Mestrado em Engenharia Informática

Dissertação

Relatório Final

Selecção de Informação para Sistemas de Transportes

Ricardo da Conceição Domingues

rcd@student.dei.uc.pt

Orientador:

Carlos Bento

Nuno Gil

Data: 20 de Julho de 2012

FCTUC DEPARTAMENTO
DE ENGENHARIA INFORMÁTICA
FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Resumo

Estamos a viver numa era em que o elevado crescimento tecnológico tem proporcionado um grande aumento da quantidade de informação disponível. Desta forma, os dispositivos que recomendem informações devem utilizar mecanismos que permitam classificar e seleccionar a informação de forma contextualizada, focada no público-alvo e nas suas necessidades.

Como resposta à necessidade de selecção de informação surgem os métodos de atenção selectiva artificial, que têm como objectivo filtrar as informações que apresentam maior interesse para cada utilizador.

Nesta dissertação é apresentado um sistema de recomendação baseado nos conceitos de relevância, novidade e diversidade.

Palavras-Chave

Atenção selectiva, Relevância, Novidade, Diversidade, Sistemas de Recomendação.

Índice

Lista de Figuras

Lista de Tabelas

Capítulo 1
Introdução

O trabalho desenvolvido enquadra-se no âmbito do projecto TICE Mobilidade[1], que tem como objectivo disponibilizar produtos tecnológicos e inovadores de forma a melhorar a mobilidade dos cidadãos em zonas urbanas.

Uma das vertentes deste projecto visa disponibilizar um serviço de recomendação de informação personalizada e contextualizada com o público-alvo e as suas necessidades. Podendo este serviço pode ser utilizado por painéis digitais de informação a bordo de sistemas de transportes ou aplicações turísticas que recomendem pontos de interesse.

O projecto TICE Mobilidade encontra-se estruturado em vários PPS's (Processos Produtos e Serviços), que fornecem serviços capazes de tornar mais eficiente a rede de mobilidade urbana. Sendo que o serviço de recomendação de informação desenvolvido pertence concretamente ao PPS2 - SEMA[2].

Um dos principais desafios dos sistemas de recomendação de informação está relacionado com a capacidade que estes apresentam no processo de sugerir informações que sejam do interesse dos utilizadores.

Como os seres humanos não são capazes de processar todas as informações recolhidas pelos sentidos, ao longo do seu processo evolutivo, desenvolveram a capacidade de atenção selectiva [1]. Esta capacidade permite que os seres humanos seleccionem as informações do meio envolvente que consideram mais relevantes. No entanto, o elevado volume de informação a que os seres humanos estão sujeitos no seu dia-a-dia, pode comprometer o seu desempenho. Nesta vertente é necessário que os sistemas de recomendação implementem técnicas que permitam evitar sobrecarregar os utilizadores com informação redundante e irrelevante.

1.1 Objectivos

O principal objectivo consistiu em desenvolver um serviço de recomendação de informação, que dado um conjunto de recursos, permita seleccionar aqueles que são mais uteis para um determinado utilizador, inserido num dado contexto e perante determinadas necessidades.

Os recursos a recomendar podem ser pontos de interesse (POIs), como por exemplo: restaurantes, museus, hotéis etc. Contudo o serviço de recomendação deve ser o mais genérico possível, de forma a permitir a adição de recursos com diferentes origens.

De forma a permitir a interacção dos utilizadores com o serviço de recomendação também existiu a necessidade de desenvolver uma aplicação cliente. Esta aplicação cliente também foi utilizada para avaliar o serviço de recomendação num cenário com utilizadores reais.

[1] http://www.tice.pt/projectos/projecto.aspx

[2] Selecção de informação baseada em mecanismos de atenção selectiva, enriquecimento semântico e interoperabilidade semântica

1.2 Abordagem

A abordagem a utilizar no processo de recomendação baseia-se na aplicação do conceito de atenção selectiva, que está fortemente ligado aos conceitos de relevância, novidade e diversidade. Numa primeira fase, são filtrados os recursos mais relevantes para o utilizador, tendo em conta as características de cada recurso, e as necessidades, perfil e contexto do utilizador.

Com o objectivo de excluir os recursos com os quais o utilizador já se encontra familiarizado, é utilizado um modelo computacional do conceito de novidade. Este modelo permite filtrar os recursos que apresentam maior novidade para o utilizador, com base no histórico de interacções que o utilizador já teve com os recursos.

Com o objectivo de excluir recursos repetidos ou muito semelhantes, é utilizado um filtro que recorra ao conceito de diversidade. O objectivo é seleccionar os recursos que apresentam maior diversidade entre si, de forma a não sobrecarregar o utilizador com informação repetida ou muito semelhante.

Através da aplicação destas três fases de filtragem, é possível seleccionar um conjunto de recursos diversos que são relevantes e apresentam novidade para o utilizador em questão.

1.3 Estrutura da Dissertação

Este documento está estruturado em seis capítulos. No Capítulo 2 é descrito o estado da arte. No Capítulo 3 é feita uma análise da arquitectura do projecto PPS2 - SEMA. No Capítulo 4 são detalhados as técnicas de filtragem, as estruturas de dados e o processo de recomendação. No Capítulo 5 são apresentados os resultados que avaliam o sistema de recomendação. Por fim, são apresentadas as conclusões no Capítulo 6.

Capítulo 2
Estado da Arte

Actualmente existem um grande número de aplicações e dispositivos que têm como objectivo recomendar informações aos utilizadores. Geralmente estes sistemas permitem recomendar determinados informações, em função do perfil, contexto e necessidades dos utilizadores. Na secção 2.1 são apresentados alguns sistemas de recomendação de informação.

O elevado crescimento tecnológico tem proporcionado um aumento drástico da quantidade de informação disponível. Desta forma, os dispositivos que recomendem informações devem utilizar mecanismos que permitam classificar e seleccionar a informação de forma contextualizada, focada no público-alvo e nas suas necessidades.

Várias abordagens têm sido propostas, indo a maior parte no sentido de diminuir a carga cognitiva. Desta forma, reduzir a quantidade de informação apresentada ao utilizador de acordo com as suas necessidades e conhecimento prévio é um desafio dos sistemas de recomendação.

2.1 Sistemas de Recomendação

O primeiro sistema de recomendação foi o Tapestry [2], desenvolvido por investigadores da Xerox Palo Alto Research Center, com a motivação de resolver a sobrecarga de e-mails existente no centro de pesquisa.

O Tapestry tinha como objectivo filtrar e arquivar os e-mails que chegavam diariamente, de acordo com as opiniões dadas pelas pessoas que já efectuaram a sua leitura. Desta forma, este sistema utilizava uma abordagem colaborativa, baseada nas opiniões da comunidade.

Na literatura é possível encontrar várias definições para sistema de recomendação. De acordo com Burke [3], um sistema de recomendação é um sistema que guia o utilizador de forma personalizada para objectos úteis ou de interesse com base num largo espaço de opções possíveis.

Para Vozalis e Margaritis [4] os sistemas de recomendação foram definidos com o objectivo de lidar com o problema da sobrecarga de informação. Para Blanco-Fernandes et al. [5] os sistemas de recomendação podem auxiliar as pessoas a escolher um plano de viagem, indicando lugares para visitar, opções de hotéis, companhias aéreas, de acordo com as preferências indicadas no seu perfil.

2.1.1 Amazon

O sistema de recomendação utilizado pela Amazon[3] é um dos mais populares. Os utilizadores desta plataforma recebem recomendações de diversos produtos, que são sugeridos com base em várias estratégias.

[3] www.amazon.com

Uma técnica utilizada é designada por *Collaborative filtering*, esta técnica que permite recomendar os itens que foram da preferência de outros utilizadores que têm um perfil semelhante ao do utilizador alvo. Deste modo, em primeiro lugar é obtida a vizinhança do utilizador alvo (com base no histórico de compras), sendo recomendado ao cliente alvo os produtos que são mais populares para este grupo de utilizadores.

Na Figura 2.1 é demostrado um exemplo em que são recomendados os itens que foram comprados pelas pessoas que também compraram um determinado item.

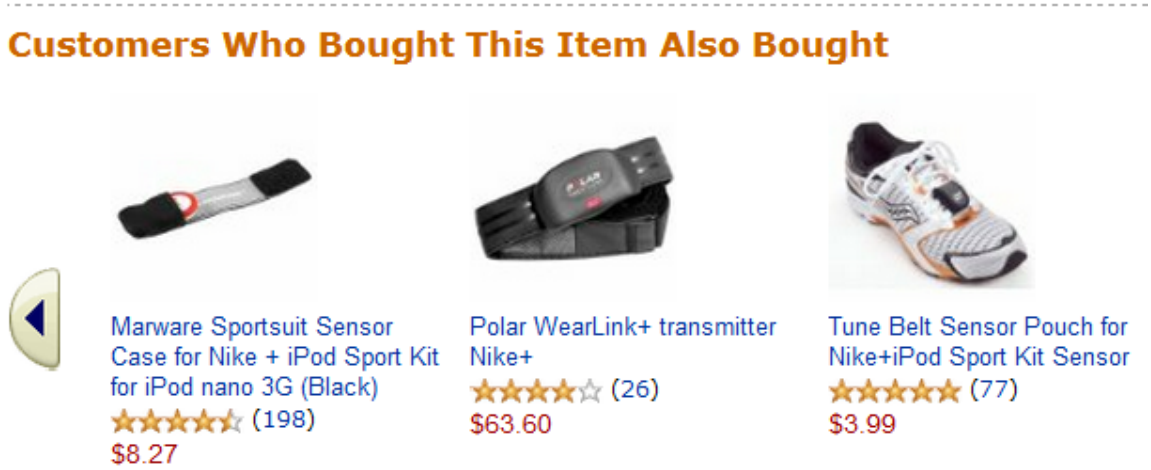

Figura 2.1: Amazon - Recomendação de itens

Outra estratégia utilizada para recomendar itens é analisar o histórico de compras de cada utilizador, e recomendar os itens que estejam associados (pertencentes à mesma categoria). No entanto o utilizador também pode filtrar as recomendações sugeridas com base na sua categoria.

2.1.2 TiVo

TiVo[4] é uma é um serviço de televisão bastante popular nos EUA, que incorpora um sistema de recomendação com o objectivo de sugerir programas que sejam do interesse dos utilizadores.

O sistema de recomendação utilizado pela TiVo utiliza a combinação das técnicas *Collaborative filtering* e *Content-based filtering*. Com a primeira técnica as recomendações são efectuadas com base nas preferências dos utilizadores que têm um perfil semelhante ao do utilizador alvo. Através da técnica *Content-based filtering* os programas recomendados são aqueles que possuem características (actores, género, tema) que o utilizador gostou no passado.

[4] http://www.tivo.com/

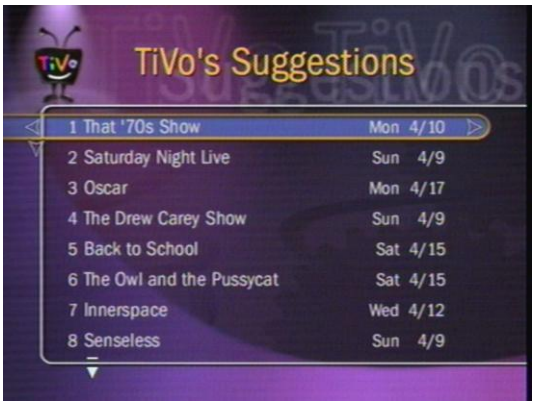

Figura 2.2: Sistema de recomendação TiVo

O processo de recomendação é efectuado com base no *feedback* dado pelos utilizadores, sendo que este *feedback* pode ser obtido de forma explícita ou implícita. O feedback explícito é fornecido pelo utilizador através dos botões "gosto" ou "não gosto" associados a cada programa.

Também é possível recolher informação sobre cada utilizador de uma forma implícita, por exemplo através da análise do comportamento de cada utilizador, verificando os programas que são frequentemente visualizados e gravados. O *feedback* proveniente de cada utilizador permite definir o seu perfil e consequentemente as suas preferências.

2.1.3 Google Hotel Finder

O Google Hotel Finder[5] permite recomendar os hotéis mais apropriados para cada pessoa de acordo com as suas preferências e necessidades. Este sistema de recomendação utiliza a técnica de filtragem *Knowledge-based*, ou seja, os utilizadores têm de especificar as características dos hotéis (número de estrelas, limite de preço) que estão interessados. Como os hotéis estão associados a um espaço geográfico, também é tido em consideração a distância a que cada hotel se encontra da localização desejada pelo utilizador.

2.1.4 Outros Sistemas de Recomendação

O TaxiMedia [6] é um sistema de informação interactivo e sensível ao contexto, tendo sido aplicado numa rede de táxis. O objectivo deste sistema é entreter os passageiros e ao mesmo tempo mostrar notícias e publicidade que sejam do interesse do utilizador.

A forma de seleccionar a informação a ser mostrada está relacionada com o contexto do utilizador. Relativamente ao contexto são tidos em conta parâmetros como: localização espacial e temporal, condições climatéricas, destino do utilizador etc.

Um dos pontos fracos deste sistema é não ser possível saber informações pessoais acerca dos passageiros, como o género, idade, etc. Uma vez que não é efectuada nenhuma autenticação dos passageiros no momento em que estes entram no veículo.

[5] http://www.google.com/hotelfinder

TouristGuide [7] é uma aplicação que tem a função de prestar um serviço de guia aos touristas da cidade de Adelaide, Austrália. Esta aplicação apenas se concentra no contexto do utilizador, sobretudo na vertente da localização. Consoante a localização do utilizador, são mostradas informações que poderão ser do seu interesse, como cabines telefónicas, casas de banho, e outros serviços que estejam localizados na sua proximidade.

O Guide [8] é uma aplicação que fornece informações turísticas com base no perfil (preferências) e contexto dos utilizadores (informação temporal e geográfica). Ao contrário do sistema anterior, existe a preocupação de fornecer informações enquadradas com o tipo de perfil, sendo que o perfil é especificado pelo próprio utilizador.

O Sightsplanner[6] é um sistema de recomendação turística, que tem como objectivo recomendar um plano turístico aos visitantes da cidade Tallinn (Estónia). Os utilizadores podem especificar as suas preferências, ou seja, indicar o tipo de pontos de interesse que desejam visitar: eventos, arte e museus, arquitectura, desporto, etc.

Como os pontos de interesse têm diferentes localizações e horários de funcionamento, o utilizador terá de fornecer informações sobre o seu contexto. O contexto é constituído pela data e hora a que se pretende iniciar a visita e o meio de deslocação, que pode ser de carro ou a pé.

O plano apresentado é constituído por uma rota de pontos de interesse tendo em conta as preferências e o contexto do utilizador. Uma das limitações deste sistema é não guardar a informação sobre os pontos de interesse já recomendados a um determinado utilizador, desta forma as novas recomendações podem apresentar pontos de interesse que já foram sugeridos anteriormente.

2.2 Atenção Selectiva

A atenção selectiva "é a habilidade que o indivíduo possui para direccionar o foco de atenção a um ponto específico no meio ambiente".

No dia-a-dia, os seres humanos estão submetidos a uma grande quantidade de informações, contidas no meio ambiente em que estes estão inseridos. Dependendo da actividade e das preferências de cada ser humano, estas informações podem ser consideradas relevantes ou completamente ignoradas. Contudo, é apenas a partir da adolescência que se adquire a capacidade de seleccionar as informações relevantes, ao mesmo tempo que se descarta o que é irrelevante.

Os estudos de Jon Driver [9] mostram que os seres humanos apresentam dificuldade em processar toda a informação que é captada pelos sentidos. Desta forma, apenas existe um foco de interesse numa pequena gama de informações relacionadas com as suas preferências pessoais ou que causem algum tipo de surpresa.

O conceito de atenção selectiva começou a ser estudado durante a década de 50, sendo que as primeiras experiências basearam-se no fenómeno "cocktail party" [10].

Estas experiências consistiam em colocar vários indivíduos numa sala a conversar ao mesmo tempo. O objectivo era analisar a capacidade que os seres humanos têm em se concentrar e

[6] http://tallinn.sightsplanner.com/maps/show

obter informação, apesar de todas as mensagens que eram geradas em simultâneo e do ruído existente.

Em 1953, Cherry [11], para conseguir entender o fenómeno "cocktail party" realizou experiencias que envolviam a escuta de áudio. Nestas experiências colocou-se um individuo a escutar uma mensagem diferente por cada ouvido, e pedia-se para se concentrar apenas num ouvido e em seguida replica-se a informação recebida.

O resultado foi que a mensagem em que não se prestava atenção era completamente perdida, mas quando esta mensagem sofria variações (idioma, tom de voz), o individuo conseguia identificar algumas das variações apesar de não saber do que se tratava.

Com o avançar das investigações de atenção selectiva surgiram os modelos de filtro. Estes modelos sugerem que todas as informações obtidas do exterior, passam por um filtro que apenas selecciona as informações mais relevantes.

Broadbent elaborou aprofundamentos nas pesquisas iniciadas por Cherry e propôs a teoria de filtro de Broadbent [12] em 1958. A sua importância é reconhecida, porque conseguiu relacionar os fenómenos psicológicos com conceitos de processamento de informação provenientes das ciências informáticas e matemáticas.

Baseando-se no trabalho de Chery que concluiu que os seres humanos apenas tinham a capacidade de prestar atenção a uma mensagem de cada vez, Broadbent propôs um filtro que opera sobre as características físicas das mensagens, como o tom de voz e tipo de som.

O processo de filtragem é feita de forma consistente com as preferências das pessoas, sendo que apenas uma mensagem é seleccionada para posterior processamento, as restantes são perdidas. A utilização do filtro previne que ocorra um excesso de mensagens, com o objectivo de evitar a sobrecarga do sistema, simulando desta forma a incapacidade que os seres humanos têm em processar todas as mensagens captadas pelos sentidos.

Em 1964 Treisman propôs o Modelo de Atenuação Treisman [13], este modelo é constituído pelo filtro de Broadbent que permite seleccionar as mensagens de acordo com as suas características físicas. A grande diferença é que este filtro apenas serve como atenuador, deste modo, os estímulos que não apresentam interesse para um determinado individuo não são complementarmente eliminados.

A atenuação realizada pelo filtro pode ser comparada ao baixar do volume, por exemplo se existir duas fontes de som numa sala (televisão e rádio) é possível atenuar uma fonte de som e assistir à outra. Como o filtro não elimina por completo as mensagens, apenas atenua a sua importância, existe a possibilidade de uma determinada informação que foi atenuada ser atendida. O que não se verifica no filtro de Broadbent, porque as mensagens não seleccionadas são descartadas.

Com a evolução das ciências cognitivas, surgiram alguns modelos computacionais de emoções. Uma dessas emoções é a surpresa, que segundo [14] está relacionado com vários processos cognitivos como a aprendizagem e a atenção selectiva. Na secção 2.2.1 está descrito o conceito de surpresa e a forma como esta se relaciona com a atenção selectiva.

2.2.1 Surpresa

O efeito de surpresa pode ser definido como uma reacção automática a um evento que não se encaixa no modelo de previsões de um individuo.

Diversos estudos têm sido apresentados, com o objectivo de determinar a importância da sensação de surpresa, quando se pretende obter o foco de atenção por parte dos seres humanos. Em [15] é apresentado um estudo que procura verificar qual o efeito que leva a uma maior mudança de atenção para um novo conjunto de estímulos, de acordo com várias métricas.

O estudo foi feito através do cálculo do número de transições ao nível ocular de vários seres humanos, quando em presença de vários estímulos visuais de diferentes origens. De acordo com as suas conclusões, a surpresa foi considerada a melhor métrica para caracterizar o efeito que mais atraía os olhares de cada pessoa.

O modelo psicológico

Os psicólogos definem surpresa como um sentimento ou uma experiência provocada pela observação de um acontecimento inesperado.

A teoria do esquema [16] representa uma visão de como a percepção humana é organizada e controlada por estruturas complexas de crenças, chamados esquemas. Um esquema representa um conjunto de crenças sobre os objectos ou eventos, e pela representação de situações passadas permitem prever situações futuras.

O efeito de surpresa, neste modelo, é definido por um processo que compara em todos os momentos, os esquemas activos com as informações adquiridas em um determinado instante. Se a informação for coerente com as crenças do observador, os esquemas não precisam de ser actualizados.

Por outro lado, se há uma diferença entre aquilo que é conhecido e o que se observa, os esquemas devem ser revistos. Nessa situação é experienciado um sentimento de surpresa e é gerado um conjunto de processos cognitivos [17].

Tais processos cognitivos consistem numa sequência de quatro etapas que se destina a preparar as acções do observador em resposta ao sentimento de surpresa.

O primeiro processo é simplesmente a avaliação da discrepância entre a informação adquirida e o esquema. Se o nível de discrepância é maior que um determinado limiar, então o segundo processo interrompe quaisquer outros processos cognitivos e centra a atenção no evento inesperado. Com esta mudança de atenção, o terceiro processo irá analisar e avaliar o evento, o quarto processo irá rever o esquema que causou a discrepância.

Esse conjunto de eventos cognitivos e eventual revisão dos esquemas prepara o observador para situações futuras, de forma a lidar melhor com novos e surpreendentes eventos.

O Modelo de Macedo-Cardoso

O modelo de surpresa proposto por Macedo e Cardoso [18], caracteriza a surpresa com base no grau de imprevisibilidade de um evento. Este modelo foi aplicado num agente artificial, cuja missão era viajar num mundo limitado, observando edifícios, e de acordo com sua forma inferir a sua função. O agente constrói o seu modelo de crenças de acordo com o que ele descobre e avaliando as verdadeiras funções de cada edifício.

O agente apenas se depara com o efeito de surpresa, se o que ele percepciona está em conflito com as suas crenças. Por exemplo, o agente detecta um edifício, e através de suas propriedades o agente determina que o edifício é uma casa com 70% de certeza, e uma loja com 10% de certeza. Se depois de visitar o edifício, o agente verificar que na verdade, era um restaurante, ele vai sentir surpresa, porque existe um conflito com as informações presentes no seu modelo de crenças.

Desta forma, sabendo qual o conhecimento do passado, e as crenças de um agente, consegue-se estimar qual o evento mais esperado, inferindo assim o grau de surpresa. Após várias iterações, estes autores apresentaram uma expressão matemática que permite modelar a surpresa humana. A fórmula é a seguinte:

$$Surprise(X) = log_2(1 + P(Y) - P(X)) \tag{2.1}$$

Nesta expressão, a surpresa do evento X é definida como uma diferença entre a probabilidade do evento mais provável, $P(Y)$, e a probabilidade do evento X ocorrer $P(X)$. Podemos constatar que quanto menor for a probabilidade de um evento ocorrer, maior será o valor de surpresa.

S. Vargas [19] utilizou o conceito de surpresa de forma a permitir que as recomendações sugeridas pelo sistema de recomendação apresentem novidade. Segundo o autor a novidade de um determinado item é condicionado pelo somatório de todos os eventos que o utilizador teve com esse item. Uma forma simplista de calcular a novidade de um item é determinar a sua popularidade, ou seja, todos os eventos que ocorreram sobre esse item. Sendo que a novidade é calculada através da seguinte fórmula:

$$novelty(i) = \begin{cases} 1, se\ p(i) = 0 \\ -log_2(p(i)) \end{cases} \tag{2.2}$$

Onde $p(i)$ representa a probabilidade do item i ser observado, sendo que esta probabilidade é calculada com base no conjunto de eventos que ocorreram sobre este item. Um valor elevado de novidade significa que poucas pessoas interagiram com o item, pelo contrário, um baixo valor indica que já ocorreram muitos eventos com o item. Contudo esta equação apenas mede a novidade de uma forma genérica, na medida em que o valor de novidade de um item será o mesmo para todos os utilizadores.

De forma a calcular a novidade de um item em relação a um utilizador específico, os autores propuseram a seguinte equação:

$$novelty(i) = \begin{cases} 1, se\ p(i|u) = 0 \\ -log_2(p(i|u)) \end{cases} \quad (2.3)$$

Em que $p(i|u)$ representa a probabilidade do utilizador u interagir com o item i, desta forma quanto maior for a probabilidade, menor será o valor de novidade.

2.3 Relevância

Com o objectivo de saber o interesse que cada objecto tem para um individuo, Furnas [20] formalizou o conceito "fish eye". Segundo o autor, para definir o interesse de um determinado objecto, é necessário saber a importância que o mesmo assume no contexto da pesquisa, assim como a distância a que se encontra do alvo.

Sendo assim, um objecto adquire maior valor de interesse, caso esteja mais perto do utilizador e a sua importância seja elevada. Caso contrário, o interesse associado ao item será menor. A fórmula proposta for Furnas é a seguinte:

$$DOI(x|y) = API(x) - D(x,y) \quad (2.4)$$

Em que DOI representa o valor de interesse de um dado ponto *x*, $API(x)$ corresponde ao valor de importância associado ao ponto e $D(x,y)$ descreve a distância entre o ponto e o valor alvo desejado.

Pombinho et al. [21] também elaboraram estudos de modo a calcular o valor de relevância que cada recurso apresenta para um determinado individuo. Segundo o seu trabalho, a relevância pode ser calcula como uma soma pesada das distâncias entre os valores alvos desejados e os valores dos atributos que constituem cada recurso.

$$Rel = \sum_{j=1}^{k} \left(1 - Dist\left(A_j, R_j\right) \times w_k, w_k \in [0,1]\right) \quad (2.5)$$

A variável k representa o número de atributos que constituiu um determinado recurso, o A_j representa o valor alvo para o atributo j e o R_j indica o valor que está associado a esse mesmo atributo.

A função $Dist$ é responsável por calcular a distância que existe entre o valor do atributo e o alvo desejado. O w_k indica o peso que está associado a cada atributo, podendo existir atributos com pesos diferentes. Desta forma é possível definir os atributos que são mais importantes para o cálculo da relevância.

2.3.1 Distância entre Atributos

Para calcular a distância entre atributos é necessário ter em consideração o seu tipo, sendo que a função de calcular a distância tem de ser adaptada a cada situação. De um modo geral, os atributos podem ser classificados como: numéricos, temporais, geográficos e nominais.

Atributos numéricos

Para calcular a distância de atributos numéricos podem ser utilizadas diversas funções, dependo do atributo em si e do resultado que se pretende obter. Por exemplo, para um atributo numérico "preço", normalmente considera-se que os valores situados abaixo do alvo são mais relevantes do que os existentes acima do valor alvo. Para esta situação pode ser utilizada a seguinte função de distância.

$$Dist(R_j) = \frac{R_j - min_j}{max_j - min_j}\ , Dist(R_j)\ \in [0,1] \tag{2.6}$$

Sendo que o R representa o valor do atributo j e o min_j e max_j referem-se aos valores máximos e mínimos conhecidos para o atributo.

Como alternativa surge a possibilidade de calcular o valor de distância, sem ter em atenção se o valor do atributo se encontra acima ou abaixo do valor alvo. Este comportamento é descrito pela seguinte equação.

$$Dist(A_j, R_j) = \left|\frac{R_j - A_j}{max_j - min_j}\right|\ , Dist(A_j, R_j)\ \in [0,1] \tag{2.7}$$

Sendo que o A_j representa o valor alvo desejado, o R representa o valor do atributo, e o max_j e min_j indicam os valores máximos e mínimos que podem ocorrer para o atributo j.

Atributos nominais (textuais)

De forma a calcular a distância entre atributos nominais podem ser utilizadas duas abordagens. Uma mais simples em que apenas é verificado se o valor do atributo e do alvo desejado é igual ou diferente.

$$Dist(A_j R_j) = \begin{cases} 0, se\ A_j = R_j \\ 1, se\ A_j \neq R_j \end{cases} \tag{2.8}$$

A outra abordagem consiste no cálculo da distância semântica entre o valor do atributo e o alvo desejado.

$$Dist(A_j, R_j) = \begin{cases} 0, se\ A_j = R_j \\ x, se\ A_j \neq R_j\ , x \in]0,1] \end{cases} \quad (2.9)$$

Sendo que A_j indica o alvo desejado e o R_j representa o valor do atributo j. O x representa a distância semântica entre A_j e R_j.

De forma a calcular a distância semântica entre dois atributos textuais, pode ser utilizado a base de dados lexical WordNet[7].

O algoritmo implementado pelo WordNet, consiste em percorrer a árvore de termos lexicais, e encontrar o caminho mínimo entre os dois termos pretendidos. A Figura 2.3 descreve a forma como os termos estão organizados.

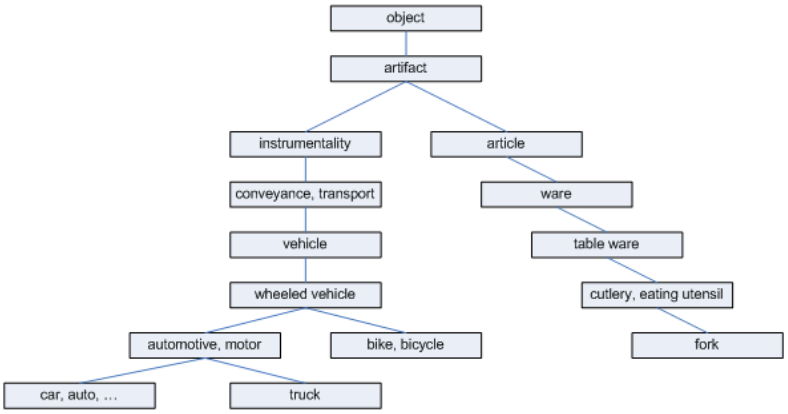

Figura 2.3: Ontologia lexical

O valor de similaridade pode ser calculado através do caminho que une dois termos, ou seja, é efectuado a contagem do saltos que é necessário realizar até chegar ao segundo termo. A equação seguinte representa o cálculo de similaridade.

$$Sim(s,t) = 1/distance(s,t) \quad (2.10)$$

Podemos constatar que a similaridade é tanto maior quanto menor for a distância entre os termos lexicais.

A biblioteca RiTa WordNet[8] permite executar o processamento sobre a base de dados lexical, de forma a calcular a distância semântica entre as palavras.

[7] http://wordnet.princeton.edu/

Atributo Temporal (Horário de funcionamento)

De forma a verificar se um determinado horário de funcionamento, é compatível com o momento temporal em que o utilizador chega ao local, poderá ser utilizado a seguinte expressão matemática:

$$Dist(x,t) = \begin{cases} 1, if\ x \geq t_{max} \vee x \leq t_{min} \\ 0, if\ x > t_{min}\ \wedge x < t_{max} \end{cases} \quad (2.11)$$

O x representa é o momento temporal em que o utilizador chega ao recurso desejado, o t_{min} representa o horário de abertura do recurso e o t_{max} representa o horário de encerramento.

Atributos geográficos

Para além dos tipos de atributos vistos anteriormente, falta ainda considerar os atributos que representam coordenadas geográficas. Como forma de calcular a distância entre atributos geográficos pode ser utilizada a função de Haversine [24].

Esta função permite determinar uma aproximação da distância física em linha recta entre duas coordenadas geográficas.

$$Haversine = hav(\phi 2 - \phi 1) + cos\ (\phi 1) cos\ (\phi 2) \times hav(\Psi 2 - \Psi 1) \quad (2.12)$$

Sendo que:

Ψ1, Ψ2: longitude do ponto 1 e longitude do ponto 2.

$\phi\ 1, \phi 2$: latitude do ponto 1, latitude do ponto 2.

$hav(\theta) = \sin\ (\theta/2)^2$

A distância entre duas coordenadas é calculada tendo em conta o raio do planeta Terra.

$$dist = r \times Haversine^{-1} \quad (2.13)$$

A limitação desta abordagem é que apenas é calculada a distância geográfica em linha recta, ou seja, é ignorado por completo a trajectória associada a estradas e percursos. Como forma de ultrapassar estas limitações pode ser utilizado a API Google Directions[9].

Através da API Google Directions, é possível especificar as coordenadas do ponto origem e destino, o modo de deslocação (carro, bicicleta ou andar), zonas a excluir (portagens,

[8] http://www.rednoise.org/rita/wordnet/documentation/index.htm
[9] http://code.google.com/apis/maps/documentation/directions

rodovias), para além de ser possível especificar as unidades métricas (quilómetros ou milhas). O resultado enviado pelo serviço encontra-se no formato JSON[10] ou XML[11], conforme o especificado no pedido.

Atributos compostos

Existem ainda atributos que podem ser constituídos por um por um conjunto ordenado de *tags*. Para se calcular a distância entre dois conjuntos de *tags* pode ser utilizada uma abordagem inspirada no algoritmo Jaro–Winkler [25].

Este algoritmo é sobretudo utilizado para calcular o grau de similaridade entre duas palavras ou sequências de caracteres.

$$Dist(s_1, s_2) = \frac{1}{3}\left(\frac{m}{s_1} + \frac{m}{s_2} + \frac{m-t}{m}\right) \quad (2.14)$$

Em que s_1 representa o tamanho do primeiro conjunto e o s_2 representa o tamanho do segundo conjunto de caracteres. A variável m corresponde ao número de caracteres que se repetem nos dois conjuntos. A componente t identifica o número de transposições, ou seja, o número de caracteres que estão presentes nos dois conjuntos mas ocupam posições diferentes.

O valor de similaridade entre dois conjuntos de caracteres é calculado a partir do número de caracteres que se repetem nos dois conjuntos. Para esses caracteres também é verificado se ocupam a mesma posição nos dois conjuntos.

No entanto, em alguns casos as *tags* podem estar associadas a pesos que indicam a sua importância. Buttler [26] propôs a equação *Weighted Tag Similarity* que permite calcular a similaridade entre dois conjuntos que contenham *tags* associadas ao respectivo peso.

$$WTS(D_i, D_j) = \frac{\sum_{k=1}^{n} 2 \times min(w_{i,k}, v_{j,k})}{\sum_{k=1}^{n}(w_{i,k} + v_{j,k})} \quad (2.15)$$

O $w_{i,k}$ representa a importância (número de vezes que aparece) da *tag* k no conjunto D_i, da mesma forma, o $v_{j,k}$ indica a importância da *tag* k no conjunto D_j.

[10] http://en.wikipedia.org/wiki/JSON
[11] http://en.wikipedia.org/wiki/XML

2.4 Diversidade

Os sistemas de recomendação convencionais adoptaram a estratégia de recomendar os itens que apresentam maior similaridade com o perfil e preferências do utilizador.

No entanto, estudos recentes mostram que o processo de recomendação baseada apenas na similaridade entre os itens e o perfil do utilizador não garante a qualidade das recomendações. Devido ao facto do conjunto de itens recomendado ser extremamente homogéneo e apresentar poucas alternativas ao utilizador. Nesta vertente, a diversidade do conjunto de recomendações é um factor que contribui para melhorar o grau de satisfação dos utilizadores.

A utilização do conceito de diversidade também é importante quando se pretende reduzir o número de dados que são seleccionados em cada pesquisa. Nesta vertente, o conceito de diversidade é utilizado para seleccionar um subconjunto de dados (de dimensão n) com base num conjunto de dados de maior dimensão. Sendo que o subconjunto de dados seleccionado deve conter os itens com informação mais diversificada entre si, de modo a oferecer a maior cobertura possível ao conjunto de dados base.

Esta prática de redução do número de itens a serem mostrados em cada pesquisa, é sobretudo importante para não sobrecarregar os utilizadores com informações iguais ou muito semelhantes. Também apresenta interesse para dispositivos de dimensões reduzidas e com pouca capacidade de memória.

Gago et al [27] desenvolveram pesquisas com o objectivo de extrair o conjunto de regras de uma base de dados que oferece-se a maior capacidade de previsão. No seu caso de estudo, cada regra representa uma instrução numa base de dados, e são definidas na forma de "IF *condições* THEN *conclusão"*. O objectivo é que o subconjunto de regras seleccionado seja o mais heterogéneo possível, de forma oferecer a maior cobertura possível ao conjunto de dados base.

De forma a avaliar a semelhança entre as várias regras é efectuada a comparação entre os atributos de cada regra, calculando a distância a que estes se encontram.

$$Diff(r_i, r_j) = \sum_{a=1}^{n} \left(Dist(r_{ia}, r_{ja}) \times w_k\right), w_k \in [0,1] \tag{2.16}$$

Se o somatório da distância entre os atributos das duas regras (r_i e r_j) for zero, significa que as regras são iguais. Com base nesta métrica de distância, os autores elaboraram o seguinte algoritmo de escolha de regras, que permite seleccionar o subconjunto de regras (de dimensão n) mais heterogéneo.

Algoritmo 1 Rule Select

R ← Recurso com maior valor de diferença
S_r ← R
while $\#S_r < n$ **do**
  for each rule R in S and not in S_r
    AV ← Distância média de R em relação S_r
  endfor
  R_{max} ← Regra com maior valor AV
  $S_r \leftarrow S_r \cup \{R_{max}\}$
endwhile
return S_r

O S corresponde ao conjunto de regras, o n indica o número de regras que se pretende seleccionar e o S_r corresponde ao subconjunto de regras seleccionado.

Este algoritmo começa por seleccionar a regra que apresente maior valor de distância em relação às restantes, sendo que para calcular a distância entre cada regra é utilizado a métrica de distância descrita na Equação 2.17.

Deste ponto em diante selecciona-se a regra que apresente o maior valor médio de distância em relação ao subconjunto S_r de regras já seleccionadas. Este processo repete-se até serem seleccionadas n regras.

Bradley et all [28] elaboraram estudos relacionados com a diversidade de itens, de forma melhorar a qualidade dos sistemas de recomendação. Basicamente o seu objectivo consistia em seleccionar um conjunto de itens que fossem semelhantes a um determinado alvo desejado, mas que apresentassem diversidade entre si. Para este efeito, os autores propuseram três estratégias: *Bounded Random Selection*, *Greedy Selection* e *Bounded Greedy Selection*.

A estratégia *Bounded Random Selection* apenas trata a questão da diversidade recorrendo a processos de selecção aleatória. Este comportamento é descrito no algoritmo que se segue.

Algoritmo 2 Bounded Random Selection

begin
  C' ← conjunto de itens de dimensão bk mais similar com t
  R ← seleccionar aleatoriamente k itens de C'
  return R
end

O t representa o item alvo, o C corresponde ao conjunto de itens, o k indica o número de itens a seleccionar e o b corresponde ao limiar de semelhança a considerar.

Este algoritmo começa por obter um conjunto C' de dimensão bk que contém os itens com maior grau de semelhança ao alvo t. Em seguida selecciona aleatoriamente k itens do conjunto C'.

Como os itens do conjunto R são seleccionados aleatoriamente, não é possível garantir que este conjunto contenha os itens que apresentem maior diversidade entre si. Com o objectivo de superar estas limitações os autores propuseram a o algoritmo *Greedy Selection*.

Algoritmo 3 Greedy Selection

begin
    $R \leftarrow \{\}$
    for i=1 to k
        Ordenar C em função de Quality (t, C, R)
        $R \leftarrow R$ + Primeiro (C)
        $C \leftarrow C$ − Primeiro (C)
    endfor
return R
end

O t representa o item alvo, o C corresponde ao conjunto de itens, o k indica o número de itens a seleccionar.

Em cada interacção k os itens são ordenados em função da sua qualidade, o item que apresentar maior valor de qualidade é adicionado ao conjunto R.

A métrica de qualidade desenvolvida pelos autores permite medir a importância de um determinado item, com base na sua similaridade com o alvo desejado e na sua diversidade em relação aos itens já seleccionados.

$$Quality(t, c, R) = (1 - a) \times similarity(t, c) + a \times diversity(c, R) \quad (2.17)$$

O parâmetro a permite ajustar a importância que é dada à similaridade do item com o alvo desejado e a importância dada à diversidade que o item apresenta em relação ao conjunto de itens já seleccionados.

O principal problema do algoritmo *Greedy Selection* está relacionado com a sua complexidade algorítmica, uma vez que em cada interacção k o conjunto de itens C tem de ser ordenado em função da sua qualidade.

Com o objectivo de diminuir a complexidade apresentada por este algoritmo, ou autores propuseram a estratégia *Bounded Greedy Selection* que é baseada nos dois algoritmos vistos anteriormente: *Bounded Random Selection* e *Greedy Selection.*

Algoritmo 4 Bounded Greedy Selection

begin
    $R \leftarrow \{\}$
    $C' \leftarrow$ conjunto de itens de dimensão bk mais similar com t
    for i=1 to k
        Ordenar C' em função de Quality (t, c, R)
        $R \leftarrow R$ + Primeiro (C')
        $C' \leftarrow C'$ − Primeiro (C')
    endfor
return R
end

O t representa o item alvo, o C corresponde ao conjunto de itens, o k indica o número de itens a selecionar e o b corresponde ao limiar de semelhança a considerar.

A estratégia deste algoritmo é reduzir o espaço de procura que é utilizado pelo algoritmo *Greedy Selection*. Deste modo, em primeiro lugar é obtido um subconjunto de itens C' que contém os itens mais similares ao alvo desejado.

Com a redução do espaço de procura consegue-se diminuir a complexidade algorítmica, mas apresenta a desvantagem de não ser aplicada a métrica qualidade (Equação 2.18) a todos os itens existentes no conjunto C, devido à exclusão que foi realizada previamente. Deste modo, existe um compromisso entre a complexidade algorítmica e a qualidade dos itens seleccionados. Este compromisso está dependente da dimensão do subconjunto C' seleccionado.

2.5 Técnicas de Recomendação

Os sistemas de recomendação apresentam uma evolução em relação a outas aplicações de selecção de informação devido à sua capacidade de fornecer informação personalizada e contextualizada com os utilizadores.

Por exemplo, enquanto motores de busca são muito susceptíveis a gerar sempre os mesmos resultados para pesquisas idênticas, os sistemas de recomendação são capazes de gerar resultados mais específicos para cada utilizador com base no perfil e contexto de cada utilizador em particular.

Os sistemas de recomendação têm como objectivo calcular a relevância que cada item apresenta para um utilizador em particular. De forma a efectuar esta previsão, podem ser utilizadas diferentes técnicas de recomendação que são: *Content-based*, C*ollaborative filtering, Knowledge-based, Demographic filtering* e *Hybrid filtering.*

A técnica *Content-based* procura recomendar os itens que sejam semelhantes aos que o utilizador já gostou no passado.

A técnica *Collaborative filtering* tem como objectivo prever os itens que são mais relevantes para um determinado utilizador, com base nas avaliações feitas por outros utilizadores que possuem um perfil (preferências) semelhante.

A técnica *Knowledge-based* efectua as recomendações com base nas necessidades de cada utilizador. Desta forma os utilizadores têm de especificar previamente o tipo de informação que desejam obter.

A técnica *Demographic filtering* caracteriza-se por efectuar as recomendações com base nas preferências de outros utilizadores que têm um perfil demográfico semelhante (idade, género, estado civil, etc.) ao do utilizador alvo.

A técnica *Hybrid filtering* surge como uma combinação das técnicas enumeradas anteriormente, tendo como objectivo ultrapassar as suas limitações de forma a proporcionar melhores recomendações.

2.5.1 Técnica de Recomendação *Content-based*

A técnica de recomendação *Content-based* [29] consiste em comparar o conteúdo de cada item que se pretende recomendar com as informações (preferências) associadas ao perfil de cada utilizador. De um modo geral, o principal objectivo desta técnica é procurar recomendar itens semelhantes com aqueles que o utilizador gostou no passado.

Por exemplo, num sistema de recomendação de filmes que utilizasse a técnica de recomendação *Content-based*, o objectivo seria tentar perceber as características dos filmes (actores, director, género, tema) que seriam do interesse de cada utilizador. Desta forma, apenas os filmes que apresentem elevado grau de semelhança com as preferências do utilizador seriam recomendados.

Os interesses e preferências dos utilizadores podem ser obtidos de forma explícita, neste caso os utilizadores têm preencher determinados formulários e questionários de forma a ser possível recolher as suas preferências.

Também é possível aprender os interesses de cada utilizador com base no seu histórico de interacções com o sistema de recomendação. Por exemplo, um utilizador ao pesquisar com frequência itens desportivos está a demostrar a sua preferência por itens relacionados a este assunto.

Outra fonte de informação importante para definir o perfil do utilizador, provém da classificação que cada utilizador atribui a cada item. De modo a permitir que os utilizadores avaliem os itens, pode ser utilizado um sistema de votação "gosto/não gosto". Neste caso é necessário verificar as características dos itens que são mais importantes para cada utilizador em concentro, porque é através desta informação que é construído o perfil do utilizador.

A técnica de recomendação *Content-based* procura classificar os itens em função da similaridade que estes apresentam em relação ao perfil do utilizador, e apenas os itens com maior classificação serão recomendados.

$$cb(c,s) = sim(profile(c), content(s)) \tag{2.18}$$

O $profile(c)$ representa o perfil do utilizador c, contendo a informações sobre os seus gostos e preferências. O $content(s)$ representa o conteúdo do item s. Comparando o perfil do utilizador com o conteúdo de cada item é possível calcular a relevância que esse item apresenta para o utilizador.

A maioria dos sistemas que implementam esta técnica são sobretudo utilizados para recomendar itens que contenham informação textual, como por exemplo: artigos, páginas Web, notícias, entre outros. Este acontecimento deve-se ao facto de ser relativamente fácil efectuar comparação entre atributos que contenham descrições textuais ou numéricas.

No entanto, a recomendação *Content-based* está limitada pelas características dos itens que se pretendem recomendar. Quando um atributo não está associado ao item, o sistema de recomendação não o pode utilizar no processo de recomendação. Por exemplo, suponhamos que o filme "Voando sobre um ninho de cucos" não era rotulado com o actor

Jack Nicholson, este acontecimento poderia originar a não recomendação deste filme a um utilizador que tenha preferência por este actor.

Deste modo, para cada item é importante ter um conjunto suficiente de atributos. No entanto, a extracção de atributos é um trabalho intensivo (quando é feito manualmente) e pode introduzir erros tanto quando é realizado manualmente como automaticamente.

Esta técnica de recomendação depende do *feedback* dado pelos utilizadores. Enquanto este *feedback* não for fornecido é impossível construir o perfil dos utilizadores, o que inviabiliza a qualidade das recomendações. Este problema é designado por *cold-start*.

Outro problema dos sistemas que utilizam técnica baseada no conteúdo é que apenas são recomendados os itens que apresentem um elevado valor de similaridade com o perfil do utilizador. Este facto implica que para um determinado utilizador sejam sempre recomendados itens semelhantes e repetidos com os que já foram mostrados anteriormente, comprometendo desta forma a diversidade e novidade das recomendações.

O Pandora[12] é um serviço de recomendação de música que utiliza a abordagem *Content-based*. Este sistema utiliza um conjunto de atributos que permitem descrever cada música (ritmo, sensações, tonalidade, etc.), também utiliza as avaliações efectuadas por cada utilizador de forma a construir o seu perfil.

2.5.2 Técnica de Recomendação *Collaborative Filtering*

A filtragem colaborativa é uma das mais eficazes técnicas de recomendação [30]. Ao contrário dos métodos de recomendação baseados no conteúdo, os sistemas que utilizam uma filtragem colaborativa têm como objectivo prever a utilidade dos itens (para um determinado utilizador) com base nas avaliações realizadas por outros utilizadores que possuam um perfil semelhante.

A ideia subjacente da técnica *Collaborative filtering* é baseada no facto dos utilizadores que classificam os mesmos itens da mesma forma (ou pelo menos de forma idêntica) serem susceptíveis de apresentarem preferências semelhantes.

Técnica Collaborative filtering
1. Calcular a similaridade entre utilizadores
2. Obter os vizinhos do utilizador alvo
3. Prever o valor de relevância de cada item

A técnica *Collaborative filtering* começa por calcular a similaridade entre o utilizador alvo e todos os restantes, de forma a recomendar os itens que sejam do interesse de outas pessoas que tem um perfil idêntico. Por exemplo, num sistema de recomendação de livros, o primeiro passo é encontrar os "vizinhos" do utilizador alvo. O termo "vizinhos" refere-se aos utilizadores que tenham gostos semelhantes, ou seja, que avaliem os mesmos livros de forma idêntica. Apenas os livros que são favoravelmente avaliados pelos "vizinhos" do utilizador alvo seriam recomendados.

De forma a aplicar a técnica *Collaborative filtering* é necessário construir uma estrutura de dados que relaciona os utilizadores com os itens. Esta estrutura de dados consiste numa

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

matriz *utilizador* × *item*, em que cada célula da matriz é composta por um valor que representa a classificação dada pelo utilizador ao item. Um exemplo desta matriz é mostrado na Tabela 1:

	Utilizador			
Item	**1**	**2**	**3**	**4**
A	5	3	4	
B	4			5
C		2		4
D	4		4	
E	3		2	2
F	2	4		
G		5	5	

Tabela 1: Matriz de avaliações (escala 1 a 5).

No contexto de redes sociais a vizinhança de utilizadores que possuem um perfil semelhante a um determinado utilizador pode ser encontrada através do cálculo do coeficiente de *Pearson Correlation*, tendo como base a matriz de avaliações.

Pearson Correlation é uma medida da correlação (dependência linear) entre duas variáveis X e Y, dando como resultado um valor que pode variar entre 1 e -1, inclusive. Quando o resultado do coeficiente de *Pearson Correlation* corresponde ao valor 1 significa que os utilizadores têm um perfil igual, o valor -1 indica o contrário. Através deste coeficiente, a similaridade $P_{a,u}$ entre dois utilizadores a e u é calculada da seguinte forma:

$$P_{a,u} = \frac{\sum_{i=1}^{m}\left(r_{a,i} - r'_a\right) \times \left(r_{u,i} - r'_u\right)}{\left(\sqrt{\sum_{i=1}^{m}\left(r_{a,i} - r'_a\right)^2}\right) \times \left(\sqrt{\sum_{i=1}^{m}\left(r_{u,i} - r'_u\right)^2}\right) + (\beta)} \tag{2.19}$$

Em que $r_{a,i}$ representa a classificação dada ao item i pelo utilizador a. A variável m indica o número de itens que foram classificados por ambos os utilizadores a e u. O r'_u representa a média de classificação em relação ao conjunto de itens avaliados pelo utilizador u. O β é uma constante que tende para zero, sendo apenas utilizada com o objectivo de evitar que o denominador tenha um valor nulo.

O algoritmo *Pearson Correlation* pode ser aplicado à Tabela 5 (descrita anteriormente), e como resultado é gerado a tabela que se segue.

	Utilizador			
Utilizador	**1**	**2**	**3**	**4**
1	1.0000	-0.9978	0.5335	0.9962
2	-0.9978	1.0000	0.8682	-1.0000
3	0.5335	0.8682	1.0000	1.0000
4	0.9962	-1.0000	1.0000	1.0000

Tabela 2: Coeficiente Pearson's Correlation

Esta tabela mostra que os utilizadores 3 e 4 são muito similares (o coeficiente de correlação é 1), enquanto os utilizadores 1 e 2 apresentam preferências opostas (coeficiente fortemente negativo). A vizinhança de dado utilizador é formada por todos utilizadores que apresentam um valor de similaridade maior do que um dado limiar que pode variar entre]0,1].

A última etapa do algoritmo *Collaborative filtering* consiste em efectuar uma previsão do interesse que cada item apresentaria para um determinado utilizador, com base nas avaliações efectuadas pelo conjunto de utilizadores vizinhos.

A previsão da avaliação que cada utilizador classificaria um determinado item é calculada a partir de uma combinação ponderada das avaliações realizadas pelo conjunto de vizinhos seleccionados.

$$w_{a,i} = r'_a + \frac{\sum_{u=1}^{n}\left(r_{u,i} - r'_u\right) \times P_{a,u}}{\sum_{u=1}^{n} P_{a,u}} \tag{2.20}$$

Onde $w_{a,i}$ é a previsão do valor que o utilizador a classificaria o item i. O $P_{a,u}$ é a similaridade entre o utilizador a e o utilizador u. O r'_a e r'_u corresponde à média das avaliações efectuadas pelo utilizador a e u, respectivamente. O $r_{u,i}$ corresponde à avaliação realizada pelo utilizador u ao item i. O n indica o número de vizinhos do utilizador a.

Os itens recomendados ao utilizador em questão são aqueles que apresentem uma classificação superior a um dado limiar. Em alternativa também podem ser recomendados os n itens que apresentem maior classificação.

Em contraste com as abordagens de recomendação baseadas no conteúdo, a técnica de filtragem colaborativa depende da disponibilidade das avaliações realizadas pelos utilizadores. No entanto, esta técnica não requer qualquer intervenção humana ou automática para extrair as características que definem cada item, porque o conhecimento destas características não é utilizado no processo de recomendação.

Apesar da técnica de recomendação *Collaborative filtering* ser uma das mais utilizadas, apresenta algumas limitações (geralmente designadas por *cold-start*):

- Quando um utilizador ainda não avaliou nenhum item torna-se impossível saber quais são as suas preferências.
- Quando existem poucos utilizadores a tarefa de encontrar os vizinhos de um determinado utilizador é dificultada, provocando uma diminuição da qualidade das recomendações.

Outra limitação está relacionada com o facto de um item que não tenha recebido avaliações por parte dos utilizadores não poder ser recomendado, este problema é designado por *long-tail*.

Alguns sistemas de recomendação, bastante populares, que implementam esta técnica de recomendação são: Netflix[13], Facebook[14] e Google+[15].

2.5.3 Técnica de Recomendação *Knowledge-based*

Todos os algoritmos de recomendação tem como objectivo seleccionar os itens que a apresentam maior relevância para cada utilizador. Como já visto nas secções anteriores os algoritmos de *Collaborative filtering* recomendam itens com base nas preferências dos utilizadores que possuem um perfil semelhante ao do utilizador alvo, ao contrário da técnica *Content-based* que consiste em comparar as descrições de cada item com as informações associadas ao perfil de cada utilizador e recomendar os itens que apresentem maior valor de semelhança.

Ao contrário das técnicas anteriores, a técnica *Knowledge-based* [31] tem como objectivo recomendar os itens que satisfazem uma determinada necessidade do utilizador, evitando recomendações irrelevantes e sem interesse para o utilizador.

A principal vantagem desta técnica está relacionada com o facto de o utilizador poder especificar de forma precisa o que deseja obter, o que proporciona recomendações mais direccionadas com o que pretende. No entanto, esta necessidade do utilizador ser obrigado especificar o que pretende, traduz-se numa sobrecarga adicional.

Por exemplo, caso se pretendesse recomendar restaurantes, o utilizador teria de especificar determinados atributos como o "tipo de cozinha", o "preço" etc.

O Google Hotel Finder[16] e o Entree[17] são alguns exemplos de sistemas de recomendação que implementam a técnica *Knowledge-based*.

[13] www.netflix.com
[14] www.facebook.com
[15] www.plus.google.com
[16] www.google.com/hotelfnder
[17] www.kdd.ics.uci.edu/databases/entree/entree.data.html

2.5.4 Técnica de Recomendação *Demographic Filtering*

A técnica de recomendação *Demographic filtering* [32] utiliza a estratégia de segmentar os utilizadores em classes demográficas. A segmentação dos utilizadores é realizada em função da sua informação demográfica como a idade, género, estado civil, profissão, etc. As recomendações são geradas conforme o estereótipo (conjunto de características encontradas na maioria das pessoas pertencentes a um grupo) a que o utilizador pertence.

A grande desvantagem destes sistemas é o esforço inicial necessário para identificar as classes de estereótipos e a margem de erro elevada que pode existir no processo de agrupar os utilizadores em função dados demográficos. Acresce ainda o problema de poderem existir utilizadores com os mesmos dados demográficos, mas apresentarem gostos completamente distintos. Devido a estas limitações, os sistemas de recomendação que utilizam apenas a informação demográfica não são muito populares.

Um exemplo desse tipo de filtragem pode ser encontrado no sistema Grundy [33]. Que é um sistema que tem como objectivo recomendar livros com base nas informações demográficas de cada utilizador. Quando um utilizador se regista no sistema é de imediato relacionado a um estereótipo e as recomendações são realizadas consoante o tipo de estereótipo.

2.5.5 Sistemas de Recomendação Híbridos

Alguns sistemas de recomendação têm como estratégia utilizar várias técnicas de recomendação em simultâneo. O objectivo é superar as limitações que cada técnica apresenta quando utilizada individualmente, melhorando desta forma a qualidade das recomendações apresentadas.

Burke [34] apresentou a taxonomia para os sistemas de recomendação híbridos. O autor classificou-os nas seguintes categorias:

- *Weighted hybrid*: consiste em combinar a pontuação obtida em cada técnica de recomendação. Neste caso as recomendações apresentadas aos utilizadores são aquelas que possuem maior pontuação média.

- *Switching hybrid*: consiste em seleccionar a melhor técnica de recomendação para cada caso em concreto. Ao contrário do *weighted hybrid* as recomendações são obtidas apenas com base numa única técnica de recomendação (a que for mais adequada).

O Netflix[18] é um bom exemplo de um sistema de recomendação hibrido, uma vez que efectua as recomendações com base nos gostos dos utilizadores que têm um perfil semelhante (*Collaborative filtering*) e com base nas características dos filmes que o utilizador avaliou positivamente no passado (*Content-based*).

[18] www.netflix.com

2.5.6 Técnicas de Recomendação - *Tradeoffs*

Todas as técnicas de recomendação abordadas anteriormente foram objecto de estudo nos meados da década de noventa, deste modo as suas capacidades e limitações são bem conhecidas actualmente.

Tradeoffs das Técnicas de Recomendação		**KB**	**CB**	**CF**	**DF**
Capacidades	Qualidade das recomendações melhora com o tempo		X	X	X
	Feedback implícito é suficiente		X	X	
	Pode identificar grupos de utilizadores			X	X
	Sensível às recentes mudanças de preferências	X			
	Capaz de fazer o mapeamento entre itens e necessidades	X			
Limitações	Novo utilizador (*cold-start*)		X	X	
	Novo item (*cold-start*)			X	X
	Problema da esparsidade			X	X
	Problema da superespecialização		X		
	Problema da generalização			X	X
	É necessária informação demográfica				X
	Não aprende as preferências dos utilizadores	X			

Tabela 3: Técnicas de recomendação - tradeoffs

As técnicas de recomendação *Content-based* (CB), *collaborative filtering* (CF), e *Demographic filtering* (DF), sofrem alguns dos problemas associados ao "arranque a frio" (*cold-start*). O *Content-based* é afectado quando um novo utilizador se regista no sistema, porque é necessário efectuar uma aprendizagem do seu perfil. O *Demographic filtering* é afectado quando se introduz um novo item no sistema, porque é necessário obter o *feedback* sobre o tipo de pessoas que gostam desse item. O *Collaborative filtering* é afectado por ambos os casos porque esta técnica depende das avaliações que os utilizadores fazem a cada item.

As técnicas *Collaborative filtering* e *Demographic filtering* sofrem ainda dos problemas de esparsidade e generalização. O primeiro refere-se ao facto de ser necessário elevadas avaliações por parte dos utilizadores para gerar recomendações de qualidade, se estas avaliações forem muito esparsas todo o processo de recomendação fica comprometido. O segundo problema é devido ao facto destas técnicas serem baseadas em generalizações, desta forma é difícil fornecer recomendações precisas para os utilizadores que tenham um interesse em particular.

A abordagem *Content-based* consegue evitar o problema da esparsidade porque as suas recomendações não dependem das avaliações de outros utilizadores. Em contrapartida apresentam o problema da superespecialização. Geralmente as recomendações tendem a ser muito semelhantes ou repetidas, porque são sempre recomendados os itens que são semelhantes aos que o utilizador gostou no passado.

Os sistemas de recomendação *Knowledge-based* conseguem evitar todos os problemas associados ao *cold-start*, uma vez que recomendações são independentes das avaliações

efectuadas pelos utilizadores. Em contrapartida, estes sistemas requerem um esforço extra dos utilizadores, porque é necessário que estes especifiquem as suas necessidades (tipo de informação que desejam obter).

2.6 Perfil do Utilizador

No contexto de sistemas de recomendação, o perfil do utilizador geralmente é representado por um conjunto de atributos que permitem representar os interesses e preferências de cada pessoa. Deste modo, o perfil é essencial para efectuar recomendações personalizadas, porque permite descriminar os itens que um determinado utilizador prefere.

A informação que constituiu o perfil do utilizador pode variar consoante o domínio do sistema de recomendação. Por exemplo, no caso de um sistema de recomendação de notícias, o perfil do utilizador é constituído pelos tópicos que gosta de ler, tópicos que não gosta de ler, jornais que habitualmente costuma ler, etc.

Não é apenas o conteúdo do perfil do utilizador que difere em função do domínio do sistema de recomendação, mas também a forma como essa informação pode ser adquirida. A informação que permite descrever o perfil do utilizador pode ser obtida de diferentes modos: explicitamente ou implicitamente. No primeiro caso a informação é fornecida directamente pelo utilizador, enquanto no segundo caso a informação provém do comportamento do utilizador.

2.6.1 Feedback Explícito

Quando se pretende obter informação sobre o perfil do utilizador de forma explícita, é comum recorrer-se a formulários ou interfaces de forma a recolher os dados introduzidos pelos utilizadores. Geralmente, a informação que se pretende obter de forma explícita é relativa a aspectos demográficos, como a idade, género, profissão, data de aniversário, estado civil, etc. No entanto, os utilizadores poderão não estar dispostos a divulgar alguns seus dados pessoais, pelo que a especificação destes atributos deverá ser preferencialmente opcional.

Para além dos dados demográficos, os interesses dos utilizadores também podem ser obtidos de forma explícita, por exemplo a Amazon[19] possuiu a funcionalidade "Favoritos" com o objectivo de permitir que o utilizador especifique as categorias de interesse (bibliografias, arte, religião, etc). Também é comum utilizar-se sistemas de votação, com o objectivo de saber se o utilizador gosta ou não de um determinado item. No MovieLens[20] e Netflix[21] os utilizadores podem especificar as suas preferências através da avaliação dos filmes que são recomendados.

No entanto existem vários problemas com a obtenção explícita do perfil do utilizador, em primeiro lugar as pessoas geralmente não estão dispostas a fornecer informações pessoais através do preenchimento de formulários. E mesmo que os utilizadores estejam dispostos a fornecer essa informação, não existe a garantia de que esta esteja correcta ou corresponda à verdade.

[19] www.amazon.com
[20] www.movielens.org
[21] www.netflix.com

2.6.2 Feedback Implícito

Existe a possibilidade de obter a informação sobre um determinado utilizador de uma forma implícita. Nesta vertente é fundamental observar as interacções do utilizador com o sistema e guardar o registo dessas interacções. Estas interacções podem ser por exemplo: itens visualizados, tempo de visualização, itens seleccionados, etc. Com base no histórico de interacções de cada utilizador é possível descobrir determinados padrões de comportamento e desta forma prever os interesses dos utilizadores.

A principal vantagem desta técnica é que não necessita de qualquer intervenção adicionar do utilizador. Uma desvantagem do feedback implícito é que tipicamente apenas é possível obter feedback positivo. Quando um utilizador clica num determinado item, é razoável supor que esse item representa algum interesse para o utilizador. No entanto, não é tão claro, que quando um utilizador deixa de examinar alguns itens, estes sejam do seu desinteresse

2.6.3 Construção do Perfil

Uma das representações mais comum do perfil do utilizador, baseia-se na utilização de palavras-chave ou *tags*. Estas *tags* podem ser fornecidas directamente pelo utilizador ou extraídas dos itens que foram do interesse do utilizador.

Cada *tag* representa um tópico de interesse para o utilizador e com base no conjunto de *tags* específicas de cada pessoa é possível deduzir as suas preferências. Geralmente cada *tag* está associado a um peso, de forma a indicar o grau de interesse que cada pessoa demostra sobre cada tópico [35].

De forma a solucionar o problema relativo à polissemia das palavras, pode ser utilizado uma estrutura de nós [36] em que as palavras com o mesmo significado estão ligadas ao mesmo nó através de um arco.

A actualização do perfil de utilizador é um requisito essencial nos sistemas de recomendação, de forma a ser possível efectuar recomendações precisas com os interesses dos utilizadores. Esta actualização pode ser feita automaticamente ou manualmente pelo utilizador, sendo que a primeira opção é menos invasiva.

Na maioria dos sistemas esta actualização é efectuada com base no *feedback* dado pelo utilizador em relação aos novos itens visualizados. A principal desvantagem deste método de actualização do perfil é que os interesses antigos não são esquecidos, causando não só um crescimento exponencial das *tags* que constituem o perfil do utilizador, mas também uma diminuição da precisão. Uma vez que o sistema de recomendação pode continuar a recomendar informações com base em interesses antigos do utilizador.

Vários mecanismos de esquecimento têm sido propostos com o objectivo de efectuar a actualização do perfil com base na alteração dos gostos dos seres humanos ao longo do tempo [37]. Uma abordagem simples é considerar uma janela de tempo e actualizar o perfil do utilizador com base nas observações que decorreram durante esse intervalo.

2.7 Privacidade

Todos os sistemas de recomendação necessitam de recolher informação sobre os utilizadores, de forma a ser possível recomendar informação contextualizada. Ao longo da última década têm sido efectuadas investigações com o objectivo de garantir a privacidade dos dados de cada utilizador.

Diversas abordagens têm sido propostas com o objectivo de proteger a privacidade dos utilizadores no contexto dos sistemas de recomendação. Canny propôs em [38] a técnica *homomorphic encryption* que tem como objectivo garantir a privacidade dos utilizadores no contexto dos sistemas de recomendação colaborativos. A técnica *homomorphic encryption* permite agregar as avaliações individuais que os utilizadores efectuam nos itens, desta forma não fica exposto as preferências pessoais de cada utilizador em particular.

Outra abordagem foi proposta por Polat and Du [39], esta técnica utiliza perturbações aleatórias (ruido) com uma distribuição conhecida. No entanto existe um *tradeoff* entre o nível de privacidade e a precisão com que é possível restaurar os dados originais

O k-anonymity [40] é um dos algoritmos mais utilizados quando se pretende anonimizar os dados pessoais do utilizador (género, idade, código postal, etc.). Para o processo de anonimizar os dados pessoais dos utilizadores, este algoritmo utiliza técnicas de generalização. A generalização consiste em substituir os valores de um determinado atributo por outros valores mais gerais. Por exemplo, o código postal pode ser generalizado mascarando os dígitos menos significativos com outros caracteres.

Mais recentemente tem existido uma evolução dos sistemas de recomendação descentralizados, estes sistemas apresentam uma arquitectura *peer-to-peer* de forma a evitar a presença de uma entidade central que reúna as informações de todos os utilizadores [41]. Deste modo, a informação encontra-se armazenada do lado do cliente e apenas existe comunicação com os outros utilizadores durante o processo de recomendação.

A nível de protocolos, o P3P[22] permite que os utilizadores tenham maior controlo sobre a utilização de informações pessoais por parte dos *sites* que visitam. A ideia básica do P3P é permitir que cada serviço descreva suas práticas de privacidade, e a forma como lida com informação pessoal do utilizador.

[22] www.w3.orgP3P

Capítulo 3
Arquitectura e Especificação

Neste capítulo é detalhada a arquitectura geral do projecto PPS2-SEMA, do qual faz parte o serviço de recomendação.

3.1 Arquitectura Geral

O PPS2-SEMA é formado pelo serviço de recomendação de informação (ASA), serviço de enriquecimento semântico (ES) e serviço de interoperabilidade semântica (IS)

Em traços gerais, o serviço de ASA permite recomendar os recursos com base nos conceitos de relevância, novidade e diversidade; o serviço de ES permite enriquecer semanticamente os recursos existentes na base de conhecimento; e finalmente serviço de IS é o responsável por todas as interacções com a base de conhecimento.

Estes três serviços estarão acessíveis entre si e para os restantes PPS's na forma de *web service* REST[23].

Na Figura 3.1 está ilustrado o modo como os diferentes serviços interagem entre si.

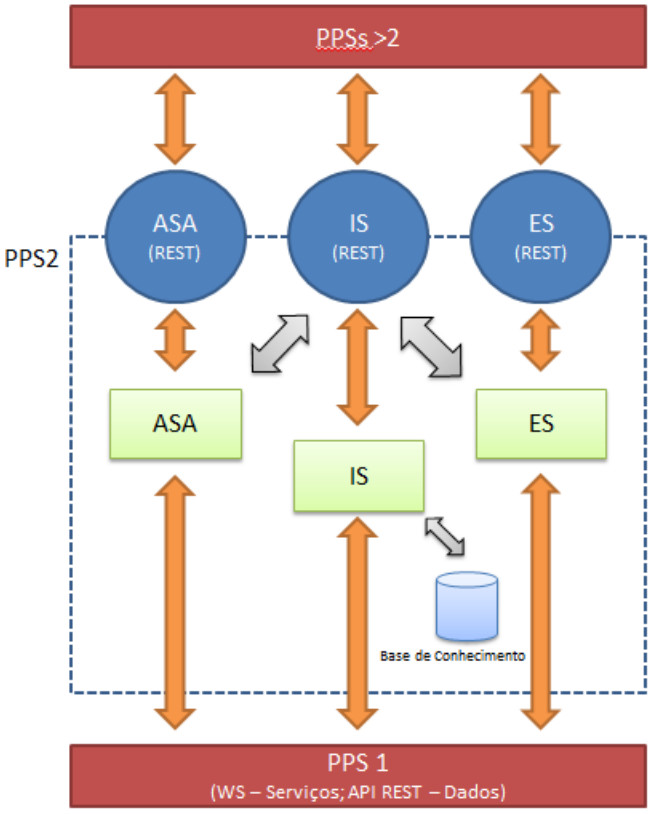

Figura 3.1: Arquitectura do PPS2-SEMA

[23] http://en.wikipedia.org/wiki/Representational_state_transfer

Como se pode constatar através da Figura 3.1 o serviço ASA pode ser acedido através do PPS 1 ou de qualquer outro PPS. A nível interno o serviço ASA comunica directamente com o serviço de interoperabilidade semântica (IS). Esta ligação permite ao serviço ASA obter uma lista de recursos a filtrar, assim como a informação sobre o histórico de interacções e informações demográficas de cada utilizador.

A comunicação entre estes dois serviços é bidireccional, mas é da responsabilidade do serviço ASA iniciar a ligação através do envio do pedido desejado (recursos de um determinado tipo, histórico de interacções de um dado utilizador). Sendo que o serviço IS é responsável por devolver como resposta a informações pedida, efectuando para isso uma consulta na sua base de dados.

3.2 Serviço de Recomendação de Informação (ASA)

O serviço ASA tem como objectivo recomendar determinados recursos aos utilizadores. Estes recursos consistem essencialmente em pontos de interesse, como por exemplo: restaurantes, museus, eventos. No entanto, o serviço tem de ser genérico ao ponto de suportar a recomendação de recursos associados a outros domínios.

Como referido anteriormente, as funcionalidades do serviço de selecção de informação estarão disponíveis na forma de um *web service* REST, podendo ser acedido mediante a execução de pedidos nos diversos verbos do método HTTP[24]. A comunicação com o serviço de recomendação deverá respeitar o formato JSON[25].

A nível do funcionamento, o serviço de recomendação de informação é composto por dois métodos públicos, uma para obter as recomendações e outro para actualizar o histórico de interacções dos utilizadores.

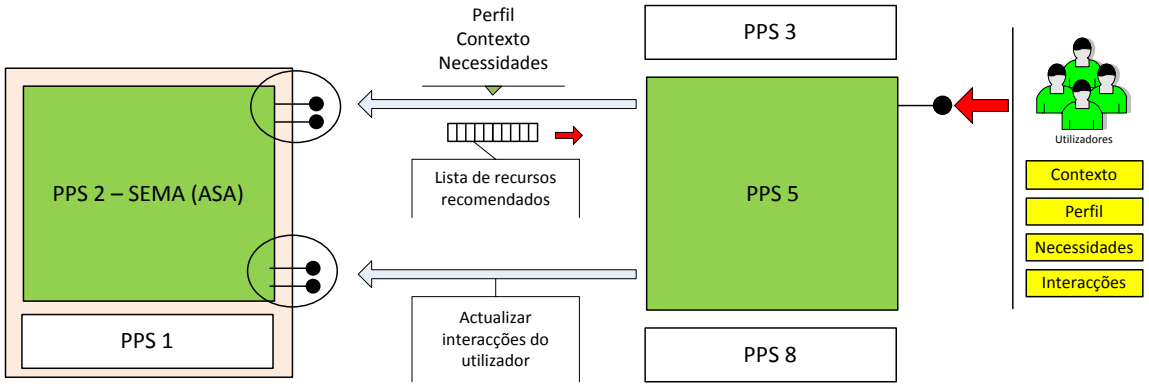

Figura 3.2: Serviço de recomendação de informação

De forma a efectuar as recomendações, o serviço recebe um conjunto de dados como: o perfil do utilizador, informação sobre o contexto actual e a informação sobre as necessidades do utilizador.

[24] http://pt.wikipedia.org/wiki/Hypertext_Transfer_Protocol
[25] http://en.wikipedia.org/wiki/JSON

Com base nesta informação, serão seleccionados os recursos mais apropriados para cada utilizador. Na secção 4.1 estão especificados os detalhes das estruturas de dados enumeradas anteriormente.

De um modo geral, o serviço de recomendação selecciona a informação a ser recomendada através de três fases de filtragem. Os filtros a aplicar estão relacionados com os conceitos de relevância, novidade e diversidade, descritos no Capítulo 2.

Numa primeira fase, são filtrados os recursos mais relevantes para o utilizador. Sendo utilizados para o processo, as características de cada recurso, os interesses, perfil e contexto do utilizador.

A segunda fase tem o objectivo de excluir os recursos com os quais o utilizador se encontra familiarizado. Para tal, recorre-se ao histórico de interacções que o utilizador já teve com o recurso.

Na terceira e última etapa são excluídos os recursos repetidos ou muito semelhantes. O objectivo é seleccionar os recursos que apresentem maior diversidade entre si, de forma a não sobrecarregar o utilizador com informação repetida ou semelhante.

Através da aplicação destes três tipos de filtragem, é possível seleccionar um conjunto de recursos diversos que são relevantes e apresentam novidade para o utilizador em questão. Estes três processos de filtragem encontram-se descritos na secção 4.3.

Como se pode verificar na Figura 3.2, para além do método de recomendação de informação, o serviço ASA possuiu outro método que permite actualizar as interacções que os vários utilizadores tiveram com os recursos. Os utilizadores podem interagir com os recursos de diferentes formas, como por exemplo: avaliar um recurso, visualizar os detalhes de um recurso, visualizar um recurso numa listagem.

3.3 Especificação de Software

Com o objectivo de especificar o serviço de recomendação de informação, foram criados os seguintes diagramas UML[26]:

- Diagrama de casos de uso
- Diagrama de actividades
- Diagrama de classes
- Diagrama de sequência

Os detalhes específicos de cada um destes diagramas podem ser consultados no Anexo D - Especificação de Software.

No diagrama de actividades é mostrado o fluxo de actividades associado ao pedido de selecção de informação por parte do utilizador. De um modo geral, podemos dizer que existem quatro fases distintas: validação dos dados recebidos (perfil, contexto e necessidades), filtragem dos recursos segundo o seu valor de relevância, filtragem dos

[26] http://pt.wikipedia.org/wiki/UML

recursos segundo o seu valor de novidade e filtragem dos recursos com base no seu valor de diversidade.

Desta forma o diagrama de actividades foi dividido em quatro partições: "Validação", "Relevância", "Novidade" e "Diversidade". Cada uma das partições corresponde a uma das fases enumeradas anteriormente.

No diagrama de classes estão representadas as classes que definem o serviço, bem como os seus atributos, métodos e forma de relacionamento. Como já foi visto anteriormente, esta aplicação pode ser dividida em quatro partes distintas: validação dos dados recebidos, cálculo do valor de relevância, novidade e diversidade dos recursos existentes. Desta forma as classes *Validation, Relevance*, *Novelty* e *Diversity* assumem o papel principal.

Neste diagrama apenas se está a considerar o cálculo da relevância e diversidade de três tipos de recursos: restaurantes, hotéis e museus. Contudo o serviço terá de suportar o cálculo da relevância e diversidade de novos tipos de recursos. Desta forma, será utilizada a *design pattern factory* que permite a utilizar objectos relacionados entre si, sem especificar as suas classes concretas.

Mais concretamente a *design pattern factory* é utilizada para criar um objecto específico para o cálculo da relevância e diversidade consoante o tipo do recurso. A especificação técnica da *design pattern factory* pode ser consultada no Anexo A – Design Pattern Factory.

Para mostrar a sequência os eventos entre as classes existentes no diagrama de classes, foi criado um diagrama de sequência. Este diagrama permite mostrar as interacções entre os objectos, para além de fornecer informações sobre a sequência temporal.

Capítulo 4
Desenvolvimento

Neste capítulo estão especificadas todas as estruturas de dados e técnicas de recomendação utilizadas pelo serviço de recomendação. Na Secção 4.3 pode ser consultada a abordagem utilizada no processo de recomendação.

De modo a permitir que utilizadores possam interagir com o serviço de recomendação de informação também foi desenvolvido uma aplicação cliente. Através desta aplicação os utilizadores podem especificar as suas necessidades e o seu contexto, de forma a obter a informação recomendada pelo serviço. Esta aplicação cliente encontra-se especificada na Secção 4.5.

4.1 Estrutura de Dados

Como já foi referido anteriormente, a selecção dos recursos a apresentar ao utilizador é feita com base no seu perfil, contexto actual, necessidades e no histórico de interacções que o utilizador teve com os recursos.

Deste modo, o processo de recomendação actua sobre as seguintes estruturas de dados:

- Recursos
- Perfil do utilizador
- Contexto actual
- Necessidades do utilizador
- Histórico de interacções.

Os recursos correspondem à informação que se pretende recomendar. Estes recursos podem pertencer a diferentes domínios. No entanto, para o processo de avaliação do sistema de recomendação, os recursos utilizados correspondem a pontos de interesse: restaurantes e hotéis.

O perfil do utilizador refere-se a um conjunto de atributos que permitem identificar e descrever um utilizador. Como por exemplo a sua informação demográfica e interesses.

O contexto do utilizador refere-se à descrição do meio ambiente em que o utilizador se encontra inserido no momento em que utiliza o sistema. De um modo geral, o contexto é definido com base na dimensão temporal e geográfica.

As necessidades correspondem à informação que o utilizador deseja obter num determinado momento. Desta forma os utilizadores podem especificar para cada atributo dos recursos os valores alvos que desejam obter.

O histórico de interacções é constituído por todas as interacções que o utilizador teve com os recursos disponibilizados pelo sistema de recomendação. Estas interacções são: "visualizar um recurso numa listagem", "ver detalhes de um recurso", "obter direcções para um recurso" e "avaliar um recurso".

Toda a informação referente às estruturas de dados enumeradas anteriormente pode ser consultada no Anexo B – Estruturas de Dados.

4.1.1 Perfil do utilizador

O perfil do utilizador permite representar toda a informação que o sistema de recomendação possuiu sobre cada utilizador.

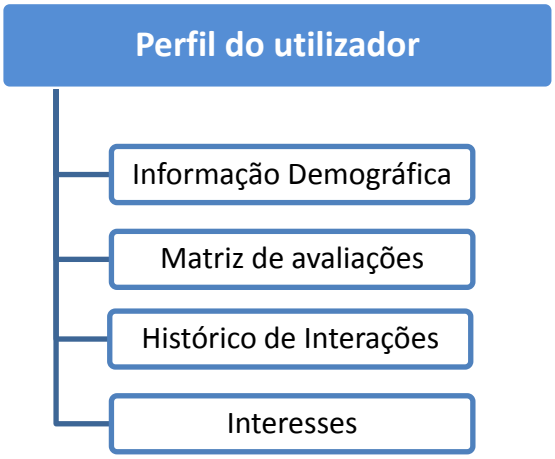

Figura 4.1: Perfil do utilizador

A informação que caracteriza o perfil de cada utilizador é a seguinte:

- **Informação Demográfica:** corresponde à informação demográfica de cada utilizador como a idade, género, estado civil, profissão etc.

- **Interesses:** conjunto de *tags* relacionadas com os recursos que o utilizador gostou no passado.

- **Histórico de interacções:** representa as interacções que cada utilizador teve com os recursos (visualizar, obter direcções, avaliar).

- **Matriz de avaliações:** matriz que relaciona as avaliações que o utilizador faz aos recursos.

4.1.2 Recursos

Os recursos que se pretendem recomendar correspondem sobretudo a pontos de interesse (POIs). Dado que pode existir uma grande variedade de recursos a recomendar e como cada tipo de recurso pode utilizar diferentes funções para calcular a relevância e diversidade foi utilizado a *design pattern Abstract Factory* (pode ser consultada no Anexo A – Design Pattern Factory). Desta forma é possível introduzir novos tipos de recursos de forma compatível com todos os algoritmos já desenvolvidos.

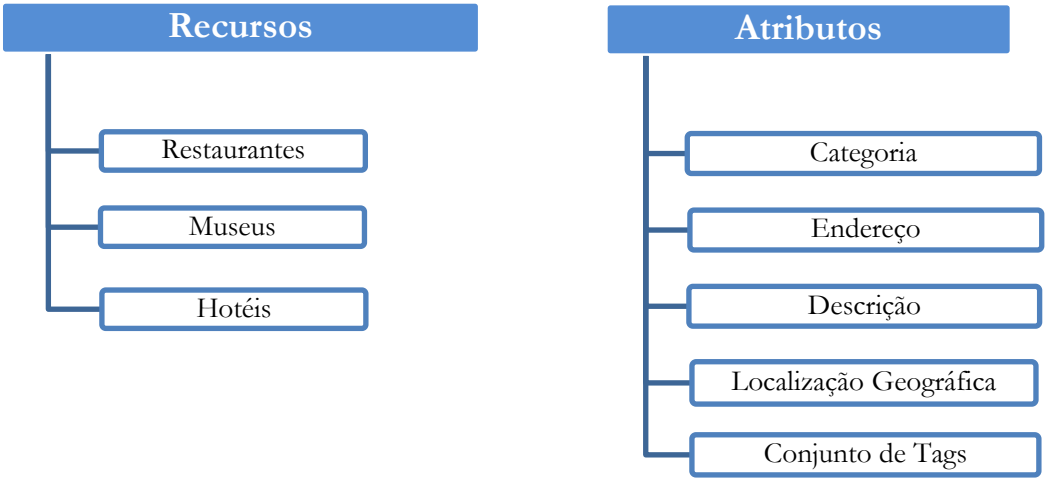

Figura 4.2: Estrutura dos recursos

Cada recurso é constituído por um conjunto de atributos base, embora possam existir outros atributos específicos de cada tipo de recurso. Os atributos específicos de cada recurso podem ser consultados no Anexo B – Estruturas de Dados.

Nesta fase foram utilizados dados sobre restaurantes e hotéis provenientes do *website* lifecooler[27]. Recorreu-se a técnicas de *screen-scrapping* para obter a informação.

4.2 Aprendizagem do Perfil do Utilizador

O perfil do utilizador é constituído com base na informação demográfica introduzida directamente pelo utilizador e pelo histórico de interacções. O histórico de interacções é constituído por todas as interacções que o utilizador teve com os recursos disponibilizados pelo sistema de recomendação. Estas interacções são: "visualizar um recurso numa listagem", "ver detalhes de um recurso", "obter direcções para um recurso" e "avaliar um recurso".

Uma parte importante de todo o processo de recomendação centra-se nas avaliações que cada utilizador efectua nos diversos recursos recomendados. Neste caso em concreto, foi utilizada uma escala de avaliação que pode varia entre 1 e 5 (5 corresponde à avaliação máxima).

	Utilizador			
Recurso	**U1**	**U2**	**U3**	**U4**
A	5	3	4	
B	4			5
C		2		4

Tabela 4: Matriz de avaliações

[27] www.lifecooler.com

Esta matriz será posteriormente utilizada para prever a relevância que cada recurso apresenta para um determinado utilizador, com base nas avaliações de outros utilizadores que tenham um perfil semelhante (filtragem colaborativa).

4.2.1 *Tags* de Interesse

A aprendizagem dos interesses de cada utilizador é efectuada com base no conjunto de avaliações realizadas pelo utilizador. Deste modo, se um utilizador avaliou positivamente um determinado recurso, podemos assumir com algum grau de confiança que o utilizador demostra interesse por todos os recursos que apresentem características idênticas.

Com o objectivo de construir os interesses de cada utilizador de forma dinâmica, foi utilizado o conceito de *tags*. Desta forma, sempre que um utilizador avalia positivamente um determinado recurso, são recolhidas as *tags* que definem esse recurso e procede-se à actualização dos seus interesses.

A actualização dos interesses é realizada através da adição das *tags* provenientes dos recursos que o utilizador mais gostou. Neste caso em concreto, foi considerado que o utilizador gostou de um determinado recurso quando efectuou uma avaliação superior a três valores (escala de 1 a 5).

Como existem *tags* que podem ter mais importância do que outras, é utilizado um campo que permite especificar o peso de cada *tag*. Os interesses do utilizador U são definidos por um conjunto de pares (*tag:* peso).

$$U = \left(t_{i,1}: v_{i,1}, t_{i,2}: v_{i,2}, \dots, t_{i,n}: v_{i,n}\right) \tag{4.1}$$

Em que $t_{x,1}$ identifica as várias *tags* e o $v_{i,x}$ representa a importância que cada *tag* apresenta no conjunto de interesses. O $v_{i,x}$ é calculado do seguinte modo:

$$v_{i,x} = \frac{N_{i,x}}{N_i} \tag{4.2}$$

Onde $N_{i,x}$ indica o número de vezes que o utilizador i gostou da *tag* x, o N_i corresponde ao número de recursos que o utilizador gostou.

As *tags* que definem os interesses de cada utilizador são segmentadas em função das categorias do recurso a que pertencem. Na Tabela 1 é mostrado um exemplo dos interesses de um determinado utilizador em relação aos recursos de tipo restaurante.

Tags	Peso
Tradicional Portuguesa	0.6
Fast Food	0.4

Tabela 5: Tags de interesse em relação aos restaurantes

Esta tabela indica que 60% dos restaurantes que o utilizador gostou no passado, estão associados à *tag* "Tradicional Portuguesa". A *tag* "Fast Food" surge apenas em 40% dos restaurantes que são do interesse do utilizador.

Todos os recursos sujeitos a serem recomendados possuem um campo que contém uma lista de *tags* específicas. Sendo que a extracção das *tags* que caracterizam cada recurso é assegurada pelo serviço de enriquecimento semântico (ES), referido anteriormente na arquitectura do sistema (Capítulo 3).

No entanto, enquanto o serviço ES se encontra em fase de desenvolvimento, a extracção das *tags* que caracterizam cada recurso foi efectuada com base na informação presente no seu conteúdo (atributos e descrição). O algoritmo de extracção de *tags* desenvolvido selecciona as cinco palavras que ocorrem com maior frequência no conteúdo de cada recurso. Como nem todos os termos são relevantes, foram utilizados filtros que excluem determinadas palavras como artigos, pronomes e alguns verbos.

De forma a calcular a importância que cada *tag* apresenta para o recurso foi utilizado o *Term Frequency*.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{4.3}$$

O *Term Frequency* é a ocorrência do termo t_i no documento j, $n_{i,j}$, dividido pelo número de ocorrências de todos os termos.

4.3 Abordagem de Recomendação

Como já foi referido anteriormente, a abordagem utilizada no processo de recomendação é baseada em três etapas: relevância, novidade e diversidade. Em cada uma das etapas é aplicado um determinado processo de filtragem.

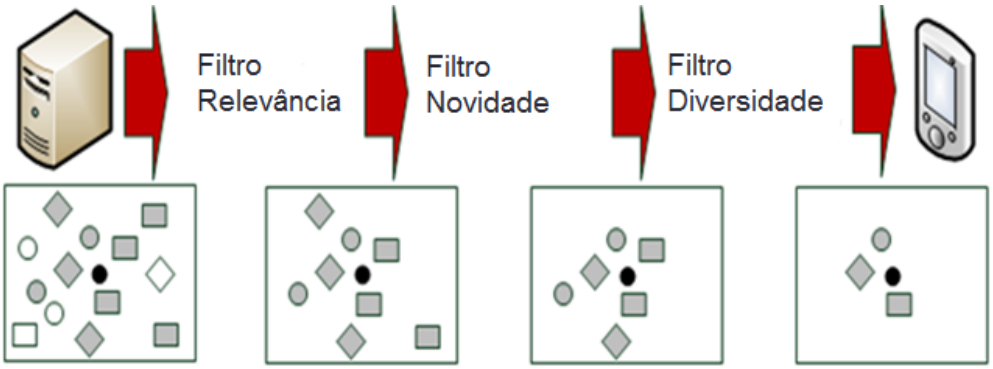

Figura 4.3: Processo de recomendação

Através da filtragem por relevância são seleccionados os recursos que apresentam maior interesse para o utilizador (valor de relevância superior a um dado limiar). A segunda fase de filtragem corresponde à filtragem por novidade, o objectivo é excluir os recursos que o utilizador melhor conhece. Também é utilizado um limiar de novidade, e todos os recursos que apresentem um valor de novidade menor do que o limiar são descartados.

Na terceira e última etapa são excluídos os recursos repetidos ou muito semelhantes. O objectivo é seleccionar os recursos que apresentem maior diversidade entre si, de forma a não sobrecarregar o utilizador com informação repetida ou semelhante.

Como resultado da aplicação destas três fases de filtragem, é possível seleccionar um conjunto de recursos diversos que são relevantes e apresentam novidade para o utilizador.

Na secção que se segue estão especificados os algoritmos associados às técnicas de filtragem enumeradas anteriormente.

4.4 Técnicas de Filtragem

Como referido anteriormente o processo de recomendação envolve três fases de filtragem, que são filtragem por relevância, filtragem por novidade e filtragem por diversidade. Cada uma destas fases de filtragem é explicada nas subsecções que se seguem.

4.4.1 Filtragem por Relevância

O processo de filtragem por relevância tem como objectivo seleccionar os recursos que apresentam um valor de relevância superior a um dado limiar.

Como referido anteriormente na Secção 2.5, podem ser utilizadas diferentes técnicas para calcular o valor de relevância que cada recurso apresenta para um determinado utilizador. Tendo em conta que cada técnica tem as suas vantagens e limitações, criou-se uma árvore de decisão para decidir qual a técnica a aplicar em cada situação.

Outro aspecto importante refere-se ao horário de funcionamento de determinados recursos, neste caso é verificado se o horário é compatível com a data e hora especificada no contexto do utilizador. Caso não exista compatibilidade, optou-se por atribuir um valor de relevância zero. Na nossa opinião não é útil recomendar recursos que não se encontrem disponíveis no período temporal desejado pelo utilizador.

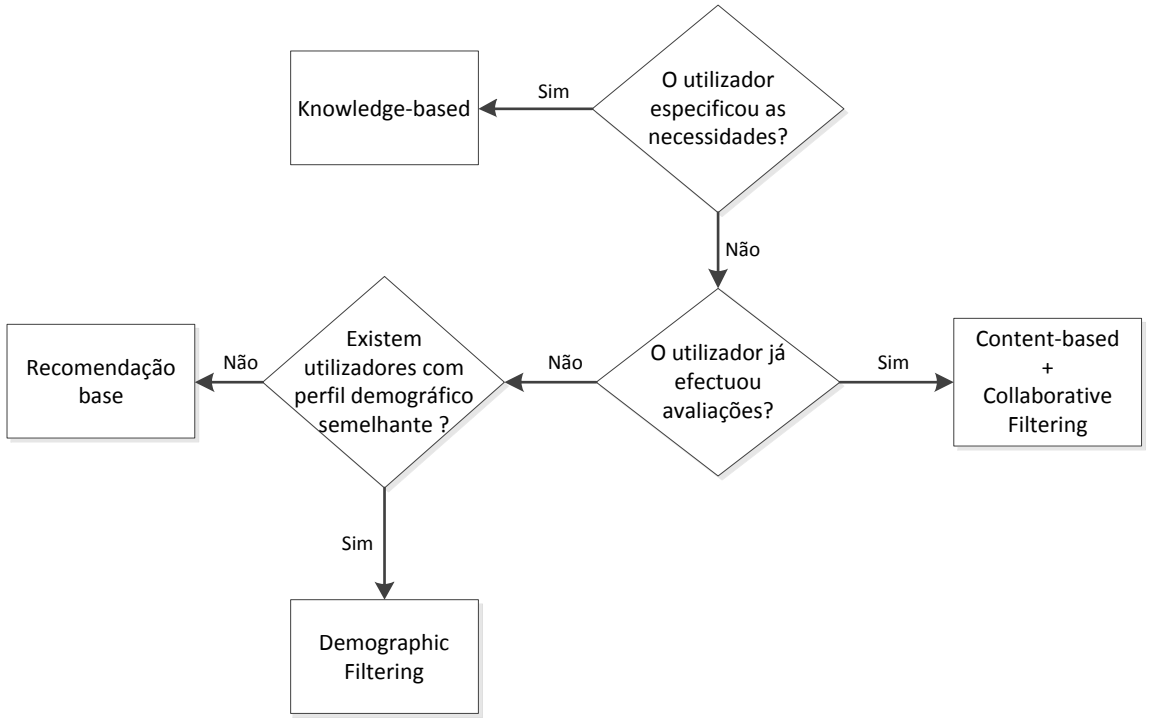

Figura 4.4: Técnica a utilizar para calcular a relevância

Através desta árvore de decisão podemos verificar que quando os utilizadores especificam as suas necessidades a relevância de cada recurso é calculada através da técnica *Knowledge-based*.

Caso o utilizador não tenha especificado as suas necessidades, será verificado se este utilizador já avaliou anteriormente alguns dos recursos. Se o utilizador já efectuou avaliações, existe informação que permite conhecer as suas preferências. Neste caso a relevância de cada recurso é calculada com base na combinação das técnicas *Content-based* e *Collaborative Filtering*, estas técnicas utilizam as preferências que o utilizador demostrou no passado.

É comum utilizar a combinação destas técnicas de forma a superar as limitações que cada uma delas apresenta quando é utilizada individualmente. Uma vez que a técnica *Content-based* apenas permite calcular a relevância dos recursos que sejam descritos por um conjunto de *tags*, e a técnica *Collaborative Filtering* apenas permite calcular a relevância dos recursos que foram previamente avaliados por outros utilizadores que tenham um perfil semelhante ao do utilizador alvo.

Como a precisão destas técnicas aumentam consoante o número de avaliações já realizadas pelo utilizador, optou-se por apenas aplicar estas técnicas despois do utilizador avaliar no mínimo três recursos.

Caso não sejam conhecidas as preferências do utilizador, e se o utilizador apresentar informação demográfica semelhante a outros utilizadores que já efectuaram avaliações, é utilizada a técnica *Demographic filtering*. No caso de nenhuma das condições apresentadas anteriormente se verificar, a relevância dos recursos é apenas calculada com base na distância geográfica a que os recursos (POIs) se encontram do utilizador.

Calcular relevância através da técnica *Knowledge-based*

Apenas é possível utilizar a técnica *Knowledge-based* quando os utilizadores especificam as suas necessidades, indicando desta forma os valores alvos que desejam obter. A representação das necessidades do utilizador está especificada na Tabela B.5 e Tabela B.6 do Anexo B – Estruturas de Dados. Por exemplo, no caso de o utilizador desejar obter recomendações sobre restaurantes, poderia especificar alguns atributos como o tipo de cozinha, preço médio, ambiente, etc.

Para calcular a relevância de cada recurso, são utilizadas as funções de distância apresentadas na secção 2.3.1. Estas funções permitem calcular a distância a que se encontra cada atributo dos recursos ao alvo desejado, sendo que para cada tipo de atributo é necessário utilizar uma determinada função de distância.

De um modo geral estas funções permitem calcular a distância entre atributos numéricos, textuais, geográficos e temporais. Para calcular o valor de relevância que cada recurso apresenta para o utilizador é utilizada a informação sobre os atributos dos recursos e a informação sobre as necessidades do utilizador.

Deste modo, o valor de relevância de cada recurso resulta do somatório das distâncias existentes entre os vários atributos de cada recurso e os alvos desejados. O valor de relevância de cada recurso é calculado através da seguinte equação:

$$Rel = \sum_{j=1}^{K} \left(1 - Dist(Aj, Rj)\right) \times w_k \, , w_k \in [0,1] \tag{4.4}$$

A variável k representa o número de atributos que constituiu um determinado recurso, o A_j representa o valor alvo para o atributo j e o R_j indica o valor que está associado a esse mesmo atributo.

A função *Dist* é responsável por calcular a distância que existe entre o valor do atributo e o alvo desejado. O w_k indica o peso que está associado a cada atributo, podendo existir atributos com pesos diferentes (os pesos utilizados podem ser consultados no Anexo C - Pesos dos Atributos). Desta forma é possível especificar os atributos que são mais importantes para o cálculo de relevância.

Nesta fase os pesos que definem a importância de cada atributo foram definidos por defeito de forma empírica. No entanto, como trabalho futuro é necessário estudar a possibilidade de efectuar uma aprendizagem da importância de cada atributo. Outra possibilidade seria permitir que os utilizadores definissem todos os pesos, no entanto esta solução não é muito atractiva, porque implica um grande esforço por parte do utilizador.

Caso o valor de algum atributo R_j não esteja disponível, assume-se que o resultado da função $Dist(Aj, Rj)$ é igual a 1, ou seja, máxima distância. Além disso, se o utilizador não especificar o valor alvo para um determinado atributo, assume-se que o utilizador está interessado em todos os valores possíveis para esse atributo, e deste modo, o resultado da

função $Dist(Aj, Rj)$ é igual a 0. Na Tabela B.6 do Anexo B é possível consultar um exemplo que corresponde a uma situação em que o utilizador especifica as suas necessidades (valores alvos).

Calcular relevância através da técnica *Content-based*

Como referido na Secção 2.5.1, a técnica *Content-based* calcula a relevância de cada através da comparação entre o perfil do utilizador e a descrição dos recursos que se pretendem recomendar.

O perfil do utilizador é descrito por um conjunto de *tags* e o respectivo peso (como referido na Secção 4.2.1), assim como cada recurso possível de recomendar. Desta forma, o processo para calcular o valor de relevância de cada recurso envolve a comparação entre o conjunto de *tags* que descrevem o perfil do utilizador e as *tags* que definem o recurso. Com este objectivo foi utilizada a solução apresentada pelo autor Yi Cai [42].

$$rel = \frac{\sum w_{c,x} \times v_{i,x}}{m} \tag{4.5}$$

Onde $v_{i,x}$ indica a importância que a *tag* x tem para o utilizador i. O m representa o número de *tags* que constituem o perfil do utilizador. O $w_{c,x}$ indica o peso que a *tag* x tem para o recurso c. Deste modo, se uma *tag* pertencer ao perfil do utilizador mas não existir no recurso, o $w_{c,x}$ é igual a zero.

Na maioria das situações os interesses do utilizador são apenas definidos por uma lista de *tags*. No entanto, devido ao facto do sistema de recomendação poder efectuar recomendações de recursos pertencentes a várias categorias, foi efectuada uma segmentação das *tags* em função das categorias dos recursos possíveis de recomendar. Desta forma, o utilizador possui um conjunto de *tags* que definem as suas preferências em relação aos restaurantes, outro conjunto de *tags* que dizem respeito aos hotéis, e assim sucessivamente.

Calcular relevância através da técnica *Collaborative filtering*

Através da técnica *Collaborative filtering* a relevância de cada recurso é calculada com base das avaliações dos utilizadores que têm um perfil semelhante ao do utilizador alvo. Neste caso os utilizadores têm um perfil semelhante quando avaliam os mesmos recursos de forma idêntica. No processo de filtragem colaborativa é utilizada uma matriz de avaliações que relaciona as avaliações que os utilizadores efectuaram nos recursos.

	Utilizador			
Recurso	**U1**	**U2**	**U3**	**U4**
A		3	4	4
B	4			5
C		2		4

Tabela 6: Matriz de avaliação Utilizador × Recurso

Como referido anteriormente, o algoritmo *Collaborative filtering* envolve três etapas:

1. Calcular a similaridade entre utilizadores
2. Obter os utilizadores semelhantes ao utilizador alvo
3. Prever o valor de relevância de cada recurso

Numa primeira fase é calculado o grau de semelhança entre um determinado utilizador alvo e todos os restantes. Para o processo é utilizado a matriz de avaliações dos utilizadores e o coeficiente de *Pearson Correlation*. Através deste coeficiente, os utilizadores que avaliam os mesmos recursos de forma idêntica possuem elevado valor de semelhança.

$$P_{a,u} = \frac{\sum_{i=1}^{m}\left(r_{a,i} - r'_a\right) \times \left(r_{u,i} - r'_u\right)}{\left(\sqrt{\sum_{i=1}^{m}\left(r_{a,i} - r'_a\right)^2}\right) \times \left(\sqrt{\sum_{i=1}^{m}\left(r_{u,i} - r'_u\right)^2}\right) + (\beta)} \tag{4.6}$$

O $P_{a,u}$ representa o grau de semelhança entre os utilizadores a e u, que pode variar entre -1 e 1 (o 1 indica que os utilizadores tem preferências iguais e o -1 o contrário). Em que $r_{a,i}$ representa a classificação dada ao recurso i pelo utilizador a. A variável m indica o número de recursos que foram classificados por ambos os utilizadores. O r'_u corresponde a média das classificações em relação ao conjunto de recursos avaliados pelo utilizador u. O β é uma constante que tende para zero, sendo utilizada para evitar que o denominador tenha um valor nulo.

Na segunda fase são seleccionados todos os vizinhos do utilizador alvo, neste contexto o termo vizinho refere-se a todos os utilizadores que apresentem um grau de semelhança superior a um dado limiar. Neste caso em concreto, definiu-se o limiar de semelhança com o valor 0.5, sabendo que 1 significa que as preferências dos utilizadores são exactamente as mesmas e -1 indica que os utilizadores têm gostos opostos. Esta opção foi tomada em conformidade com a avaliação realizada na Secção 5.2.

Na terceira e última etapa é realizada uma previsão da importância que cada recurso, ainda não classificado pelo utilizador, apresentaria para o utilizador em questão. Esta previsão tem como base as avaliações realizadas pelo conjunto de utilizadores vizinhos obtidos na fase anterior.

A previsão do valor que cada utilizador classificaria um determinado recurso é calculada a partir de uma combinação ponderada das avaliações realizadas pelos vizinhos seleccionados.

$$w_{a,i} = r'_a + \frac{\sum_{u=1}^{n}(r_{u,i} - r'_u) \times P_{a,u}}{\sum_{u=1}^{n} P_{a,u}} \quad (4.7)$$

Onde $w_{a,i}$ é a previsão do valor que o utilizador a classificaria o recurso i. O $P_{a,u}$ é a semelhança entre o utilizador a e o utilizador u (Equação 4.6). O n corresponde ao número de vizinhos do utilizador a.

Como resultado obtém-se o valor de relevância que cada recurso apresenta para utilizador a, com base nas avaliações dos utilizadores que constituem a sua vizinhança. No entanto, como referido anteriormente na Secção 2.5.2, para os recursos que não receberam avaliações não é possível efectuar a previsão do seu valor de relevância.

Calcular relevância através da técnica *Demographic filtering*

A técnica *Demographic filtering* tem como objectivo prever a relevância de cada recurso com base nas avaliações de todos os utilizadores que possuem informação demográfica semelhante (género, idade, estado civil, grau académico, etc.) à do utilizador alvo. Na tabela que se segue está representado um exemplo da informação demográfica utilizada.

	Género	Idade	Estado civil	Nível Financeiro	Grau académico
Utilizador 1	M	20	Solteiro	Baixo	12º
Utilizador 2	M	35	Casado	Médio	Licenciatura

Tabela 7: Exemplo da informação demográfica dos utilizadores

Para o atributo "idade" foi considerado que a idade coincide quando os utilizadores pertencem à mesma faixa etária (jovens, adultos, idosos). Em relação ao atributo "nível financeiro" foram consideradas três categorias: "baixo", "médio", "alto".

O algoritmo *Demographic filtering* envolve duas etapas:

- Seleccionar o conjunto de utilizadores semelhantes ao utilizador alvo.
- Calcular relevância de cada recurso com base no conjunto de vizinhos.

A relevância de cada recurso é calculada com base na média das avaliações realizadas pelo conjunto de utilizadores que têm um perfil demográfico semelhante ao do utilizador alvo. Uma das limitações é o facto de não ser possível calcular a relevância para os recursos que não foram previamente avaliados pelo conjunto de utilizadores que definem a vizinhança.

O limiar de semelhança utilizado para definir o conjunto de utilizadores vizinhos é de extrema importância. Se este limiar for demasiado elevado ocorre o risco da cobertura das previsões ser reduzida, devido ao facto de existirem poucos utilizadores em comum. No

entanto para limiares baixos, a previsão do valor de relevância tende a ser menos precisa. Para avaliar a precisão desta técnica foram realizados testes que podem ser consultados na Secção 5.2.1.

Com base nos testes realizados, definiu-se que os utilizadores são semelhantes quando apresentam três atributos demográficos em comum. Porque para este caso conseguiu-se melhorar a capacidade de prever o valor de relevância que cada recurso apresenta para o utilizador, sem que a cobertura das previsões seja demasiado prejudicada.

4.4.2 Filtragem por Novidade

Como referido na secção 2.2.1, a novidade está relacionada com as expectativas do utilizador. De um modo geral, podemos considerar que as expectativas do utilizador dependem das experiências anteriores, que lhe permite deduzir o que vai acontecer num determinado contexto.

Com o objectivo de suportar a modelação das expectativas dos utilizadores, é utilizado o conceito de *grau de apropriação*. Este conceito define a relação que o utilizador tem com um determinado recurso, com base no seu histórico de interacções.

As interacções que o utilizador pode ter com cada recurso podem ser de diferentes tipos. Deste modo, é necessário definir um peso para cada tipo de interacção, de forma a ser possível dar maior importância às interacções mais marcantes.

Tipo de interacção	**Pesos**
Avaliar um recurso	1
Pedir direcções para a localização do recurso	0,75
Ver informações sobre os detalhes do recurso	0,5
Visualizar o recurso	0,25

Tabela 8: Peso associado aos tipos de interacções

Outro aspecto considerado diz respeito ao tempo que os utilizadores conseguem lembrar das interacções passadas. Sendo que a tendência é lembrar os acontecimentos mais recentes e esquecer os que já aconteceram há mais tempo. Esta característica é designada por *decaimento de memória*.

Como não existem modelos matemáticos que traduzam o *decaimento de memória*, optamos por desenvolver o nosso próprio modelo. No entanto, os valores deste modelo foram definidos de forma empírica.

Data da interacção	Peso
Na última semana	1
Há mais de uma semana e menos de um mês	0,75
Há mais de um mês e menos de um ano	0,5
Há mais de um ano	0,25

Tabela 9: Decaimento de memória em função do tempo

Com base no tipo e na data dos contactos que o utilizador teve com um determinado recurso R_i é possível calcular o grau de apropriação, através da seguinte expressão matemática:

$$GA(R_i) = \sum_{i=1}^{k} w_i \times p_i, w_i, p_i \in [0,1] \tag{4.8}$$

O valor do grau de apropriação do utilizador com cada recurso resulta do somatório dos pesos associados a cada uma das k interacções que o utilizador teve com o recurso. Sendo que o w_i indica o peso associado ao tipo de interacção e o p_i diz respeito ao peso associado à data da interacção.

Um grau de apropriação elevado significa que o utilizador está familiarizado com o recurso, deste modo, quanto menor for o grau de apropriação maior será a probabilidade do recurso apresentar novidade para o utilizador.

Depois de calcular o grau de apropriação de todos os recursos, é necessário aplicar uma normalização de escala (entre 0 e 1).

$$NGA(R_j) = \frac{GA(R_J) - min(GA(R_k))}{max(GA(R_k)) - min(GA(R_k))} \tag{4.9}$$

Onde o $NGA(R_j)$ é o *grau de apropriação* normalizado para o recurso R_j. O $min(GA(R_k))$ e $max(GA(R_k))$ correspondem aos valores de *grau de apropriação* mínimo e máximo existentes na lista de recursos.

Os recursos que apresentem um valor normalizado do *grau de apropriação* inferior a um dado limiar (entre [0,1]), são seleccionados para a próxima e última fase de filtragem (filtragem por diversidade).

4.4.3 Filtragem por Diversidade

A última etapa corresponde à filtragem por diversidade, em que é seleccionado um conjunto de recursos diversos de forma a não sobrecarregar o utilizador com informação repetida ou muito semelhante. Nesta fase é recebida uma lista de recursos que foram previamente seleccionados nas filtragens anteriores (filtragem por relevância e novidade).

Se a lista for constituída por recursos de várias categorias, por exemplo, hotéis e restaurantes, procede-se à divisão da lista em função da categoria.

Despois de efectuada a separação dos recursos em função da sua categoria, o algoritmo de diversidade procede à selecção de um recurso de cada lista. Sendo que este processo é repetido até serem seleccionados o número de recursos desejados, ou não existirem mais recursos candidatos.

Figura 4.5: Sequência de selecção

O primeiro recurso a ser seleccionado de cada lista é sempre o que apresentar maior valor de relevância. Para seleccionar os restantes recursos, é necessário calcular a diversidade que estes apresentam em relação aos recursos do mesmo tipo já seleccionados.

A diversidade entre dois recursos é calculada através do somatório da distância existente entre os atributos de ambos os recursos. Como podem existir atributos que são mais importantes do que outros, cada atributo esta associado a um determinado peso. Os pesos utilizados encontram-se especificados no Anexo C - Pesos dos atributos.

$$Diff(R_i, R_j) = \sum_{a=1}^{K} \Big(Dist\big(R_{ia}, R_{ja}\big)\Big) \times w_k \,, w_k \ \in [0,1] \tag{4.10}$$

O R_i e R_j correspondem aos recursos a comparar. O w_k indica a importância do atributo k. A função *Dist* é utilizada para calcular a distância entre os atributos dos recursos que podem ser numéricos, textuais ou um conjunto de *tags*. Estas funções de distância encontram-se especificadas na Secção 2.3.1.

Apenas os recursos que apresentem um valor de diversidade superior a um limiar β são seleccionados. No entanto, como é importante manter os recursos mais relevantes, se um recurso candidato for muito semelhante a outro que já tenha sido seleccionado, optamos por escolher o que apresenta maior relevância para o utilizador.

4.5 Aplicação Cliente

De forma a ser possível os utilizadores interagirem com o serviço de recomendação de informação, foi desenvolvida uma aplicação cliente. Esta aplicação também foi utilizada com o objectivo de avaliar o serviço de recomendação num cenário com utilizadores reais.

A aplicação cliente é um *website* que permite efectuar a comunicação entre os utilizadores e o serviço de recomendação. Através desta aplicação, os utilizadores podem especificar o seu perfil, contexto e necessidades, de forma a obter as recomendações associadas ao pedido em questão. Na Figura 4.6 está ilustrado o processo de comunicação.

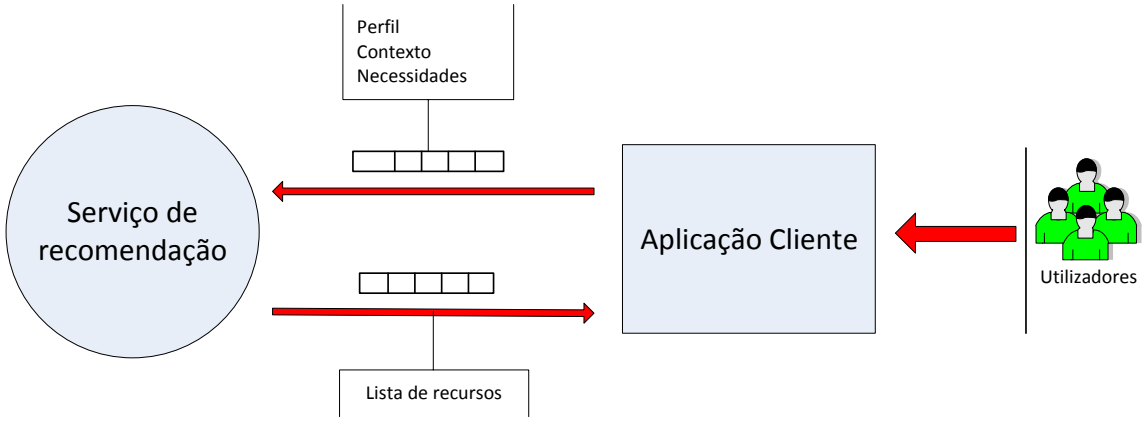

Figura 4.6: Interacção entre a aplicação cliente e o serviço de recomendação

A comunicação entre a aplicação cliente e o serviço de recomendação é efectuada através dos métodos GET e POST pertencentes ao protocolo HTTP. As mensagens de comunicação são estruturadas no formato JSON.

A nível de interface aplicação cliente é composta por três secções distintas:

- **Definições:** Permite que o utilizador especifique o seu contexto (localização geográfica e data/hora do dia), as necessidades e o número de itens que deseja obter.

- **Recomendações:** nesta secção é mostrada a lista de recomendações devolvida pelo serviço de recomendação.

- **Detalhes:** nesta secção é possível verificar os detalhes de um determinado recurso recomendado e a sua localização no mapa.

Na imagem que se segue está representada a interface da aplicação cliente.

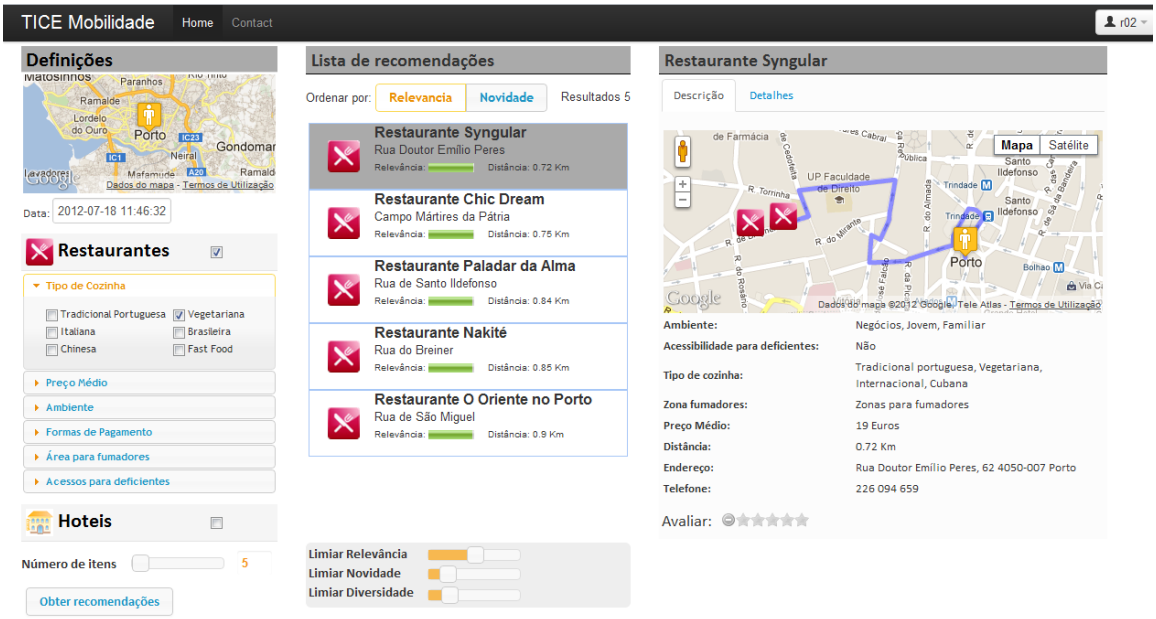

Figura 4.7: Aplicação cliente

Os limiares de relevância, novidade e diversidade são especificados pelo utilizador, no entanto é feita uma aprendizagem com base na média dos valores de limiar definidos nas várias interacções com o sistema. Desta forma, quando os utilizadores entram no sistema, os limiares de relevância, novidade e diversidade são definidos por defeito com base na média dos limiares já especificados pelo utilizador em interacções passadas.

Para o desenvolvimento da aplicação cliente foram utilizadas tecnologias como: HTML[28], JSP[29], AJAX[30], CSS[31] e Javascript[32].

[28] http://pt.wikipedia.org/wiki/HTML
[29] http://pt.wikipedia.org/wiki/JavaServer_Pages
[30] http://pt.wikipedia.org/wiki/AJAX
[31] http://pt.wikipedia.org/wiki/Cascading_Style_Sheets
[32] http://pt.wikipedia.org/wiki/JavaScript

Capítulo 5
Resultados

Neste capítulo são analisados os resultados provenientes da avaliação que foi efectuada ao serviço de recomendação. Na Seção 5.1 é apresentada a experiência de avaliação realizada, assim como a sua discussão. Na Secção 5.2 são discutidos os resultados que avaliam a precisão das técnicas de recomendação *Collaborative filtering* e *Demographic Filtering*.

5.1 Avaliação do Serviço de Recomendação

Um dos objectivos do trabalho consistia em avaliar até que ponto os filtros de novidade e diversidade contribuíam para melhorar a qualidade das recomendações.

O processo de avaliação consistiu em comparar dois sistemas recomendação. Um sistema de recomendação base que apenas utiliza a filtragem por relevância e um outro sistema que implementava a nossa abordagem (filtragem por relevância, novidade e diversidade).

Para o estudo realizado foi utilizada uma base de dados constituída por um conjunto de 300 restaurantes que se encontram localizados no Porto. Como o estudo envolvia a participação de utilizadores reais, foi utilizada a aplicação cliente referida anteriormente.

Figura 5.1: Aplicação utilizada na avaliação

Foi construído um cenário de avaliação, em que se pedia aos utilizadores: *"Imagine que você está no Porto, em férias, e quer comer alguma coisa. Para economizar tempo e dinheiro você decide escolher restaurantes fast-food com um preço alvo de 5 euros."*.

Era pedido aos utilizadores que executassem três tarefas em cada um dos sistemas (sistema base e sistema que introduzia a novidade e diversidade). A primeira tarefa consistia em analisar a listagem apresentada e escolher o seu restaurante preferido. A segunda e terceira

tarefas consistiam em fazer a mesma coisa, no entanto era pedido aos utilizadores para imaginarem que eram executadas no dia seguinte e um ano mais tarde, respectivamente.

No entanto, para avaliar melhor o efeito da novidade as tarefas deviam ser realizadas efectivamente em dias diferentes. Mas como tal processo exigia um longo período de avaliação e devido a restrições de tempo, optou-se por pedir aos utilizadores que realizassem todas as tarefas de uma única vez (sem intervalo temporal). No entanto o sistema actua como se as tarefas tenham sido realizadas em datas diferentes.

De forma a garantir a qualidade da avaliação, é importante que os utilizadores não tenham conhecimento prévio da forma como os dois sistemas de recomendação se comportam. Deste modo, os sistemas de recomendação foram apenas identificados como "Sistema A" e "Sistema B". Os nomes são atribuídos de forma aleatória, o que implica que para um determinado utilizar o "Sistema A" corresponda ao sistema de recomendação base, mas para os outros utilizadores pode corresponder ao sistema que implementa a nossa abordagem.

Depois de realizadas as três tarefas os utilizadores foram convidados a responder a um pequeno questionário relacionado com a experiência que acabaram de ter.

Q1: *O sistema apresentou recomendações uteis?*

Q2: *Em cada tarefa, o sistema apresentou muitas recomendações semelhantes (poucas alternativas de escolha)?*

Q3: *Nas três tarefas o sistema apresentou sempre as mesmas recomendações?*

Q4: *Qual o sistema que considera melhor?*

O teste foi realizado com 43 pessoas, sendo 22 do género masculino e 21 do género feminino. A idade média da população era de 32 anos, variando entre 17 anos e 52 anos.

Na primeira questão 100% das pessoas responderam que ambos os sistemas apresentavam recomendações relevantes. O que está de acordo com o esperado porque ambos os sistemas implementam a filtragem por relevância. Nesta avaliação a relevância dos restaurantes foi apenas calculada através da técnica *Knowledge-based*, que consiste em efectuar recomendações que satisfação as necessidades do utilizador.

Através da questão Q2 pretendeu-se avaliar a diversidade dos restaurantes apresentados, ou seja, até que ponto os participantes achavam que lista de restaurantes recomendados oferecia alternativas de escolha.

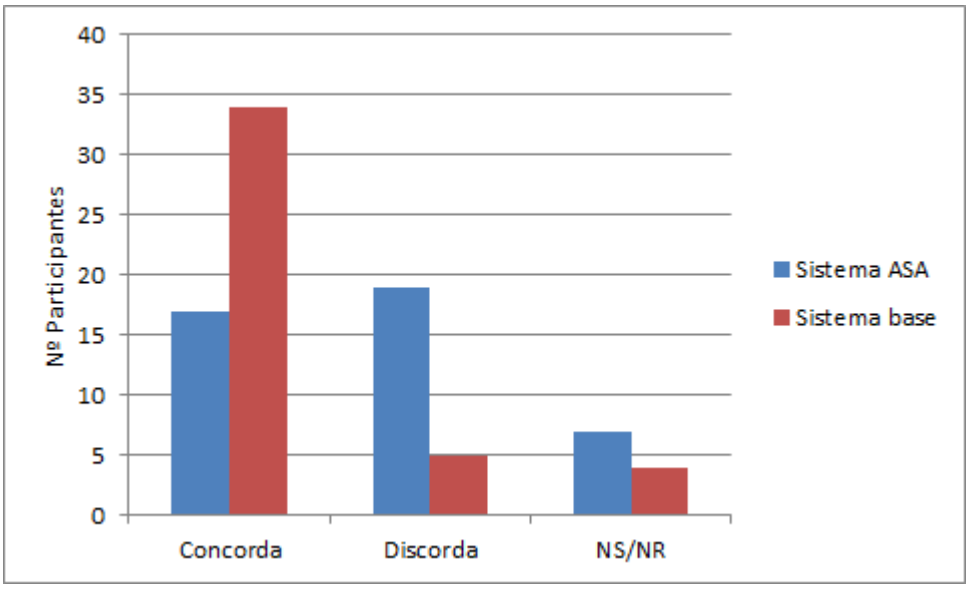

Figura 5.2: Respostas à segunda questão

Em comparação com o sistema base, as respostas a esta questão foram favoráveis ao nosso sistema (sistema ASA). No entanto ainda existiram algumas pessoas a considerar que a lista de recomendações apresentava restaurantes idênticos. Este resultado é um pouco discutível, porque o sistema ASA apresentava efectivamente restaurantes diversificados: Pans & Company, McDonalds, Pizza Hut, A Cascata, KFC. Por sua vez, o sistema base repetia sugestões da mesma marca. Leva-nos a concluir que diferentes pessoas podem percepcionar a questão da diversidade de forma distinta, deste modo torna-se fundamental efectuar uma aprendizagem do limiar de diversidade óptimo para cada utilizador.

Na terceira questão pretendeu-se avaliar se o sistema apresentava novidade na lista de recomendações apresentada.

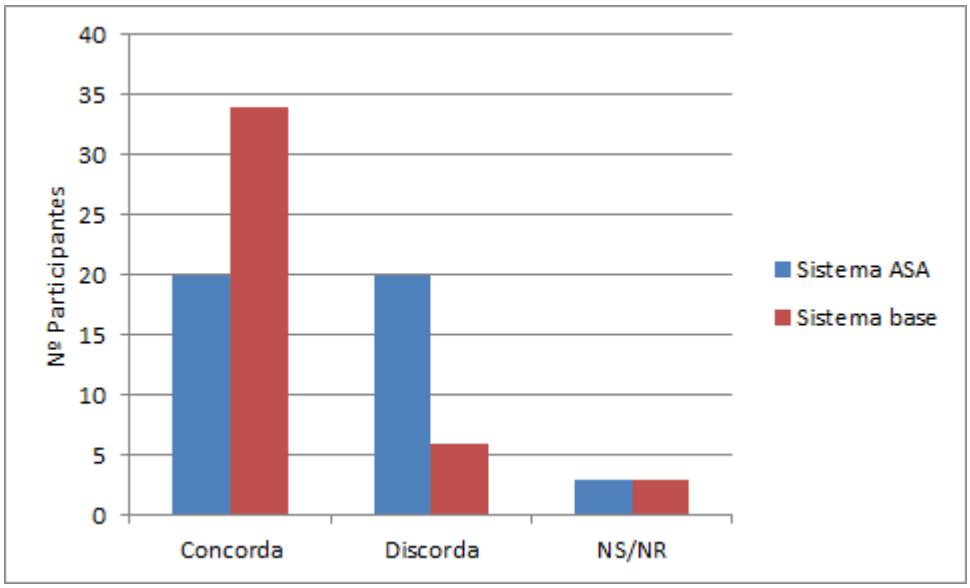

Figura 5.3: Respostas à terceira questão

A grande maioria dos participantes concorda que o sistema base apresenta sempre a mesma lista de recomendações nas três tarefas realizadas. Em relação ao sistema ASA achamos que o número de pessoas a considerar que as recomendações são sempre as mesmas é demasiado elevado.

No entanto, este resultado pode ser explicado com base nos *logs* de interacções de cada participante. Como no geral estes participantes interagiram pouco com a lista de recomendações (clicar nos restaurantes, ver restaurante no mapa, ver descrições do restaurante), a lista de restaurantes recomendados repetiu-se com frequência nas várias interacções com o sistema. Porque de acordo com o filtro de novidade, todos os recursos que o utilizador não interagiu são susceptíveis de apresentar novidade para o utilizador.

Na quarta questão pretendeu-se avaliar se os participantes preferiam o sistema base (apenas filtragem por relevância), ou o sistema que implementava a nossa abordagem (filtragem por relevância, novidade e diversidade).

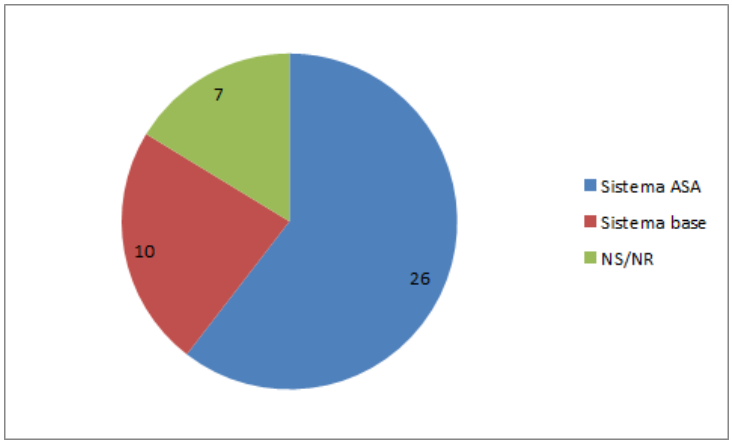

Figura 5.4: Resultados da questão "Qual o sistema que prefere?"

A maioria dos participantes preferiu o sistema que implementava a nossa abordagem. Outro aspecto importante é que os participantes que preferiram a nossa abordagem foram sobretudo aqueles que se aperceberam da novidade e diversidade das recomendações. Também existiram 7 participantes que optaram por não responder, justificando que ambos os sistemas pareciam idênticos, apesar de serem efectivamente diferentes. Existiram ainda 10 participantes que preferiram o sistema base, sendo que alguns deles justificaram que o nosso sistema eliminava algumas recomendações relevantes.

De facto é verdade, porque os restaurantes que os utilizadores se encontravam mais familiarizados não eram recomendados (devido ao filtro de novidade). Este facto leva-nos a concluir que algumas pessoas, em determinadas situações, podem desejar obter os recursos mais relevantes apesar de já os poderem conhecer. Desta forma, torna-se fundamental que os utilizadores possam decidir se desejam utilizar o filtro de novidade, podendo optar por maximizar ou anular o seu efeito.

5.2 Avaliação das Técnicas *Collaborative filtering* e *Demographic Filtering*

Para avaliar as técnicas de recomendação *Collaborative filtering* e *Demographic Filtering*, foi utilizado um *dataset* público da MovieLens[33] (sistema de recomendação de filmes).

O *dataset* utilizado contém 943 utilizadores que efectuaram 100,000 avaliações (escala de 1 a 5) em 1682 filmes, sendo que 80% da informação foi utilizado para treino e 20% para teste.

Para o estudo realizado foram consideradas como métricas a precisão e a cobertura das previsões. A cobertura das previsões corresponde à percentagem de itens para o qual o sistema de recomendação consegue efectuar previsões. Relativamente à precisão foi utilizada a métrica Mean Absolute Error (MAE), que permite calcular a proximidade a que as previsões das avaliações se encontram dos dados reais.

$$MAE = \frac{\sum_{u,i} |r_{u,i} - r'_{u,i}|}{N}$$

Onde $r_{u,i}$ indica a avaliação que o utilizador u atribuiu ao item i, enquanto que o $r'_{u,i}$ corresponde à previsão da avaliação do utilizador u em relação ao item i. O N indica o número de casos de teste.

5.2.1 Avaliação da Técnica de Recomendação *Demographic Filtering*

Como referido anteriormente, a técnica *Demographic Filtering* prevê a relevância de cada recurso com base nas avaliações de todos os utilizadores têm informação demográfica semelhante (género, idade, estado civil, profissão, etc.). Deste modo o valor de relevância é calculado através a média de todas avaliações realizadas pelo conjunto de utilizadores que têm um perfil demográfico semelhante.

No *dataset* da MovieLens a informação demográfica existente sobre os utilizadores é apenas a idade, género e profissão. Embora esta informação seja um pouco limitada, foi realizada a avaliação da técnica *Demographic Filtering* com base nestes três atributos.

É ainda importante referir que o atributo idade encontra-se segmentado em 6 categorias e existem 20 profissões diferentes. O género pode ser masculino ou feminino.

Para avaliar a precisão da técnica *Demographic Filtering* foi utilizada a métrica Mean Absolute Error (MAE) referida anteriormente. Também foi verificado a influência do limiar que define se os utilizadores têm um perfil semelhante.

Deste modo foram realizados quatro testes, no primeiro teste foi considerado que os utilizadores tinham um perfil semelhante se apresentarem pelo menos 1 atributo em comum, no segundo teste foram considerados pelo menos 2 atributos em comum e no terceiro teste foram considerados os 3 atributos em comum.

Para termos um valor de referência foi efectuado outro teste que calculava a relevância com base na média das avaliações realizadas por todos os utilizadores, que neste caso consiste em

[33] www.movielens.org

considerar que os utilizadores têm um perfil semelhante quando apresentam 0 ou mais atributos em comum.

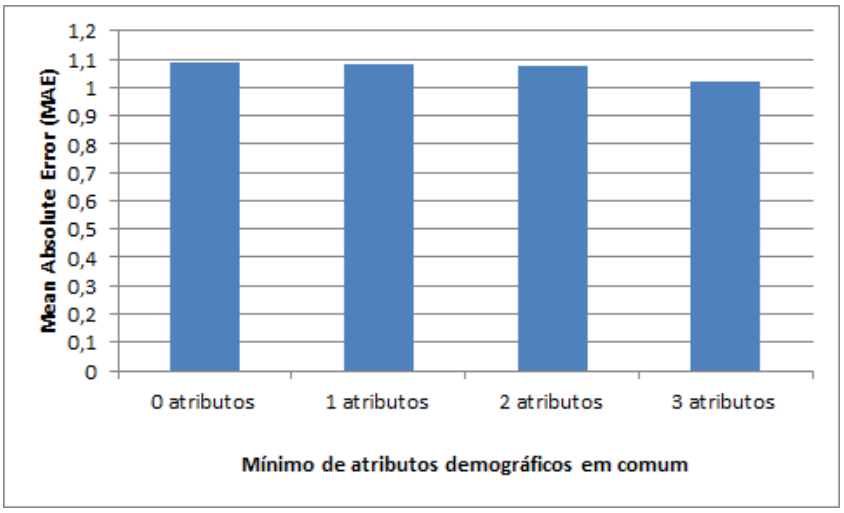

Figura 5.5: Precisão da técnica *Demographic Filtering*

Podemos verificar que quando se efectua a previsão da relevância com base em todos os utilizadores que avaliaram os filmes (corresponde à barra "0 atributos") o MAE é de 1,1. Ou seja, em média a previsão da relevância de cada filme apresenta um erro de 1,1 em relação à verdadeira classificação efectuada pelos utilizadores (escala de 1 a 5).

Quando se considerou que os utilizadores têm um perfil semelhante quando apresentam um ou dois atributos em comum, não se obteve resultados muito significativos. Apenas se nota uma ligeira melhoria do valor de precisão (menor MAE) nos casos em que se considerou que os utilizadores têm um perfil semelhante quando apresentam três atributos demográficos em comum.

No entanto estes resultados são muito específicos do *dataset* utilizado, sendo perfeitamente possível que noutras situações a técnica *Demographic Filtering* apresente resultados mais significativos. A nível da cobertura das previsões obteve-se o resultado detalhado na Figura 5.6.

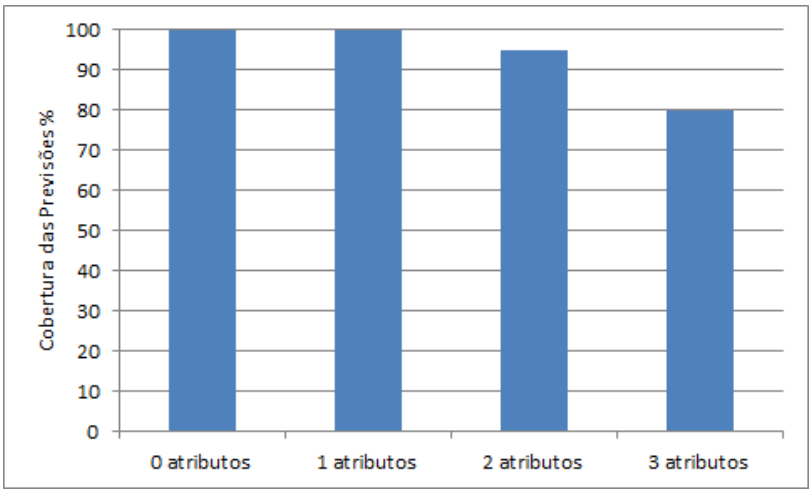

Figura 5.6: Cobertura da previsão da técnica *Demographic Filtering*

Como se pode verificar, quando se considera que os utilizadores apenas têm um perfil semelhante quando apresentam os três atributos demográficos em comum, a cobertura das previsões é de 80%. Isto significa que apenas para 80% dos filmes existentes no *dataset* é possível efectuar a previsão do valor de relevância.

Tendo em consideração os resultados anteriores, verifica-se que existe um *tradeoff* entre a precisão das previsões e a taxa de cobertura das previsões. Isto significa que para melhorar a precisão das previsões está-se a comprometer a percentagem de itens para os quais é possível efectuar a previsão do valor de relevância.

5.2.2 Avaliação da Técnica de Recomendação *Collaborative Filtering*

Como referido na Secção 2.5.2 a técnica *Collaborative filtering* efectua previsão do valor de relevância que cada recurso apresenta para um determinado utilizador, com base das preferências de todos os utilizadores que têm um perfil semelhante (vizinhança).

Neste caso os utilizadores têm um perfil semelhante quando avaliam os mesmos recursos de forma idêntica. Desta forma a precisão das recomendações depende do limiar de semelhança considerado para definir o conjunto de vizinhos do utilizador alvo.

Como foi utilizado o coeficiente de *Pearson Correlation* para calcular a semelhança dos utilizadores, o limiar de semelhança pode variar entre -1 e 1, em que o valor 1 corresponde aos utilizadores que tem um perfil igual e o valor -1 significa o oposto.

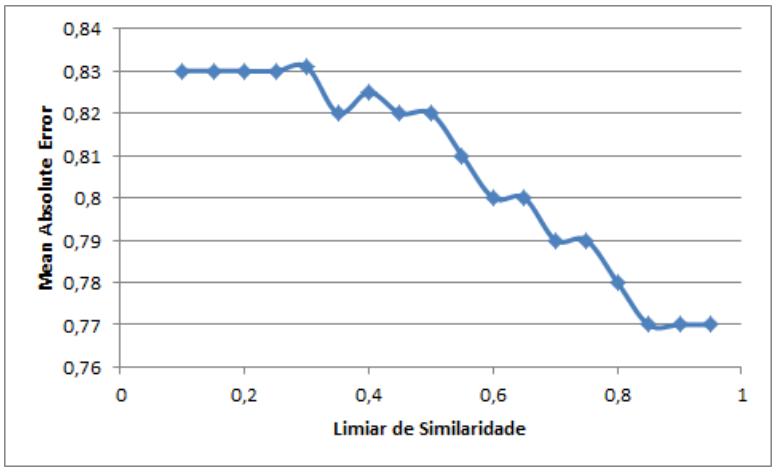

Figura 5.7: Precisão da técnica *Collaborative filtering*

Como se pode constatar no gráfico, a precisão das recomendações melhora com o aumento do limiar que define se os utilizadores têm perfil semelhante. No entanto, quando o limiar é demasiado elevado existe o problema de não ser possível efectuar previsões para todos os recursos, devido ao facto de reduzir a vizinhança de utilizadores.

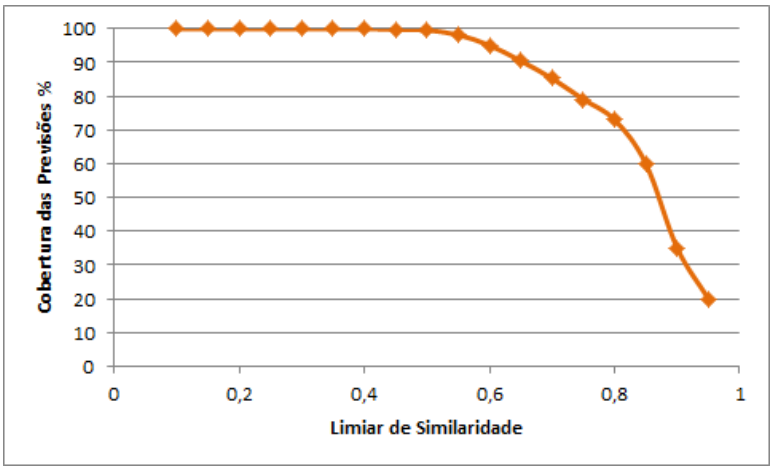

Figura 5.8: Cobertura da previsão da técnica *Collaborative filtering*

Para valores de similaridade superiores a 0,5 deixa de ser possível efectuar previsões para a totalidade dos recursos. À medida que o limiar de similaridade se aproxima de 1, o número de recursos para o qual é possível efectuar previsões é reduzido drasticamente.

Com base nos dois gráficos anteriores o valor de similaridade óptimo ronda os 0,5. Porque a precisão das previsões não fica comprometida (apesar de não ser a melhor) e consegue-se efectuar a previsão para a totalidade dos recursos. Em termos de performance do algoritmo, também é importante que a vizinhança de cada utilizador seja a mais reduzida possível.

Com base no *dataset* utilizado, verificamos a técnica de recomendação *Collaborative Filtering* apresenta maior precisão do que a técnica *Demographic Filtering*. No entanto este resultado é muito específico do *dataset* utilizado, deste modo, não é possível afirmar que este resultado se verifique em todos as situações.

Capítulo 6
Conclusões

Deste estágio resultou como produto final um serviço de recomendação e uma aplicação cliente. Através da aplicação cliente os utilizadores podem especificar o seu perfil, contexto e necessidades de forma a obter as recomendações provenientes do serviço de recomendação.

Em relação ao serviço de recomendação, existiu a preocupação de tornar o sistema mais genérico possível, de forma a possibilitar o cálculo da relevância e diversidade de novos recursos que sejam introduzidos no sistema. Este objectivo foi conseguido através da utilização de padrões de *software* (*Design Pattern Factory*), que permite adicionar novos recursos de forma compatível com o processo de recomendação já desenvolvido.

Foi proposta uma abordagem para o processo de recomendação que se baseia na aplicação de três fases de filtragem:

- Relevância: o objectivo é seleccionar os recursos em função da relevância que apresentam para o utilizador em questão.

- Novidade: pretende-se excluir os recursos que o utilizador melhor conhece.

- Diversidade: o objectivo é reduzir a redundância da lista de recomendações, de forma a não sobrecarregar o utilizador com informação repetida ou muito semelhante.

Para avaliar se a abordagem de recomendação proposta permitia melhorar a qualidade das recomendações, foi definido um cenário que possibilitou a participação de utilizadores reais.

O cenário criado permitiu comparar um sistema de recomendação que implementava a nossa abordagem, em relação a um sistema mais tradicional que apenas efectuava as recomendações com base na relevância.

Apesar de alguns resultados não serem tão bons como esperávamos inicialmente, ainda assim a maioria dos participantes preferiram o sistema de recomendação que implementava a nossa abordagem.

Como os testes não foram realizados em contexto real, os resultados que pretendiam avaliar a novidade das recomendações podem não corresponder totalmente à realidade. A principal limitação está relacionada com o facto da passagem do tempo ser difícil de simular, o que implica um esforço extra por parte dos utilizadores durante a realização do cenário de teste.

Outro aspecto importante é o facto da novidade das recomendações apresentar maior importância quando os utilizadores interagem frequentemente com o sistema de recomendação. Para as pessoas que utilizem o sistema de recomendação pela primeira vez, todos os recursos possíveis de recomendar apresentam novidade, pelo que não é possível sentir o efeito do filtro de novidade.

Analisando o histórico de interacções foi possível verificar que os participantes interagiram de forma mais frequente com os recursos que se encontravam no topo da lista de recomendações. Este comportamento dos utilizadores permite-nos concluir que é benéfico

o sistema de recomendação apresentar um número reduzido de recomendações, que sejam diversificadas e apresentem alternativas ao utilizador. Deste forma o filtro de diversidade acrescenta qualidade às recomendações, porque permite seleccionar um subconjunto de recursos diversificados a partir de um conjunto de maior dimensão.

Com base nos resultados foi possível concluir que os filtros de novidade e diversidade permitem melhorar a qualidade das recomendações. O filtro de novidade é útil porque promove a descoberta de novos recursos, por parte dos utilizadores. O filtro de diversidade é sobretudo importante para reduzir a redundância da lista de recomendações, com o objectivo de maximizar o número de alternativas que são apresentadas aos utilizadores.

6.1 Trabalho Futuro

O sistema de recomendação implementa várias técnicas para calcular a relevância dos recursos, apesar destas técnicas terem sido avaliadas através de um *dataset* da MovieLens (sistema de recomendação de filmes), é necessário realizar uma avaliação mais profunda tendo em conta o âmbito do sistema de recomendação.

A integração com as redes sociais (Facebook, Twitter, Google+, etc.) é uma possibilidade para melhorar a qualidade do sistema de recomendação. Uma vez que é possível obter informações adicionais sobre os utilizadores, como por exemplo os seus gostos pessoais e características do seu conjunto de amigos.

A informação proveniente das redes sociais permite desta forma resolver o problema designado por *cold start*, que acontece quando um novo utilizador é introduzido no sistema e ainda não existe informação sobre as suas preferências ou interesses.

As redes socais enumeradas anteriormente incorporam uma interface através da qual aplicações exteriores podem aceder aos dados dos utilizadores (se estes o permitirem). Para este processo é utilizado o protocolo OAuth[34], porque é um protocolo seguro e eficiente, que permite o acesso de terceiras aplicações sem que os utilizadores tenham de fornecer as suas credencias de acesso.

Outro aspecto importante, que deverá ser considerado, diz respeito à definição dos vários pesos utilizados durante cálculo da relevância e diversidade. Num caso extremo os utilizadores poderiam configurar os pesos de cada atributo, no entanto esta opção requer grande intervenção por parte dos utilizadores, pelo que não é do todo praticável. Desta forma, deverá ser estudada a possibilidade do sistema aprender a importância que cada atributo tem para cada utilizador.

Arslan et al. [43] calcula a importância de cada atributo em função do número de vezes que este é utilizado na *query* de pesquisa pelo utilizador. Transpondo para o sistema de recomendação, seria verificar o número de vezes que o utilizador especifica uma determinada necessidade. Sendo que as necessidades que ocorrem com maior frequência são susceptíveis de assumir maior importância. Técnicas de *machine learning* [44] também podem ser utilizadas para determinar os melhores valores para cada peso. Neste caso o histórico de interacções dos utilizadores deve ser analisado.

[34] www.oauth.net

Referências

[1] Kahneman, D., Attention and effort. Englewood Cliffs, NJ: Prentice-Hall, 120-130.

[2] Goldberg, D., Nichols, Terry, D. Using Collaborative Filtering to Weave an Information Tapestry. In: Communications of The ACM, v.35, n.12, 1992.

[3] Burke, R. Hybrid Recommender Systems: Survey and Experiments. In: User Modeling and User-Adapted Interaction Journal, v.12, n.4, p.331-370, 2002.

[4] Vozalis, E.; Margaritis, K. G. Analysis of Recommender Systems' Algorithms In: - Proceedings of the 6th Hellenic European Conference on Computer Mathematics - HERCMA, Athens, Greece, p.732-745, 2003.

[5] Blanco Fernandez, Y. Propuesta Metodológica para el Razonamento Semántico en Sistemas de Recomendación Personalizada y Automática. Aplicación Al Caso de Contenidos Audiovisuales. 2007. 320 p. Tese (Doutorado) - Departamento de Enxeneria Telemática E.T.S.E. de Telecomunicación, Universidade de Vigo, 2007.

[6] TaxiMedia, pervasive.wiwi.uni-due.de/uploads/tx_itochairt3/publications/TaxiMedia-camera-ready_01.pdf, Setembro 2011.

[7] T. Simcock, S. Hillenbrand, and B. Thomas. Developing a location based tourist guide application. In Proc. of the CRPITS '21 at ACSW frontiers, 2003.

[8] K. Cheverst, K. Mitchell, and N. Davies. The role of adaptive hypermedia in a context-aware tourist guide. Commun. ACM, 45(5):47-51, 2002.

[9] Driver, J. A selective review of selective attention research from the past century British Journal of Psychology. (2001)

[10] Haykin, S., Chen, Z. The Cocktail Party Problem. Neural Computation, (2005).

[11] Cherry, E. C. (1953). "Some Experiments on the Recognition of Speech, with One and with Two Ears". The Journal of the Acoustical Society of America 25 (5): 975.

[12] Broadbent, D.: Perception and communication. Oxford: Oxford University Press, (1958).

[13] Treisman, A., & Gelade, G., 1980. A feature integration theory of attention. Cognitive Psychology, 12, 97-136.

[14] Meyer, W., Reisenzein, R. and Schutzwohl, A.: Towards a process analysis of emotions: The case of surprise. Motivation and Emotion, 21, 251-274.

[15] L.Itti and P.Baldi. Bayesian surprise attracts human attention.In Advances in Neural Information Processing Systems(NIPS),pages1–8,Cambridge,MA,2006.MITPress.

[16] Rumelhart, D. E. (1984). Schemata and the cognitive system. Handbook of Social Cognition, pp. 161-188.

[17] Meyer, W. Reisenzein, R., & Niepel, M. Uberraschung [Surprise].Emotionspsychologie: Ein Handbuch, pp. 253-263.

[18] Macedo, L. and Cardoso, A.: Modeling Forms of Surprise in an Artificial Agent. Proceedings of the 23rd Annual Conference of the Cognitive Science Society, 588-593. Mahwah, NJ: Erlbaum. (2001).

[19] S. Vargas, P. Castells. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. 5th ACM Conference on Recommender Systems (RecSys 2011).

[20] G. Furnas. Generalized Fisheye Views. In Proceedings of the ACM CHI 86 Human Factors in Computing Systems Conference, pp.16-23, 1986.

[21] Pombinho, P. Carmo, M. B. and Afonso, A. P.: 'Evaluation of Overcluttering Prevention Techniques for Mobile Devices'. In: IV 2009, (2009).

[22] O'Connell, N. Interruption overload. Strategic Direction, Volume 24, Issue 10, 3-5. Emerald Group Publishing Limited.

[23] Fonseca, N. and Bento, C. and Rente, L.: Artificial selective attention for in-vehicle information systems: European Conference on Human Centred Design for Intelligent Transport Systems, Berlim, Alemanha, Abril, 2010.

[24] Haversine, http://en.wikipedia.org/wiki/Law_of_haversines, Junho 2012.

[25] www.en.wikipedia.org/wiki/Jaro-Winkler_distance, Junho 2012

[26] D. Buttler. The 5th International Conference on Internet Computing Las Vegas, NV, United States June 21, 2004 through June 24, 2004

[27] Gago, P. and Bento, C. A Metric for Selection of the Most Promising Rules, in Principles of Data Mining and Knowledge Discovery, Zytkow J. and Quafafou M. (Eds.), LNAI 1510, Springer, Berlin, pp.19-27. (1998)

[28] Bradley, K, Smyth, B. Improving Recommendation Diversity, Dublin, Ireland, (2001).

[29] M. Balabanovic and Y. Shoham, "Fab: content-based, collaborative recommendation," In: Communications of the ACM, vol. 40, 1997, pp. 66-72.

[30] Breese, J. S., Heckerman, D., and Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the14th Annual Conference on Uncertainty in Artificial Intelligence (UAI98), pages 43–52.

[31] Robin Burke. Integrating Knowledge-Based and Collaborative-Filtering Recommender Systems. In Proceedings of the AAAI Workshop on AI in Electronic Commerce, pages 69–72, 1999.

[32] Pazzani, M.J. A framework for collaborative, content-based and demographic filtering. Artificial Intelligence Review 13, 393–408 (1999)

[33] RICH, E. User Modeling via Stereotypes. Cognitive Science vol. 3, no. 4, p. 329-354, 1979.

[34] Burke, R. Hybrid Recommender Systems: Survey and Experiments. In: User Modeling and User-Adapted Interaction Journal, v.12, n.4, p.331-370, 2002.

[35] Xu, S., Bao, S., Fei, B., Su, Z., Yu, Y. 2008. Exploring Folksonomy for Personalized Search. In Proceedings of SIGIR 2008, 155-162.

[36] F. Asnicar, C. Tasso, "ifWeb: A Prototype of User Model-Based Intelligent Agent for Documentation Filtering and Navigation in the World Wide Web," In: Proceedings of the 6th International Conference on User Modeling, Chia Laguna, Sardinia, Italy, 1997, pp. 3-11.

[37] G.I. Webb, M. Kuzmycz, "Feature Based Modelling: A methodology for producing coherent, consistent, dynamically changing models of agents' competencies.," Methodology, vol. 5, 1996, pp. 1-32.

[38] J. F. Canny, "Collaborative filtering with privacy," in IEEE Symposium on Security and Privacy. IEEE Comp. Soc., 2002, pp. 45–57.

[39] H. Polat, W. Du, "Privacy-preserving collaborative filtering using randomized perturbation techniques," in Proc. of ICDM'03. IEEE Comp. Soc., 2003, pp. 625–628.

[40] P. Samarati, L. Sweeney. Protecting privacy when disclosing information: k- Anonymity and its enforcement through generalization and suppression," SRI Int., Tech. Rep., 1998.

[41] S. Berkovsky, Y. Eytani, T. Kuflik, F. Ricci. Enhancing privacy and preserving accuracy of a distributed collaborative filtering. In RecSys'07: Proceedings of the 2007 ACM conference on Recommender systems, pages 9–16, New York, NY, USA, 2007. ACM.

[42] Yi Cai, Qing Li. Personalized search by tag-based user profile and resource profile in collaborative tagging systems, pages 969-978, 2010.

[43] B. Arslan, F. Ricci, N. Mirzadeh, A. Venturini, A dynamic approach to feature weighting, Management Information Systems 6 (2002).

[44] D. Wettschereck, and D.W. Aha. Weighting features. In Veloso, M.M., Aamodt, A., eds.: Proceedings of the 1st International Conference on Case-Based Reasoning, Springer-Verlag. 347-358.1995.

Anexo A

Design Pattern Factory

Com o objectivo de tornar as aplicações compatíveis com a adição de novos tipos de objectos ou produtos, é comum utilizar-se a *design pattern Factory*[35].

Uma vez que esta *design pattern* define um padrão que fornece uma interface para criação de famílias de objectos, sem especificar suas classes concretas.

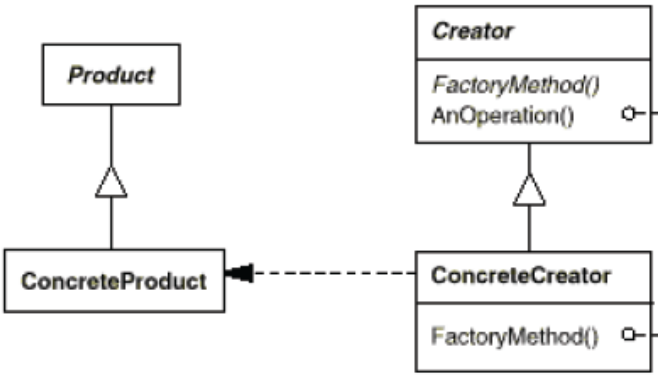

Figura A.1: Design Pattern Factory

Com base na imagem anterior, podemos verificar que esta *design pattern* é formada pelos seguintes elementos: o *Product* que define uma interface comum a todos os objectos que podem ser criados; o *ConcreteProduct* que implementa a interface do *Product* e o *ConcreteCreator* que é responsável por criar um objecto específico.

Esta *pattern* é de elevada importância para o serviço de recomendação de informação, uma vez que permite introduzir novos objectos (restaurantes, museus, hotéis, etc.) de forma compatível com todos os algoritmos já desenvolvidos.

[35] http://en.wikipedia.org/wiki/Factory_method_pattern

Anexo B

Estruturas de Dados

Como já foi referido anteriormente, a selecção dos recursos a apresentar ao utilizador é feita com base no seu perfil, contexto actual, necessidades e no histórico de interacções do utilizador com os recursos.

Nesta secção é feita uma análise detalhada das estruturas:

- Recursos
- Perfil do utilizador
- Contexto do utilizador
- Necessidades do utilizador
- Histórico de interacções.

Recursos

Os recursos são descritos por uma lista de atributos que corresponde à informação que se pretende filtrar. Estes recursos podem ser por exemplo: restaurantes, museus, hotéis etc. Consoante o tipo de recurso podem existir atributos diferentes, na Tabela B.1 é mostrado um exemplo associado a um recurso do tipo restaurante.

Atributo	**Valor**
Tipo	Restaurante
Latitude	Número decimal
Longitude	Número decimal
Horário Funcionamento	Hora abertura/Hora fecho
Tipo de cozinha	Comida italiana, *fastfood*…
Zona fumadores	S/N
Preço médio	Número
Ambiente	Romântico, Familiar…
Pagamento	Dinheiro, Cartão

Tabela B.1: Atributos associados a um recurso do tipo restaurante

Na Tabela B.2 estão especificados os atributos do recurso do tipo hotel.

Atributo	Valor
Tipo	Hotel
Latitude	Número decimal
Longitude	Número decimal
Preço médio	Número
Decoração	Moderna, clássica
Número de estrelas	[1,2,3,4,5]

Tabela B.2: Atributos associados a um recurso do tipo hotel

Perfil demográfico do utilizador

O conceito de perfil do utilizador refere-se a um conjunto de atributos que permitem identificar e descrever um utilizador. Na Tabela B.3, estão especificados os atributos que poderão fazer parte do perfil demográfico do utilizador.

Atributo	Valor
Género	Masculino/Feminino
Profissão	Nome Profissão
Nível financeiro	Baixa, Media, Alta
Data Nascimento	AAAA/MM/DD
Estado civil	Solteiro, casado, divorciado, viúvo.
Nacionalidade	Portuguesa, Francesa…

Tabela B.3: Perfil do utilizador

Contexto

O contexto do utilizador refere-se à descrição do meio ambiente em que o utilizador se encontra inserido no momento em que utiliza o sistema. De um modo geral, o contexto é definido com base na dimensão temporal e espacial.

Na Tabela B.4 está especificada a lista de atributos que definem o contexto.

Atributo	Valor
Data	AAAA/MM/DD
Hora	HH:MM:SS
Latitude	Número decimal
Longitude	Número decimal
Temperatura	Celsius (ºC)
Estado do tempo	Chuva, sol, nublado
Feriado	Sim/Não

Tabela B.4: Atributos do contexto

Necessidades do utilizador

As necessidades do utilizador são definidas por um conjunto de tuplos: atributo, operador e valor. O atributo refere-se a uma característica dos recursos que o utilizador pretende seleccionar e o conjunto operador/valor permite descrever quais os recursos que o utilizador pretende obter.

Na Tabela B.5 encontram-se detalhados os tipos de operadores que poderão ser utilizados para especificar as necessidades do utilizador.

Operador	Descrição
>	Maior
>=	Maior ou igual
<	Menor
<=	Menor ou igual
=	Igual

Tabela B.5: Operadores permitidos na definição das necessidades

Através deste conjunto de atributo/operador/valor é possível especificar as necessidades do utilizador de uma forma bastante simples e intuitiva.

Na Tabela B.6 é apresentado um exemplo de uma lista de necessidades de um determinado utilizador.

Atributo	**Operador**	**Valor**
Tipo	=	Restaurante
Preço	<	50€
Distância	<=	2Km

Tabela B.6: Exemplo de definição de necessidades

Com base na informação existente na tabela, sabemos que o utilizador está interessado em restaurantes que tenham um preço menor a 50€ e que se encontrem a uma distância menor ou igual a 2 Km.

Histórico de interacções

O histórico de interacções de um utilizador com um determinado recurso tem uma enorme importância para o correcto funcionamento do serviço de recomendação, principalmente na componente de cálculo do grau de apropriação que permite classificar os recursos em função do seu valor de novidade.

Na Tabela B.7 apresentam-se os atributos que permitem definir o histórico de interacções de um utilizador com um recurso num determinado contexto.

Histórico de interacções
Identificador do utilizador
Identificador do recurso
Tipo de interacção
Data da interacção

Tabela B.7: Atributos do histórico de interacções

Através da tabela anterior é possível verificar que para cada interacção, existe um identificador que permite relacionar um utilizador com o recurso.

O atributo tipo de interacção refere-se ao tipo de contacto que o utilizador teve com o recurso em questão. Este contacto pode ser apenas visualização de um recurso no âmbito de uma listagem, ou algo mais especifico como visualização dos detalhes de um recurso, pedir direcções para um recurso e avaliar um recurso.

Anexo C

Pesos dos atributos

Neste anexo estão especificados os pesos utilizados nos atributos que definem os recursos do tipo "Restaurante" e "Hotel". Estes pesos permitem definir a importância que cada atributo apresenta para o cálculo do valor de relevância, numa situação em que o utilizador especifica as suas necessidades.

Restaurantes	
Atributo	**Peso**
Tipo cozinha	0.3
Ambiente	0.05
Zona fumadores	0.1
Formas de Pagamento	0.05
Preço médio	0.3
Distância geográfica	0.2

Tabela C.1: Pesos dos atributos que definem os restaurantes

Hotéis	
Atributo	**Peso**
Número de estrelas	0.3
Decoração	0.1
Preço médio	0.3
Distância geográfica	0.3

Tabela C.2: Pesos dos atributos que definem os hotéis

Nesta fase os pesos dos atributos foram definidos por defeito de forma empírica. No entanto, como trabalho futuro é necessário estudar a possibilidade de efectuar uma aprendizagem da importância que cada atributo apresenta para cada utilizador em concreto.

Anexo D

Especificação de Software

Neste anexo estão especificados os seguintes diagramas UML: casos de uso, diagrama de actividades, diagrama de classes, diagrama de sequência.

Diagrama de actividades

O diagrama que se segue representa o fluxo que é realizado pelo serviço ASA no processo de recomendação de recursos.

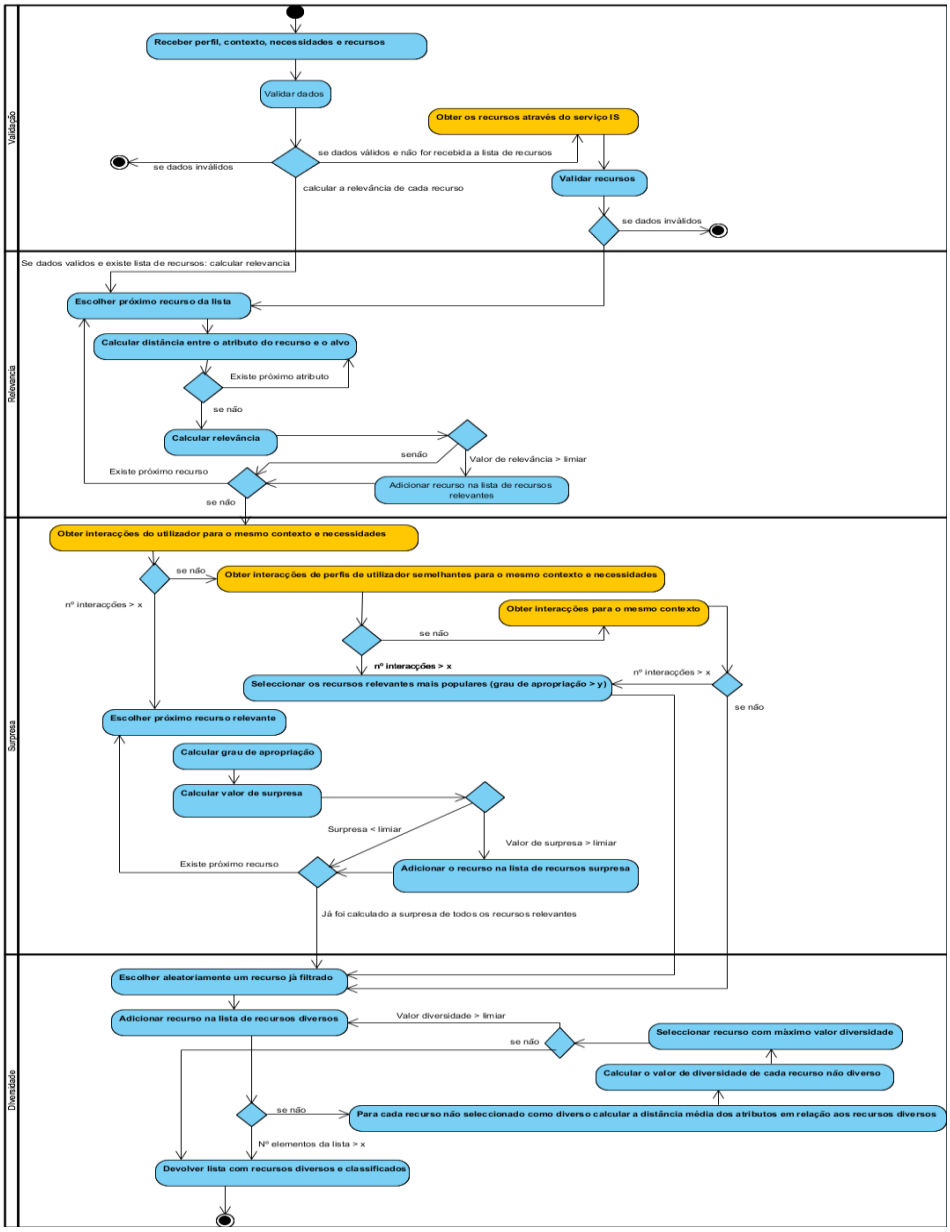

Figura D.1: Diagrama de actividades do serviço ASA

Este diagrama descreve as actividades que compõem o serviço, mostrado a sua relação e ordem de execução. De um modo geral, podemos dizer que existem quatro fases distintas: validação dos dados recebidos, filtragem dos recursos segundo o seu valor de relevância, filtragem dos recursos segundo o seu valor de novidade e filtragem dos recursos com base no seu valor de diversidade.

Na fase inicial, a aplicação recebe os dados necessários, como o perfil do utilizador, o contexto actual, as necessidades do utilizador.

Depois dos dados serem recebidos, é necessário proceder à sua validação. Caso estes dados não passem no teste de validação o pedido de recomendação de recursos será ignorado; caso contrário avança-se para a etapa do cálculo do valor de relevância de cada recurso.

A relevância de cada recurso resulta da distância entre o valor dos seus atributos e o valor desejado pelo utilizador. Apenas os recursos que apresentem um valor de relevância maior que um limiar previamente definido, serão adicionados à lista de recursos relevantes.

Após a construção de uma lista com os recursos relevantes, surge a fase de seleccionar os recursos que apresentam maior valor de novidade. Deste modo, é necessário calcular o valor de novidade de cada um dos recursos existentes na lista de relevância, e seleccionar aqueles que possuem um valor de novidade superior a um dado limiar.

O cálculo da novidade é efectuado através do grau de apropriação, que define a relação que o utilizador tem com um determinado recurso com base no seu histórico. O histórico de interacções que cada utilizador teve com um determinado recurso é obtido através do serviço IS (que é responsável por gerir o histórico de interacções de cada utilizador). Como resultado, obtém-se uma lista contendo apenas os recursos mais surpreendentes para um dado utilizador.

Depois de obtida a lista de recursos filtrados segundo a sua relevância e novidade, procede-se à selecção dos recursos que são mais distintos entre si, desta forma obtém-se um conjunto de recursos diversos.

Para se obter o valor de diversidade de um recurso, é necessário calcular a diferença entre os valores médios dos atributos dos recursos já seleccionados e o respectivo recurso, sendo que o primeiro recurso seleccionado é o mais relevante.

O recurso que apresente maior diversidade em relação aos recursos já seleccionados será adicionado à lista de recursos diversos, apenas se o seu valor de diversidade for maior que um limiar previamente definido. De modo a evitar a selecção de muitos recursos, é estipulado previamente o número máximo de recursos que a lista pode conter.

A última tarefa corresponde à devolução de uma lista que possuiu os recursos diversos, classificados segundo o seu valor de relevância e novidade.

Diagrama de classes

Nesta secção está especificado o diagrama de classes que representa o serviço a desenvolver, bem como uma breve explicação do seu funcionamento.

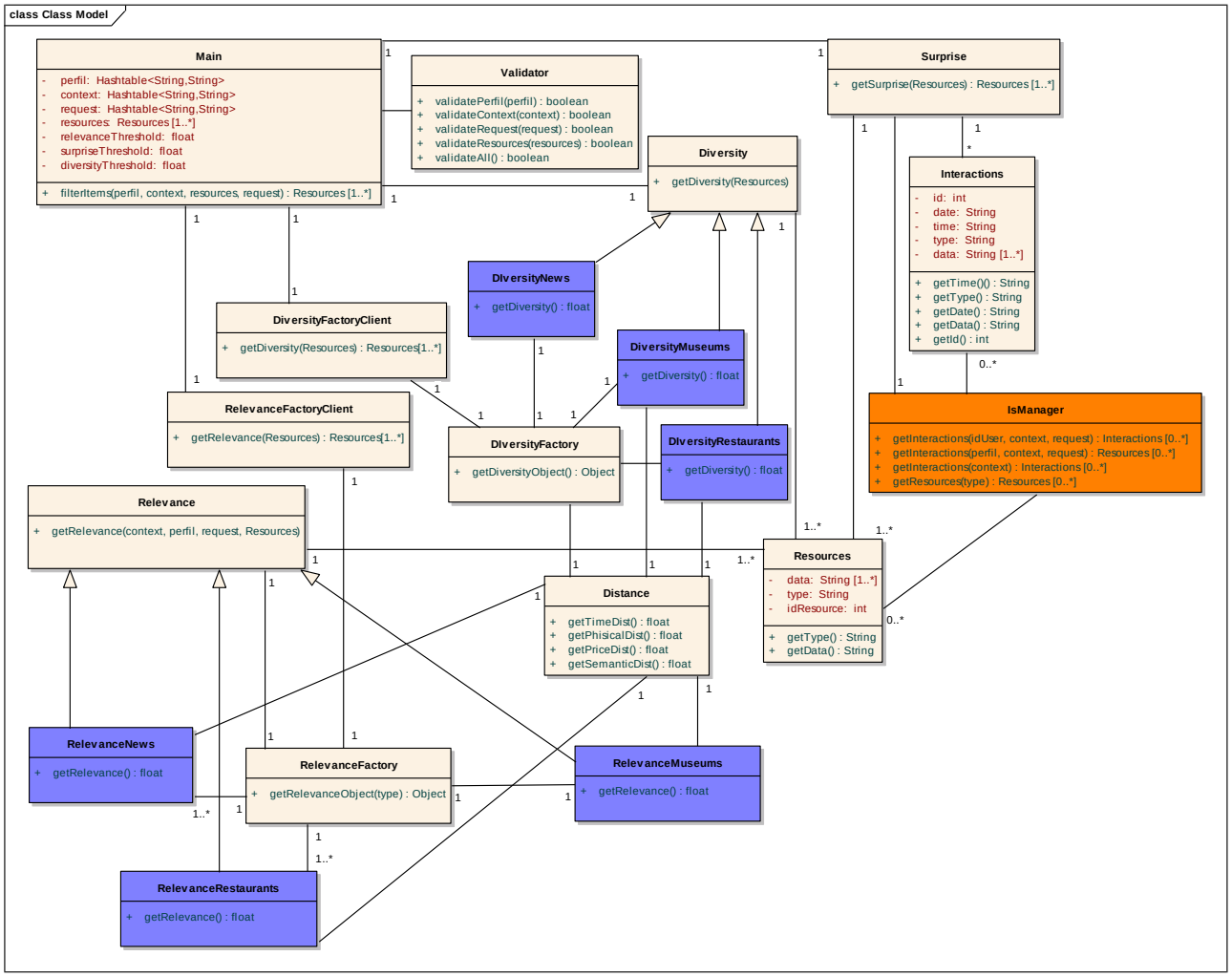

Figura D.2: Diagrama de Classes

Neste diagrama de classes estão representadas as classes que definem o serviço, bem como os seus atributos, métodos e forma de relacionamento. Como já foi visto anteriormente, esta aplicação pode se dividir em três partes distintas: cálculo do valor de relevância, novidade e diversidade dos recursos existentes. Desta forma as classes *Relevance*, *Novelty* e *Diversity* assumem o papel principal.

O serviço vai receber informação sobre o perfil, contexto, necessidades do utilizador e ainda a possibilidade de receber uma lista de recursos a classificar, deste modo, é necessário garantir que os dados recebidos são válidos. Para efectuar todo o processo de validação será utilizado a classe *Validator*.

Caso se verifique que os recursos a classificar não foram enviados no pedido, o serviço ASA terá de comunicar com o serviço IS, de forma a obter os recursos existentes na base de conhecimento. Todo o processo de comunicação com o serviço IS é da responsabilidade da classe *IsManager*.

Como este serviço pode ser utilizado em diversas situações, é necessário tornar as classes e métodos o mais genérico possível. Desta forma, será possível acrescentar numa fase posterior, outras classes que permitam representar novos tipos de recursos.

Como se pode ver no diagrama, existe a classe *Relevance*, que é uma generalização das classes específicas que permitem calcular a relevância de cada tipo de recurso. Estas classes são as seguintes: *RelevanceNews*, *RelevanceMuseums*, *RelevanceRestaurants*.

Com o objectivo de tornar a aplicação genérica, criou-se a classe *RelevanceFactory*. Esta classe representa uma *Factory Pattern* e define uma interface que permite criar objectos de acordo com o tipo de recurso, desta forma é fácil adicionar novas classes que permitam calcular a relevância de diferentes tipos de recursos.

De forma a ser possível calcular a relevância de um conjunto de recursos, criou-se a classe *RelevanceFactoryClient*. Esta classe comunica com a classe *RelevanceFactory*, de modo a obter o objecto que representa cada tipo de recurso, e é ainda responsável por devolver uma lista ordenada contendo os recursos mais relevantes.

Apesar de poderem existir diversos de tipos de recursos, a forma de calcular o valor de relevância é semelhante. Desta forma, foi criada uma classe *Distance* que permite auxiliar no cálculo do valor de relevância e é partilhada por todos os tipos de recursos.

A classe responsável por calcular o valor de novidade dos recursos é a *Novelty*, esta classe tem um método *getNovelty* que recebe uma lista de recursos e devolve o conjunto de recursos mais surpreendentes, aqueles que possui um valor de novidade superior a um dado limiar.

O cálculo da novidade está relacionado com grau de apropriação, este grau define a relação que um utilizador tem com um recurso com base no seu histórico de interacções. Para obter o histórico de interacções de cada utilizador é utilizado a classe *IsManager*, que tem a responsabilidade de efectuar um pedido ao serviço IS e este é encarregue de devolver a lista de interacções existentes (na base de conhecimento) para o utilizador especificado.

Para o calcular o valor de diversidade dos recursos, é necessário criar classes específicas que permitam comparar recursos do mesmo tipo, deste modo, estamos perante uma situação idêntica ao do cálculo da relevância. Por isso, também se utilizou a *pattern Factory*, a *DiversityFactory*, que permite criar objectos consoante o tipo de recurso.

Como a forma de calcular o valor de diversidade dos vários tipos de recursos, contém partes idênticas, a classe *Diversity* também irá utilizar a classe *Distance*, uma vez que esta classe possuiu métodos que permitem auxiliar no cálculo do valor de diversidade.