



Sistemas de personalização em comércio electrónico

© Artur Marques, 2005

Índice

Sumário e introdução.....	3
Categorias de sistemas de personalização: conceitos e exemplos	4
Sistema de personalização: componentes e exemplos	6
Sistema de personalização: dados, processos, arquitectura, exemplos	9
Ética e privacidade	12
Conclusões.....	12
Referências:.....	13

Sumário e introdução

As tecnologias para comércio electrónico na Internet (TCEI) podem permitir um relacionamento personalizado, ao extremo de um-para-um, entre web sites e os seus utilizadores.

Disponer de informação, assim detalhada, sobre os utilizadores é valorizá-los, desde que aconteça a articulação devida entre modelos de negócio, tecnologia e marketing (Krishnamurthy 2002). Neste sentido, a Internet pode ser vista como um meio desmassificado, pois a tecnologia – ao contrário do que se passa, por exemplo, em Televisão e Rádio – tem o potencial de *singularizar* entre «espectadores» ou, mais correctamente no caso de experiências de comércio electrónico, entre «actores» (Laurel 1993).

A personalização da experiência é um processo baseado em dados, que determinam decisões sobre preferências (Mobasher et al., 2000b). Técnicas como **filtragem baseada em regras**, vão explicitamente adaptando a experiência ao «actor», por exemplo, por questionários. Técnicas como **filtragem colaborativa**, baseiam a personalização, nas preferências de outras pessoas tidas como do mesmo segmento. Técnicas de **personalização implícita**, ou não intrusiva, são comumente implementadas por agentes inteligentes.

Este documento estuda um sistema de personalização genérico, materializado em sistemas concretos, referidos oportunamente.

Quando levada à escala de centenas ou de milhares de utilizadores, a personalização diz-se «em massa» (Mobasher et al., 2000a). O seu objectivo derradeiro é aproximar «actores» e fornecedores da experiência de comércio electrónico, desde que isso se traduza em valor para ambos.

Apresenta-se uma visão **por componentes**, que considera o sistema de personalização genérico dividido nas metades **offline** – dedicada à preparação e exploração de dados – e **online**, correspondente a um motor de recomendações.

Apresenta-se também uma visão arquitectural, baseada em dados e em processos: os dados podem ser de **conteúdo**, de **estrutura**, de **utilização** e de **perfil** (Srivastava et al., 2000), entretanto envolvidos em processos de **colecta** e de **análise**, para a **gestão** e para a **publicação** de conteúdos (Eirinaki and Vazirgiannis 2003).

Por fim, abordam-se questões de ética e de privacidade, inseparáveis da tecnologia e dos modelos de negócios, principalmente nas situações de adaptação implícita, que podem considerar-se situações de espionagem dissimulada (Martin et al., 2003).

Categorias de sistemas de personalização: conceitos e exemplos

Embora muito de uma experiência de comércio «tradicional» possa ser transposto para uma experiência estritamente electrónica, incluindo sons, cheiros (ver digiscents.com) e sentido de comunidade, no limite, porque o todo pode ser mais do que a soma das partes, a experiência electrónica deverá procurar balançar a sua potencial desvantagem sensorial global, criando laços com o utilizador pelo fomento da sua participação e interacção: ao arquitectar-se um sítio de Comércio Electrónico, deverá pensar-se nos utilizadores como «actores» e não como meros observadores passivos (Laurel 1993): se os actores, por definição, actuam, o ambiente deve adaptar-se-lhes.

A personalização em web sites (Web Personalization) consiste nas acções que apropriam a experiência a um «actor», ou conjunto de actores, em particular.

Nos sistemas com **filtragem baseada em regras**, a personalização faz-se respeitando perfis de utilizadores e/ou históricos de sessão, recolhidos explicitamente; isto é, por entrada «manual» dos próprios visitantes do web site, por exemplo, via formulários.

Um exemplo muito simples, onde não se mantêm perfis estáticos, encontra-se em worldsbestbars.com, onde o visitante pode condicionar o conteúdo, escolhendo tão somente o tipo de estabelecimento que lhe interessa: com ou sem restaurante.

Um exemplo mais elaborado, que pode envolver a manutenção de perfis, mas que os dispensa para algumas filtrações, encontra-se em restaurant.com, onde pode procurar-se por restaurantes nos EUA, numa sucessão de etapas, que começam por condicionar o conteúdo a um só Estado do país ou a uma distância máxima de certo código postal. Os refinamentos possíveis incluem ainda a região, a cidade e o tipo de cozinha.

Nos sistemas com **filtragem colaborativa**, as preferências explicitamente expressas por utilizadores, são tratadas por um **motor de correlações**, que adequa os conteúdos, consoante os perfis (Mobasher et al., 2000a).

Um dos maiores exemplos de personalização em massa com filtragem colaborativa, é Amazon.com.

Um outro exemplo é musiciansfriend.com, onde se podem comprar equipamentos musicais. O conteúdo do web site adapta-se às preferências que o visitante manifesta em perfil: quem procura guitarras não deverá ter uma página de entrada dedicada a teclados ou a material de gravação em estúdio.

Muitas lojas electrónicas, incluindo os exemplos acima, *não* fazem o desenvolvimento do seu próprio sistema de filtragem colaborativa, optando por recorrer a soluções comerciais, ao contrário do que acontece para as filtrações baseadas em regras.

A consequência é que (partes de) web sites com personalização por filtragem baseada em regras, tendem a distinguir-se claramente uns dos outros, na aplicação desse mecanismo, pela especificidade das suas regras, enquanto que a expressão da filtragem colaborativa tende a ser mais homogénea.

Insistindo nos exemplos Amazon.com e musiciansfriend.com, a forma como se classificam os itens e a própria expressão da classificação que lhes está atribuída, são muito próximos, porque na base está o mesmo produto: o Net Perceptions NetP. Outros exemplos de empresas recorrendo ao NetP são a 3M, JC Penney, Great Universal Stores e a Half.com, entretanto adquirida pela eBay.

Nos sistemas com **filtragem baseada em conteúdos**, os perfis são recolhidos explícita ou **implicitamente**.

As técnicas para personalização implícita, pela sua transparência, têm um potencial singular para a recolha de informação sem intrusão, mas podem levantar questões éticas, pelo que, num cenário de auto-regulação, recomendam que se denuncie a sua utilização (Baron 2001). Tipicamente, essa denúncia acontece nas condições de utilização dos web sites, mas a leitura pode não ser fácil. Por exemplo, no caso de A9.com, as condições de utilização referem patentes que Amazon.com conseguiu, relativamente à extracção de preferências; uma das patentes diz respeito a tecnologia que permite a recolha de informação sobre as pessoas às quais se oferecem presentes, pelo que já não é necessário ser-se utilizador directo de certos serviços, para que, anónimas ou não, pessoas que nunca os utilizaram passem a constar da lista de perfis.

A Net Perceptions tem evoluído o seu produto de forma a tornar possível a classificação implícita de artigos; isto ilustra a fronteira ténue entre as diferentes categorias de filtragem.

Uma vantagem significativa da recolha implícita de informação, a partir de dados de utilização, como a sequência de acções do utilizador (clickstream), processados por algoritmos próprios (web mining algorithms), é a sua maior actualidade e menor subjectividade, relativamente a perfis estáticos, que correspondem a uma fotografia de preferências, algures no tempo, da autoria do próprio fotografado, logo dificilmente neutra. Um outro problema com a informação estática, é a sua natureza condicionante; por exemplo: lá porque alguém tem um histórico de aquisição de filmes franceses, isso não significa que não esteja receptivo(a) a outros filmes, pelo que assumir a linearidade do seu comportamento pode ser condicionante.

Sistema de personalização: componentes e exemplos

A forma geral de um sistema para a personalização em web sites, implica a modelação dos objectos do web site (produtos e páginas, por exemplo), a categorização desses objectos, a modelação de utilizadores, a categorização desses utilizadores e o *matching* de objectos com utilizadores; isto é, fazer o encontro dos conteúdos certos, com os utilizadores certos.

Por modelação ou desenho, entenda-se a identificação dos atributos relevantes em classes de objectos e de utilizadores; por categorização, entenda-se a segmentação, consoante os valores dos atributos, dos objectos e dos utilizadores.

A personalização ocorre suportada por dois componentes: **offline** e **online** (Mobasher et al., 2000a).

A nível **offline**, (1) **preparam-se os dados** e (2) **exploram-se esses dados**.

Preparar os dados é produzir um ficheiro de sessão, que regista a sequência de páginas visitadas.

Explorar os dados é descobrir utilizadores próximos (perfis de utilizadores, identificados por análise de sequências de sessões), páginas próximas (identificadas por análise de sequências de páginas visitadas), padrões de utilização (que se traduzem em preferências de utilização) e regras que permitam associar objectos e utilizadores.

Os padrões identificados na exploração dos dados, conduzem à adequação do conteúdo ao utilizador, pelo componente online. A forma particular como essa adequação se expressa, depende de sistema para sistema, mas um cenário habitual é a produção de sugestões de produtos e/ou conteúdos, dentro do mesmo web site.

Assim, o requisito mínimo para um sistema de personalização é a capacidade de distinguir entre sessões, considerando-se uma sessão o conjunto das interacções entre uma aplicação cliente (por exemplo, um browser) e o servidor. Quando a aplicação cliente é encerrada, a sessão termina.

No cenário ideal, cada registo de sessão permitirá identificar quem acedeu, a que páginas acedeu, por que ordem foram acedidas e a duração de cada acesso. Na prática, estes dados podem não estar todos disponíveis, ao menos com fidelidade; por exemplo, estão generalizadas ferramentas para a navegação com relativo anonimato que, no mínimo, eliminam o «http referrer» (em rigor deveria escrever-se referer, mas este erro propagou-se até ao presente), que regista o URL que conduziu ao URL actual, dificultando a identificação de sequências.

A identificação de perfis de utilização, toma como input o ficheiro de sessão e sujeita-o a algoritmos de exploração (web mining). Alguns desde algoritmos são de tal modo singulares que a ideia que lhes está subjacente consegue protecção por patente. Assim se compreende que o orçamento para R&D (Research and Development), em empresas para as quais a Internet assume um papel central, contemple estas actividades.

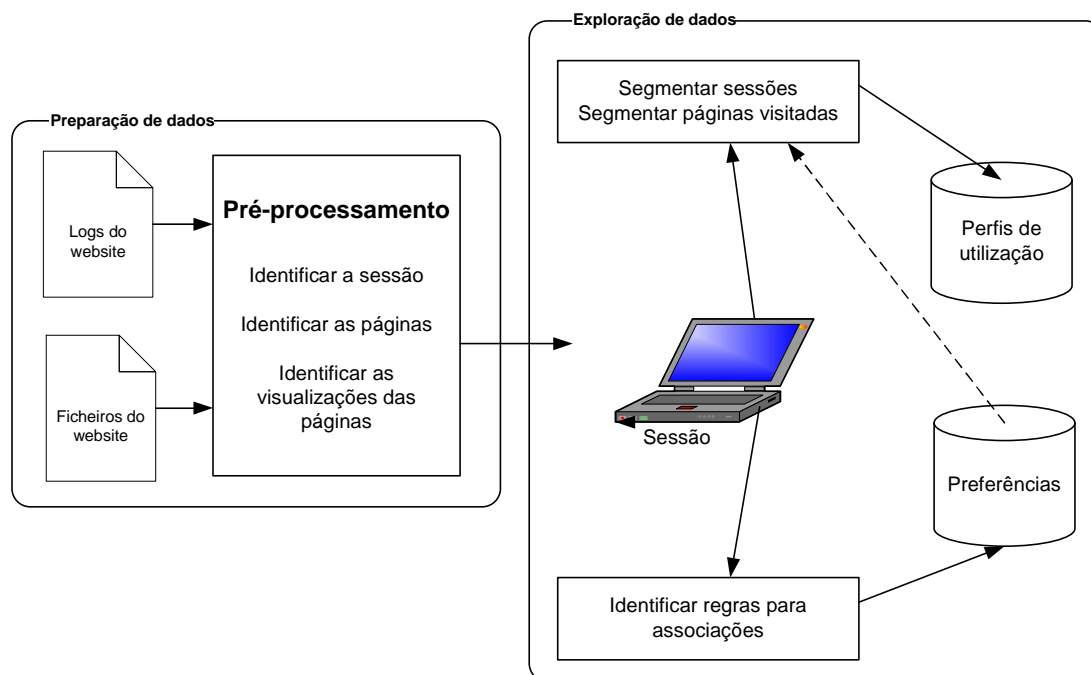


Figura 1 - Componente offline de um sistema de personalização genérico.

Uma forma de responder à informação produzida na fase de exploração de dados, é através de assistentes que proponham mudanças aos web masters. Um caso particular destes assistentes dá pelo nome de **IndexFinder** (Perkowitz and Etzioni 2000) e sugere páginas candidatas a índices de conteúdos, em função das navegações que os utilizadores efectivamente fazem. A premissa do IndexFinder é que muitas vezes são apresentadas hierarquias de navegação, que não ajudam a navegar-se até àquilo que se procura... O IndexFinder implementa um algoritmo de «síntese de índices» que categoriza no mesmo tópico as páginas que (julga que) certo visitante considera próximas.

Outro trabalho relacionado, com resultados transpostos para a componente online do sistema de personalização, é o **Web Utilization Miner** (Spiliopoulou 2000), que objectiva a descoberta de trajectos de navegação «interessantes», conforme medidas estatísticas, sintácticas e comportamentais. Alguns exemplos: (1) se um visitante esgota sempre, primeiro, os caminhos em profundidade, é provavelmente um robot e o seu comportamento não deve ser considerado para efeitos de qualquer perfil; (2) se uma certa sequência de navegação é percorrida por uma fracção estatisticamente significativa de utilizadores, é provável que seja «interessante», para essa fracção; (3) se ao longo de uma certa sequência de páginas visitadas vai aumentando o contador de ocorrências da palavra X, é provável que o assunto a que X diz respeito seja de interesse para esses utilizadores.

A identificação de padrões e de preferências não é, por si só, suficiente para a personalização da experiência. O importante é a extrapolação de perfis válidos, que fundamentem recomendações online.

O conjunto dos perfis deverá, idealmente, permitir (1) capturar interesses comuns; isto é, regiões de intersecção entre perfis distintos; (2) distinguir entre visitas a páginas, em termos da sua significância por perfil; e (3) ter uma representação uniforme, para facilitar a implementação do motor de recomendações.

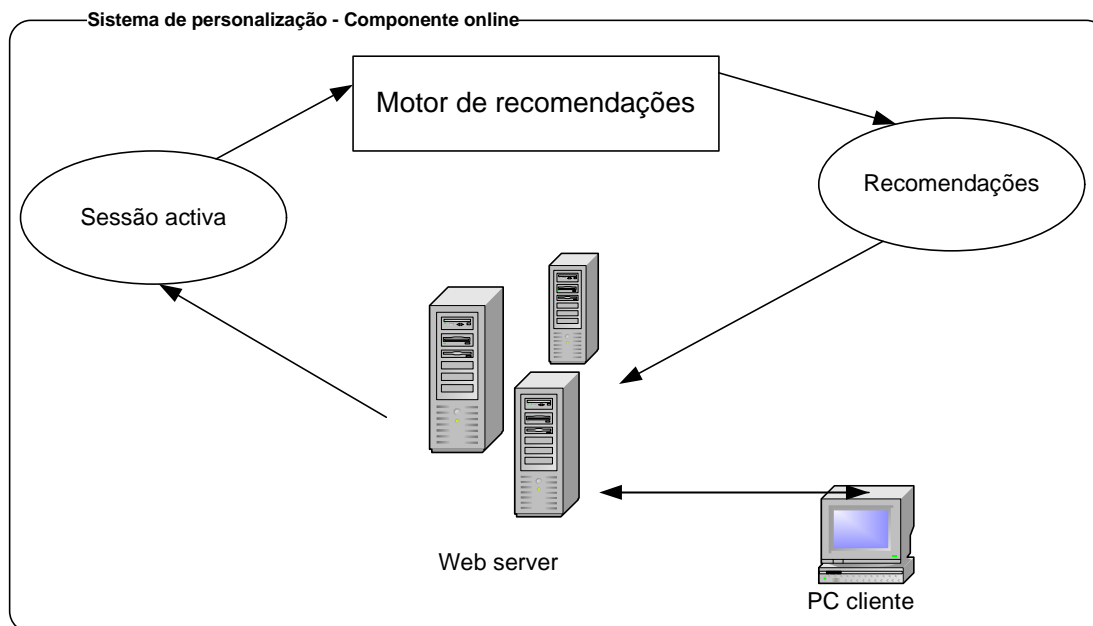


Figura 2 - Componente online de um sistema de personalização genérico.

Numa perspectiva funcional, o motor de recomendações consiste numa ferramenta que, perante informação de sessão pós-processada, calcula um conjunto de recomendações para essa sessão. O conjunto de recomendações consiste em objectos (conteúdos) que, espera-se, satisfarão o utilizador. Por outras palavras, compete ao motor de recomendações fazer o encontro (matching) entre o utilizador corrente e um dos perfis agregados.

Sistema de personalização: dados, processos, arquitectura, exemplos

De notar que a designação de «componente offline» pretende significar que o processamento de objectos e a determinação de perfis é uma actividade que não acontece em «tempo real», ao contrário da manifestação das recomendações, que deverá ser célere o suficiente para não perturbar a experiência do utilizador.

A componente offline trabalha sobre diferentes tipos de dados, classificáveis em quatro categorias (Srivastava et al., 2000):

È **Dados de conteúdo** – correspondentes ao que o utilizador efectivamente acede; por exemplo na forma de textos, sons, imagens e vídeos.

È **Dados de estrutura** – correspondentes à organização dos dados de conteúdo; por exemplo na forma de marcas HTML ou, a um nível mais abstracto, na forma de objectos, mantidos por algum sistema de hipermedia, como hiperligações, âncoras e páginas.

È **Dados de utilização** – representativos daquilo que o web site vai registando explicitamente, tipicamente com origem em logs feitos pelo servidor de http; por exemplo indicando o endereço IP dos visitantes, os seus URLs de origem e a hora de início dos acessos.

È **Dados de perfil** – dizendo respeito aos utilizadores do web site, obtidos de forma explícita (por exemplo, por formulários) ou implícita (por exemplo, pelos processos que inferem preferências e perfis), na fase de exploração de dados.

O papel de cada uma destas categorias de dados, pode ser sistematizado por processos (Eirinaki and Vazirgiannis 2003):

È **Colecta de perfis de utilizador** (user profiling) – correspondente à obtenção de dados pessoais que servirão para a edificação de perfis.

Um perfil diz-se **estático** quando, depois de obtido, não volta a ser actualizado – estes perfis são adequados para alguma informação demográfica, como data de nascimento, nome e números sociais, mas não interessam para a representação de preferências, conforme já justificado.

Um perfil **dinâmico** é actualizado com frequência, por métodos explícitos ou implícitos. Os utilizadores são agregados em grupos (perfis agregados), mas há situações em que se justificam perfis individuais, dependendo da escala e da natureza do relacionamento. Tipicamente, a personalização «em massa» faz sentido com perfis agregados e esse é o cenário em que também se justificam as técnicas de exploração de dados, para a segmentação.

Uma tecnologia comum para identificação unívoca de utilizadores é a utilização de **cookies** (Kristol 2001): frases, correspondentes a uma http header, que o servidor envia ao cliente, para que este a carregue em memória, e que pode conter parâmetros que servem para manter informação de estado sobre a sessão.

Por exemplo, com a frase

Set-Cookie: visitante=artur; version=1; domain=.books.com; path=/books/ecommerce/

um servidor assinala, enquanto a sessão persistir, que do lado do cliente está o visitante *artur*, acedendo à path */books/ecommerce* do web site. Só web sites no domínio *.books.com* poderão

ler/escrever a frase, pela versão 1.0 do sistema de cookies, sendo este o único atributo obrigatório na comunicação.

A esta frase, um cliente aceitante responderia

Cookie: \$visitante=artur; \$version=1; \$path=/books/ecommerce/; \$domain=.books.com

Na forma geral, o servidor escreve atributos com uma sintaxe *nome=valor*, e o cliente responde com uma sintaxe *\$nome=valor*.

Desde o seu surgimento, em 1995, que o sistema de cookies incorpora mecanismos de segurança razoáveis, como (1) a exigência de dois *dots* no nome de domínio, evitando fraudes do estilo *.com, (2) a recusa de acesso à cookie a domínios diferentes do original, e (3) a recusa de acesso à cookie, a localizações remotas diferentes da referida em path.

Para além das cookies, o protocolo *identd* (RFC 1413) e o endereço IP, permitem identificar o cliente, mas com limitações, pois é frequente um só IP acolher muitos utilizadores e mesmo um só computador manter várias sessões.

È **Análise de registos e exploração de dados** (logs & mining) – quando os dados reunidos são sujeitos a exploração (data mining) para (1) extracção de informação estatisticamente relevante e identificação de padrões de utilização; (2) segmentação de utilizadores, conforme as suas preferências; e (3) descoberta de correlações (as regras para associação) entre objectos do web site e grupos de utilizadores.

A forma mais simples de um registo mantido pelo servidor de http é

computador-remoto rfc931 utilizador data pedido estado bytes

onde *computador-remoto* representa o nome ou o IP do cliente; *rfc931* representa o nome com que o utilizador se autentica, quando isso é necessário; *data* é a data e a hora, no servidor, aquando do pedido; *pedido* é a frase exacta com que o cliente fez o acesso; *estado* é o retorno http ao pedido; e *bytes* é a quantidade de informação respondida.

O formato W3Clog da W3C inclui mais campos, entre os quais se destacam *referrer* (o URL de proveniência do cliente) e *user_agent* (uma string que é suposto identificar o software cliente).

Durante a exploração dos registos, o objectivo é identificar padrões, por técnicas como (1) análise estatística simples, (2) regras de associação e (3) descoberta sequencial. As regras de associação procuram correlações entre páginas acedidas numa mesma sessão e a «descoberta sequencial» é uma variante que incorpora a noção de tempo.

O mercado tem diversas soluções, algumas gratuitas e em regime de *open source*, para o processo de análise de registos e exploração de dados: weblog.com, STstat e Follow2.

è **Gestão de conteúdos** – classificação dos objectos do web site, para que o seu encontro (matching) com os clientes da informação possa acontecer a um nível semântico, mais natural do que o nível sintáctico.

è **Publicação dos conteúdos** (web site publishing) – que objectiva uniformidade na publicação dos objectos do web site, independentemente da sua proveniência, que pode ser heterogénea, em termos de localização física e/ou de tecnologia de suporte, como no caso dos portais.

Um exemplo de produto comercial, disponível para download, é o Microsoft [SharePoint Portal Server](#).

è **Aquisição e procura de informação** – no caso de web sites que dependam de informação externa, como motores de procura, genéricos ou especializados, os conteúdos externos deverão ser categorizados de forma compatível com o processo de publicação, para garantir a uniformidade

Uma representação das relações destes dados e processos, resume a arquitectura do sistema de personalização genérico.

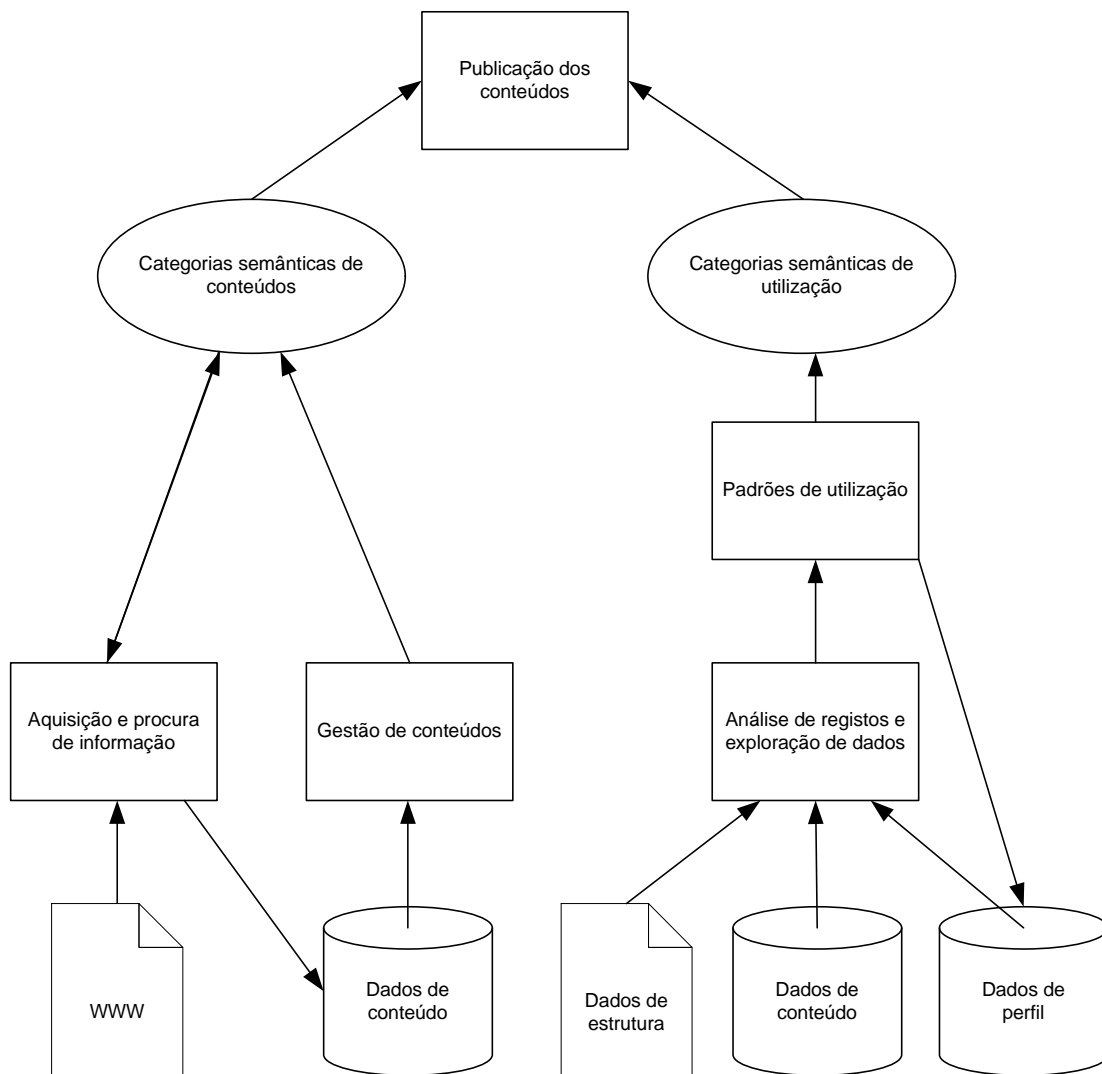


Figura 3 - Arquitectura de um sistema de personalização genérico

Ética e privacidade

A personalização em massa tem associada questões de ética e de privacidade (and Warren 2000); assim, o desenvolvimento e a prática destes sistemas não deve estar restrita a um vértice tecnológico, sob pena da criação de valor poder transformar-se em quebra de valor, como aconteceu com [DoubleClick](#) (NASDAQ:DCLK) quando, em 2000, depois de adquirir a empresa [Abacus](#), planeou complementar a informação anónima que conseguia online, com informações de compras por cartão de crédito, provenientes da Abacus Direct Database (Baron 2001). Apesar do vazio legal, a reacção a estes planos foi de tal modo negativa e mereceu tal divulgação, que a empresa comprometeu-se a não avançar, criando para si mesma um problema de integração imperfeita das suas unidades de negócio.

Um outro caso, que em 2005 começou a receber divulgação na óptica da potencial violação de privacidade (Linn 2005), é a articulação que [Amazon.com](#) faz entre as **preferências** que conhece dos consumidores que frequentam a loja, com as **procuras** que os mesmos executam utilizando o motor [A9.com](#), que ficam todas registadas e associadas ao perfil de utilizador, e os seus **objectivos** de vida, conforme admitidos em [43things.com](#). De notar que A9.com e 43things.com são propriedade de Amazon.com.

No que toca a tecnologias específicas, os receios «massificados» de violação de privacidade começaram com as cookies, novamente no caso da DoubleClick, na sequência do qual, 17 estados passaram a aplicar legislação relativa à privacidade na Internet (Baron 2001). Nos restantes estados funciona a chamada auto-regulação da indústria, conforme defendida por grupos representativos, como a Online Privacy Alliance (AOL, AT&T, Dell, IBM, Microsoft, Network Solutions, Time Warner, and Visa U.S.A.).

Com a emergência de novas técnicas de observação dos utilizadores, passaram a designar-se de web bugs os elementos HTML que pretendem passar despercebidos (como imagens com 1 pixel de área...) e cujo objectivo é precisamente a recolha de dados de comportamento. Não é fácil automatizar a detecção destes elementos, tipicamente associados a cookies com origem em web sites diferentes daquele que está a ser visitado, pois a concepção de páginas em HTML recorre-lhes com frequência, de forma a conseguir certos efeitos estéticos (Martin et al., 2003).

Uma a solução está na identificação das strings abusivas, correspondentes às web bugs: um software open source especialmente configurável para esse efeito, está disponível em [privoxy.org](#).

Conclusões

Os sistemas de personalização têm potencial para conferir valor às experiências de comércio electrónico. Estes sistemas alicerçam-se hoje sobre teoria madura, conforme as categorizações, componentes e arquitectura genérica apresentadas, pretendem ilustrar.

Ainda assim, os algoritmos envolvidos na recolha implícita e na exploração de dados são uma área de intensa investigação académica e industrial, que não pode abstrair-se de questões éticas e de privacidade, principalmente num contexto em que a indústria assumiu a auto-regulação.

Referências:

- Baron, D. (2001) Stanford University.
- Eirinaki, M. and Vazirgiannis, M. (2003) "Web mining for web personalization", In: *13th international world wide web conference*
- Krishnamurthy, S. (2002) *E-Commerce Management: Text and Cases*, South-Western College Pub.
- Kristol, D. (2001) In *ACM Transactions on Internet Technology*, Vol. **1**, pp. 151-198.
- Laurel, B. (1993) *Computers as Theatre*, Addison Wesley.
- Linn, A. (2005) *Amazon.com Knows, Predicts Shopping Habits*
http://biz.yahoo.com/ap/050328/your_amazon_dossier.html?.v=2
- Martin, D., Wu, H. and Alsaïd, A. (2003) In *Communications of the ACM*, Vol. **46**, pp. 258-264.
- Mobasher, B., Cooley, R. and Srivastava, J. (2000a) In *Communications of the ACM*, Vol. **43**, pp. 142-151.
- Mobasher, B., Honghua Dai, T. L., Nakagawa, M., Sun, Y. and Wiltshire, J. (2000b) "Discovery of Aggregate Usage Profiles for Web Personalization", In: *ACM Conference on Knowledge Discovery and Data Mining*, Boston, EUA
- Perkowitz, M. and Etzioni, O. (2000) In *Communications of the ACM*, Vol. **43**, pp. 152-158.
- S. L. and Warren, M. (2000) "Ethics and Electronic Commerce", In: *CRIPTS* 56-59
- Spiliopoulou, M. (2000) In *Communications of the ACM*, Vol. **43**, pp. 127-134.
- Srivastava, J., Cooley, R., Deshpande, M. and Tan, P.-N. (2000) "Web usage mining: discovery and applications of usage patterns from web data", In: *SIGKDD Explorations*