação, pelo que este atributo na fonte poderá ter valores nulos. O atributo "PageView" indica se a operação efetuada pelo utilizador corresponde a uma visualização de página, o que será importante a nível de cálculo de estatísticas de acesso e visualização. Por último, temos o atributo "Course" que indica qual a

unidade curricular acedida. Este atributo também pode assumir valores nulos, uma vez que poderão ocorrer acessos que não estejam no âmbito de uma unidade curricular.

Como já foi referido, o *data warehouse* trata do armazenamento dos dados transformados e filtrados provenientes da base de dados de histórico do sistema de *e-learning*. O processo de carregamento de dados visa, essencialmente, colocar no destino os dados novos que entretanto foram surgindo na fonte, para as dimensões que já referimos. O processo de carregamento é parametrizável, tendo que ser fornecido as datas de início e de fim dos registos que pretendemos importar. Para o desenvolvimento deste processo foi utilizado o *Integration Services do Microsoft Visual Studio 2008*, utilizando maioritariamente *querys* SQL diretas sobre a fonte de dados. Na figura 4.3 apresentamos sumariamente as diversas partes deste processo.

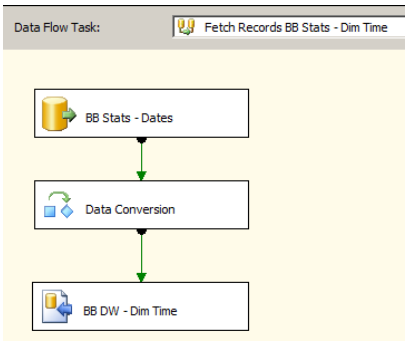


Figura 4.3: Esquema do processo de povoamento da dimensão tempo – "DimTime"

O processo de carregamento da dimensão tempo faz o povoamento da tabela "DimTime", com todas as datas que foram utilizadas nos registos da tabela de factos. Neste processo incorporaram-se várias funções de conversão de dados necessárias para a correção e homogeneização de alguns dados provenientes das fontes. A dimensão tipo de utilizador - "DimUserType" - foi construída *a priori* com os tipos de utilizador que seriam alvos de análise. Assim sendo, não foi incluído o processo de carregamento desta dimensão no sistema de ETL.

UserType	Description
O	Other
P	Professor
S	Student
U	Unknown

Figura 4.4: Conteúdo da tabela de dimensão "DimUserType"

Finalmente, foi tratado do processo de carregamento da tabela de factos "TFActivity". Nas instruções T-SQL (*Transact-SQL*) presentes nesta tarefa são realizadas várias operações de conversão e mapeamento, como sendo a geração das chaves de substituição para os utilizadores, o mapeamento do tipo de acesso e a operação efetuada, bem como a indicação se o tipo de operação feita pelo utilizador corresponde a uma visualização de página.

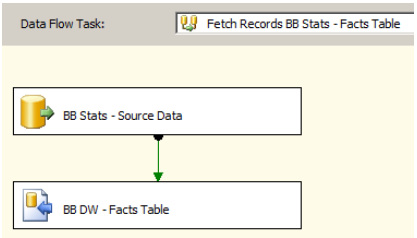


Figura 4.5: Carregamento da tabela de factos

Na sequência de execução do processo de ETL, ilustrada na Figura 4.6, vemos que a primeira tarefa a ser executada é a tarefa de carregamento da dimensão tempo – "DimTime". Caso ocorresse um erro nesta tarefa, o ETL seria terminado, senão seria executada a tarefa de carregamento da tabela de factos – "TFActivity". Como já foi referido a dimensão tipo de utilizador - "DimUserType"- não foi incluída no processo de carregamento ETL, tendo sido previamente carregada. No ETL desenvolvido foram utilizadas conexões diretas através de clientes nativos para o SQL Server e utilizados componentes, como os de *OLE DB* e *SQL Server Destination*, que se revelaram eficazes na diminuição dos tempos de execução dos processos de recolha e transformação dos dados.

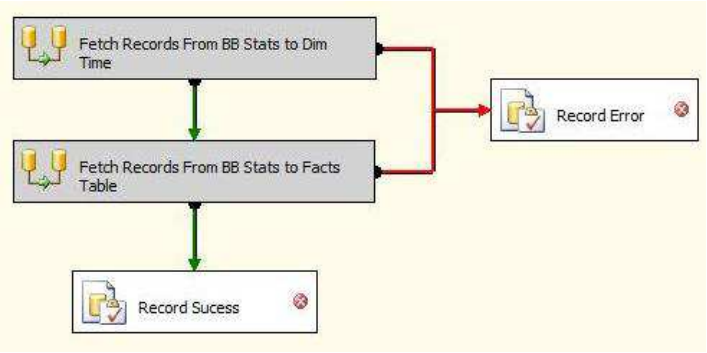


Figura 4.6: Esquema dos processos incluídos no sistema de ETL

4.4 Definição da assinatura a analisar

O conjunto de atributos da assinatura que foi escolhido foi determinado com base no objetivo final da análise que se pretendia obter: determinar variações no padrão de utilização dos utilizadores na plataforma de *e-learning.* Os atributos teriam, assim, que ser sensíveis o suficiente para detetar mudanças no comportamento de utilização dos utilizadores de forma o mais precisa possível. Uma variação no padrão de utilização pode ter à partida dois motivos, mais ou menos evidentes: uma mudança real no perfil desse utilizador ou uma situação anómala ocorrida que possa provocar essa variação de comportamento. A forma de deteção de variações e a sua consequente interpretação seria uma análise a efetuar posteriormente, com algoritmos específicos depois de ter os dados devidamente preparados. Como sabemos, a assinatura de um utilizador deve ser única. As possíveis variações da sua própria assinatura serão guardadas na tabela de assinaturas (figura 4.7) ao longo do tempo. A chave da tabela "TFSignature" é constituída pela identificação temporal da chave: ano, mês, semana e dia e pela identificação do utilizador o que irá permitir, como referido, guardar a evolução da assinatura de um determinado utilizador.

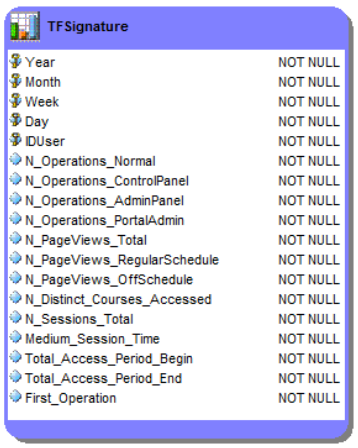


Figura 4.7: Estrutura da tabela de factos "TFSignature"

O atributo "N_Operations_Normal" é um atributo numérico que guarda o número de operações que são consideradas normais. Como operações normais são consideradas todos os eventos internos capturados e guardados pela plataforma e que não tenham nenhuma origem que possa ser considerada de acesso privilegiado, isto é cuja origem não seja uma área reservada a determinados tipos de utilizadores da plataforma. Como exemplo deste tipo de operações temos o acesso aos diversos conteúdos das unidades curriculares ou o acesso de utilização das áreas de discussão ou de anúncios. O atributo "N_Operations_ControlPanel", também um atributo numérico, é aquele que guarda as operações cuja origem seja do painel de controlo de uma unidade curricular. Estes eventos internos são tipificados de forma especial, pelo que podem ser distinguidos dos restantes eventos, o que é uma forma importante de diferenciação. Teoricamente, somente utilizadores considerados como docentes podem despoletar este tipo de eventos. São exemplos deste tipo de eventos os que tenham origem nas áreas de avaliação, de classificações ou de ferramentas, como o envio de correio eletrónico. O atributo "N_Operations_AdminPanel" é também um atributo numérico que armazena os eventos que sejam gerados através do painel de controlo de administração. Entre estes eventos estão, por exemplo, a listagem de utilizadores de uma determinada unidade curricular. Temos também o atributo "N_Operations_PortalAdmin". Este guarda o número de eventos que sejam tipicamente atribuídos a funções de administração do próprio portal. Serão portanto operações executadas por utilizadores que tenham permissões de administração. Operações de gestão e configuração estão entre o leque de eventos que são abrangidos e guardados por este atributo.

Atributo	Descrição
Year	Ano do registo.
Month	Mês do registo.
Week	Semana do registo.
Day	Dia do registo.
IDUser	Identificação do utilizador.
N_Operations_Normal	Número de operações consideradas de utilização normal, sem acesso através do painel de controlo, funções de administração ou operações consideradas de administração do portal.
N_Operations_ControlPanel	Número de operações efetuadas através do painel de controlo de uma unidade curricular.
N_Operations_AdminPanel	Número de operações associadas com funções de administração da plataforma.
N_Operations_PortalAdmin	Número de operações efetuadas através do painel de controlo de administração da plataforma.
N_PageViews_Total	Número total de registos de atividade considerados como visualização de página.
N_PageViews_RegularSchedule	Número total de registos de atividade considerados como visualização de página efetuados em horário normal. O horário normal definido foi o de um horário mais comum de trabalho, das 09:00 às 17:30 de cada dia.
N_PageViews_OffSchedule	Número total de registos de atividade considerados como visualização de página efetuados fora do horário normal definido, das 09:00 às 17:30.
N_Distinct_Courses_Accessed	Número distinto de unidades curriculares que o utilizador acedeu durante o dia.
N_Sessions_Total	Número total de sessões.
Medium_Session_Time	Tempo médio de cada sessão.
Total_Access_Period_Begin	Inicio do período de acesso do utilizador nesse dia.
Total_Access_Period_End	Fim do período de acesso do utilizador nesse dia.
First_Operation	Primeira operação efetuada pelo utilizador.

Tabela 4.4: Descrição dos atributos da tabela de assinaturas "TFSignature"

Na tabela 4.4 também se apresentam os atributos numéricos que guardam a informação relativa à visualização das páginas - *page views*. A base de dados de histórico do sistema de *e-learning* possui um atributo "EVENT_TYPE", tipo de evento, que permite verificar se o evento despoletado corresponde ou não a uma visualização de uma página. Foi este atributo que foi usado para gerar o valor do atributo "PageView" na tabela de factos "TFAactivity" do *data warehouse* que armazenou os dados tratados e transformados da fonte original. Uma visualização de página, por vezes também chamada de *hit*, ocorre num momento específico, quando um navegador apresenta uma determinada página *web*. Todavia, como sabemos, nem todos os eventos despoletados correspondem a uma visualização de página. Nesta tabela figuram três atributos que guardam informação sobre diferentes níveis de visualização das páginas, nomeadamente:

1) "N_PageViews_Total", que corresponde ao total de visualizações efetuadas por um utilizador.

2) "N_PageViews_RegularSchedule", que regista o número de visualizações realizadas num horário considerado normal - 09:00 às 17:30 -, e que permite detetar possíveis variações de padrões de utilização durante o dia para os diversos dias em análise.
3) "N_PageViews_OffSchedule", que, com o mesmo objetivo do atributo anterior, contabiliza o número de visualizações fora do horário considerado como normal. De referir que o número de visualizações total menos o número de visualizações em horário normal terá que ser igual ao número de visualizações fora do horário normal.

De seguida temos o atributo "N_Distinct_Courses_Accessed", que guarda informação sobre o número distinto de unidades curriculares acedidas durante o dia. Num comportamento de um utilizador considerado como normal espera-se que este número não sofra grandes variações e se mantenha estável para esse utilizador, e o atributo "N_Sessions_Total", que guarda informação sobre o número total de sessões, valor este bastante relevante já que permite dar uma perspetiva da quantidade de vezes que um utilizador acede à plataforma durante o dia, já que, teoricamente, cada sessão corresponde a uma visita do utilizador à plataforma de *e-Learning*. Ainda relacionado com as sessões dos utilizadores é guardada também a informação do tempo médio de sessão do utilizador no atributo "Medium_Session_Time", o que nos permite ter uma indicação do tempo médio da duração de todas as sessões do utilizador durante o dia, contabilizando dessa maneira o tempo médio de cada acesso que um dado utilizador teve na plataforma. Na tabela 4.4 também se guarda a informação sobre a mancha total de acesso do utilizador à plataforma durante o dia. Para suportar essa informação temos dois atributos: o "Total_Access_Period_Begin", que guarda o momento do dia em que o utilizador registou o seu primeiro acesso à plataforma e o "Total_Access_Period_End", que regista o último acesso feito pelo utilizador nesse dia. Por último temos o atributo "First_Operation", que indica qual a primeira operação que o utilizador efetuou quando acedeu à plataforma, isto no caso quando existe alguma operação elegível e guardada pela própria plataforma. Este atributo dá uma boa perspetiva daquilo que, normalmente, o utilizador faz em primeiro lugar quando acede à plataforma. O valor deste atributo poderia ser usado para, por exemplo, implementar uma funcionalidade em que fossem disponibilizadas ao utilizador as suas operações mais recentes ou mesmo colocar em destaque as operações que ele mais efetua na plataforma de *e-learning*.

4.5 Carregamento e Atualização dos Dados das Assinaturas

Tendo a definição da assinatura já sido realizada, foi necessário tratar da implementação do processo de carregamento dos dados necessários para a geração e atualização dos dados das assinaturas. O processo de carregamento da informação das assinaturas foi efetuado recorrendo a instruções T-SQL sobre o repositório dos dados provenientes da fonte de informação e construído anteriormente. Foi definida uma única tarefa de recolha e processamento de dados, desde a tabela de factos do *data warehouse* intermédio até ao *data warehouse* de armazenamento dos dados finais das assinaturas dos utilizadores.

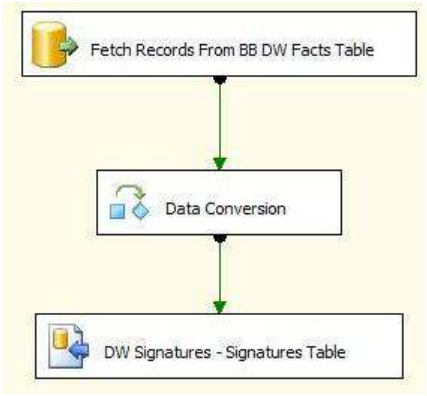


Figura 4.8: Carregamento da tabela de assinaturas

O processo ilustrado na figura 4.8 tem incluído o processo de recolha de dados e o cálculo necessário para determinar o valor final dos atributos que compõem a assinatura definida. Para efeitos da realização deste trabalho, foi definido uma atualização diária para o registo de assinaturas, ou seja, a compilação e cálculo do valor das assinaturas foi feita com um intervalo de um dia. Mas, este intervalo temporal pode ser alterado em processos de trabalho futuros. Para efeitos de flexibilidade do processo de atualização a data de processamento das assinaturas foi colocada como parâmetro do processo ETL.

Mais tarde, aquando do agendamento dos processos de extração, transformação e carga em ambiente produtivo, e para que tudo corra segundo o previsto, teremos que garantir as seguintes condições:

1) O primeiro processo de carregamento dos dados terá que ser executado sobre a base de histórico da plataforma de *e-learning*, que não sendo uma base de dados operacional recebe também ela dados do sistema operacional por um processo interno da plataforma.
2) O segundo processo de carregamento procederá ao preenchimento dos dados necessários para a assinatura que será executado depois, com um intervalo suficiente para a execução do primeiro, tendo também em consideração o processo de extração de informação das assinaturas. Isto porque as assinaturas só fornecem a fonte para que processos analíticos possam ser executados, neste caso processos que detetem alterações de perfis ou que possam gerar perfis baseados nesta informação.

Nesta implementação, a atualização dos dados necessários para as assinaturas foi feita para um determinado intervalo de tempo pré-definido, neste caso um processo diário. No entanto e tal como também já foi referido, tudo iria depender, aquando da entrada em ambiente produtivo, de saber qual a periodicidade de execução dos processos de análise das assinaturas e qual o desfasamento que seria aceitável verificar-se entre os diversos processos de atualização de dados. Tudo isto definirá a janela temporal dos diversos processos de atualização de dados.

4.6 O Cálculo da Variação de uma Assinatura

Tendo a tabela de assinaturas definida e os respetivos dados recolhidos passámos então para a fase de cálculo da variação das assinaturas dos utilizadores. O cálculo dessa variação permite verificar alterações comportamentais dos utilizadores, no caso concreto deste trabalho, o objetivo foi determinar variações nos perfis de utilização da plataforma de *e-learning*. Aparte do processo de deteção em si, esta verificação de alterações de perfis de utilização poderá servir no futuro vários propósitos, como a implementação de alterações dinâmicas a nível de estrutura do portal em si ou mesmo a deteção de acessos indevidos.

A assinatura de um utilizador, tal como já foi referido, depende do objetivo da análise que se pretende efetuar. Para este trabalho trabalhou-se, essencialmente, em duas vertentes, a nível dos atributos da assinatura e na definição de assinatura como um todo, agregando vários atributos. Os atributos escolhidos para análise e presentes na tabela de assinatura, foram os atributos que se conseguiam extrair dos dados da fonte original e que se consideraram relevantes para o processo

de análise de perfis de utilização. Para além desta definição de atributos também teria que se definir mais dois aspetos importantes:

1) Qual a fórmula a utilizar para o cálculo da assinatura de um utilizador, para um determinado dia de análise.
2) Qual a fórmula a utilizar para calcular a variação das assinaturas dos utilizadores ao longo dos diversos dias de análise.

A nível da definição global de assinatura trabalhou-se com os atributos numéricos da tabela de assinatura, atribuindo a cada um deles um peso específico no cálculo da assinatura. Sucintamente, o valor de uma assinatura x corresponde ao somatório do peso w dos seus atributos multiplicado pelo seu respetivo valor v.

$$x = \sum w \times v$$

O valor do peso do atributo w foi determinado pela relevância do atributo na assinatura, tendo sofrido alguns ajustes à medida que foram sendo obtidos os resultados sobre a variação das assinaturas dos utilizadores. Na tabela 4.5 são mostrados os pesos dos diversos atributos numéricos que foram utilizados para o cálculo do valor da assinatura.

Atributo	**Peso (*weight*)**
N_Operations_Normal	5
N_Operations_ControlPanel	100
N_Operations_AdminPanel	100
N_Operations_PortalAdmin	100
N_PageViews_Total	3
N_PageViews_RegularSchedule	3
N_PageViews_OffSchedule	5
N_Distinct_Courses_Accessed	2
N_Sessions_Total	1
Medium_Session_Time	1

Tabela 4.5: Peso dos atributos na assinatura

O peso dos atributos espelha a importância do atributo na assinatura e também a importância que uma variação de valores nesse atributo representa. Por exemplo, uma variação no número de operações de administração é bastante significativa, pois são operações de acesso reservado. Se um utilizador passar a contar na sua assinatura com operações deste género quando normalmente não tem pode indiciar uma clara alteração no perfil de comportamento. Da mesma forma pode ser interpretado que uma variação do número de visualizações de página fora do horário normal tem um significado mais relevante a nível de assinatura do que as visualizações dentro do horário regular pois este último teoricamente será o comportamento normal de um utilizador na plataforma de *e-learning*.

Para além da referida fórmula que permitiu efetuar o cálculo do valor numérico de uma assinatura para um determinado utilizador e para um determinado dia de análise, seria ainda necessário analisar a variação da sua assinatura e verificar se essa variação seria razão suficiente para determinar uma alteração no perfil do utilizador. Para a determinação da fórmula de cálculo para a variação da assinatura foram estudados alguns trabalhos realizados, nomeadamente na área das telecomunicações (Ferreira et al., 2007) e na área de deteção de fraude (Phua et al., 2010).

Neste trabalho, optou-se por utilizar o conceito de média e de desvio padrão. No cálculo da assinatura de um utilizador, para um determinado dia, é feita a comparação com a média e desvio padrão das assinaturas desse utilizador para os dias anteriores ao dia de cálculo da assinatura. Se o valor diário da assinatura não excedesse a sua média e duas vezes o seu desvio padrão, calculados com base nos dados das assinaturas dos dias anteriores então essa variação, seria considerada normal, senão seria considerada uma variação anómala e um alerta seria emitido. Para isso utilizámos a seguinte fórmula:

$$\bar{x} - 2 \times S < x < \bar{x} + 2 \times S$$

Em que x representa o valor da assinatura do dia, $\bar{x}$ a média do valores das assinaturas anteriores e S o desvio padrão dos valores das assinaturas anteriores.

No entanto, saber se um determinado valor anómalo na assinatura para um determinado instante pode indiciar uma mudança no perfil de um utilizador foi outro ponto que teve que ser ponderado

e esta foi também uma questão sensível de analisar. O facto do valor de uma assinatura sair do intervalo definido indica que um alerta tem que ser emitido e uma análise ao porquê desta variação tem que ser feita, mas não havia forma de saber se a distância do valor da assinatura relativamente aos valores limites definidos é alta ou baixa. Essa informação seria útil para priorizar a análise das situações anómalas detetadas. Se a variação do valor da assinatura fosse muito alta então a probabilidade de essa variação indiciar uma alteração de perfil seria maior.

Por isso decidimos implementar níveis - verde, amarelo e vermelho - para os alertas emitidos. O nível verde indicaria que o valor da assinatura estaria fora do intervalo definido mas não estaria muito afastado das zonas limites. O nível amarelo indicaria já um afastamento considerável das zonas limites e o nível vermelho indicaria que o valor da assinatura estaria muito afastado nas zonas limites definidas. Os níveis de alerta definidos e os respetivos valores mínimos e máximos estão apresentados na tabela 4.6 – recorda-se aqui que, x representa o valor da assinatura, $\bar{x}$ a média do valores das assinaturas anteriores e S o desvio padrão dos valores das assinaturas anteriores.

Nível alerta	**Valor mínimo**	**Valor máximo**
Verde	$x > \bar{x} \pm 2 \times S$	$x < \bar{x} \pm 3 \times S$
Amarelo	$x > \bar{x} \pm 3 \times S$	$x < \bar{x} \pm 4 \times S$
Vermelho	$x > \bar{x} \pm 4 \times S$	–

Tabela 4.6: Níveis de emissão de alertas

Os valores limites de intervalo estabelecidos para o nível verde foram a média dos valores das assinaturas anteriores e três vezes o seu desvio padrão. Para o nível amarelo a média dos valores das assinaturas anteriores e quatro vezes o seu desvio padrão e por último, caso o valor da assinatura ultrapasse os limites definidos para o nível amarelo, então entraria no nível vermelho.

Após a definição dos intervalos dos alertas a serem emitidos o próximo passo seria proceder ao cálculo das variações para a amostra de dados recolhida.

4.7 Aplicação para o Cálculo de Variação de Assinaturas

A aplicação para cálculo do valor da variação das assinaturas teria como propósito detetar variações consideradas anómalas, que saíssem fora dos limites normais definidos e emitir o respetivo nível de alerta. Depois seria feita uma análise para determinar se essa variação poderia indicar (ou não) uma mudança no seu perfil.

Quando um utilizador é colocado na plataforma de *e-learning* é-lhe atribuído um determinado perfil, assim, numa primeira nível de análise seria necessário verificar se ao longo do intervalo de tempo que a amostra continha, a sua assinatura se teria mantido fiel ao perfil que lhe fora atribuído.

Anteriormente, neste trabalho foi determinado que uma assinatura se manteria em valores aceitáveis se a sua variação estivesse enquadrada dentro do intervalo da sua média e duas vezes o seu desvio padrão. Caso isso não ocorresse deveria ser despoletado um alerta para posterior tratamento com um determinado nível, verde, amarelo ou vermelho.

Para além da determinação do valor de variação da assinatura do utilizador, foram calculados também o valor da variação dos diversos atributos numéricos da assinatura (tabela 4.5). O intuito de efetuar o cálculo da variação dos atributos da assinatura foi facilitar e auxiliar o processo de análise. Quando o valor da assinatura excedesse o intervalo definido seria necessário determinar a razão do valor obtido. Sabendo a variação dos valores dos atributos poder-se-ia, numa análise mais minuciosa, verificar qual ou quais os atributos da assinatura que mais contribuíram para que a variação da assinatura fosse tão elevada. Assim, para cada dia de análise o valor de cada atributo foi também comparado com a sua média e desvio padrão. Tal como no cálculo de variação da assinatura, se o valor do atributo nesse dia excedesse a média e duas vezes o seu desvio padrão, calculados com base dos dados dos dias anteriores, então a própria variação do valor do atributo seria vista como anómala.

Para determinar o valor da variação da assinatura dos utilizadores foi desenvolvida uma aplicação em C#, utilizando o *Microsoft Visual Studio* versão 2010. Esta aplicação utilizou uma conexão à base de dados que contém as assinaturas dos utilizadores para os diversos dias e executou um algoritmo para determinação das variações de assinatura. Esta aplicação teve duas funcionalidades distintas:

1) Uma funcionalidade ("DETECT ANOMALIES" - figura 4.9), que executa um procedimento para um determinado dia e que verifica em todas as assinaturas desse dia aquelas que estão fora do intervalo definido como razoável, e apresenta essa informação. Caso o valor da assinatura exceda os valores desse intervalo mostra um alerta cuja severidade (verde, amarelo, vermelho) será tanto maior quanto maior a variação da assinatura dos valores pré-definidos.
2) Uma segunda funcionalidade ("GET SIGNATURE VARIATION" - figura 4.9), que foi utilizada na fase de análise mais minuciosa, que calcula as variações de assinatura e dos atributos numéricos que compõem a assinatura para um determinado dia e para um determinado utilizador.

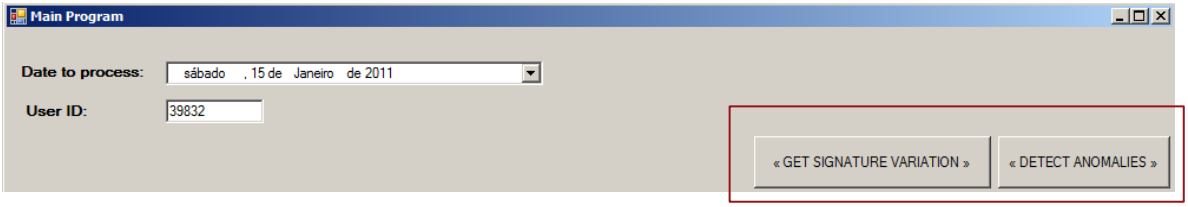


Figura 4.9: Funcionalidades principais da aplicação de cálculo de variação assinaturas

De uma forma simplificada o procedimento algorítmico implementado pode ser descrito da seguinte maneira. A função "LoadUsers" irá para uma determinada data ("Date") carregar todos os utilizadores para essa mesma data. A função "Load_User_Signature_Attributes" irá para uma determinada data ("Date") e para determinado utilizador ("User") efetuar o carregamento de todos os atributos da assinatura. Esta função, como já foi referido, neste trabalho, só considerou os atributos numéricos da assinatura anteriormente mencionados.

A função de cálculo da assinatura "CalculateUserSignature" foi responsável por, com os atributos já carregados, efetuar o cálculo da assinatura, para essa data, para o utilizador em questão, utilizando a fórmula de cálculo mencionada anteriormente para a assinatura, recorrendo aos atributos da assinatura e considerando também o peso de cada um deles.

```
List  User_List = LoadUsers(Date);
Foreach (User in User_List)
Begin
       List List_User_Signature_Attributes = Load_User_Signature_Attributes(User,Date);
       User_Daily_Signature = CalculateUserSignature(List_User_Signature_Attributes);

       List List_Previous_User_Signatures = CalculateUserPreviousSignatures(User,Date);

       User_Signature_Average= CalculateAverage(List_Previous_User_Signatures);
       User_Signature_StandardDeviation =
       CalculateStandardDeviation(List_Previous_User_Signatures);

       LowerLimit = User_Signature_Average – 2* User_Signature_Standard_Deviation;
       UpperLimit = User_Signature_Average + 2* User_Signature_Standard_Deviation;

       If (LowerLimit  < User_Daily_Signature > UpperLimit) Issue_Alert();
End
```

Figura 4.10: O algoritmo do cálculo de assinaturas

Seguidamente a função "CalculateUserPreviousSignatures" calcula o valor de todas as assinaturas anteriores à data em análise, armazenando-as numa lista para que a nossa assinatura atual possa ser comparada com as anteriores assinaturas desse mesmo utilizador. Esta função considera também o valor da assinatura atual do utilizador. A comparação é feita com base na média e desvio padrão, valores que serão calculados pelas funções "CalculateAverage" e "CalculateStandardDeviation". Como referido, estas funções tiveram em conta além das assinaturas anteriores, o valor da assinatura do dia para o cálculo da média e desvio padrão.

Neste trabalho, os valores fixados para os limites inferior e superior, respetivamente "LowerLimit" e "UpperLimit", são os apresentados na fórmula, ou seja o intervalo compreendido entre a média dos valores das assinatura e duas vezes o seu desvio padrão. Valores fora desse intervalo foram considerados fora do normal, anomalias, e um alerta foi lançado.

De seguida é apresentada (figura 4.11) a execução da aplicação de deteção de variações no padrão de assinatura dos utilizadores, que também podem ser designadas como anomalias, e que poderiam indicar uma mudança de comportamento de um determinado utilizador face à plataforma

de *e-learning*. É necessário fornecer à aplicação uma data para processamento. Para esta data a aplicação efetua o cálculo dos valores de assinaturas para todos os utilizados com entradas registadas nesse dia e compara o valor dessa assinatura com a média e desvio padrão das assinaturas desse utilizador para os dias anteriores. Caso o valor da assinatura esteja fora do intervalo definido como normal, ou seja a sua média mais duas vezes o seu desvio padrão, um alerta é emitido com o nível de severidade verde, amarelo ou vermelho.

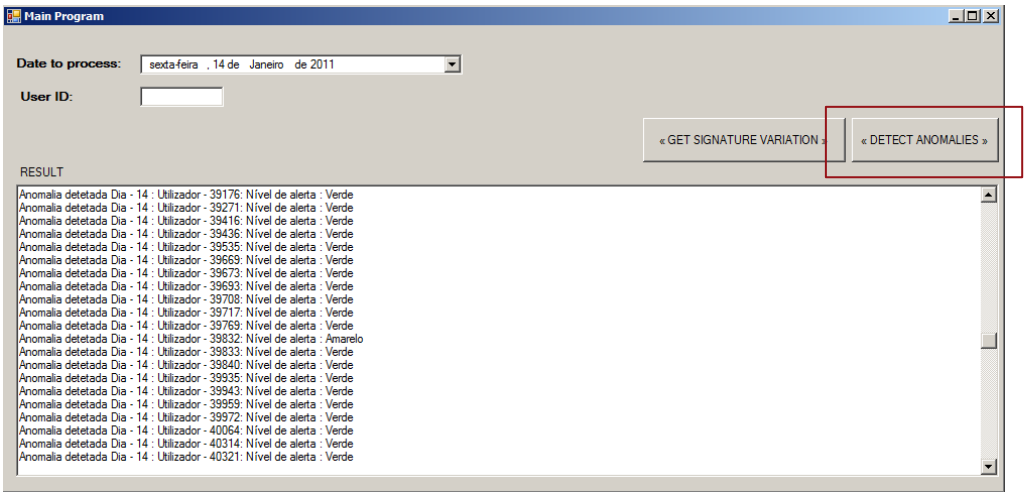


Figura 4.11: Aplicação com o resultado da execução da deteção de anomalias

Passamos à análise de um caso concreto de variação anómala no valor de uma assinatura. Podemos constatar na Figura 4.11 que ilustra a execução da funcionalidade "DETECT ANOMALIES" para o dia 14 de Janeiro de 2011, no qual foi emitido um alerta amarelo para o utilizador com a identificação 39832. O próximo passo seria por isso analisar o porquê da emissão deste alerta, com este nível de severidade.

É essa mesma verificação que irá agora ser demonstrada, recorrendo a esta mesma aplicação e à opção "GET SIGNATURE VARIATION". Nesta opção obtemos os valores de variação da assinatura e respetivos atributos para todos os dias até uma data que é definida pelo campo "Date to process". É necessário ainda fornecer a identificação do utilizador, no campo "User ID". Como a nossa amostra de dados está limitada aos primeiros quinze dias do mês de Janeiro de 2011, foi utilizada essa data para obtenção dos dados de variações de assinatura e respetivos atributos.

Na figura 4.12 são apresentados os resultados decorrentes da execução da aplicação para o dia 15 de Janeiro de 2011, calculando, por isso, as variações para cada dia, desde o dia 1 ao dia 15 para o utilizador com a identificação 39832, responsável pela emissão do alerta.

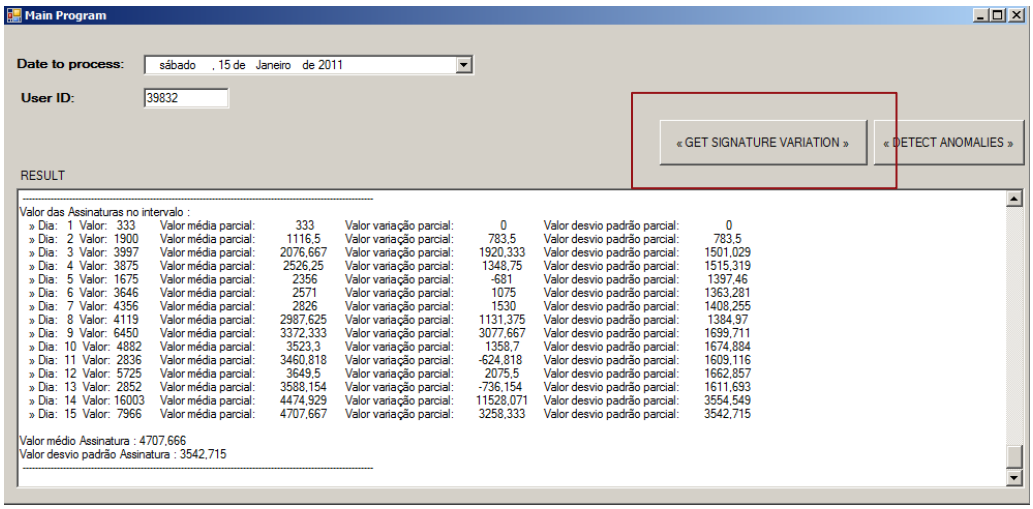


Figura 4.12: Aplicação do cálculo da variação de uma assinatura

Na figura 4.12 são apresentados os valores de variação da assinatura deste utilizador para os primeiros quinze dias de 2011. É ilustrado, ainda nesta figura, que a aplicação efetua também o cálculo dos valores parciais de média e desvio padrão para cada um dos dia, tendo-se usado, contudo, para o estabelecimento de limites finais, o valor final médio da assinatura e do desvio padrão. Ao traçar um gráfico com os valores resultantes, conseguimos obter uma melhor visão dos valores da amostra (figura 4.13).

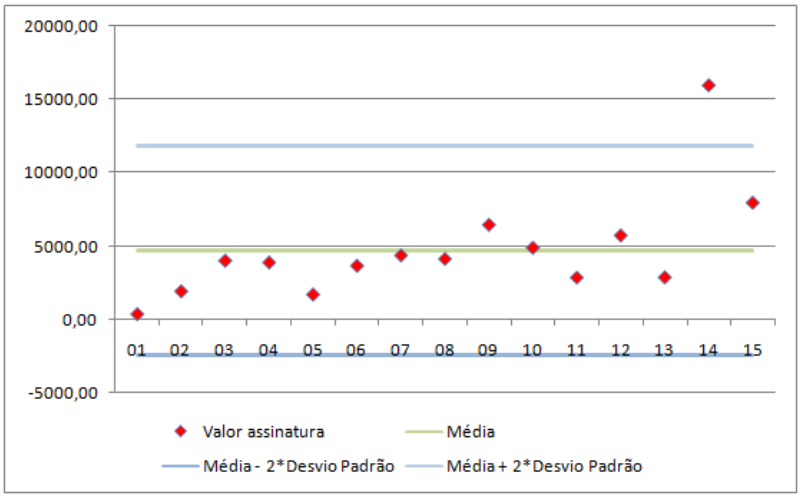


Figura 4.13: Análise da variação de uma assinatura

Pelo gráfico da figura 4.13 podemos ver mais uma vez o alerta emitido no dia 14, dado o valor da assinatura estar claramente fora do intervalo definido. O porquê deste valor foi analisado em maior detalhe, estudando-se os valores dos atributos da assinatura individualmente de modo a que fosse possível fazer a interpretação do padrão considerado invulgar. Como já mencionado, a aplicação calcula também as variações dos atributos da assinatura. Assim, iremos agora, a título de exemplo, demonstrar o processo de análise que foi seguido para analisar esta anomalia e também outras que foram detetadas.

Para o dia 14 de Janeiro, os três atributos de maior peso da assinatura, os atributos que se referem a operações de administração("N_Operations_ControlPanel", "N_Operations_AdminPanel" e "N_Operations_PortalAdmin") revelaram valor zero para este utilizador, pelo que foram excluídos do processo de análise. Seguiu-se a análise dos restantes atributos, mas como o processo de análise é algo repetitivo, são apresentados de seguida os valores de análise para três atributos em particular, o número de operações normais "N_Operations_Normal", número de visualizações de página fora do horário normal, "N_PageViews_OffSchedule" e o tempo médio de sessão "Medium_Session_Time".

Começaremos, então, pela análise do atributo "N_Operations_Normal", que representa o número de operações normais para um dado dia. Analisando os valores obtidos, vemos que existiu uma saída fora dos padrões normais nos dias 01, 02, 03, 10 e 15. Mas, no dia 14, o valor registado foi

de 25, um valor um pouco acima da média, que denota maior atividade na plataforma, mas perfeitamente dentro dos parâmetros considerados normais.

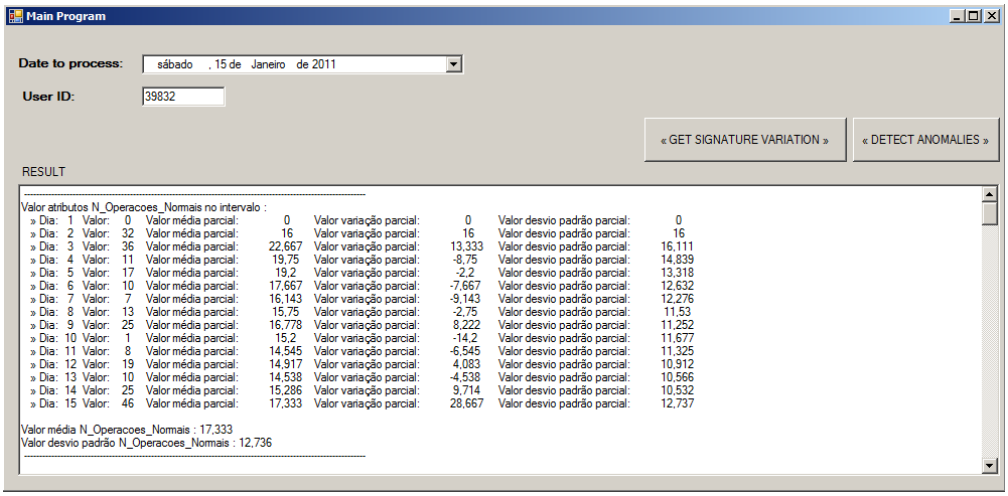


Figura 4.14: Cálculo aplicacional da variação dos valores do atributo "N_Operations_Normal"

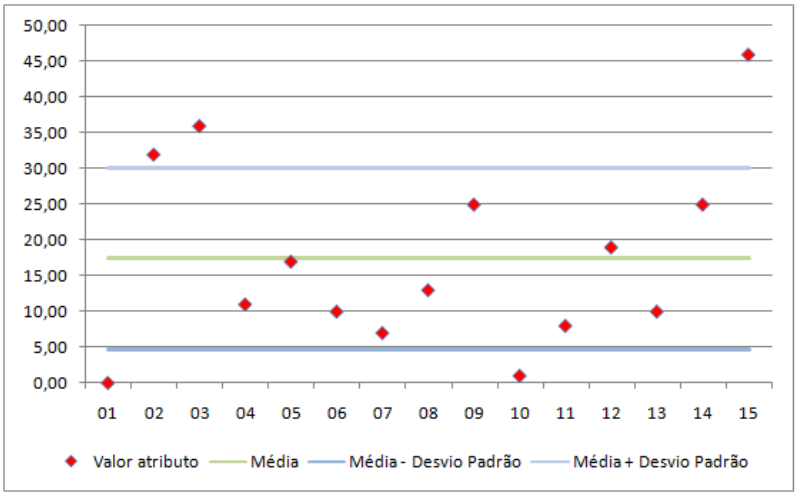


Figura 4.15: Análise da variação do atributo "N_Operations_Normal"

No gráfico apresentado na figura 4.15 podemos observar os valores para o caso particular do atributo "N_Operations_Normal", para o intervalo de tempo considerado. É possível visualizar um aumento de atividade considerável para o dia 15 mas não para o dia 14. Pela análise feita este

atributo não poderia ser o responsável pelo valor anormal no dia 14. Seria, assim, necessário continuar o processo de análise.

De seguida, é apresentada a análise de um outro atributo que tem, também, um peso considerável na determinação do valor de uma assinatura: o atributo "N_PageViews_OffSchedule", que acolhe o número de visualizações de página feitas por um utilizador, fora do horário considerado normal.

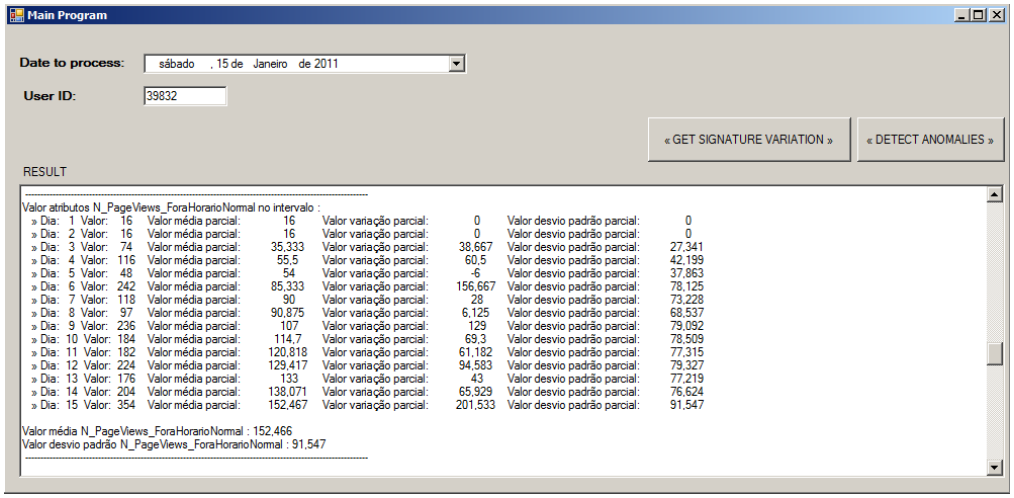


Figura 4.16: Aplicação para cálculo da variação atributo "N_PageViews_OffSchedule"

Na figura 4.16 conseguimos ver que os valores registados nos dias 01, 02 e 05 estão um pouco fora do intervalo definido como normal, enquanto que o valor de dia 15 está claramente fora desse intervalo, confirmando que no dia 15 este utilizador esteve particularmente ativo na plataforma e fora do horário normal, indiciando trabalho fora de horas. Contudo, no dia 14 o valor registado é de 204, registando portanto 204 visualizações de páginas, o que está perfeitamente dentro do intervalo entre 60,92 e 244,01

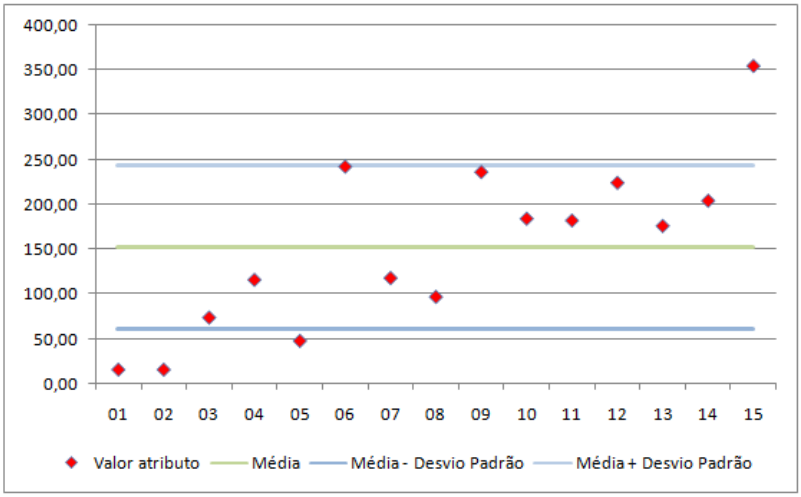


Figura 4.17: Análise da variação do atributo "N_PageViews_OffSchedule"

Na figura 4.17 podemos ver um gráfico refletindo uma análise dos valores relativos ao atributo "N_PageViews_OffSchedule". Pela análise destes valores vemos que também não seria este atributo o responsável pelo valor anormal da assinatura no dia 14.

Continuando para a análise do atributo tempo médio de sessão, "Medium_Session_Time", ilustrada na figura 4.18, chegamos a um valor que indicia que foi este o atributo responsável pela variação anormal da assinatura do utilizador para o dia 14. Como podemos verificar, nesse dia, temos um valor registado de tempo médio de sessão de 11 939 - valor medido em segundos -, o que equivale a um tempo médio de sessão de cerca de 3 horas e meia, valor este claramente invulgar, se tivermos em consideração que o valor médio para este atributo registado neste período de 15 dias foi de 2 166 segundos, o que equivale a um pouco mais de 1 hora de sessão.

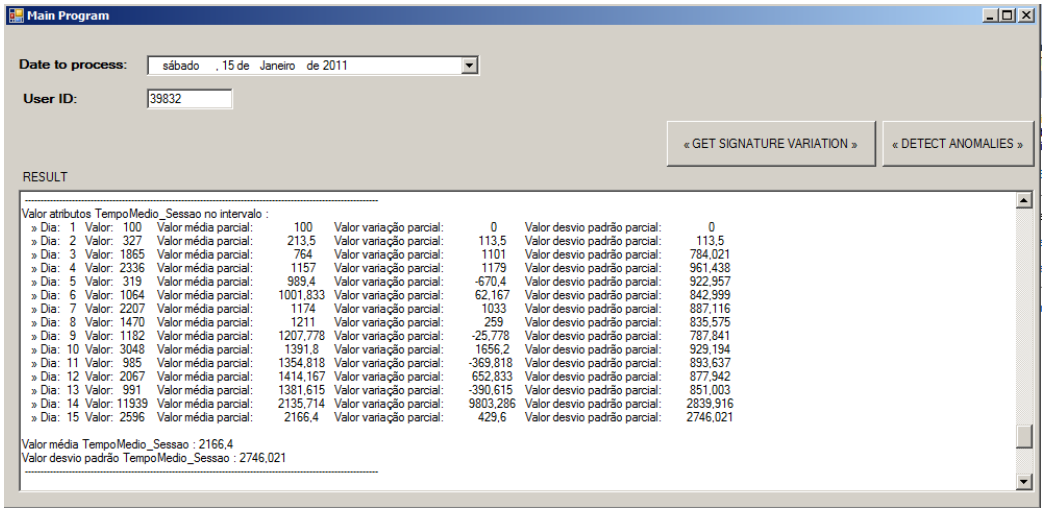


Figura 4.18: Cálculo aplicacional de variação atributo "Medium_Session_Time"

Esta variação, bastante invulgar, e que contrasta com os valores registados nos restantes dias (figura 4.19) foi suficiente para provocar a variação registada no dia 14 na assinatura deste utilizador. Esta variação do valor da assinatura, causada pela variação anormal no atributo que diz respeito ao tempo médio de sessão poderia indiciar uma alteração de perfil? Pela análise feita vemos que não.

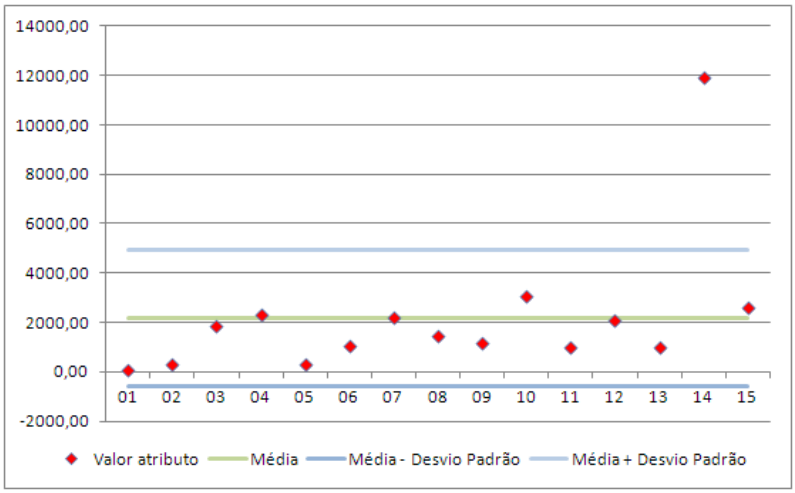


Figura 4.19: Análise da variação do atributo "Medium_Session_Time"

A aplicação detetou corretamente a variação anómala mas se analisarmos cuidadosamente os dados vemos que para o dia 14 a atividade do utilizador na plataforma não esteve muito afastada

da normalidade. Todavia, no dia 15 sim. Neste caso, vemos que existe um aumento efetivo de atividade, mas no dia 14, aparte do valor anómalo do tempo médio de sessão, os outros atributos revelam normalidade. Um tempo médio de sessão elevado com pouca atividade indicia que este utilizador, neste dia, poderá ter estado com o seu navegador com sessão iniciada na plataforma de *e-learning* mas sem efetuar qualquer tipo de operação durante largos períodos de tempo, provocando este tempo médio de sessão elevado.

4.8 Análise e Considerações sobre os Resultados Obtidos

O período da amostra recolhida - primeiros quinze dias do mês de Janeiro, do ano de 2011 – foi um período algo reduzido, que necessitaria de ser alargado de forma a melhor validar os resultados obtidos. A deteção de variação de assinaturas realizada neste período revelou que as para as variações anómalas detetadas, nenhuma indiciava claramente uma alteração de perfil. Essas variações eram maioritariamente causadas por valores elevados nos tempos médios de sessão, número de sessões e visualizações de páginas. Isto indicia alterações na atividade dos utilizadores na plataforma de *e-learning* ou sessões com pouca atividade por parte dos utilizadores, ou seja, são alterações comportamentais mas não suficientes para estabelecer um perfil diferente do estabelecido. Dentro das variações detetadas e que foram consideradas como anomalias, estas enquadraram-se sempre dentro dos níveis verde e amarelo, não tendo sido detetadas variações muito invulgares, de nível vermelho, que poderiam indiciar alterações anormais nos vários perfis de comportamento estudados.

Mesmo as variações detetadas não eram coerentes ao longo do tempo, ou seja, um determinado utilizador era alvo de um alerta para um determinado dia mas nos dias seguintes esse nível de alerta não se mantinha. Isto revela que a alteração de comportamento que teve foi pontual, não constituindo por si só uma variação de perfil perante a aplicação. Como já foi referido, os atributos da assinatura que mais contribuíram para as variações detetadas a nível das assinaturas foram os atributos relacionados com as visualizações de página e os atributos relacionados com a componente de sessão, o número de sessões e o tempo médio destas. Não foi detetada nenhuma variação anómala nos três atributos de maior peso da assinatura, os atributos que se referem a operações de administração("N_Operations_ControlPanel", "N_Operations_AdminPanel" e "N_Operations_PortalAdmin"), reforçando a ideia de que as anomalias detetadas não revelaram mais que aumentos pontuais de atividade na plataforma por parte dos utilizadores.

Pelos resultados obtidos e para a amostra recolhida, podemos depreender também que o comportamento dos utilizadores na plataforma se adequou ao comportamento esperado para o seu perfil, ou seja, não foram detetadas situações em que por exemplo um aluno regista-se operações de acesso restrito, que poderia indiciar acessos indevidos à plataforma. Aparte da análise mais relacionada com o propósito deste trabalho, a deteção de alterações de perfis, outros indicadores interessantes poderiam ser extraídos com a informação recolhida, podendo ser feitas comparações aos tempos médios de atividade, visualizações de páginas em horário normal e fora do horário normal por perfil ou mesmo definir manchas horárias de maior atividade para detetar períodos de maior constrangimento para o sistema por intervalos que poderiam ser períodos horários, diários, semanais ou mesmo mensais.

Capítulo 5

Conclusões e Trabalho Futuro

5.1 Síntese

O enquadramento de um utilizador num dado sítio é uma questão que tem sido muito abordada pelos investigadores segundo as mais variadas vertentes. Muitas vezes esse enquadramento é feito de uma forma estática, em que o utilizador é classificado de acordo com um determinado perfil e com esse perfil tem acesso a determinadas funcionalidades - o próprio sítio apresenta-se a esse utilizador de acordo com esse mesmo perfil. Porém, o perfil de um utilizador pode sofrer evoluções significativas ao longo do tempo. O seu comportamento e apresentação perante um sítio pode mudar, sendo algo que, como sabemos, pode não ser estático ao longo do tempo. A forma como navegamos nos diversos sítios é diferente de dia para dia. Mesmo para um dado sítio específico mudamos a forma como interagimos com ele, mudando as nossas opções, ou escolhendo algo diferente em outras alturas.

Como todos sabemos, os nossos hábitos variam, como tal as nossas preferências também. Logo, a nossa assinatura também varia. O intuito deste trabalho foi esse, perante um caso de estudo concreto selecionar um conjunto de atributos que pudessem caracterizar um utilizador, definir a sua assinatura e com isso determinar a sua variação ao longo do tempo por forma a verificar se essa variação influenciaria a forma como o utilizador se apresenta perante o sítio e se essa

variação seria suficiente para indiciar uma mudança no seu perfil, bem como a mudança das suas preferências. Do trabalho realizado sobressai também um outro aspeto, as variações no modo de comportamento de um utilizador não podem ser determinadas por um único instante no tempo, uma mudança consistente no padrão de utilização de uma determinada aplicação tem que ser acompanhada em vários instantes temporais. No entanto, esta constatação tem que levar em consideração a análise do caso de estudo que temos em mão e no objetivo que temos. Em casos de deteção de intrusão ou de fraude surgem situações em que uma única anomalia no comportamento do utilizador determina por si só uma ação de reação a essa ocorrência. No próprio caso de estudo exposto nesta dissertação, uma mudança na variação de determinados atributos da assinatura, como sendo os atributos que caracterizam operações de administração, poderia indiciar uma brusca alteração de perfil que poderia desencadear uma ação preventiva de verificação. Tudo depende da análise e do objetivo que temos e as áreas de aplicação deste tipo de análise são vastas.

5.2 Análise ao Trabalho Efetuado

Do trabalho efetuado à que referir que é um trabalho evolutivo, ou seja, os resultados obtidos adequam-se à amostra e objetivos pretendidos, no entanto é necessário um aperfeiçoamento que só será conseguido alargando o espectro de análise temporal e pondo em prática as ações correspondentes às deteções de variação de assinaturas aqui descritas. Tal irá permitir refinar o processo de cálculo de assinatura desenvolvido, mas, para isso, terá que haver uma nova definição de objetivos. Por exemplo, definir o que será feito quando uma anomalia de comportamento for detetada. Esta redefinição de objetivos permitirá, entre outros aspetos, verificar a consistência e aplicabilidade da forma de cálculo de variação de assinatura com vista ao seu melhoramento.

O processo de extração, carregamento e transformação dos dados, desde a sua fonte até ao destino final, é algo que necessitaria de ser testado num ambiente mais similar ao ambiente produtivo. Isto serviria, por exemplo, para testar os tempos dos diversos passos, dependendo contudo da frequência de análise pretendida aquando da colocação do sistema em funcionamento. Caso se pretenda uma análise que seja mais parecida com deteções realizadas em tempo real então todo o processo necessitaria de ser revisto.

As anomalias detetadas, variações anormais nos padrões de variação de assinaturas, estiveram de acordo com o objetivo de verificação pretendido, enquadrando-se naquilo que se pretendia analisar e validar. Esta análise foi feita por amostragem das variações detetadas. Para cada uma dessas variações foram analisados quais os atributos da assinatura que contribuíram para essa variação e foi verificado também se esses dados eram coerentes com os dados guardados na base de dados intermédia, confrontando-os também com os dados da fonte original verificando se esses dados estavam de acordo com a transformação efetuada. Além disso, foi necessário verificar se o perfil do utilizador sujeito a essa análise se adequava aos dados constantes da sua assinatura e se a variação da sua assinatura também se apropriava ao seu perfil na aplicação Não foram detetadas situações que a variação anómala da assinatura revela-se dados que permitissem concluir uma variação efetiva no perfil do utilizador.

De uma forma mais pormenorizada e sistematizada podemos resumir o trabalho realizado nos seguintes pontos:

- Análise de um caso de estudo - o sítio de *e-learning* da Universidade do Minho -, a análise realizada direcionou-se essencialmente para a identificação e definição de perfis de utilização bem como para a identificação de características relevantes de utilização que seriam utilizadas para a identificação de um utilizador.
- Identificação e definição das diversas fontes de informação do sistema e dos atributos que foram utilizados para a construção da assinatura de um utilizador.
- Definição e caracterização dos atributos que constituíram a assinatura do utilizador perante o sítio analisado.
- Construção e implementação do modelo de dados para os sistemas de informação que albergaram os dados provenientes da fonte e das assinaturas dos utilizadores.
- Construção e implementação dos processos de povoamento da base de dados que albergaram os dados provenientes da fonte e para a base de dados que guardou os dados das assinaturas dos utilizadores.
- Construção e implementação de uma aplicação para o cálculo da variação dos valores das assinaturas dos utilizadores e para a identificação de anomalias na sua variação.

O trabalho realizado teve também as suas limitações, que se deveram sobretudo a questões de logística, nomeadamente a nível de equipamento utilizado. O volume de dados disponível era bastante grande e mesmo só considerando a pequena amostra de registos de *log* retirados da

aplicação, o tempo gasto no seu processamento foi muito grande. O tempo consumido a visualizar, trabalhar e processar os dados revelou-se uma tarefa mais complicada precisamente porque a capacidade de processamento ficava muito aquém do que seria necessário para abarcar a quantidade de informação disponível. De seguida, enumera-se em alguns pontos as principais limitações detetadas:

- O volume de informação processado na amostra e o intervalo de tempo foi reduzido face à globalidade da informação disponível na base de dados da plataforma de *e-learning*.
- O equipamento utilizado não possuía a capacidade necessária para responder às necessidades do trabalho realizado, o que acabou por ter impacto no tempo gasto para a realização do mesmo.
- O método de análise às variações de assinaturas anómalas revelou-se uma tarefa bastante morosa, sendo necessária a implementação de um mecanismo mais automático para análise de anomalias.
- O alargamento do período de amostragem permitiria uma melhor validação aos algoritmos utilizados para cálculo das variações das assinaturas.

5.3 Trabalho futuro

Relativamente ao trabalho futuro a realizar e que surja em complemento ao trabalho já realizado este poder-se-á prender, sobretudo, em eliminar as limitações detetadas neste trabalho. A principal componente de desenvolvimento será sobretudo dar seguimento ao trabalho já realizado, implementando, por exemplo, mecanismos de resposta automáticos às anomalias detetadas a nível de variação de assinaturas dos utilizadores e do seu impacto perante as características presentes no seu perfil. Isto fará com que todas as variações de assinatura sejam validadas o que permitirá anuir os algoritmos de deteção utilizados. Podemos enumerar algumas das tarefas que poderiam ser executadas numa primeira fase de melhoramentos ao trabalho já efetuado:

- Abranger um conjunto de dados maior, alargando o intervalo temporal contemplado na amostra.
- Implementar algum tipo de mecanismo de validação automático para as variações anormais de assinaturas detetadas, confrontando essas variações com o perfil do utilizador.

Após esta primeira fase, seguir-se-ia a implementação de outras tarefas, que teriam como base os resultados obtidos na primeira fase de melhoramentos, incluindo, por exemplo:

- O teste de novos algoritmos para deteção de alterações nas assinaturas dos utilizadores, no qual poderão ser incluídos algoritmos de *clustering* para agrupamento das assinaturas dos utilizadores de forma a complementar a análise já efetuada.
- O acoplamento no sítio do caso de estudo de alguns mecanismos de recomendação que assentem na informação recolhida a partir das assinaturas dos utilizadores e da sua natural variação. Este sistema de recomendação poderia verificar, por exemplo, quais as operações mais realizadas para os diversos perfis de utilizador e colocar essas operações em destaque aquando da sua entrada no sistema.
- A implementação de mecanismos de otimização do funcionamento do sítio, também recorrendo à informação presente nas assinaturas, nomeadamente a obtida a partir da observação de quais os períodos temporais de maior utilização da plataforma, podendo prever possíveis estrangulamentos ao sistema.

Estas são, somente, algumas possíveis melhorias na implementação do sistema atual, dentro do conceito de identificação e reação a alterações nos perfis de comportamento dos utilizadores com base em variações das suas assinaturas - podendo mesmo este conceito ser alargado. O trabalho realizado pode ter outras aplicações, podendo ser por exemplo aprofundada a identificação de acessos indevidos à plataforma, onde, para além dessa identificação poderiam ser implementados mecanismos preventivos, recorrendo à informação que está presente nas assinaturas recolhidas.

> "*Pouco conhecimento faz com que as pessoas se sintam orgulhosas. Muito conhecimento, que se sintam humildes. É assim que as espigas sem grãos erguem desdenhosamente a cabeça para o Céu, enquanto que as cheias as baixam para a terra, sua mãe.*"
>
> **Leonardo da Vinci**

Bibliografia

Adomavicius, G. e Tuzhilin, A. 2005. *Toward the next generation of recommender systems: a survey of the stateof-the-art and possible extensions*. IEEE Transactions on Knowledge and Data Engineering 17, 734-749.

Agrawal, R. e Srikant, R. 1995. *Mining sequential patterns*. In Proc. 1995 Int. Conf. Data Engineering (ICDE'95), pages 3–14, Taipei, Taiwan, Mar.

Alcalá, J. del Jesús, M.J., Garrell, J.M., Herrera, F., Hervás, C., Sánchez, L. 2004. *Proyecto KEEL: Desarrollo de una Herramienta para el Análisis e Implementación de Algoritmos de Extracción de Conocimiento Evolutivos*. Tendencias de la Minería de Datos en España. Eds. Giradles, J., Riquelme,J.C., Aguilar, J.S. (pp. 413-423).

Ardissono L., Console L., Torre I. 2001. *An adaptive system for the personalized access to news*. AI Communications, 14(3):129–147.

Au Yeung, C.M., Gibbins, N., & Shadbolt, N. 2009. *Contextualising Tags in Collaborative Tagging Systems*. Proceedings of the 20th ACM Conference on Hypertext and Hypermedia, 251-260

Baeza-Yates, R. e Ribeiro-Neto, B. 2011 *Modern Information Retrieval* (Second edition). Pearson, 2011.

Baldi, P., Frasconi, P., e Smyth, P. 2003. *Modelling the Internet and the Web*. ACM Trans. Inter. Tech. 3, 1, 171-209.

Bolton, R. e Hand, D. 2001. *Unsupervised Profiling Methods for Fraud Detection*. Credit Scoring and Credit Control VII.

Borges J. e Levene M. 2007: *Evaluating Variable-Length Markov Chain Models for Analysis of User Web Navigation Sessions*. IEEE Trans. Knowl. Data Eng. 19(4): 441-452

Buchner, G., Baumgarten, M., Anand, S., Mulvenna, D. e Hughes, G. 1999. *Navigation Pattern Discovery From Internet Data*. In Proceedings of theWeb Usage Analysis and User Profiling Workshop (WEBKDD99). Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Diego, August), 25-30.

Burge, P. e Shawe-Taylor, J. 2001. *An Unsupervised Neural Network Approach to Profiling the Behaviour of Mobile Phone Users for Use in Fraud Detection*. Journal of Parallel and Distributed Computing 61: 915-925.

Burke, R. 2002. *Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction*, 12(2002), 331-370

Brusilovsky P. 1996. *Methods and techniques of adaptive hypermedia. User Modelling and User-Adapted Interaction, 6(2-3):87–129*, July 1996.

Brusilovsky P. 1997. *Efficient techniques for adaptive hypermedia. Intelligent Hypertext: Advanced Techniques for the World Wide Web*, 1326:12–30.

Brusilovsky P. 2004 *Adaptive Navigation Support: From Adaptive Hypermedia to the Adaptive Web and Beyond*, volume 2, pagina 7 23. PsychNology Journal.

Brusilovsky P., e Nejdl W. 2004. *Adaptive Hypermedia and Adaptive Web*, © 2004 CSC Press LLC.

Cahill, M., Lambert, D., Pinheiro, J., e Sun, D. 2000. *Detecting fraud in the real world*. Technical Report, Bell Labs, Lucent Technologies, 25-30.

Ceglar, A. e Roddick, J. 2006. *Association Mining*. ACM Computing Surveys, 38(2).1-42. ACM.

Chakrabarti, S. 2000. *Data mining for hypertext: A tutorial survey*. SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM 1, 2, 1-11.

Chandola V., Banerjee A., e Kumar V. 2007, *Anomaly Detection: A Survey.* University of Minnesota, Tech. Rep., August 2007.

Chapman, P. et al, 2000. CRISP-DM 1.0 - *Step-by-step data mining guide*. Disponível em: <http://www.crisp-dm.org/CRISPWP-0800.pdf>. [20 de Junho de 2011].

Chen, R., Chiu, M., Huang, Y. e Chen, L. 2004. *Detecting Credit Card Fraud by Using Questionaire-Responded Transaction Model Based on Support Vector Machines*. Proc. of IDEAL2004,800-806.

Cios K., Pedrycz W., Swiniarski R., Kurgan L. 2007, *Data Mining: A Knowledge Discovery Approach*, Springer

Cios K., e Kurgan, L. 2005. *Trends in data mining and knowledge discovery*. In Pal, N.R., and Jain L.C. (Eds.), Advanced Techniques in Knowledge Discovery and Data Mining, 1–26, Springer Verlag, London.

Clark, R. C. e Mayer, R. E. 2007*. E-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning (2nd ed.)*. Pfeiffer & Company.

Cortes C, Pregibon D. 2001. *Signature-based methods for data streams. Data Mining and Knowledge Discovery* 5:167–82.

Dai, Honghua, Mobasher, e Bamshad. 2003. *A road map to more effective Web personalization: Integrating domain knowledge with Web usage mining*. In Proceedings of the International Conference on Internet Computing (IC), Las Vegas, Nevada

De Bra P. e Calvi L. 1998. *AHA: a generic adaptive hypermedia system. In Proceedings of the 2nd Workshop on Adaptive Hypertext and Hypermedia, HYPERTEXT´98*, Pittsburgh, USA, June 20–24.

Delerablée C. e Pointcheval D. 2006. *Dynamic fully anonymous short group signatures*. In VIETCRYPT 2006, pages 193–210.

Donoho, S. 2004. *Early Detection of Insider Trading in Option Markets*. Proc. of SIGKDD04, 420-429.

Duda, R.O., Hart, P.E., Stork, D.G. 2000. *Pattern Classification*. Wiley Interscience.

Edge M., Sampaio e P., Pedro R. F. 2009. *A survey of signature based methods for financial fraud detection*. Manchester Business School, University of Manchester, Booth Street East, Manchester M16 6PB, United Kingdom;

Eirinaki, M. e Vazirgiannis, M. 2003. *Web mining for web personalization*. ACM Trans. Inter. Tech. 3, 1, 1-27.

Esprit, the EU information technologies programme, 1999.
Disponível em: < http://cordis.europa.eu/esprit/home.html > . [25 de Janeiro de 2013]

Etzioni, 0. 1996. *The World Wide Web: Quagmire or gold mine*. Communications of the ACM, 39(ll), 65-68.

Extended Log File Format, 1995. Disponível em: < http://www.w3.org/TR/WD-logfile.html> .
[25 de Janeiro de 2013]

Fayyad, U. M., Piatetsky-Shapiro, G., e Smyth, P. 1996 *From data mining to knowledge discovery: An overview*. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors, Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press

Ferreira, P., Alves, R., Belo, O., e Cortesão, L. 2006 *Establishing Fraud Detection Patterns Based on Signatures*. In Proceedings of Industrial Conference on Data Mining. 526-538.

Ferreira, P., Alves, R., Belo, O., Ribeiro, J. 2007. *Detecting Telecommunications Fraud based on Signature Clustering Analysis*. In: 13th Portuguese Conference in Artificial Intelligence - EPIA'07, 2007, Guimaraes. Workshop on Business Intelligence

Fu, Y. , Sandhu, K. e Shih, M. 2000. *A generalization-based approach to clustering of Web usage sessions*. In Brij M. Masand and Myra Spiliopoulou, editors, Web Usage Analysis and User Profiling, Lecture Notes in Artificial Intelligence, Springer 1836, 21-38.

Gabrilovich, E. e Markovitch, S. 2005. *Feature generation for text categorization using world knowledge*. In International 11th Joint Conference on Artificial Intelligence. Edinburgh, UK, pp. 1048–1053, 2005.

Ghosh, S. e Reilly, D. 1994. *Credit Card Fraud Detection with a Neural Network*. Proc. of 27th Hawaii International Conference on Systems Science 3: 621-630.

Goldberg, H., Kirkland, J., Lee, D., Shyr, P. e Thakker, D. 2003. *The NASD Securities Observation, News Analysis & Regulation System (SONAR).* Proc. of IAAI03.

Han, J., Pei, J., Yan, X., 2005. *Sequential Pattern Mining by Pattern-Growth: Principles and Extensions*. StudFuzz. Springer. (pp. 183-220).

Han, J. e Kamber, M. 2001, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers.

Handl, J. e Meyer, B. 2007. *Ant-based and Swarm-based clustering*. Swarm Intelligence 1,95-113.

Heymann, P., Ramage, D., e Garcia-Molina, H. 2008. *Social Tag Prediction*. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 531-538

Hou, J. e Zhang, Y. 2003. *Effectively finding relevant web pages from linkage information*. IEEE Trans. Knowl. Data Eng. 15, 4, 940-951.

Internet World Stats, Internet Usage Statistics, 2012.
Disponível em:<http://www.internetworldstats.com/stats.htm> . [25 de Janeiro de 2013].

Jian C., Jian Y., Tung, A.K.H. e Bin L. 2004. *Discovering Web usage patterns by mining cross-transaction association rules*, International Conference on Machine Learning and Cybernetics,Vol.5,Pp.2655-2660.

Links, 2013. Disponível em: < http://www.w3.org/TR/html401/struct/links.html >.
[25 de Janeiro de 2013]

Logging Control In W3C, 1995.
Disponível em: <http://www.w3.org/Daemon/User/Config/Logging.html>.
[25 de Janeiro de 2013]

Khader, D. 2007. *Attribute based group signatures*. Cryptology ePrint Archive, Report 2007/159

Kilfoil M., Ghorbani A., Xing W., Lei Z., Lu J., Zhang J., e Xu X., 2003. *Toward an Adaptive Web: The State of the Art and Science*. In Proc. Of CNSR 2003, pages 119–130, Moncton, Canada.

Kim, H., Pang, S., Je, H., Kim, D. e Bang, S. 2003. *Constructing Support Vector Machine Ensemble*. Pattern Recognition 36: 2757-2767.

Kimball, R. e Richard, M. 2000. *The Data Webhouse Toolkit*: Building the Web-Enabled Data Warehouse.Wiley Hill, T., Lewicki, P. 2006. STATISTICS Methods and Applications. StatSoft.

Kirkland, D., Senator, T., Hayden, J., Dybala, T., Goldberg, H. e Shyr, P. 1999. *The NASD Regulation Advanced Detection System*. AAAI 20(1): Spring, 55-67.

Klosgen, W., e Zytkow, J. 2002. *Handbook of data mining and knowledge discovery*. New York: Oxford University Press.

Kobsa A., Koenemann J., e Pohl W. 2001 *Personalized hypermedia presentation techniques for improving online customer relationships. The Knowledge Engineering Review*, 16(2):111–155.

Koch N.P. 2000 *Software Engineering for Adaptive Hypermedia Systems: Reference Model, Modeling Techniques and Development Process*. Tesis de Doutorado, Ludwig Maximilians University Munich., Munich.

Kopka J., Reves M. e Giertl, J. 2010: *Anomaly Detection Techniques for Adaptive Anomaly Driven Traffic Engineering*, 10th Scientific Conference of Young Researchers (SCYR 2010, Kosice, Slovakia), I. Edition, pp. 254-257, May 19, 2010.

Kosala R. e Blockeel H. 2000. *Web Mining Research: A Survey*, ACM SIGKDD Explorations Newsletter, Volume 2 Issue 1.

Leckie, T. e Yasinsac, A. 2004. *Metadata for Anomaly-Based Security Protocol Attack Deduction*. IEEE Transactions on Knowledge and Data Engineering 16(9): 1157-1168.

Lee, W. e Xiang, D. 2001. *Information-theoretic Measures for Anomaly Detection*. Proc. of 2001 IEEE Symposium on Security and Privacy.

Li, X., Guo, L., e Zhao, Y.E. 2008. *Tag-based Social Interest Discovery*. Proceeding of the 17th International Conference on World Wide Web, 675-684

Magdalena G., Tadeusz L., e Bogdan Trawiński, 2009. *Comparative Analysis of Premises Valuation Models Using KEEL, RapidMiner, and WEKA*. In Proceedings of the 1st International Conference on Computational Collective Intelligence., Ngoc Thanh Nguyen, Ryszard Kowalczyk, and Shyi-Ming Chen (Eds.). Springer-Verlag, Berlin, Heidelberg, 800-812

Maty´aˇs, V. e Cvrˇcek, D. 2004. *On the Role of Contextual Information for Privacy Attacks and Classification*. In Privacy and Security Aspects of Data Mining Workshop, Brighton, UK, November 2004. IEEE ICDM.1.1, 2.3

Melody, Y., Ivory, e M., H. 2002. *Improving Web site design. IEEE Internet Computing 6, 2* (March-April), 56-63.

Melville, P., Mooney, R. J., e Nagarajan, R. 2002*. Content-boosted collaborative filtering for improved recommendations*. In Proceedings of the 18th National Conference on Artificial Intelligence (AAAI02), 187-192.

Mobasher, B., Cooley, R. e Srivastava, J. 1999. *Creating adaptive web sites through usage-based clustering of URLs*, In: Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), 143-153.

Mulvenna M. D., Anand S. S., e Buchner A. G. 2000. *Personalization on the net 23using web mining*, Communications of the ACM, 43, 8 (August), 123–125.

Neill, D. e Moore, A. 2004. *Rapid Detection of Significant Spatial Clusters*. Proc. of SIGKDD04, 256-265

Nielsen J. 1999. *User interface directions for the web*. Communications of the ACM, 42(1):65–72.

Otey, M., Parthasarathy, S., Ghoting, A., Li, G., Narravula, S. e Panda, D. 2003. *Towards NIC-based Intrusion Detection*. Proc. of SIGKDD03, 723-728.

Pazzani, M. J. e Billsus, D. 2007. *Content-based Recommender Systems*. Lecture Notes In Computer Science, The Adaptive Web: Methods and Strategies of Web Personalization, 2007,325-341

Pei, J., Han, J., B., M.A., e Zhu, H. 2000. *Mining access patterns efficiently from web logs*. PADKK 00: Proceedings of the 4th Pacific- Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications. London, UK: Springer-Verlag, 396-407.

Perkowitz M. e Etzioni O. 1998. *Adaptive web sites: Automatically synthesizing web pages*. In Proceedings of the Fifteenth National Conference on Artificial Intelligence, Madison, WI, USA.

Perkowitz, M. e Etzioni, O. 2000: *Towards adaptive Web sites: Conceptual framework and case study*, Artificial Intelligence 118(1-2)

Phua, C., Alahakoon, D. e Lee, V. 2004. *Minority Report in Fraud Detection: Classification of Skewed Data*, SIGKDD Explorations 6(1): 50-59.

Phua, C., Lee, V.C.S., Smith-Miles, K., e Gayler, R. 2010 *A Comprehensive Survey of Data Mining-based Fraud Detection Research*. In Proceedings of CoRR.

Piatetsky-Shapiro, G. 1991. *Knowledge discovery in real databases*: a report on the IJCAI-89 workshop. AI Magazine, 11(5):68–70

Pierrakos, D., Paliouras, G., Papatheodorou, C., Spyropoulos ,C. D. 2003. *Web Usage Mining as a Tool for Personalization:A Survey. User Modeling and User-Adapted Interaction*, 13(4):311–372.

Ribeiro-Neto, B., Cristo, M., Golgher, P., e Moura, E. 2005. *Impedance coupling in content-targeted advertising*. In Proceedings of the 28th International ACM SIGIR Conference on Research and Development of Information Retrieval. Salvador, Brazil, pp. 496–503, 2005.

Rosenstein, M. 2003. *What is Actually Taking Place in Web Sites*: E-Commerce Lessons from Web Server Logs. ACM Conference on Electronic Commerce.

Sarwar, B., Karypis, G., Konstan, J., e Riedl, J. 2000. *Analysis of recommender algorithms for ecommerce*. In Proceedings of the ACM Electronic Commerce Conference,158-167.

Schafer, J., Ben, Konstan, J., e Riedl, J. 2000. *Electronic commerce recommender applications.* Journal of Data Mining and Knowledge Discovery 5, 1-2, 115-152.

Senator, T., Goldberg, H., Wooton, J., Cottini, M., Khan, U., Klinger, C., Llamas, W., Marrone, M. e Wong, R. 1995. *The Financial Crimes Enforcement Network AI System* (FAIS). AAAI 16(4): Winter, 21-39

Sen, S., Vig, J., e Riedl, J. 2009. *Tagommenders: Connecting Users to Items through Tags*. Proceedings of the 18th International Conference on World Wide Web, 671-680

Shamir A. 1984. *Identity-based cryptosystems and signature schemes*. In CRYPTO'84, pages 47-53,1984.

Song, X.; Wu, M., Jermaine, C., e Ranka S. 2007. *Conditional anomaly detection*. IEEE Transactions on Knowledge and Data Engineering, 19(5):631–645.

Spiliopoulou, M. 1999. *The laborious way from data mining to web mining*. International Journal of Comp. Sys., Sci. & Eng., Special Issue on Semantics of the Web 14, 2 (Mar), 113-126.

Spiliopoulou, M. 2000. *Web usage mining for site evaluation: Making a site better fit its users*. Communications of ACM 43, 8 (August), 127-134.

Spiliopoulou, M., Faulstich, L. C. e Wilkler, K.: 1999, *A data miner analyzing the navigational behavior of Web users*, In: Proceedings of the Workshop on Machine Learning in User Modeling of the ACAI99, Chania, Greece, 54-64.

Srivastava, J., Cooley, R., Deshpande, M., e Tan, P. 2000. *Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations* 1, 2, 12-23.

Steinwart, I., Hush, D. e Scovel, C. 2005 *A classification framework for anomaly detection*. Journal of Machine Learning Research, 6:211–232, 2005.

Universal Resource Identifiers, 2013.
Disponível em: <http://www.w3.org/Addressing/URL/URI_Overview.html>.
[25 de Janeiro de 2013].

Vander Wal, Thomas. *Folksonomy*, 2007.
Disponível em:< http://www.vanderwal.net/folksonomy.html> . [25 de Janeiro de 2013].

Virvou M., Tsiriga V. e Moundridou M. 2001. *Adaptive navigation support in a web-based software engineering course*. In Proceedings of the 2nd International Conference on Technology in Teaching and Learning in Higher Education, pages 333–338.

Wang, H., Fan, W., Yu, P. e Han, J. 2003. *Mining Concept-Drifting Data Streams Using Ensemble Classifiers*. Proc. Of SIGKDD03, 226-235.

Wang J., Huang Y., Wu G., Zhang F. 1999 Web mining: *knowledge discovery on the Web*. Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference - on Volume 2, Page(s):137 - 141 vol.2 - 12-15 Oct. 1999

Wang, Y. 2000. *Web Mining and Knowledge Discovery of Usage Patterns*. In: Proceedings of the 1st International Web-age Information Management Conference (WAIM'2000), pp. 227-232, Shanghai, China, Jun.

Weng, L.T., Xu, Y., Li, Y., e Nayak, R. 2008. *Web Information Recommendation Making based on Item Taxonomy*. Proceedings of the Tenth International Conference on Enterprise Information Systems, 20-28

Witten, I.H. e Frank, E., 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann.

Wong, W., Moore, A., Cooper, G. e Wagner, M. 2003. *Bayesian Network Anomaly Pattern Detection for Detecting Disease Outbreaks*. Proc. of ICML03, 217-223.

World Wide Web Size, the size of the World Wide Web, 2013.
Disponível em: < http://www.worldwidewebsize.com>. [25 de Janeiro de 2013]

Yamanishi, K., Takeuchi, J., Williams, G. & Milne, P. 2004. *On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms*. Data Mining and Knowledge Discovery 8: 275-300.

Yeung, D., Ding, e Yuxin. 2002. *User Profiling for Intrusion Detection Using Dynamic and Static Behavioral Models*. In M.-S. Chen, P.S. Yu, and B. Liu, editors, 6th Pacific-Asia Conference, PAKDD 2002, Taipel, Taiwan, Springer-Verlag 2336, 1 (May), 25-30.

Ziegler, C. N., Lausen, G., e Schmidt-Thieme, L. 2004. *Taxonomy-driven Computation of Product Recommendations*. The 13th ACM International Conference on Information and Knowledge Management, 406-415