Instituto Superior de Engenharia do Porto
Departamento de Engenharia Electrotécnica
Rua Dr. António Bernardino de Almeida 431, 4200-072 Porto

Recomendação de Conteúdos Multimédia

Dissertação de Mestrado em Engenharia Electrotécnica e de Computadores
Área de Especialização de Telecomunicações

Fátima Manuela da Silva Leal

Orientação científica
Professora Doutora Maria Benedita Campos Neves Malheiro

Ano Lectivo: 2012/2013

Resumo

A quantidade e variedade de conteúdos multimédia actualmente disponíveis constituem um desafio para os utilizadores dado que o espaço de procura e escolha de fontes e conteúdos excede o tempo e a capacidade de processamento dos utilizadores. Este problema da selecção, em função do perfil do utilizador, de informação em grandes conjuntos heterogéneos de dados é complexo e requer ferramentas específicas. Os Sistemas de Recomendação surgem neste contexto e são capazes de sugerir ao utilizador itens que se coadunam com os seus gostos, interesses ou necessidades, *i.e.*, o seu perfil, recorrendo a metodologias de inteligência artificial.

O principal objectivo desta tese é demonstrar que é possível recomendar em tempo útil conteúdos multimédia a partir do perfil pessoal e social do utilizador, recorrendo exclusivamente a fontes públicas e heterogéneas de dados. Neste sentido, concebeu-se e desenvolveu-se um Sistema de Recomendação de conteúdos multimédia baseado no conteúdo, *i.e.*, nas características dos itens, no historial e preferências pessoais e nas interacções sociais do utilizador. Os conteúdos multimédia recomendados, *i.e.*, os itens sugeridos ao utilizador, são provenientes da estação televisiva britânica, British Broadcasting Corporation (BBC), e estão classificados de acordo com as categorias dos programas da BBC.

O perfil do utilizador é construído levando em conta o historial, o contexto, as preferências pessoais e as actividades sociais. O YouTube é a fonte do historial pessoal utilizada, permitindo simular a principal fonte deste tipo de dados - a *Set-Top Box* (STB). O historial do utilizador é constituído pelo conjunto de vídeos YouTube e programas da BBC vistos pelo utilizador. O conteúdo dos vídeos do YouTube está classificado segundo as categorias de vídeo do próprio YouTube, sendo efectuado o mapeamento para as categorias dos programas da BBC. A informação social, que é proveniente das redes sociais Facebook e Twitter, é recolhida através da plataforma Beancounter. As actividades sociais do utilizador obtidas são filtradas para extrair os filmes e séries que são, por sua vez, enriquecidos semanticamente através do recurso a repositórios abertos de dados interligados. Neste caso, os filmes e séries são classificados através dos géneros

da IMDb e, posteriormente, mapeados para as categorias de programas da BBC. Por último, a informação do contexto e das preferências explícitas, através da classificação dos itens recomendados, do utilizador são também contempladas.

O sistema desenvolvido efectua recomendações em tempo real baseado nas actividades das redes sociais Facebook e Twitter, no historial de vídeos Youtube e de programas da BBC vistos e preferências explícitas. Foram realizados testes com cinco utilizadores e o tempo médio de resposta do sistema para criar o conjunto inicial de recomendações foi 30 s. As recomendações personalizadas são geradas e actualizadas mediante pedido expresso do utilizador.

Abstract

The amount and diversity of multimedia contents currently available poses a challenge to users since the search space, *i.e.*, choice of sources and contents, exceeds largely their available time and processing capacity. This problem of selecting information from large heterogeneous data sets based on the user profile is complex and requires specific tools. This is the context where Recommender Systems make a difference since they are able to suggest items inline with the user tastes, interests or needs, *i.e.*, based on the user profile, using artificial intelligence methodologies.

The main objective of this thesis is to demonstrate that it is possible to recommend in real time multimedia contents based on the personal and social profile of the user, using only public heterogeneous data sources. In this sense, a Recommendation System based on the content, *i.e.*, based on the features of the items, the personal history, context and preferences as well as the social interactions of the user was designed and implemented. The content recommended, *i.e.*, the items suggested to the user, are from British Broadcasting Corporation (BBC) and are classified according to the BBC programmes categories.

The user profile is built taking into account the personal history, context and preferences as well as the social activities. YouTube acts as the source of personal history, substituting the traditional source of this type of data - the Set-Top Box (STB). The history is composed of the set of YouTube videos and BBC programmes seen by the user. The content of YouTube videos, which is classified according to the YouTube categories, is mapped into the BBC programmes categories. The social data, which is derived from Facebook and Twitter social networks, is collected through the Beancounter platform. The user social activities collected by the Beancounter platform are filtered to extract the references to films and series which are, in turn, semantically enhanced through linked open data repositories. In this case, the films and series are classified according to the IMDb genres and, subsequently, mapped into the BBC programmes categories. Finally, the personal context and preferences, obtained from the user ratings, are

also considered.

The developed system provides recommendations in real time based on the user social and personal data. Tests were conducted with five users and the average time system response to provide an initial set of recommendations was 30 s. The personalized recommendations are created and updated upon user request.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Equações

Glossário

Abreviatura	Descrição	Página
AJAX	*Asynchronous JavaScript and* XML	59
API	*Application Programming Interface*	44
BBC	British Broadcasting Corporation	iii
B2B	*Business-to-Business*	1
B2C	*Business-to-Consumer*	5
CSS	*Cascading Style Sheets*	105
DAML	DARPA *Agent Markup Language*	60
DES	*Data Enrichment Service*	47
DVB	*Digital Video Broadcasting*	26
EPG	*Electronic Programme Guide*	23
ESA	*Explicit Semantic Analysis*	15
FQL	*Facebook Query Language*	79
FOAF	*Friend Of A Friend*	34
HTTP	*HyperText Transfer Protocol*	34
HTML	*HyperText Markup Language*	63
IDE	*Integrated Development Environment*	60
IMDb	*Internet Movie Database*	24
ISEP	Instituto Superior de Engenharia do Porto	xvii
JAWS	*Java* API *for WordNet Searching*	93
JDK	*Java Development Kit*	65
JRE	*Java Runtime Environment*	65
JSON	*JavaScript Object Notation*	34
JSP	*JavaServer Pages*	62
JVM	*Java Virtual Machine*	64
kNN	*k-Nearest-Neighbour*	54
LinkedMDB	*Linked Movie Database*	46
LOD	*Linked Open Data*	34
MPEG	Moving Picture Experts Group	26
MQL	*Metaweb Query Language*	44
OKBC	*Open Knowledge Base Connectivity*	62
OWL	*Web Ontology Language*	15

Abreviatura	Descrição	Página
RDF	*Resource Description Framework*	34
REST	*Representational State Transfer*	61
SMS	*Short Message Service*	29
SPARQL	SPARQL *Protocol and* RDF *Query Language*	43
SQL	*Structured Query Language*	62
STB	*Set-Top Box*	iii
SUMO	*Suggested Upper Merged Ontology*	45
TF-IDF	*Term Frequency-Inverse Document Frequency*	13
URI	*Uniform Resource Identifier*	42
URL	*Uniform Resource Locator*	38
EUA	Estados Unidos da America	22
VSM	*Vector Space Model*	13
VOD	*Video On Demand*	24
WAR	*Web Application Archive*	62
W3C	World Wide Web Consortium	43
XML	*eXtensible Markup Language*	34
XMPP	*eXtensible Messaging and Presence Protocol*	64
YAML	*Yet Another Multicolumn Layout*	59

Agradecimentos

No culminar desta etapa de formação, são muitos os agradecimentos que são devidos e que eu sinto necessidade de aqui deixar expressos, pois foram igualmente muitos os que contribuíram para que todo o meu percurso académico fosse possível. Assim, muito reconhecida, agradeço:

- A todos os colegas que se cruzaram comigo, pela entreajuda, companheirismo e bons momentos passados nesta Escola.
- Aos colegas com quem fiz trabalhos de grupo, pois sempre tive a sorte de encontrar pessoas aplicadas e que, assim, contribuíram para o sucesso obtido em algumas disciplinas curriculares.
- À Professora Doutora Benedita Malheiro por toda a ajuda recebida, ajuda que resultou da grande sensibilidade, da incondicional disponibilidade, do permanente entusiasmo e dedicação que põe na transmissão dos seus conhecimentos e que são qualidades que a destacam como professora e como orientadora. Para ela vai o meu profundo obrigado. Este agradecimento é extensivo ao Bruno Veloso, pelo apoio que me deu no desenvolvimento do Sistema de Recomendação, pois foi o responsável pelo Projecto que originou esta tese.
- A todos os funcionários, professores e colaboradores do Instituto Superior de Engenharia do Porto (ISEP), com quem muito aprendi e que trabalham diariamente para fazer do ISEP uma das mais conceituadas Escolas de Engenharia do país.
- A meus pais e irmãos, pelos sacrifícios passados e pela confiança que sempre depositaram em mim. Aos meus amigos, por acreditarem que eu era capaz. A António Correia, pelo incentivo na minha caminhada académica.

Por tanto carinho, apoio e amizade a todos expresso aqui o meu grande e sincero Obrigado.

Capítulo 1

Introdução

Neste capítulo de introdução é apresentada a contextualização do projecto, o problema que se pretende resolver, os objectivos a atingir, o planeamento do projecto e, por fim, a estrutura desta dissertação.

1.1 Contextualização

Esta tese foi proposta no âmbito de um projecto mais abrangente em curso – o desenvolvimento de uma plataforma *Business-to-Business* (B2B) de transacção automática de componentes multimédia. Esta plataforma, que modela produtores e distribuidores de conteúdos multimédia, tem como finalidade a transacção automática em tempo útil de conteúdos entre produtores e distribuidores baseada no perfil dos espectadores em linha. Este projecto engloba diversos componentes, designadamente, a Plataforma Electrónica de Transacção de Componentes Multimédia suportada por agentes desenvolvida por Sousa *et al.* [10] e Veloso *et al.* [11] e os Serviços de Criação do Perfil do Utilizador e de Recomendação que são o objecto desta tese.

1.2 Problema

O desafio deste projecto consiste na criação do perfil do utilizador e na geração de recomendações consentâneas a partir de um conjunto heterogéneo de fontes de dados em tempo útil. As recomendações devem filtrar a elevada quantidade de informação envolvida e apresentar ao utilizador o conjunto de conteúdos multimédia disponíveis que se adeqúem ao seu perfil. O perfil do utilizador deve, por sua vez, incluir dados pessoais e sociais relevantes para a sugestão de conteúdos multimédia. Os serviços a desenvolver – Serviço de Criação do Perfil do Utiliza-

dor e Serviço de Recomendação – devem permitir uma integração transparente com a plataforma B2B em curso.

1.3 Motivação

A escolha desta tese deveu-se à natureza do problema – um problema do mundo real, ao interesse pelo domínio – a personalização de conteúdos multimédia, e ao desafio colocado – a criação do perfil do utilizador e a recomendação em tempo útil de conteúdos multimédia, *i.e.*, programas (filmes, vídeos, séries, *etc.*) baseados no perfil construído. O desafio envolveu o estudo e aprendizagem de um conjunto de novos conceitos, a pesquisa, escolha e aplicação de tecnologias desconhecidas e a concepção e realização dos serviços referidos.

Do ponto de vista empresarial, esta abordagem deverá contribuir para melhorar a experiência, aumentar a satisfação e, consequentemente, fidelizar o espectador.

1.4 Objectivos

O principal objectivo desta tese é demonstrar que é possível recomendar em tempo útil conteúdos multimédia a partir do perfil pessoal e social do utilizador, recorrendo exclusivamente a fontes públicas e heterogéneas de dados. Neste sentido, pretende-se conceber e desenvolver um Sistema de Recomendação de programas baseado nas características dos itens e no perfil do utilizador. Este sistema deve ser suportado pelos serviços de criação do perfil do utilizador e de recomendação.

O desenvolvimento deste sistema foi organizado nas seguintes etapas:

- Estudo do estado da arte dos Sistemas de Recomendação em geral e dos sistemas congéneres em particular;
- Estudo e escolha das fontes de dados e de enriquecimento para a construção do perfil do utilizador;
- Definição do perfil do utilizador;
- Estudo das bibliotecas de interface das fontes de dados e de enriquecimento a utilizar;
- Uniformização de categorias para a construção dos vectores de pesos;
- Determinação da similaridade entre os itens e o perfil do utilizador;
- Geração da recomendações;
- Desenvolvimento da interface do utilizador.

1.5 Planeamento do Projecto

Embora os prazos originais não tenham sido cumpridos devido à complexidade do projecto, o planeamento inicial serviu de referência para a gestão e organização do trabalho. Este projecto foi dividido num conjunto de doze tarefas representadas na Figura 1.1.

ID	Task Name	Start	Finish	Duration
1	Levantamento e estudo dos sistemas de recomendação e personificação de TV	05-11-2012	11-12-2012	5,4w
2	Familiarização com as tecnologias envolvidas	07-11-2012	07-12-2012	4,6w
3	Escrita dos capítulos 1 e 2	05-11-2012	01-01-2013	8,4w
4	Estudo das fontes de dados e de enriquecimento	01-01-2013	31-01-2013	4,6w
5	Escolha das linguagens, tecnologias e ferramentas de desenvolvimento	12-02-2013	07-03-2013	3,6w
6	Escrita dos capítulos 3, 4 5 e 6	24-01-2013	25-03-2013	8,6w
7	Interacção com APIs	26-03-2013	10-04-2013	2,4w
8	Mapeamento de Categorias	08-04-2013	17-05-2013	6w
9	Construcção Perfil Utilizador	17-05-2013	05-07-2013	7,2w
10	Construcção Recomendação	26-06-2013	03-09-2013	10w
11	Análise e precisão da Recomendação	21-10-2013	24-10-2013	,8w
12	Escrita dos capítulos 7 e 8	29-08-2013	06-11-2013	10w

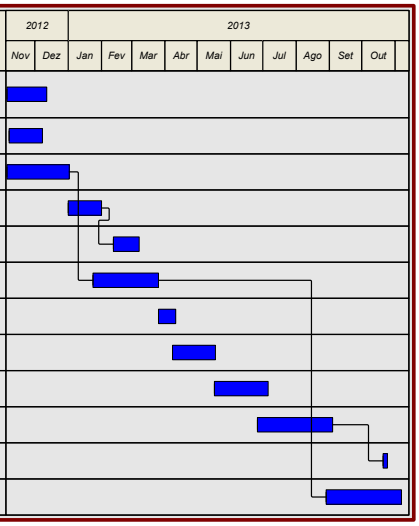

Figura 1.1: Calendarização.

O primeiro conjunto de tarefas consistiu no estudo do estado da arte dos Sistemas de Recomendação e envolveu o estudo dos diferentes tipos de Sistemas de Recomendação e das técnicas utilizadas assim como os sistemas congéneres de recomendação de conteúdos multimédia. Em segundo lugar, procedeu-se ao levantamento das fontes de dados e de enriquecimento semântico a adoptar. O terceiro conjunto de tarefas consistiu na selecção e instalação das linguagens, tecnologias e do ambiente de desenvolvimento. Prosseguiu-se com o estudo, escolha e instalação das bibliotecas de interface com as fontes de dados adoptadas. De seguida, desenvolveu-se o mapeamento entre os sistemas de classificação de conteúdos (categorias) das diferentes fontes e procedeu-se à construção do perfil do utilizador e à geração de recomendações. Por fim, testou-se e depurou-se o sistema e analisaram-se os resultados. Em paralelo, no final de cada tarefa, efectuou-se a escrita do capítulo correspondente da dissertação.

1.6 Estrutura da Dissertação

A dissertação é composta por sete capítulos. No Capítulo 1 é feita a contextualização e a apresentação do projecto, incluindo os objectivos, o planeamento e a

estrutura deste documento. O Capítulo 2 contém o estado da arte dos Sistemas de Recomendação, incluindo os diferentes tipos de Sistemas de Recomendação, as técnicas utilizadas, os Sistemas de Recomendação *online* e personalização de conteúdos multimédia mais representativos. No Capítulo 3 são apresentadas as fontes de dados e de enriquecimento semântico estudadas para a criação e enriquecimento do perfil do utilizador. O Capítulo 4 apresenta as componentes do perfil do utilizador. No Capítulo 5 descrevem-se, não só, as ferramentas, tecnologias e ambiente de desenvolvimento, mas também a respectiva instalação e configuração. O Capítulo 6 é dedicado ao desenvolvimento do Sistema de Recomendação. Por fim, o Capítulo 7, apresenta as conclusões, incluindo uma análise crítica do trabalho efectuado e sugestões de desenvolvimento futuro. Nos Anexos A, B, C, D, E, F e G apresentam-se as ontologias construídas e respectivos mapeamentos e que permitem a integração do Sistema de Recomendação com a plataforma B2B.

Cada capítulo começa com um pequeno resumo e termina com uma breve conclusão, contendo a comparação/análise do estudo/desenvolvimento efectuados.

Capítulo 2

Sistemas de Recomendação

Neste capítulo apresenta-se o estudo realizado acerca dos Sistemas de Recomendação, incluindo alguns dos exemplos mais representativos dos sistemas online e dos projectos de recomendação personalizada de conteúdos multimédia existentes.

2.1 Definição

Dada a grande quantidade e diversidade de informação actualmente disponível, a comunidade científica da inteligência artificial tem vindo a conceber desde a década de noventa do século passado metodologias de recomendação de produtos, bens ou serviços baseadas no perfil do utilizador [12]. Os Sistemas de Recomendação são ferramentas que produzem recomendações personalizadas diante de uma grande variedade de opções [3]. Existem três metodologias de recomendação: (*i*) a Recomendação Pessoal; (*ii*) a Recomendação Colaborativa; e (*iii*) a Recomendação Híbrida. A Recomendação Pessoal agrega recomendações baseadas no Conteúdo (características dos produtos, bens ou serviços) e no Contexto (demográfico, geográfico ou temporal) do utilizador. A Recomendação Colaborativa suporta-se nas classificações previamente atribuídas pelo utilizador a produtos, bens ou serviços, *i.e.*, requer interacção do utilizador, para gerar as recomendações. Por fim, a Recomendação Híbrida agrega a Recomendação Pessoal e a Recomendação Colaborativa. Estas técnicas de recomendação permitem gerar sugestões compatíveis com o comportamento passado e o contexto actual do utilizador.

Os Sistemas de Recomendação, apesar de serem frequentemente invisíveis, fazem parte do quotidiano da actualidade, recomendando ao utilizador os artigos, locais, eventos, serviços ou conteúdos que mais se adequam ao seu perfil. Enquanto as empresas de *Business-to-Consumer* (B2C) incluem este tipo de sis-

temas nos seus *sites* para aumentar as vendas, cooperando com o utilizador na filtragem da informação e guiando-o na pesquisa e aquisição de produtos [13], a Google, no seu sítio de partilha de vídeos YouTube, para fidelizar os utilizadores recomenda vídeos com base no historial, localização e preferências de cada utilizador como se mostra na Figura 2.1.

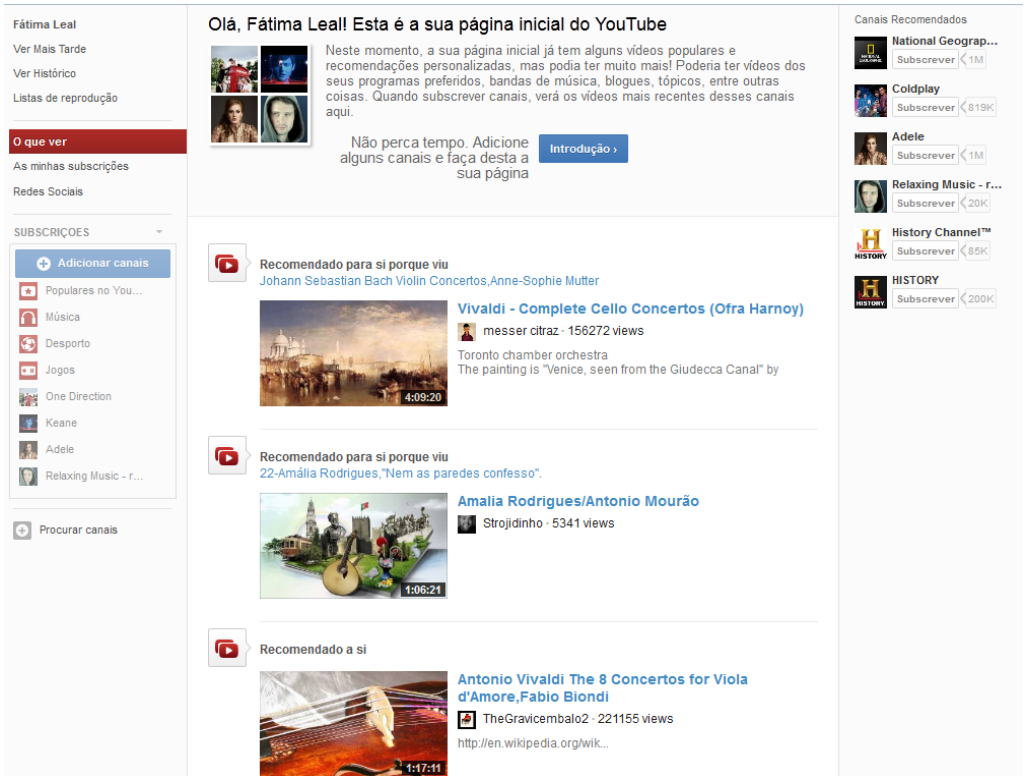

Figura 2.1: Recomendação do YouTube.

Neste domínio, quando o número de utilizadores é elevado, os utilizadores com perfis idênticos são organizados em grupos ou *clusters* para reduzir o esforço computacional de geração de recomendações. De um ponto de vista arquitectónico, os sistemas da recomendação dividem-se em:

- Sistemas baseados em Modelos ou *Model-based* - Criam em tempo diferido modelos através de aprendizagem automática a partir de conjuntos de dados de elevada dimensão;
- Sistemas baseados em Memória ou *Memory-based* - Aplicam em tempo útil técnicas de determinação de similaridade a partir dos dados disponíveis.

Os Sistemas de Recomendação podem ser adoptados do lado do utilizador (computador pessoal, *smartphone*, *etc.*) ou do lado das empresas (Amazon, eBay,

etc.) para gerar recomendações baseadas nas actividades, comportamentos, preferências e necessidades do utilizador [13]. Os Sistemas de Recomendação são incontornáveis no domínio do comércio electrónico onde o utilizador pode ser conduzido através das recomendações que lhe são propostas.

2.2 Recomendação Pessoal

A recomendação pessoal é baseada no conjunto de informação disponível associada a cada indivíduo, *i.e.*, o seu perfil, que inclui o contexto, o historial de interacção com o sistema, as preferências explícitas, *etc.*

2.2.1 Recomendação Baseada no Conteúdo

A recomendação baseada no conteúdo foi a primeira técnica a ser explorada, combinando a informação dos itens com as preferências dos utilizadores a fim de sugerir os itens apropriados. Assim sendo, os Sistemas de Recomendação baseados no conteúdo tomam em consideração as características do item e tentam sugerir ao utilizador itens de acordo com as escolhas efectuadas no passado. Este tipo de recomendação pode ser usado em vários domínios como a recomendação de páginas *Web*, artigos, restaurantes, programas de televisão e produtos [14].

Esta abordagem filtra e selecciona os itens com características idênticas às do perfil do utilizador, que é baseado na informação fornecida pelo próprio utilizador e nas suas acções. A filtragem da informação baseada no conteúdo recorre a técnicas específicas, não só, de representação dos itens, mas também para construir o perfil do utilizador. A recomendação baseada no conteúdo pode ser baseada na recuperação de informação ou na filtragem de informação. Os sistemas de recuperação de informação apresentam sugestões utilizando palavras-chave para efectuar uma pesquisa numa base de dados. O sistema apresenta ao utilizador os itens que contêm a palavra-chave no seu conteúdo. Os sistemas de filtragem de informação são mais complexos e os resultados obtidos são de curto prazo.

O sistema gera recomendações baseadas no perfil corrente do utilizador através de um procedimento composto por três etapas:

- Análise do conteúdo - extrai as características relevantes dos itens;
- Análise do perfil - constrói o perfil do utilizador de acordo com as características dos itens que ele seleccionou no passado;
- Filtragem de componentes - agrupa e recomenda os itens compatíveis com o perfil do utilizador.

Após a construção dos perfis do item e do utilizador é efectuada a filtragem dos componentes utilizando um dos sistemas referidos: filtragem de informação ou

recuperação de informação). Por último, para ordenar as recomendações obtidas por relevância, é determinada a similaridade entre os itens seleccionados e o perfil do utilizador. A Figura 2.2 apresenta a arquitectura de uma sistema de recomendação.

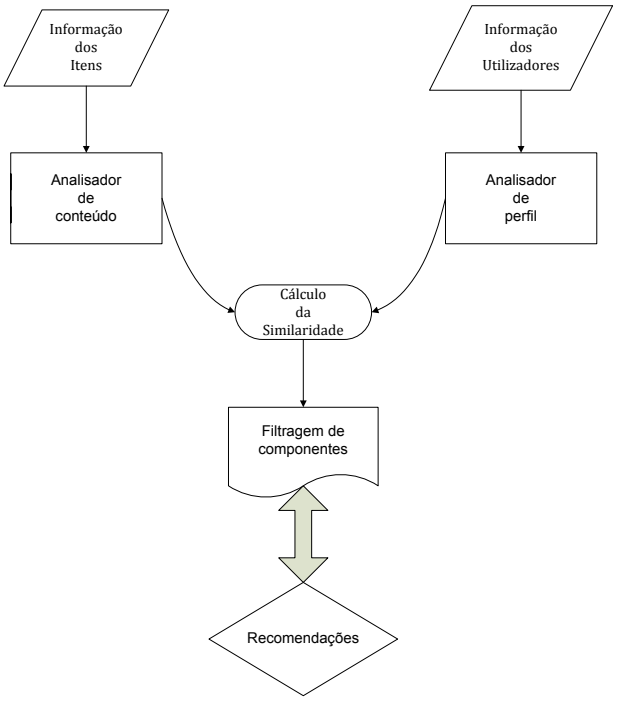

Figura 2.2: Recomendação baseada no Conteúdo.

2.2.2 Recomendação Baseada no Contexto

O contexto está presente de forma intrínseca em qualquer actividade e interacção humana, sendo natural e implicitamente usado para melhorar a qualidade da comunicação. A percepção do contexto também permite aos indivíduos avaliar, tomar decisões e adaptar o seu comportamento.

No âmbito dos Sistemas de Recomendação também se pode utilizar a informação contextual do utilizador para modular as recomendações. Na Figura 2.3 representa-se o contexto do utilizador como um agregado das componentes cultural, demográfica, temporal e espacial.

Para recomendar um pacote de férias, um sítio *Web* ou um filme pode ser necessário considerar, para além do perfil dos utilizadores e das características dos itens, o contexto do utilizador. Por exemplo, para um Sistema de Recomendação de viagens sugerir destinos de férias distintos em função da estação do ano, *e.g.*, praia no verão e neve no inverno, é necessário conhecer o contexto temporal

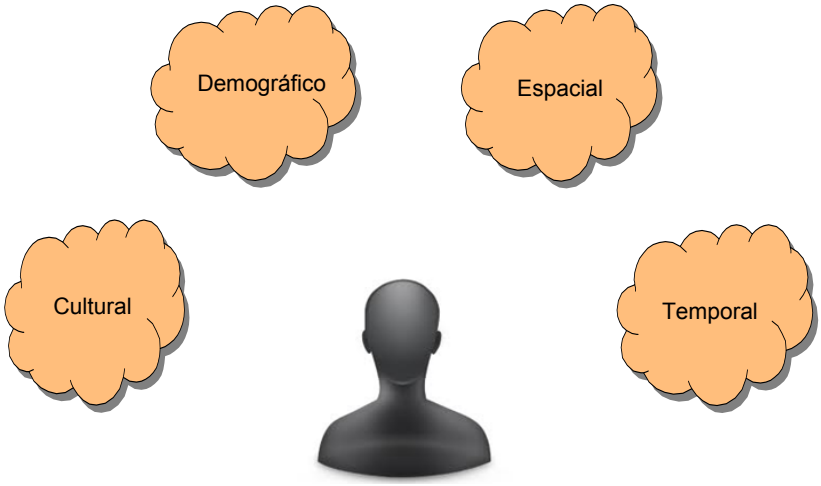

Figura 2.3: Contexto do Utilizador.

do utilizador [15]. Os contextos cultural, demográfico e espacial são igualmente importantes porque permitem efectuar recomendações conforme a localização, o género, a idade ou a religião do utilizador. A utilização de *smartphones* tem permitido aceder ao contexto espácio-temporal do utilizador para recomendar serviços, espectáculos, promoções e informações diversas em função da localização, estação do ano, dia da semana e período do dia do utilizador.

2.3 Recomendação Colaborativa

A recomendação colaborativa tem por objectivo sugerir novos itens com base no historial das classificações do utilizador e da sua rede de contactos [16]. Isto é, agrega classificações ou recomendações de objectos, identifica utilizadores com gostos idênticos e gera novas recomendações utilizando inter-comparações. Nesta abordagem não é necessário conhecer as características dos itens e o perfil do utilizador, existindo apenas um vector de itens e as suas respectivas classificações.

Os Sistemas de Recomendação colaborativos definem a utilidade de cada item através da atribuição de uma classificação numa escala pré-definida, *e.g.*, a utilizadora "Fátima Leal" atribuiu ao filme "A Vida é Bela" a classificação de "8" numa escala de 1 a 10. Estas classificações são armazenadas numa matriz como a representada na Equação 2.1. Neste exemplo, os utilizadores avaliaram os músicos numa escala de 1 a 5. O X significa que o utilizador não classificou o músico em causa, devendo a aplicação ser capaz de estimar essa classificação recorrendo às classificações utilizador-músico existentes para gerar recomendações adequadas.

Equação 2.1: Matriz de classificação colaborativa.

$$\begin{pmatrix} & \textit{Luís Reprezas} & \textit{Rui Veloso} & \textit{Pedro Abrunhosa} & \textit{Jorge Palma} \\ \textit{Daniel} & 4 & 3 & 2 & 4 \\ \textit{António} & X & 4 & 5 & 5 \\ \textit{Luís} & 2 & 2 & 4 & X \\ \textit{Sandra} & 3 & X & 5 & 2 \end{pmatrix} \tag{2.1}$$

Ao invés dos sistemas baseados no conteúdo, os sistemas colaborativos tentam prever a utilidade dos itens para um dado utilizador com base nos itens previamente avaliados pelos restantes utilizadores. O processamento dos dados armazenados nas matrizes de classificação é baseado no cálculo da vizinhança entre itens e utilizadores, *i.e.*, procura-se identificar a similaridade entre itens ou entre utilizadores.

2.4 Recomendação Híbrida

A Recomendação Híbrida combina as técnicas da recomendação baseada no conteúdo e da recomendação colaborativa, apresentando melhor desempenho do que cada uma das técnicas *per si*.

Existem diferentes formas de combinar as duas abordagens de recomendação. A Tabela 2.1 apresenta algumas das metodologia de combinação mais utilizadas. Deste modo, a filtragem híbrida consegue reunir os pontos fortes das técnicas de base e minimizar as limitações do uso de apenas uma abordagem.

Sendo este tipo de recomendação uma combinação de técnicas é, por isso, o que apresenta maior fiabilidade no que diz respeito às recomendações que gera. Contudo, a sua implementação requer um nível de complexidade elevado pois exige um grande domínio das técnicas usadas tanto na recomendação baseada no conteúdo como na recomendação colaborativa.

2.5 Técnicas de Recomendação

Existem metodologias de geração de recomendações baseadas na construção de modelos (*Model-based*), que aplicam técnicas de aprendizagem automática em tempo diferido para construir os modelos de recomendação, e baseadas em memória (*Memory-Based*), que determinam em tempo útil a similaridade entre as escolhas dos utilizadores e os itens candidatos. Estas metodologias incluem a representação dos itens e a construção do perfil do utilizador. A Tabela 2.2 contém os algoritmos mais utilizados em cada metodologia de recomendação.

[1] *Term Frequency-Inverse Document Frequency* (TF-IDF)

Tabela 2.1: Métodos híbridos [3].

Método	Descrição
Ponderado	A pontuação de várias técnicas de recomendação é combinada de forma a gerar uma única recomendação.
Alternado	O sistema alterna as técnicas de recomendação dependendo da situação actual.
Misto	Recomendações de diferentes aplicações de recomendação apresentadas ao mesmo tempo.
Combinação de características	Características de diferentes fontes de dados de recomendação são combinadas para formar um único algoritmo de recomendação.
Cascata	Um sistema de recomendação refina as recomendações dadas por outro sistema.
Aumento de características	As recomendações geradas por uma técnica são utilizadas como entrada para outra técnica.
Meta-nível	O modelo aprendido por uma técnica é usado como entrada para outro.

Tabela 2.2: Técnicas de recomendação [4].

Recomendação	*Memory-Based*	*Model-based*
Baseada no conteúdo	TF-IDF[1] *Clustering*	Classificadores Bayesianos *Clustering* Árvores de decisão Redes neuronais artificiais
Colaborativa	Vizinho mais próximo *Clustering* Teoria dos grafos	Redes Bayesianas *Clustering* Redes neuronais artificiais e regressão linear Modelos probabilísticos
Híbrida	Combinação das técnicas anteriores	

2.6 Análise Comparativa

Esta análise comparativa foca-se essencialmente na Recomendação baseada no Conteúdo e na Recomendação Colaborativa. A Recomendação Híbrida faz uso de uma ou mais técnicas de recomendação, geralmente combinando uma abordagem colaborativa com outra técnica, para ultrapassar as limitações de cada abordagem e, desta forma, melhorar a qualidade das recomendações.

2.6.1 Recomendação Pessoal

A Recomendação Pessoal é uma Recomendação baseada no Conteúdo. As vantagens da Recomendação baseada no Conteúdo são: (*i*) a capacidade de recomendar qualquer item do conjunto de itens disponível; e (*ii*) não sofrer do problema do primeiro avaliador. Esta abordagem não carece, ao invés da abordagem colaborativa, que um item tenha sido previamente classificado por algum utilizador para poder ser recomendado. Todos os itens que constam na base de dados são comparados com o perfil do utilizador, tendo, à partida, idêntica possibilidade de serem recomendados.

Os pontos fracos da Recomendação baseada no Conteúdo são: (*i*) a sintaxe; *(ii)* a semântica; e (*iii*) a super-especialização [4]. Dado que este tipo de recomendação se baseia na análise do conteúdo do item, a sintaxe e a semântica do texto são bastante importantes. A super-especialização decorre do facto de os Sistemas de Recomendação baseados no Conteúdo se basearem na comparação de palavras-chave e, consequentemente, todas as recomendações geradas satisfazerem as palavras-chave utilizadas. Neste caso, o sistema pode nunca recomendar um livro de C a um utilizador que se interesse por livros de Java e linguagem C, apesar de ambos os interesses serem relativos a linguagens de programação [17].

2.6.2 Recomendação Colaborativa

A Recomendação Colaborativa apresenta três vantagens: (*i*) geração de recomendações baseadas nas preferências dos utilizadores; (*ii*) independência do conteúdo; e (*iii*) possibilidade de produzir recomendações inesperadas e de alta qualidade. A filtragem colaborativa, ao basear-se essencialmente na opinião de outros utilizadores, assegura a independência do conteúdo dos itens, podendo recomendar itens desconhecidos em termos de conteúdo. A terceira vantagem consiste na recomendação de produtos inesperados, *i.e.*, serendipidade.

As desvantagens inerentes à filtragem colaborativa são: (*i*) o problema da falta de avaliador; (*ii*) a dispersão da base de dados; e (*iii*) o custo de processamento. Dado que a filtragem colaborativa utiliza o conjunto de classificações de produtos efectuadas pelos outros utilizadores para efectuar as recomendações, um produto não classificado jamais será recomendado. O problema da dispersão da base de dados resulta da grande diversidade de itens disponíveis face à reduzida dimensão e elevada dispersão dos itens avaliados pelos utilizadores, tornando difícil a recomendação de itens conformes com os gostos e preferências de cada utilizador. Estas dificuldades são minimizadas através de um custo de processamento elevado, *e.g.*, aplicando algoritmos de aprendizagem automática para modelar o comportamento do utilizador [17].

2.7 Representação de Itens

Os itens devem ser representados através de um conjunto de atributos ou propriedades relevantes. Por exemplo, numa aplicação de recomendação de filmes, um filme pode ser descrito através do título, data de estreia, género(s), lista de actores, realizador, *etc.* Neste cenário, cada filme/item passa a ser representado por este conjunto de características que constitui o seu perfil. Existem ainda sistemas onde os itens são caracterizados pela informação textual extraída directamente de páginas *Web*, *e-mails*, artigos ou notícias.

Para a representação dos itens são essencialmente utilizadas as técnicas baseadas em: (*i*) *keyword-based vector space model*; (*ii*) ontologias; e (*iii*) conhecimento enciclopédico.

2.7.1 *Keyword-based Vector Space Model*

A maioria dos Sistemas de Recomendação baseados no conteúdo usa modelos relativamente simples que consistem na procura de palavras-chave utilizando um *Vector Space Model* (VSM). O VSM foi primeiramente introduzido por Gerard Salton [18, 19] e tornou-se uma técnica popular na investigação e na indústria. Trata-se de um modelo algébrico que representa objectos em vectores, *i.e.*, representa documentos textuais num modelo vectorial. Isto significa que, por exemplo, uma colecção de N documentos pode ser representada num modelo vectorial por uma matriz de T termos e D documentos – ver Equação 2.2.

Equação 2.2: Matriz VSM.

$$\begin{pmatrix} & T1 & T2 & ... & Tt \\ D1 & W11 & W21 & ... & Wt1 \\ D2 & W12 & W22 & ... & Wt2 \\ ... & ... & ... & ... & ... \\ Dn & W1n & W2n & W3n & Wtn \end{pmatrix} \tag{2.2}$$

Uma entrada na matriz corresponde ao peso do termo no documento. O zero indicará que o termo não é significativo ou que simplesmente não existe. Este peso é na maioria das aplicações dado por *Term Frequency-Inverse Document Frequency* (TF-IDF) que reflecte a importância de uma palavra num conjunto de documentos [20, 19]. A importância da palavra aumenta proporcionalmente ao número de vezes que aparece no documento. Contudo, se a palavra é muito frequente num documento, especialmente no corpo, vai obter uma pontuação mais baixa. É o caso de palavras como “a”, “e”, “que” que aparecem com elevada frequência num texto, mas que têm pouca importância no contexto destas aplicações.

O TF representa o quociente entre a frequência de ocorrência de um termo t num documento d e a frequência de ocorrência do termo mais comum no mesmo documento d. O TF é calculado através da Equação 2.3. O IDF representa a frequência com que o termo ocorre em todos os documentos analisados, *i.e.*, um termo que ocorra com frequência elevada num documento mas raramente no resto da colecção tem peso maior. Este parâmetro obtém-se dividindo o número total de documentos N pelo número de documentos que contêm o termo n_t, logaritmicamente. O TF-IDTF é o obtido através da Equação 2.4.

Equação 2.3: *Term Frequency.*

$$TF(t,d) = (\frac{f_{td}}{max f_{td}}) \tag{2.3}$$

Equação 2.4: *Term Frequency-Inverse Document Frequency.*

$$TF - IDF(t,d) = TF(t,d) \times log(\frac{N}{n_t}) \tag{2.4}$$

Por último, para se normalizar os pesos obtidos através da Equação 2.4 aplica-se a Equação 2.5.

Equação 2.5: Normalização

.

$$w(t,d) = \frac{TF - IDF(t,d)}{\sqrt{\sum_{s=1}^{|T|} TF - IDF(t,d)^2}} \tag{2.5}$$

Na recomendação de conteúdos baseada em VSM, representam-se os perfis dos utilizadores e os itens através de vectores de termos ponderados. O interesse de um utilizador num item pode ser calculada recorrendo a técnicas de determinação similaridade, comparando ambos os vectores.

2.7.2 *Tag Clouds*

As *tag clouds* tornaram-se numa técnica popular de apresentação da frequência da ocorrência de expressões. São muitas vezes incorporadas nas plataformas *Web* para representar visualmente o conjunto de expressões mais utilizadas pelos utilizadores. Nesta representação a cor ou o tamanho identificam a relevância pois quanto mais viva for a cor ou maior o tamanho mais destaque terá a palavra na plataforma [21]. As *tag clouds* podem ser utilizadas para guiar o utilizador na sua pesquisa ou para representar os perfis dos utilizadores ou de itens para efeitos de recomendações. Por exemplo, um Sistema de Recomendação baseado no Conteúdo que recomende ao utilizador produtos semelhantes aos que ele preferiu no passado, pode adoptar *tags* para descrever os itens. Num cenário de recomendação de filmes, um utilizador que goste do filme "Matrix" obteria recomendações de filmes que envolvessem as *tags*: acção, ficção, *etc.*

2.7.3 Ontologias

A ontologia é uma forma de representar o conhecimento. Um domínio pode ser descrito hierarquicamente através de uma ontologia [**?**]. Para essa representação geralmente é utilizada a linguagem *Web Ontology Language* (OWL) que permite instanciar todas as classes do domínio a representar. No âmbito dos Sistemas de Recomendação existem vários domínios que podem ser representados através de ontologias: (*i*) representação do contexto; *(ii)* representação do perfil do utilizador e *(iii)* representação do item. Assim, os dados são divididos e organizados eficazmente permitindo uma melhor visão sobre a representação em questão.

Para a representação dos itens como programas, filmes e vídeos, existem também ontologias já desenvolvidas para a representação conhecimento. A título ilustrativo apresentam-se as seguintes ontologias que representam programas, tempo e relações entre utilizadores:

- Ontologia dos programas da BBC[2] - Esta ontologia fornece um vocabulário simples para descrever programas de TV e Rádio;
- Ontologia Time[3] - Ontologia que se centraliza em torno da noção de espaço temporal.
- Ontologia GoodRelations[4] - Ontologia que anota as ofertas e outros aspectos do comércio na Web. É desenvolvida em OWL e é a única ontologia oficialmente apoiada pela Google e Yahoo.

2.7.4 Conhecimento Enciclopédico

O conhecimento enciclopédico fornece suporte à análise semântica de textos. A análise semântica de um texto é uma tarefa complexa dada a ambiguidade, *i.e.*, a polissemia intrínseca a qualquer língua. No caso deste projecto a polissemia dificulta a determinação da similaridade entre um dado item (descrito através de um texto) e o perfil de um utilizador, podendo resultar na recomendação de itens pouco relevantes.

A *Explicit Semantic Analysis* (ESA) é uma metodologia de análise semântica que recorre a conhecimento enciclopédico, *e.g.*, a Wikipedia, para determinar o significado dos elementos de um texto [22]. A adopção da Wikipedia como fonte de conhecimento apresenta várias vantagens, nomeadamente, a sua permanente disponibilidade, actualização, crescimento e precisão. Adicionalmente, avaliações empíricas mostraram que a análise semântica através de técnicas como a ESA

[2] `http://www.bbc.co.uk/ontologies/programmes/2009-09-07.shtml`
[3] `http://www.w3.org/TR/owl-time/`
[4] `http://www.w3.org/wiki/GoodRelations`

introduzem melhoramentos substanciais nos resultados [23]. Este tipo de metodologias associam a cada conceito um vector de pesos do tipo da Equação 2.2.

Outra forma de minimizar os problemas associados à análise semântica é a adopção de ontologias de representação estruturada do conhecimento do domínio [24]. Um Sistema de Recomendação com representação de conhecimento por ontologias e análise semântica de conteúdos suportada na recolha da informação complementar de fontes de conhecimento externas consegue construir perfis mais exactos. Na implementação de uma aplicação deste tipo são considerados diversos aspectos:

- As características da fonte de conhecimento envolvido (ontologia);
- As técnicas adoptadas para a anotação ou a representação dos itens;
- O conteúdo incluído no perfil do utilizador [23].

A Tabela 2.3 apresenta exemplos de Sistemas de Recomendação com representação de conhecimento por ontologias e análise semântica suportada por conhecimento enciclopédico.

Tabela 2.3: Sistemas suportados por ontologias e análise semântica.

Sistema	Descrição
Semantic Enhancement for Web Personalization [25]	Recomendação de páginas *Web*; Taxonomia das categorias usadas na anotação automática de páginas da *Web*; baseada em palavras-chave e no historial de navegação do utilizador.
Quickstep e Foxtrot [26]	Recomendação de artigos académicos; Representação dos perfis através de uma ontologia.
Informed Recommender [27]	Recomendação de itens; Representação da opinião dos consumidores através de uma ontologia.
News@hand [28]	Recomendação de notícias; Representação de notícias e preferências do utilizador através de uma ontologia.
Sistema de Recomendação de TV [29]	Recomendação de programas televisivos; Representação de conteúdos e espectadores através de ontologias.

A análise semântica pode também ser realizada recorrendo a tesauros que fornecem os conjuntos de palavras com significado idêntico. O uso de tesauros é essencial no contexto das redes sociais dada a diversidade da informação publicada. A WordNet e o Big Huge Thesaurus são dois exemplos de tesauros

que disponibilizam uma biblioteca de interface para interacção com aplicações de terceiros [30, 31].

2.8 Similaridade

As funções de determinação da similaridade permitem estabelecer o grau de semelhança entre dois objectos [32]. Assim, o processo de recomendação consiste em três etapas consecutivas: (*i*) determinação da similaridade entre as características dos itens candidatos e o perfil do utilizador; (*ii*) ordenação dos candidatos por ordem decrescente de similaridade; e (*iii*) selecção do(s) candidato(s) do topo da lista.

2.8.1 Distância Euclidiana

A distância euclidiana calcula a distância entre dois pontos x e y através da fórmula apresentada na Equação 2.6.

Equação 2.6: Distância Euclidiana.

$$d(x, y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \tag{2.6}$$

Quanto mais próximo de zero for o resultado, mais próximos estão x e y.

No caso dos Sistemas de Recomendação a distância euclidiana é calculada entre objectos (x,y) com igual número de atributos k, onde x e y representam o objecto candidato e o perfil do utilizador. No final, depois de ordenados os pares (x,y) por ordem decrescente de similaridade, são geradas as recomendações.

2.8.2 Similaridade dos Cossenos

A semelhança entre o vector de características dos objectos candidato e o perfil do utilizador pode ser alternativamente determinada através da similaridade dos cossenos. A Equação 2.7 apresenta a fórmula da similaridade do cosseno entre os vectores A e B.

Equação 2.7: Similaridade dos cossenos.

$$\cos\alpha = \frac{\hat{A}\cdot\hat{B}}{|\hat{A}\|\hat{B}|} \equiv \frac{\sum_{j=1}^{n}\hat{A}_j\hat{B}_j}{\sqrt{\sum_{j=1}^{n}\hat{A}_j^2}\sqrt{\sum_{j=1}^{n}\hat{B}_j^2}} \tag{2.7}$$

A similaridade do cosseno corresponde ao cosseno do ângulo formada pelo par de vectores A e B representado na Figura 2.4.

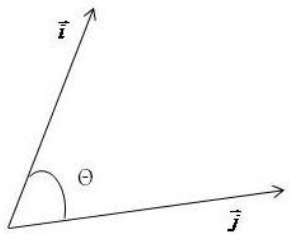

Figura 2.4: Ângulo entre dois vectores.

2.8.3 Coeficiente de Correlação de Pearson

A correlação de Pearson (r) é uma métrica que determina o coeficiente de correlação entre duas amostras é descrito através da Equação 2.8.

Equação 2.8: Coeficiente de Pearson.

$$r(x,y) = \frac{\sum_i (r_{x,i} - \overline{r}_x)(r_{y,i} - \overline{r}_y)}{\sqrt{\sum_i (r_{x,i} - \overline{r}_x)^2}\sqrt{\sum_j (r_{y,i} - \overline{r}_y)^2}} \tag{2.8}$$

Onde:

- x e y representa o par de utilizadores a ser correlacionado;
- i corresponde ao item;
- $r_{x,i}$ e $r_{y,i}$ são as classificações que cada utilizador atribuiu ao item i;
- $\bar{r}_x$ e $\bar{r}_y$ são as médias aritméticas do conjunto de classificações de cada utilizador.

Este método de cálculo de similaridade é o mais utilizado nos Sistemas de Recomendação Colaborativos.

2.8.4 Coeficiente de Similaridade de Jaccard

O coeficiente de similaridade de Jaccard (J), que é um índice estatístico da similaridade entre conjuntos de amostras, é representado através da Equação 2.9.

Equação 2.9: Coeficiente de Jaccard.

$$J(A,B) = \frac{A \cap B}{A \cup B} = \frac{c}{a+b} \tag{2.9}$$

Onde:

- A e B representam os dois conjuntos de amostras;

- c é o subconjunto de amostras comuns a A e B;
- a é o subconjunto de amostras que só ocorre em A;
- b é o subconjunto de amostras que só ocorre em B.

Este índice varia entre os valores 0 e 1, representando a disjunção e a identidade dos dois conjuntos, respectivamente. Quanto mais próximo de 1 for o coeficiente de similaridade de Jaccard, maior é a semelhança entre os dois conjuntos de amostras [33].

2.9 Avaliação de Sistemas de Recomendação

Para se determinar a fiabilidade das recomendações efectuadas através das diferentes metodologias, é necessário proceder a uma avaliação, determinando a precisão (*precision*), a abrangência (*recall*) e a Medida-F (*F-Measure*) dos resultados produzidos [34].

2.9.1 Precisão

A precisão (Pr) dos resultados representa a probabilidade de um item recomendado ser relevante e calcula-se através do quociente entre o número de itens relevantes recomendados e o número total de itens recomendados – Equação 2.10.

Equação 2.10: Precisão.

$$Pr = \frac{P_v}{P_v + P_f} \tag{2.10}$$

Onde:

- P_v - Representa os itens relevantes recomendados pelo sistema, *i.e.*, os resultados positivos verdadeiros;
- P_f - Representa os itens irrelevantes recomendados pelo sistema, *i.e.*, os resultados positivos falsos.

2.9.2 Abrangência

A abrangência (Ab) dos resultados representa a probabilidade de um item relevante ser recomendado e obtém-se através do quociente entre o número de itens relevantes recomendados e o número total de itens relevantes – Equação 2.11.

Equação 2.11: Abrangência.

$$Ab = \frac{P_v}{P_v + N_f} \tag{2.11}$$

Onde:

- P_v - Representa os itens relevantes recomendados pelo sistema, *i.e.*, os resultados positivos verdadeiros;
- N_f - Representa os itens relevantes não recomendados pelo sistema, *i.e.*, falsos negativos.

2.9.3 Medida-F

A Medida-F (F) representa a média harmónica da Precisão e Abrangência apresentada na Equação 2.12. Esta medida permite combinar numa única medida os valores da Precisão e a Abrangência de um sistema de recomendação.

Equação 2.12: Medida-F.

$$F = 2\frac{Pr * Ab}{Pr + Ab} \tag{2.12}$$

No domínio das recomendações, a Medida-F indica a utilidade da recomendação gerada, *i.e.*, quanto mais próximo o valor estiver de 0 menos relevante é a recomendação.

2.10 Recomendações *Online*

Os Sistemas de Recomendação estão implícita ou explicitamente presentes em inúmeros sítios *Web*, desde motores de busca ao comércio electrónico de bens ou serviços. No primeiro caso, o utilizador é guiado, sem se aperceber, através da apresentação de sugestões de produtos, serviços ou sítios consentâneas com o seu historial de interacção. No segundo caso, o utilizador tem de se registar, inserindo os seus dados pessoais e preferências, autenticar e classificar os resultados para a aplicação refinar o perfil do utilizador e efectuar recomendações personalizadas cada vez mais exactas.

2.10.1 Amazon

A Amazon é um exemplo de uma empresa de comércio electrónico que utiliza algoritmos de recomendação para personalizar os produtos que apresenta a cada cliente. No momento do registo, a Amazon solicita ao cliente que colaborativamente declare os interesses e gostos conforme se apresenta na Figura 2.5.

Estes dados fornecidos *a priori* por todos os utilizadores servem para construir o perfil inicial do utilizador. Este perfil é posteriormente refinado à medida que o utilizador interage com a aplicação e vai expressando os seus interesses e

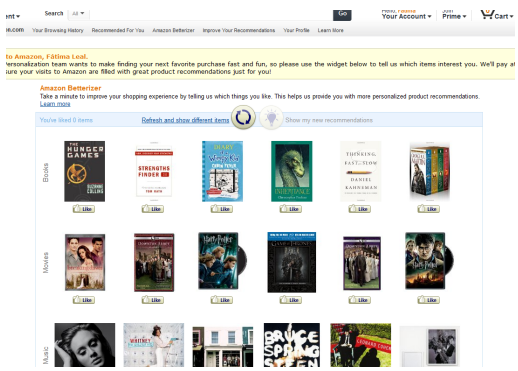

Figura 2.5: Registo das preferências na Amazon.com.

classificando as sugestões apresentadas. A Figura 2.6 apresenta uma recomendação baseada no interesse que o utilizador demonstrou noutro produto da mesma categoria.

Figura 2.6: Recomendação da Amazon.

2.10.2 YouTube

O YouTube apresenta dois níveis de utilização: anónimo e autenticado. Os utilizadores anónimos podem pesquisar e ver vídeos e os utilizadores autenticados podem adicionalmente carregar e classificar vídeos. O YouTube gera recomendações personalizadas em função do historial da actividade do utilizador. Estas recomendações são baseadas no conteúdo já que os vídeos recomendados são das categorias e possuem características idênticas aos das visualizações mais recentes do utilizador. Estas recomendações guiam o utilizador num universo de milhões

de vídeos, pesquisando e propondo ao utilizador aqueles cujo conteúdo que mais adequa ao seu perfil.

2.10.3 MovieLens

O MovieLens é um sítio dedicado à recomendação de filmes que requer o registo e autenticação dos utilizadores e pode ser acedido em `http://movielens.org/login`. Esta aplicação foi desenvolvida no âmbito de um projecto do Departamento de Ciência e Engenharia da Computação da Universidade de Minnesota dos Estados Unidos da América (EUA). O utilizador, quando efectua o registo, classifica um conjunto de filmes apresentados, sendo estes dados usados para criar o perfil inicial do utilizador. O registo só tem sucesso se o utilizador fornecer a sua opinião acerca de pelo menos 15 filmes - Figura 2.7

Figura 2.7: Registo das preferências no MovieLens.

O utilizador, uma vez autenticado, pode ver a classificação dos filmes já pontuados, pesquisar novos filmes e gerar automaticamente a previsão de alguns itens. Esta aplicação disponibiliza conjuntos de dados para teste e validação de Sistemas de Recomendação em diferido.

2.10.4 Netflix

A Netflix é uma empresa norte americana de distribuição comercial de filmes a partir do sítio `https://signup.netflix.com/global` destinada apenas ao mercado dos EUA. A Netflix oferece um serviço de recomendação de filmes baseado nas classificações atribuídas pelos utilizadores registados. As aplicações e projec-

tos de recomendação de filmes desenvolvidas nos EUA adoptam frequentemente a Netflix como fonte de informação.

2.10.5 Last.fm

O Last.fm é um sítio de distribuição comercial e personalização de música disponível em `http://www.lastfm.pt/` que foi fundado no Reino Unido e tem mais de 30 milhões de utilizadores activos. A recomendação de conteúdos musicais personalizados é efectuada pela aplicação Audioscrobbler. O Last.fm constrói um perfil detalhado dos gostos musicais de cada utilizador, reunindo e apresentando as músicas e artistas favoritos numa página gerada a partir dos dados coligidos e armazenados. Esta informação, que consiste no repertório musical pessoal do utilizador, é proveniente de duas fontes:

- *Plugin* do Audioscrobble instalado na aplicação de reprodução musical (*player*) do utilizador;
- Rádio da Last.fm que apresenta e permite aos utilizadores registados classificar (*scrobbling*) a lista das músicas recentemente tocadas.

As recomendações apresentadas ao utilizador incluem, não só, a lista de artistas do seu perfil, mas também listas de perfis com gosto musical idêntico. A aplicação Last.fm efectua o agrupamento de utilizadores com interesses e gostos em comum e permite que o utilizador recomende manualmente artistas, álbuns, *etc.* a outros utilizadores. Dada a sua grande popularidade, editoras e artistas são encorajados a lançarem as suas músicas através da aplicação para obterem uma maior visibilidade do seu trabalho.

2.11 Recomendação de Conteúdos Televisivos

O elevado número de canais de televisão disponíveis aumentou a oferta de conteúdos televisivos. Esta quantidade e variedade de programas disponíveis impede que o telespectador conheça as múltiplas grelhas de programação. A personalização dos conteúdos televisivos através da criação de *Electronic Programme Guide* (EPG) personalizados, *i.e.*, baseados no perfil do telespectador, é uma solução atraente. O perfil do utilizador é construído com base no historial da interacção com o sistema, armazenando os programas vistos ou gravados assim como as preferências explícitas. Esta construção pode obrigar ao registo e acesso a sítios de recomendação personalizada de conteúdos ou à utilização de sistemas de fornecimento automático de conteúdos personalizados.

2.11.1 Jinni

O Jinni é um motor de busca semântica e de recomendação de filmes, programas de televisão e séries disponível em `http://www.jinni.com/`. O utilizador regista-se previamente e autentica-se para usufruir do serviço. As recomendações são geradas através das preferências do utilizador, *i.e.*, através das classificações que o utilizador atribui aos programas apresentados. Este serviço envolve uma taxonomia de classificação de conteúdos criada por profissionais do cinema e aplica aprendizagem automática para analisar as opiniões e os metadados dos programas ou filmes em análise.

Figura 2.8: Jinni.

2.11.2 Matcha

O Matcha é um agregador de vídeos disponível em `http://www.matcha.tv/login.html` que oferece aos utilizadores um serviço de procura, acompanhamento e visualização dos filmes e programas de televisão favoritos provenientes da Netflix, Hulu, iTunes, Amazon e a *Internet Movie Database* (IMDb). Este serviço está também disponível para navegadores *Web* e iPads. A autenticação pode ser efectuada através da conta do Facebook, permitindo, logo à partida, utilizar o perfil do utilizador nesta rede social para gerar recomendações assim como apresentar as preferências dos amigos - Figura 2.9.

2.11.3 Red Bee Media

A Red Bee Media é uma empresa de multimédia sediada em Londres que tem como clientes, não só, estações de televisão como a BBC, mas também marcas como a Nike, a Lonely Planet, *etc.* A Red Bee Media transmite mais de 120 canais de televisão analógica, digital terrestre, satélite digital, cabo, *Internet Protocol TV* (IPTV), incluindo todos os canais nacionais e internacionais da BBC, oferece maioritariamente serviços de *Video on Demand* (VOD) e está disponível em `http://www.redbeemedia.com/`. Em particular, no domínio da

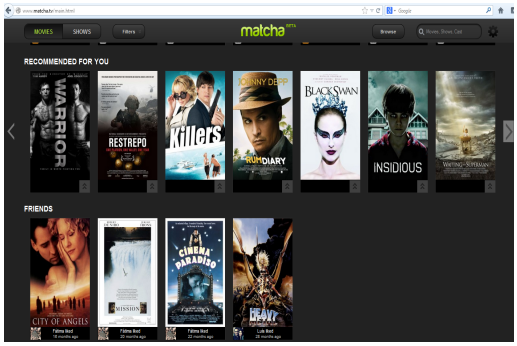

Figura 2.9: Matcha TV.

descoberta de conteúdos, oferece o serviço RedDiscover que inclui um conjunto de soluções para as plataformas de televisão, incluindo pesquisa e recomendação de conteúdos. O mecanismo de recomendação RedDiscover determina, com base nos metadados dos programas e dos telespectadores (historial e preferências), a probabilidade de um programa interessar a um utilizador. Os resultados assim obtidos incluem:

- Recomendações baseadas num programa televisivo frequentemente visto;
- Recomendações activas com base nas preferências explícitas do utilizador;
- Recomendações passivas com base na actividade recente ou no perfil do utilizador;
- Recomendações baseadas nas preferências explícitas de outros utilizadores.

Este sistema de recomendação destina-se a conteúdos televisivos VOD e integra múltiplas fontes de dados a fim de oferecer as melhores sugestões de conteúdo a cada espectador. Os metadados dos conteúdos televisivos provêm de fontes como a IMDb, Rotten Tomatoes ou iTunes. As redes sociais também são utilizadas como fontes de informação. Tecnicamente, as recomendações geradas pelo Red Bee Media são colaborativas e baseiam-se no conteúdo e nas redes sociais [35, 36].

2.11.4 NoTube Beancounter

O projecto NoTube decorreu entre Fevereiro de 2009 e Janeiro de 2012 e dedicou-se genericamente à integração entre conteúdos televisivos e *Web* e à sua difusão através de vários dispositivos. Uma das componentes deste projecto financiado pela União Europeia foi a personalização de conteúdos de TV através da modelação do utilizador e do tratamento de informação contextual. Um dos protótipos

resultantes do projecto NoTube foi o conjunto de serviços *Web* dedicados à criação do perfil social dos utilizadores designado Beancounter.

O Beancounter obtém dados da actividade do utilizador através das redes sociais como o Facebook ou o Twitter e de outras fontes como a Last.fm, a IMDb, o RottenTomatoes, *etc.*, que oferecem conteúdos e dados sobre as interacções dos utilizadores. O perfil do utilizador é representado através de um conjunto de interesses ponderados descritos através de ontologias específicas do domínio televisivo como TV-Anytime ou a Ontologia de Programas da BBC. A informação acerca do projecto encontra-se em `http://notube.tv/` [37, 38].

2.11.5 LinkedTV

O *Television Linked To The Web* (LinkedTV) é um projecto financiado pela União Europeia que pretende interligar os conteúdos *Web* e multimédia. Neste âmbito, os vídeos são interligados a conteúdos *Web* complementares, aumentando assim a informação acerca dos conteúdos audiovisuais. O LinkedTV utiliza as preferências, o contexto (espácio-temporal e a envolvente (empresarial, familiar, *etc.*)) e o estado de espírito do utilizador. O LinkedTV recorre a tesauros e ontologias, incluindo a DBpedia e a Ontologia de Programas da BBC. O utilizador é modelado através da ontologia dedicada *LinkedTV User Model Ontology* (LUMO) que combina o modelo psicológico, o contexto e as preferências do utilizador. A informação sobre o projecto está disponível em `http://www.linkedtv.eu/` [39, 40, 41].

2.11.6 IndexTV

O objectivo deste projecto é desenvolver, não só, ferramentas de descrição de conteúdos, mas também mecanismos de interacção entre os utilizadores e os conteúdos a visualizar. O IndexTV, desenvolvido pela Ramon Llull University em Barcelona, é um sistema de personalização de televisão digital ponto-a-ponto com base no padrão Moving Picture Experts Group 7 (MPEG7) de transmissão sincronizada de conteúdos e a aplicação de metodologias de inteligência artificial para construir um perfil de utilizador fiável com base nas suas preferências. O sistema resultante é compatível com normas como MPEG2 e *Digital Video Broadcasting* (DVB). A geração de recomendações está a cargo de um módulo designado assistente de programas televisivos representado na Figura 2.10.

Este módulo gera sugestões baseadas na descrição dos programas e nas preferências do utilizador recorrendo a um sistema baseado em conhecimento composto por uma base de dados, um motor de inferência e uma interface de entrada e saída de dados. Os dados necessários à criação do perfil são introduzidos através do preenchimento de um formulário que inclui questões acerca das preferências do utilizador, podendo o utilizador ainda adicionar géneros, canais ou personagens

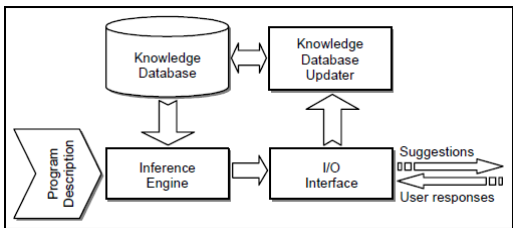

Figura 2.10: Assistente de Programas Televisivos [1].

favoritos. Uma vez armazenados estes dados na base de dados, o motor de inferência compara-os com a descrição dos programas e produz sugestões/recomendações baseadas na similaridade. Estas recomendações textuais aparecem em rodapé como ilustra a Figura 2.11 [1].

Figura 2.11: Mensagem de recomendação [1].

Após a visualização da mensagem o utilizador pode mudar para o canal sugerido, gravar instantaneamente o conteúdo sugerido ou recusar a sugestão.

2.11.7 AIMED

O projecto AIMED foi desenvolvido pela National Chiao Tung University de Taiwan e apresenta um Sistema de Recomendação para televisão personalizada. O mecanismo de recomendação é baseado nas Actividades, Interesses, Estado de Espírito, Experiências e Informação Demográfica. Como se trata de um Sistema de Recomendação híbrido contém métodos baseados no conteúdo e de filtragem colaborativa. As recomendações são efectuadas, não só com base na informação pessoal dos utilizadores, *e.g.*, o estilo de vida, dados demográficos e preferências explícitas, mas também são baseadas no estado de espírito e nas visualizações

prévias. A informação do estado de espírito é dada pelo utilizador através do controlo remoto, como se ilustra na Figura 2.12 [42].

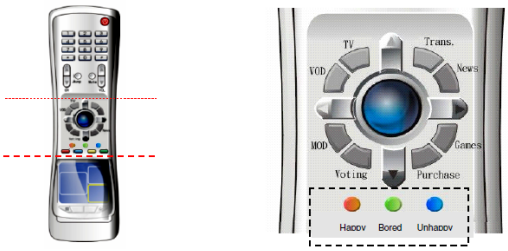

Figura 2.12: Indicação do estado de espírito.

O utilizador preenche para a construção do perfil um questionário quando se regista na aplicação. O questionário encontra-se dividido em três partes: (*i*) dados demográficos; (*ii*) estilo de vida (interesses e actividades); e (*iii*) preferências de categorias de programas. Esta informação é armazenada na STB e é enviada ao mecanismo de recomendação sempre que o utilizador solicita recomendações de programas. As redes neuronais artificiais são a técnica utilizada para construir a previsão e a recomendação, de acordo com as preferências do utilizador.

2.11.8 MiSPOT

MiSPOT é um sistema para televisão digital interactiva que personaliza conteúdos publicitários. O sistema faz uso de técnicas de raciocínio semântico para seleccionar anúncios mais adaptados às preferências, interesses e necessidades de cada espectador. Os anúncios publicitários são concatenados com o programa de televisão que o espectador está a assistir, sendo possível ver o anúncio publicitário e o programa televisivo como se representa na Figura 2.13 [43].

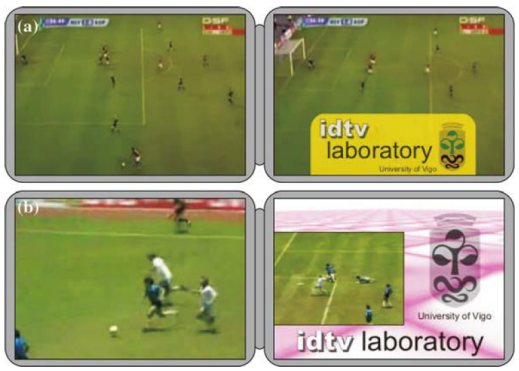

Figura 2.13: Projecto MiSPOT.

2.11.9 ImTV

O ImTV é um projecto de investigação financiado pela Fundação para a Ciência e Tecnologia que envolve universidades e parceiros dos EUA e de Portugal. O estudo foca-se nos novos métodos de trabalho e distribuição de dados multimédia na indústria do entretenimento de forma a acomodar novos utilizadores e respectivos interesses. Um dos seus objectivos é desenvolver mecanismos de interacção mais ricos entre o mundo televisivo e o espectador. Os modelos tradicionais, que se baseiam nas audiências dos programas televisivos, fornecem um quadro incompleto. Um quadro mais rico requer mecanismos que captem dados adicionais como *Short Message Service* (SMS), fóruns de Internet, comentários de redes sociais, *etc.* [44]. O protótipo desenvolvido permite a criação, acesso e partilha de ambientes *Web* personalizados através de diferentes dispositivos. Ao assistir a um programa de televisão ou a um vídeo, os telespectadores podem seleccionar os tópicos que são apresentados ou pedir informações sobre o programa. Existe ainda a possibilidade de aceder, adicionar e compartilhar conteúdos com outros utilizadores a qualquer hora e em qualquer lugar. É utilizado um Sistema de Recomendação que processa a informação das redes sociais para determinar o estado de espírito do utilizador, bem como as avaliações que um determinado programa recebe na respectiva rede. Adicionalmente, recorre à IMDb e à TV.com, entre outros sites, para obter metadados suplementares para caracterizar o conteúdo e permitir melhores recomendações. Para a construção das recomendações é utilizado um mecanismo de aprendizagem automática [45].

2.12 Conclusão

Neste capítulo apresentaram-se os diversos tipos de Sistemas de Recomendação e as respectivas técnicas de recomendação e efectuou-se um estudo comparativo das vantagens e desvantagens de cada técnica. Apresentaram-se algumas das fontes de enriquecimento semântico disponíveis com conhecimento enciclopédico. Referiram-se as técnicas de similaridade que se utilizam para estabelecer a similaridade entre utilizadores, características ou categorias de programas. Foi efectuado um estudo sobre a análise e precisão dos Sistemas de Recomendação pelo que se concluiu que este tipo de análise apenas pode ser efectuada em Sistemas Colaborativos onde se conhece na totalidade os dados envolvidos, por exemplo, o conjunto de dados da MovieLens. Para fazer comparações ou avaliar sistemas sem se ter conhecimento da colecção de dados é comum adoptar-se uma metodologia de *pooling*. Neste capítulo foram também descritos alguns dos Sistemas de Recomendação de conteúdos *online* existentes e, por fim, efectuou-se o levantamento de alguns dos projectos de televisão personalizada mais representativos tendo resultado a construção da Tabela 2.4.

Após este estudo verifica-se que as diferentes metodologias são aplicadas em

Tabela 2.4: Sistemas de Recomendação.

Sistema	Fontes	Mecanismo
JINNI	Netflix;	Colaborativo; Baseado no Conteúdo;.
MATCHA	Facebook; Netflix; Hulu; iTunes; Amazon; IMDb.	Colaborativo.
Red Bee Media	BBC; Rotten Tomatoes; IMDb; iTunes; Facebook.	Baseado no Conteúdo.
NoTube	LastFM; IMDb; Rotten Tomatoes; Ontologia de programas da BBC; Ontoologia da TVAnytime.	Baseado no Conteúdo; Baseado no Contexto.
LinkedTV	DBPedia; WordNet 3.0; Ontologia de programas da BBC; Media Resources; Ninsuna; Open Annotation.	Baseado no Conteúdo; Baseado no Contexto.
IndexTV	Preenchimento de um questionário.	Colaborativo.
AIMED	Preenchimento de um questionário; STB.	Baseado no Conteúdo; Colaborativo.
ImTV	IMDb; TV.com; Facebook.	Baseado no Conteúdo; Colaborativo; Aprendizagem automática.
MiSPOT	Anúncios publicitários.	Baseado no Conteúdo; Colaborativo; Raciocínio Semântico.

função da finalidade do Sistema de Recomendação. A Recomendação Colaborativa é bastante utilizada, mas requer uma interacção inicial que pode ser desmotivadora. O Sistema de Recomendação baseado no Conteúdo, que é a técnica mais utilizada, requer aprendizagem automática e análise semântica. A Recomendação Híbrida apresenta-se como a abordagem mais completa pois anula as desvantagens da Recomendação baseada no Conteúdo e da Recomendação Colaborativa.

No capítulo seguinte são descritas as fontes de dados e de enriquecimento semântico que serão utilizadas no desenvolvimento do Sistema de Recomendação deste trabalho.

Capítulo 3

Fontes de Dados e de Enriquecimento Semântico

Neste capítulo são descritas as fontes de dados de carácter pessoal, social e as fontes de enriquecimento semântico estudadas. Estas fontes de dados incluem: (i) a BBC, a IMDb, o YouTube e os conjuntos de dados MovieLens para os dados pessoais do utilizador; (ii) o Facebook e o Twitter para dados da interacção social; e (iii) a WordNet, os repositórios de dados interligados DBpedia, Freebase, Yago, Linked Movie Database e os serviços de enriquecimento Lupedia e Data Enrichment Service para o enriquecimento semântico.

3.1 Fontes de Dados

As fontes de dados foram divididas em Fontes de Dados Pessoais e Fontes de Dados Sociais. Estas fontes armazenam a informação acerca da actividade pessoal (historial) e social (partilhas, gostos ou comentários) do utilizador que permitirá gerar o perfil do utilizador e, posteriormente, efectuar a recomendação personalizada de conteúdos multimédia.

3.1.1 Fontes de Dados Pessoais

As Fontes de Dados Pessoais permitem obter e caracterizar o comportamento do utilizador. O historial do comportamento é construído a partir das interacções do utilizador com o serviço de difusão de conteúdos e inclui dados como os programas visualizados, o tempo de visualização do programa; o horário de visualização do programa e, caso existam, preferências ou classificações explícitas. O género dos programas e o período do dia em que foram vistos podem, por

exemplo, constituir um ponto de partida para a caracterização do utilizador ou espectador desse período do dia. Se, por exemplo, o historial contém durante o período da manhã maioritariamente programas de desenhos animados poder-se-á então concluir que o espectador naquele período do dia é uma criança. Estes dados são normalmente armazenados pelo fornecedor do serviço televisivo através de uma STB que, por vezes, inclui uma aplicação interactiva. Os dados provenientes desta funcionalidade são também relevantes para as recomendações. Estas aplicações interactivas permitem que o utilizador manifeste as suas preferências, *e.g.*, atribuindo classificações, procedendo a gravações e visualizando a descrição de filmes, séries, *reality shows*, *etc.*

3.1.1.1 BBC

A BBC é a estação de rádio e televisão pública do Reino Unido e detém múltiplos canais de televisão e de rádio. A estação criou uma ontologia de representação dos programas designada Ontologia dos Programas da BBC constituída por um vocabulário simples para descrever os programas de rádio e televisão, incluindo marcas, séries, episódios, *etc.* A ontologia, que é baseada nos documentos de especificação do vocabulário *Friend of a Friend* (FOAF) e da *Music Ontology* [46], permite representar qualquer programa através de atributos como a Identificação do programa, o Título, o Tipo (série ou episódio), a Sinopse, o Horário, a Duração ou Categoria. Para a BBC, a Categoria de um programa consiste no conjunto formado pelo Género e Formato do programa. A Tabela 3.1 apresenta a lista de categorias dos programas que engloba todos os géneros e formatos.

A BBC disponibiliza ainda um EPG que constitui a interface com a sua grelha de programas, permitindo o acesso à programação dos diferentes canais e também ao repositório de programas da BBC através de pedidos *HyperText Transfer Protocol* (HTTP), sendo o resultado fornecido nos formatos *eXtensible Markup Language* (XML), *JavaScript Object Notation* (JSON) e *Resource Description Framework* RDF.

A BBC está envolvida em diversos projectos de personalização e recomendação de conteúdos multimédia como o NoTube ou o LinkedTV e de descrição e interligação de dados como a *Linked Open Data* (LOD). Em Outubro de 2013, a estação apresentou um serviço *on-demand* pago de personalização de conteúdos designado "My BBC" iPlayer destinado a utilizadores registados residentes no Reino Unido [47].

No âmbito deste trabalho adoptou-se a Ontologia dos Programas da BBC para representar os conteúdos (filmes, séries, episódios ou vídeos). A organização e filtragem dos conteúdos por categorias, *i.e.*, géneros e formatos, permitirá realizar recomendações genéricas por categoria.

Tabela 3.1: Categorias BBC [5] [6].

Dados BBC	Descrição
Géneros	Children's; Comedy; Drama; Entertainment; Factual; Learning; Music; News; Religion & Ethics; Sport; Weather.
Formatos	Animation; Appeals; Bulletins; Discussion & Talk; Docudramas; Documentary; Film; Games & Quizzes; Makeovers; Performances & Events; Phone-ins; Readings; Reality; Talent Shows.

3.1.1.2 IMDb

A IMDb[1] é uma base de dados sobre filmes, programas de televisão e jogos, incluindo informações sobre actores, realizadores, data de estreia, localização, *etc.*[48]. Na Figura 3.1 apresenta-se um exemplo da informação retornada.

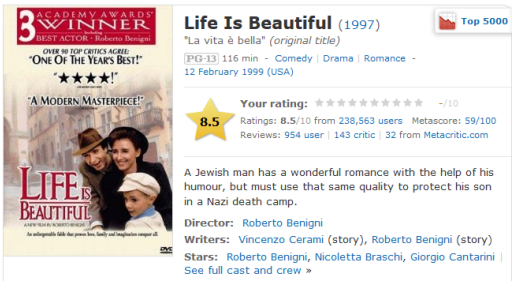

Figura 3.1: Descrição do filme "Life is Beautiful".

[1] `http://www.imdb.com/`

A IMDb caracteriza os filmes através da lista de géneros que a Tabela 3.2 apresenta.

Tabela 3.2: Géneros IMDb [7].

Action;
Comedy;
Family;
History;
Mystery;
Sci-fi;
War;
Adventure;
Crime;
Fantasy;
Horror;
News;
Sport;
Western;
Animation;
Documentary;
Film-Noir;
Music;
Reality-TV;
Talk Show;
Biography;
Drama;
Game-Show;
Musical;
Romance;
Thriller.

A IMDb é uma das Fontes Factuais utilizadas neste projecto para caracterizar filmes, séries ou celebridades, sendo utilizada a sua interface *Web* pública. O Sistema de Recomendação deste projecto utiliza também os géneros da IMDb.

3.1.1.3 YouTube

O YouTube[2] é um serviço da Google de partilha de vídeos com milhões de utilizadores. Milhares de vídeos são carregados diariamente, sendo a fonte de multimédia mais acedida da *Web*. Este serviço permite a criação de um canal pessoal onde cada utilizador pode, de uma forma privada ou pública, disponibilizar os seus vídeos. Os vídeos são catalogados segundo as categorias apresentadas na Tabela 3.3.

Os utilizadores podem partilhar os vídeos nas redes sociais e manifestar o seu interesse ou desinteresse pelo vídeo visto, colocando colaborativamente um

[2] `http://www.youtube.com/`

Tabela 3.3: Categorias YouTube [8].

Autos & Veicles;
Comedy;
Education;
Entertainment;
Film & Animation;
Gaming;
Howto & Style;
Music;
News & Politics;
Non profits & Activism;
People & Blogs;
Pets & Animals;
Science & Tecnology;
Sport;
Travel & Events.

like. Enquanto fonte de dados para efeitos de recomendação, o YouTube pode ser inserido tanto como uma fonte de dados pessoal como social. A partilha de um vídeo ou o interesse manifestado por um vídeo é automaticamente contabilizado pelas redes sociais se o utilizador associar a conta do YouTube à respectiva conta na rede social. Pode também constituir-se como uma fonte de dados pessoais porque mantém o historial das visualizações do utilizador. Este historial contém informação acerca da categoria do vídeo visualizado, o país associado ao vídeo, *etc*. Neste projecto estes dados do historial do utilizador vão permitir a simulação de uma STB.

3.1.1.4 Conjunto de Dados MovieLens

Os *datasets* da MovieLens foram obtidos através do projecto de pesquisa do GroupLens da Universidade de Minnesota e podem ser descarregados a partir de `http://www.grouplens.org/node/73`. O grupo disponibiliza conjuntos de dados de diversos tamanhos. O conjunto MovieLens 100k relaciona 1682 filmes e 943 utilizadores, tendo cada utilizador avaliado pelo menos 20 filmes. Por razões de organização, o conjunto de dados está dividido em diversos ficheiros que incluem a seguinte informação:

- Ocupação (u.occupation) - Lista de profissões dos utilizadores (Figura 3.2);
- Utilizador (u.user) - Informação demográfica do utilizador, incluindo *ID* do utilizador, idade, género, profissão e código postal (Figura 3.3);
- Género (u.genre) - Lista dos géneros dos filmes (Figura 3.4), correspondendo a cada género uma posição no vector de géneros que descreve os filmes;

- Item (u.item) - Informação sobre os filmes, incluindo o título, a data de estreia, o *Uniform Resource Locator* (URL) da descrição na IMDb e o género do filme (Figura 3.5);
- Dados (u.data) - Classificações atribuídos pelos utilizadores aos filmes, tendo cada utilizador classificado pelo menos 20 filmes (Figura 3.6).
- Informação (u.info) - Informação sobre a dimensão do conjunto de dados, *i.e.*, o número de utilizadores, filmes e de classificações.

```
administrator
artist
doctor
educator
engineer
entertainment
executive
healthcare
homemaker
lawyer
librarian
marketing
none
other
programmer
retired
salesman
scientist
student
technician
writer
```

Figura 3.2: Lista de profissões dos utilizadores.

```
1|24|M|technician|85711
2|53|F|other|94043
3|23|M|writer|32067
4|24|M|technician|43537
5|33|F|other|15213
6|42|M|executive|98101
7|57|M|administrator|91344
8|36|M|administrator|05201
9|29|M|student|01002
```

Figura 3.3: Dados dos utilizadores.

Este conjunto de dados, representando o historial da visualização de filmes e respectiva classificação de diferentes utilizadores, pode, no âmbito deste projecto, ser utilizado para simular o acesso a múltiplas STB, permitindo gerar recomendações personalizadas para um conjunto de utilizadores.

3.1.2 Fontes Sociais

As Fontes Sociais permitem obter a actividade social do utilizador. Segundo a eBizMBA, uma empresa que mantém uma base de conhecimento sobre sítios *Web*, o Facebook tem cerca de 750 mil milhões de visitas mensais e o Twitter

```
unknown|0
Action|1
Adventure|2
Animation|3
Children's|4
Comedy|5
Crime|6
Documentary|7
Drama|8
Fantasy|9
Film-Noir|10
Horror|11
Musical|12
Mystery|13
Romance|14
Sci-Fi|15
Thriller|16
War|17
Western|18
```

Figura 3.4: Lista de géneros dos filmes.

```
1|Toy Story (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Toy%20Story%20(1995)
|0|0|0|1|1|1|0|0|0|0|0|0|0|0|0|0|0|0|0|0|0
2|GoldenEye (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?GoldenEye%20(1995)
|0|1|1|0|0|0|0|0|0|0|0|0|0|0|0|1|0|0
3|Four Rooms (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Four%20Rooms%20(1995)
|0|0|0|0|0|0|0|0|0|0|0|0|0|0|0|1|0|0
4|Get Shorty (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Get%20Shorty%20(1995)
|0|1|0|0|0|1|0|0|1|0|0|0|0|0|0|0|0|0
```

Figura 3.5: Dados dos filmes.

user id	item id	rating	timestamp
196	242	3	881250949
186	302	3	891717742
22	377	1	878887116
244	51	2	880606923
166	346	1	886397596
298	474	4	884182806
115	265	2	881171488

Figura 3.6: Classificação dos filmes.

aproximadamente 250 mil milhões [49]. Nestas duas redes sociais a actividade do utilizador é caracterizada através do conjunto de *shares*, *likes* ou *tweets* realizados pelo utilizador. A inclusão da dimensão social na construção do perfil do utilizador permitirá obter um perfil mais próximo do real.

3.1.2.1 Facebook

O Facebook é a rede social mais popular da actualidade [50], contando com cerca de mil milhões de utilizadores em todo o mundo [51]. Este serviço permite que os utilizadores criem uma lista de amigos, coloquem informações sobre a sua personalidade, comentem *posts* de amigos, exprimam a sua concordância ou apoio através de *likes* (Figura 3.7) a um comentário, fotografia ou página, entre muitas outras funcionalidades. Com estes dados das interacções realizadas pelo utilizador, o Facebook constrói um perfil público consoante o nível de privacidade definido pelo utilizador. Por outro lado, o Facebook é também um meio publicitário e de divulgação importante.

Adicionalmente, o Facebook definiu um conjunto de categorias e subcategorias para caracterizar as páginas de acordo com a sua finalidade. Assim, quando um utilizador cria um página no Facebook escolhe qual a categoria principal ("*Local Business or Place*", "*Company Organization or Institution*", "*Brand or Product*", "*Artist, Band or Public Figure*", "*Entertainment*" ou "*Cause or Community*") e as subcategorias que se aplicam à página.

Figura 3.7: Representação do *Like* no Facebook.

Esta rede social utiliza alguns mecanismos de recomendação para recomendação de novos amigos e de páginas. Os *likes*, *posts* e *shares* são dados bastante relevantes para a construção do perfil do utilizador pois espelham os seus gostos e interesses.

3.1.2.2 Twitter

O Twitter permite trocar pequenas mensagens de texto entre utilizadores designadas *tweets*. O Twitter é representado pelo símbolo representado na Figura 3.8, podendo os tweets ser integrados num sítio *Web*. Apesar de não ser tão popular como o Facebook, continua a ser uma das redes sociais com mais utilizadores em todo o mundo. Em 2012 esta rede social contava com cerca de 500 milhões de utilizadores inscritos gerando mais de 340 milhões de *tweets* diários [52].

A informação contida nos *tweets* pode ser analisada e enriquecida de forma a extrair informação relevante para a caracterização do utilizador.

Figura 3.8: Twitter.

3.2 Enriquecimento Semântico

Os dados do utilizador obtidos das fontes de dados pessoais e sociais necessitam de ser processados de forma sistemática. Em particular, os dados provenientes das redes sociais, *e.g.*, os comentários publicados nas redes sociais, necessitam de ser enriquecidos semanticamente através do recurso a tesauros, repositórios abertos de dados interligados e serviços de enriquecimento semântico. Este é um dos principais desafios dos Sistemas de Recomendação - a utilização e exploração da *Web* semântica [53] - devido à importância da Internet e das redes sociais na vida dos utilizadores.

3.2.1 Tesauros

Os tesauros são ferramentas que agrupam as palavras do léxico de uma língua em conjuntos de significados. A linguagem corrente é composta por frases e palavras com significados diversos de difícil compreensão para os sistemas computacionais. Para uma aplicação interpretar o significado de uma frase é necessário mais que um dicionário pois a linguagem corrente é polissémica, *i.e.*, a mesma palavra ou frase pode adquirir diversos significados de acordo com o contexto em que se insere. A ferramenta da língua inglesa mais completa é a WordNet.

3.2.1.1 WordNet

A WordNet[3] é uma base de dados lexicográfica pública com uma interface específica para aplicações computacionais desenvolvida pela Universidade de Princeton. A WordNet organiza as palavras (nomes, adjectivos, verbos e advérbios) em conjuntos de acordo as seguintes relações semânticas [54]:

- Sinonímia - Relação semântica entre duas ou mais palavras com o mesmo significado[4].
  - Triste, infeliz;

[3] `http://wordnet.princeton.edu/`
[4] `http://www.priberam.pt/dlpo/`

  - Rápido, veloz.

- Antonímia - Relação semântica entre palavras com significação oposta[4].
  - Triste, feliz;
  - Rápido, lento.

- Hiperonímia - Relação semântica entre uma palavra de significado mais geral ou abrangente e outra ou outras com significado mais específico em relação à primeira[4].
  - Maçã, Macieira;
  - Macieira, Árvore.

- Meronímia - Relação semântica entre uma palavra que tem o significado de uma parte e outra com significado de um todo em relação à primeira[4].
  - Volante, Carro;
  - Barco, Frota.

Estas relações entre as palavras do léxico de uma língua permitem efectuar a desambiguação semântica.

3.2.2 Repositórios Abertos de Dados Interligados

Os repositórios abertos de dados interligados - *Linked Open Data*[5] - foram desenvolvidos com o intuito de publicar, partilhar e interligar dados e conhecimento da Rede Semântica. A Figura 3.9 representa a nuvem LOD [55]. Os dados armazenados são considerados recursos e são representados de acordo com o modelo proposto pelo RDF. O RDF define um modelo de metadados para a descrição de recursos baseado em declarações do tipo sujeito-predicado-objecto designadas triplas RDF. O sujeito representa o recurso, identificado através de um *Uniform Resource Identifier* (URI), o predicado especifica a relação entre sujeito e objecto e, por último, o objecto pode ser um valor literal ou outro recurso identificado pelo respectivo URI. Desta forma, dados dos mais diversos domínios, representados através de triplas RDF, são interligados. Uma ligação entre dados de diferentes repositórios resume-se a uma tripla RDF em que o sujeito e o objecto pertencem a *namespaces* de diferentes repositórios [56], *e.g.*, à DBpedia e à Freebase.

A caracterização dos dados provenientes das redes sociais, que é baseada na análise do conteúdo, recorre aos repositórios abertos de dados interligados para aceder à descrição de pessoas, lugares, eventos, organizações, filmes ou programas

[5] `http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`

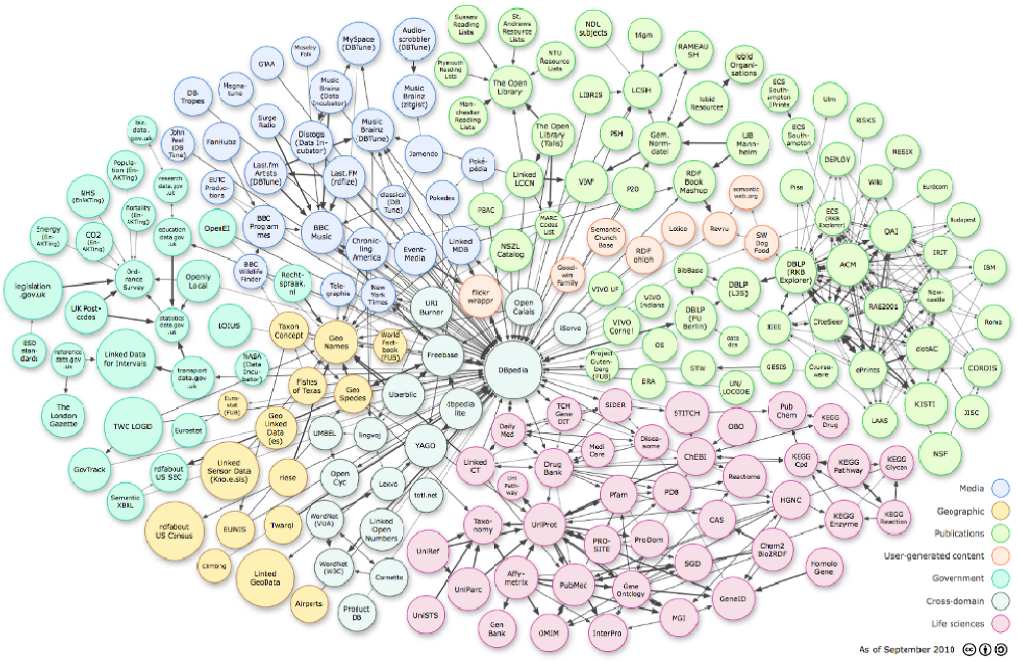

Figura 3.9: Nuvem LOD [2].

televisivos. A DBpedia, a Freebase, a Yago e a Linked Movie Database são exemplos de bases de conhecimento que permitem obter a descrição e caracterização dos dados em questão.

3.2.2.1 DBpedia

A DBpedia[6] extrai informação estruturada da Wikipedia. A Wikipedia, enquanto fonte de informação disponível na *World Wide Web*, consiste num conjunto de descrições de conceitos. Este conhecimento contido na Wikipedia é extraído pela DBpedia na forma de triplas RDF conforme definido pelo World Wide Web Consortium (W3C), permitindo a aplicação da tecnologia da Rede Semântica para explorar e extrair qualquer informação contida na Wikipedia [57]. A informação pode ser consultada através de um navegador ou através de *queries* SPARQL *Protocol and RDF Query Language* (SPARQL). A Figura 3.10 apresenta o resultado de pedido efectuado através do navegador relativo ao recurso "Porto" (`http://dbpedia.org/resource/Porto`).

A DBpedia ajuda a enriquecer e categorizar os dados provenientes da interacção social do utilizador, que são maioritariamente constituídos por textos e partilhas, para refinar o perfil do utilizador.

[6] `http://dbpedia.org`

Figura 3.10: DBpedia.

3.2.2.2 Freebase

A Freebase[7] é uma base de conhecimento aberta construída, à semelhança da Wikipedia, com a colaboração dos utilizadores [58]. A Freebase conta com mais de 39 milhões de tópicos que abrangem diversos domínios e entidades e coloca à disposição dos utilizadores uma biblioteca de interface - *Application Programming Interface* (API)[8] - que permite aceder à informação armazenada através de pedidos HTTP, *e.g.*, `https://www.googleapis.com/freebase/v1/search?query=ISEP`. A Figura 3.11 apresenta o conjunto de dados em formato JSON devolvidos pela Freebase cerca do acrónimo ISEP.

A Freebase pode também ser interrogada utilizando uma linguagem análoga ao SPARQL designada *Metaweb Query Language* (MQL). As *queries* MQL são reescritas directamente no URL do pedido HTTP. Por exemplo, para se obter a lista de álbuns do grupo musical Madredeus especifica-se o seguinte URL: `https://www.googleapis.com/freebase/v1/mqlread?query={"type":"/music/artist","name":"Madredeus","album":[]}`. A Figura 3.12 apresenta o resultado.

3.2.2.3 Yago

A Yago[9] é uma grande base de conhecimento semântico que engloba a DBpedia, a WordNet e a GeoNames. É composta por mais de 10 milhões de entidades representando pessoas, organizações, cidades, *etc.* Segundo [59], as principais vantagens da Yago face aos demais repositórios são:

- Apresentar uma elevada precisão dos resultados obtidos ($\approx$ 95 %);

[7] `http://www.freebase.com/`
[8] `https://developers.google.com/freebase/`
[9] `http://www.mpi-inf.mpg.de/yago-naga/yago/`

```
{
   status: "200 OK",
 - result: [
    - {
         mid: "/m/04jmftk",
         id: "/en/isep",
         name: "Isep",
       - notable: {
            name: "City/Town/Village",
            id: "/location/citytown"
         },
         lang: "en",
         score: 20.549786
      },
    - {
         mid: "/m/084z2_",
         id: "/en/international_students_exchange_program",
         name: "International Student Exchange Programs",
       - notable: {
            name: "Non-profit organization",
            id: "/organization/non_profit_organization"
         },
         lang: "en",
         score: 8.976775
      },
    - {
         mid: "/m/06814j",
         id: "/en/instituto_superior_de_engenharia_do_porto",
         name: "Instituto Superior de Engenharia do Porto",
         lang: "en",
         score: 6.118175
      },
```


Figura 3.11: Exemplo de dados obtidos da Freebase.


```
{
  - result: {
      - album: [
            "Os Dias Da Madredeus",
            "Lisboa",
            "O Espírito da Paz",
            "Ainda",
            "O Paraíso",
            "O Porto (disc 1)",
            "Antologia",
            "Movimento",
            "Electronico",
            "Euforia",
            "Um Amor Infinito",
            "Faluas do Tejo",
            "Palavras Cantadas",
            "Existir",
            "Electrónico"
         ],
         name: "Madredeus",
         type: "/music/artist"
      }
}
```


Figura 3.12: Exemplo de dados obtidos por *query* MQL.

- Ser suportada por uma ontologia espácio-temporal;
- Especificar domínios temáticos, *e.g.*, música e ciência, provenientes dos domínios da WordNet.

O acesso ao conteúdo da Yago é igualmente efectuado através de um *endpoint* SPARQL. Em termos ontológicos possui ligações com a ontologia da DBpedia e a ontologia *The Suggested Upper Merged Ontology* (SUMO).

Figura 3.13: Yago.

3.2.2.4 Linked Movie Database

Nos Sistemas de Recomendação de conteúdos multimédia todos os dados interligados acerca de filmes e séries são relevantes. A *Linked Movie Database*[10] (LinkedMDB) destaca-se na qualidade de primeiro repositório semântico aberto sobre filmes. Os dados podem ser obtidos através de um navegador ou de um *endpoint* SPARQL. Na Figura 3.14 apresenta-se uma parte da descrição do filme "The Shinnig".

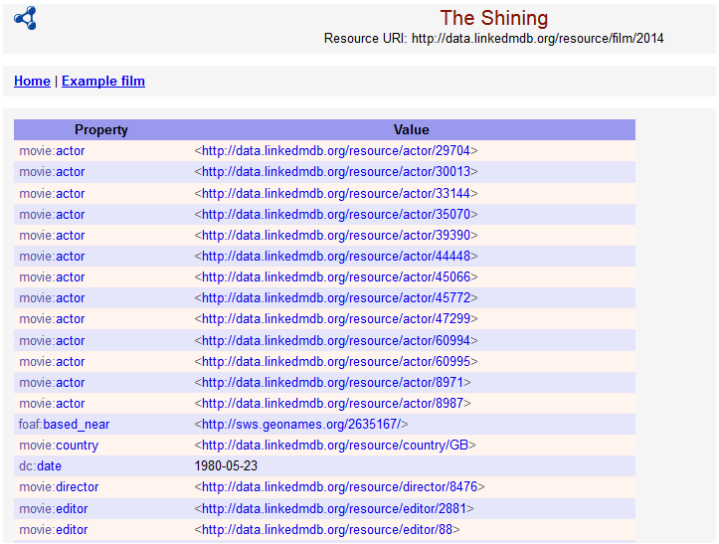

Figura 3.14: Linked Movie Database.

A LinkedMDB inclui nos resultados múltiplas interligações e referências a páginas *Web* relacionadas, *e.g.*, da IMDb, Rotten Tomatoes e Freebase.

3.2.3 Serviços de Enriquecimento

Para a análise de conteúdo de texto são necessários serviços de enriquecimento de texto que identifiquem expressões relevantes que possam descrever o utilizador

[10] `http://linkedmdb.org/`

e os seus interesses e as enriqueçam com as respectivas interligações na nuvem LOD. Dois exemplos de serviços de enriquecimento semântico de texto são a Lupedia e o *Data Enrichment Service* (DES).

3.2.3.1 Lupedia

O projecto NoTube desenvolveu um serviço de enriquecimento de texto denominado Lupedia[11] que usa a Ontotext para procurar palavras que possam ser descritas e enriquecidas pela DBpedia e LinkedMDB. Suporta diversas línguas como o Inglês, Italiano e o Francês. Por exemplo, submetendo ao serviço de enriquecimento Lupedia a frase "*Luis de Camões was a good portuguese writer*", resultam os seguintes URI:

- `http://dbpedia.org/ontology/Work`
  - `http://dbpedia.org/resource/Camoes_(film)`
  - `http://dbpedia.org/resource/Was_(novel)`
- `http://dbpedia.org/ontology/Writer`
  - `http://dbpedia.org/resource/Luis_de_Camoes`
- `http://dbpedia.org/ontology/Person`
  - `http://dbpedia.org/resource/Luis_de_Camoes`

A partir destes URI pode-se aceder aos repositórios da nuvem LOD para o enriquecimento detalhado das palavras encontradas.

3.2.3.2 Data Enrichment Service

O DES[12] é uma infra-estrutura para extracção de informação importante. Muitos documentos contêm dados não estruturados que podem ser relevantes para a construção do perfil de um utilizador. O DES permite extrair um conjunto palavras-chave de um texto, enriquecê-las através das interligações da nuvem LOD para, posteriormente, serem utilizadas na construção do perfil do utilizador. A Figura 3.15 mostra um exemplo em que o serviço de enriquecimento detectou que New York é um lugar e forneceu uma ligação à DBpedia para a descrição e caracterização de New York.

[11] `http://lupedia.ontotext.com/`
[12] `http://openup.tso.co.uk/des`

Figura 3.15: Sistema de enriquecimento de dados.

3.3 Conclusão

Neste capítulo foram apresentadas, não só, as fontes de dados utilizadas para a criação do perfil do utilizador, mas também as fontes que permitem efectuar o seu enriquecimento semântico. Os dados que contribuem para a construção do perfil do utilizador são provenientes de dois tipos de fontes: pessoais e sociais.

As fontes de dados pessoais analisadas incluem: (*i*) a IMDb; (*ii*) o YouTube; e (*iii*) os conjuntos de dados da MovieLens. A IMDb fornece os géneros que classificam as séries e filmes. Para o historial do utilizador, e apesar de ter sido realizado o estudo do conjunto de dados da MovieLens, vai-se utilizar o YouTube. Esta escolha baseou-se no facto de o presente trabalho se focar no mundo da multimédia e os conjuntos de dados disponibilizados pela MovieLens serem apenas acerca de filmes, o que provocaria no resultado final uma centralização apenas na área cinematográfica. Dada a heterogeneidade da representação das categorias dos filmes, séries, programas e vídeos pelas diferentes fontes de informação, adoptou-se o conjunto de categorias da BBC. As fontes sociais utilizadas são o Facebook e o Twitter.

O enriquecimento semântico é realizado através de tesauros, repositórios abertos de dados interligados e serviços de enriquecimento. Os serviços de enriquecimento permitem processar texto e extrair referências a recursos de potencial interesses. Esta funcionalidade está embutida na plataforma Beancounter que é reutilizada neste projecto. O Beancounter utiliza o serviço de enriquecimento semântico Lupedia, que é baseado na DBpedia e no LinkedMDB, para determinar os recursos de interesse provenientes da interacção social do utilizador. Os interesses assim identificados são apresentados através da ligação ao respectivo recurso DBpedia. Na Tabela 3.4 está uma comparação da interacção, interligação e exportação das diferentes fontes do repositório aberto de dados interligados.

Tabela 3.4: Repositórios Abertos de Dados Interligados [9].

		DBpedia	Freebase	Yago	LinkedMDB
Interacção	*Queries* SPARQL	✓		✓	✓
	Pedidos HTTP	✓	✓		
Interligação	Wikipedia	✓			✓
	Freebase	✓			✓
	Yago	✓			✓
	Geonames	✓		✓	✓
	LinkedMDB	✓	✓		
	DBpedia		✓	✓	✓
	Rotten Tomatoes				✓
	IMDb				✓
Exportação	Descarga de *datasets*	✓	✓	✓	✓

Pela comparação dos repositórios de dados interligados e tendo em conta que a maior parte dos dados utilizados serão relativos a filmes e séries, optou-se por se utilizar sobretudo a DBpedia e a Freebase dada a variedade de ligações com fontes de dados relacionados que fornecem. No próximo capítulo serão analisadas e descritas as componentes do Perfil do Utilizador.

Capítulo 4

Perfil do Utilizador

Neste capítulo descreve-se a representação do Perfil do Utilizador e as suas componentes e são referidas as fontes utilizadas para a construção de cada componente.

4.1 Representação do Perfil do Utilizador

O perfil do utilizador constitui o modelo do utilizador no sistema e inclui todas as características disponíveis, nomeadamente, dados pessoais, incluindo contexto, historial e preferências, e sociais do utilizador. O perfil do utilizador permite ao sistema de recomendação gerar recomendações, antecipando o comportamento do utilizador.

4.1.1 Contexto

A definição de contexto é ambígua pois pode incluir diferentes níveis de detalhe de uma entidade. A sua caracterização e representação passa pela resposta a questões do tipo quem, como, o quê, onde e quando. Os dados pessoais, contexto espácio-temporal e o historial de interacções são os dados que caracterizam num dado momento uma entidade. Estes tipos de contexto não só respondem às perguntas referidas como também funcionam como índices para aceder a outras fontes de informação contextual.

4.1.1.1 Dados Pessoais e Preferências

Os dados pessoais e respectivas preferências são especialmente relevantes para caracterizar o utilizador e incluem a identidade, idade, género, naturalidade, nacionalidade, língua materna e as preferências. As preferências podem ser expressas pelo próprio utilizador, especificando gostos, interesses ou passatempos.

4.1.1.2 Contexto Físico

O contexto físico do utilizador encontra-se associado à sua localização. Os parâmetros que fazem parte deste contexto são a região geográfica e a caracterização do espaço. O país, continente e ilhas são exemplos de regiões geográficas e cidade, vila, aldeia, campo e praia são exemplos de caracterização do espaço. Este tipo de informação permite efectuar recomendações consentâneas com a envolvente externa do utilizador.

4.1.1.3 Contexto Temporal

Os dados temporais são essenciais em muitos Sistemas de Recomendação. O conhecimento da hora do dia, dia da semana, estação, época festiva é obrigatório para recomendar uma actividade ou sugerir um filme ou uma série. Por exemplo, para se recomendar a aquisição de um pacote de férias é necessário conhecer o contexto temporal do utilizador.

4.1.2 Historial de Interacções Pessoais

O historial de interacção do utilizador com o sistema é uma fonte de informação essencial e fiável para qualquer sistema de recomendação. No contexto das recomendações televisivas os dados recolhidos incluem tipicamente o tempo de visualização de um programa, a hora a que foi visto, o tipo do programa, *etc.*

4.1.3 Historial de Interacções Sociais

As redes sociais são actualmente uma fonte de informação incontornável dada a sua natureza: são montras de opiniões, interesses e preferências dos utilizadores. Neste contexto, os Sistemas de Recomendação recolhem os dados da actividade social do utilizador de redes sociais como o Facebook e o Twitter e utilizam-nos para gerar recomendações mais precisas. Para além dos dados sociais do utilizador, podem ainda obter informação acerca de utilizadores com gostos e preferências idênticos. É de destacar a informação representada nos *likes* e *tweets* dado que foram inseridos expressamente pelo utilizador. Do ponto de vista de um dado produto, *e.g.*, um filme, a sua relevância pode ser aferida através do conjunto global de *likes* atribuídos ou através de *sites* como a IMDb.

4.1.4 Aprendizagem Automática

A aprendizagem automática é utilizada nos Sistemas de Recomendação para a construção do perfil do utilizador, tentando extrair automaticamente das actividades e interacções recolhidas as preferências do utilizador. Os algoritmos mais utilizados neste contexto são:

- Métodos probabilísticos;
- Método do vizinho mais próximo [23];
- Árvores de decisão;
- Redes neuronais artificiais [14].

4.1.4.1 Métodos Probabilísticos

Um método probabilístico recorre a uma função que descreve a probabilidade de uma variável aleatória assumir certos valores. Nos Sistemas de Recomendação o uso das probabilidades facilita o processo de construção dos perfis. Thomas Bayes propôs um conjunto de métodos probabilísticos designados métodos bayesianos que são importantes por dois motivos:

1. Fornecem algoritmos práticos de aprendizagem:
   - Classificador Naïve Bayes;
   - Combinam o conhecimento *a priori* com os dados observados *a posteriori*;
   - Requerem probabilidades *a priori*.
2. Fornecem uma estrutura conceptual útil:
   - Avaliação de outros algoritmos de aprendizagem.

O classificador Naïve Bayes é baseado no teorema de Bayes representado através da Equação 4.1 [60].

Equação 4.1: Teorema de Bayes.

$$P(c|d) = (\frac{P(c)P(d|c)}{P(d)}) \tag{4.1}$$

A probabilidade de um documento d estar na classe c – $P(c|d)$ – é definida através da probabilidade *a posteriori* do documento d pertencer à classe c. Em contraste, a probabilidade *a priori* $P(c)$ é independente de d. Esta estimativa é baseada na probabilidade *a priori*, $P(c)$, e na probabilidade de observar um documento na classe c, $P(d|c)$. $P(d)$ é a probabilidade de observar um documento d [23]. Para a classificação de textos vários investigadores têm adoptado o algoritmo Naïve Bayes para os seus trabalhos dada a grande qualidade dos resultados [14]. O classificador Naïve Bayes é bastante utilizado em Sistemas de Recomendação baseados em conteúdo [61].

4.1.4.2 Método do Vizinho mais Próximo

O algoritmo de interpolação baseado no método do vizinho mais próximo processa todos os dados do perfil, sejam eles preferências implícitas ou explícitas. Para classificar um perfil, o algoritmo compara-o com todos os perfis armazenados usando uma função de similaridade que determina o "vizinho mais próximo". A função de similaridade utilizada depende do tipo de dados. Para dados estruturados utiliza frequentemente a métrica da distância euclidiana – ver Secção 2.8. Quando se trata de vectores de características, adopta a similaridade dos cossenos – ver Secção 2.8 [14]. O *k-Nearest-Neighbour* (kNN) é uma variante deste algoritmo que utiliza os k vizinhos mais próximos, em vez de usar o vizinho mais próximo como propunha o algoritmo original [62].

4.1.4.3 Árvores de Decisão

As árvores de decisão são constituídas por nós internos etiquetados através de termos e em que os ramos da árvore que partem desses termos estão etiquetados com o peso que esse termo tem, por exemplo, num documento. O objectivo do algoritmo é percorrer a estrutura hierárquica da árvore fazendo comparações de acordo com as acções do utilizador. As decisões do sistema são tomadas com base numa sequência de testes [63].

4.1.4.4 Redes Neuronais Artificiais

Uma rede neuronal artificial é um conjunto de nós interligados através de *links* ponderados inspirados na arquitectura do cérebro biológico. As unidades de processamento são compostas por redes que têm a capacidade de aprender e solucionar problemas quando o conjunto de dados disponível é elevado. Uma rede neuronal artificial pode ter centenas ou milhares de unidades de processamento [64]. Este tipo de sistema é constituído por diversas camadas:

- Camada de Entrada - Recebe os dados;
- Camadas Intermédias ou Escondidas - Processam os dadoa através das conexões ponderadas,extraindo as principais características;
- Camada de Saída - Devolve o resultado.

4.1.4.5 Factorização de Matrizes

A factorização de matrizes é uma operação da álgebra linear para decompor matrizes. No contexto dos Sistemas de Recomendação a factorização é aplicada às matrizes dos utilizadores *versus* itens, *i.e.*, que contêm as classificações atribuídas pelos utilizadores aos produtos. Cada item i é associado a um vector $q_i \in \mathbb{R}$ e cada utilizador u é associado a um vector $p_u \in \mathbb{R}$. A factorização de matrizes

consiste no cálculo do produto escalar dos dois vectores referidos como ilustra a Equação 4.2 [65].

Equação 4.2: Produto vectorial na factorização das matrizes.

$$r_{ui} = p_u.q_i \tag{4.2}$$

A factorização de matrizes é uma das técnicas mais utilizadas nos Sistemas de Recomendação colaborativos dada a sua precisão e escalabilidade [66].

4.1.5 Ontologias para Representação do Perfil do Utilizador

Existem diversas propostas de modelos e/ou ontologias para a representação do perfil do utilizador. Neste domínio destacam-se as seguintes propostas:

- *General User Model Ontology* (GUMO) - Esta ontologia OWL foi desenvolvida na Saarland University para representar o utilizador através do seu estado emocional, características e estado fisiológico [67].
- *LinkedTV User Model Ontology* (LUMO) - Esta ontologia OWL foi desenvolvida no âmbito do projecto LinkedTV para representar semanticamente o utilizador e, assim, permitir a personalização via contextualização semântica [68].

Estes modelos serviram de inspiração para a ontologia do perfil do utilizador desenvolvida no âmbito desta tese.

4.2 Construção das Componentes do Perfil

O perfil do utilizador vai ser dividido em quatro vertentes: (*i*) perfil contextual; (*ii*) perfil social; (*iii*) perfil baseado no historial; e (*iv*) preferências explícitas. Cada uma destas vertentes terá um peso associado de acordo com a sua importância. A junção de todos os dados que definirão o utilizador resultará numa ontologia.

4.2.1 Perfil Contextual

O contexto do utilizador engloba não só os seus dados pessoais como também o contexto físico e temporal em que está envolvido. Os primeiros dados a serem atribuídos ao perfil do utilizador serão de acordo com o seu contexto. Como analisado, o contexto do utilizador envolve tanto os seus dados pessoais como também o meio onde este está inserido. Os dados pessoais como o nome, a morada, a idade e o género são as principais fontes para identificar o utilizador. Estes dados podem ser fornecidos pelo próprio no momento da adesão ao serviço

de recomendação. Neste primeiro contacto com o utilizador normalmente é pedida a introdução de algumas preferências, resolvendo desta forma o problema da primeira recomendação. Ainda no âmbito do perfil contextual, as informações sobre o meio físico, cultural e temporal também podem contribuir para a criação de recomendações. A informação da época do ano, do meio em que o utilizador se situa, bem como o local, são os dados que se podem ter em conta. A época do ano pode ser facilmente identificada pela análise do calendário. A determinação do local exacto do utilizador e a descrição do ambiente em que está inserido podem ser conseguidas com o uso de sensores integrados nos dispositivos electrónicos. Neste trabalho a construção do contexto é realizada através do nome do utilizador, data e local.

A definição de modelos ou estereótipos baseados no contexto do utilizador permite criar grupos de utilizadores com contextos idênticos e preparar uma grelha de recomendação de programas diária por omissão. Estes estereótipos podem ser gerados com base no contexto temporal, no género do utilizador, dado que tipicamente os gostos femininos e masculinos diferem, e no contexto espacial. Poderiam, então, ser criados estereótipos para o dia-de-semana, fim-de-semana, épocas festivas ou férias por área geográfica e género. Um estereótipo consiste numa grelha de categorias de programas a instanciar com a programação do dia em que for aplicado. Um exemplo de um estereótipo está apresentado na Tabela 4.1 e foi criado a partir dos formatos e géneros da BBC.

Tabela 4.1: Esteriótipos.

Semana	Fim-de-semana
Bulletins;	Animation;
Discussion and Talk;	Appeals;
Docugramas;	Films;
Documentaries;	Makeovers;
Magazines and Reviews;	Performances and Events;
Readings;	Talent Shows;
Phone-ins.	Reality.

4.2.2 Perfil Social

No perfil social são integradas todas as interacções do utilizador nas redes sociais Facebook e Twitter. Do Facebook virão todas as informações dos *likes* e das partilhas efectuadas, utilizando para tal as ferramentas de enriquecimento descritas. O mesmo acontecerá com os dados do Twitter. A manipulação destes dados dará origem às categorias que mais se adeqúem e descrevam o utilizador. A informação que adverte dos *likes* inclui as categorias associadas à página para a qual o utilizador manifestou interesse. Na Tabela 4.2 estão descritas algumas dessas categorias.

Tabela 4.2: Categorias Facebook.

TV Programme;
Musician;
Technology and telecommunications service;
Film;
Financial Service;
Education;
Religious organisation;
Music;
Health and beauty;
Museum/attraction;
Sports/athletics.

Dado que o obejctivo é recomendar conteúdos multimédia, a informação dos *shares* e dos *tweets* será filtrada pára identificar e extrair referências a filmes, séries e programas televisivos. O processamento do texto contido nos *shares* e *tweets* será efectuado pelo serviço de enriquecimento semântico Lupedia. Os dados obtidos serão classificados através de fontes LOD, nomeadamente a Freebase para obter as respectivas referências IMDb.

4.2.3 Perfil baseado no Conteúdo

O historial é normalmente construído pela agregação de todas as visualizações de séries, filmes, *reality shows*, entre outros. Esta informação é a mais importante e a que deverá ter maior peso na criação da recomendação, dado que as visualizações informam das preferências reais do utilizador. Num operador de serviço televisivo, por exemplo o MEO, o historial é armazenado na STB. Dada a inacessibilidade de dados reais é utilizada outra solução. Como historial do utilizador e como alternativa aos dados da STB é utilizado um historial construído a partir dos dados do YouTube e pelos programas da BBC que o utilizador acede após as recomendações. Este conjunto de informação é também transposto para categorias já disponíveis no YouTube.

4.2.4 Preferências Explícitas

Neste projecto as preferências explícitas correspondem às classificações que o utilizador atribuiu aos vídeos YouTube e aos programas da BBC vistos. A classificação dos programas da BBC vistos é importante para perceber se as recomendações foram do agrado do utilizador. Na Figura 4.1 apresenta a interface desenvolvida para a classificação das recomendações.

A escala da classificação varia de um a cinco pontos, permitindo ao utilizador descriminar o grau de adequação da recomendação.

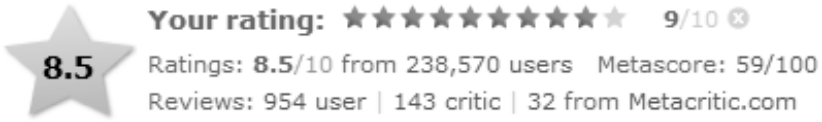

Figura 4.1: Classificação de uma Preferência Explícita.

4.3 Conclusão

Neste capítulo foi efectuada uma análise sobre as características do perfil do utilizador. Esta análise, que foi dividida em várias componentes, permitiu identificar as diferentes fontes de informação assim como a sua natureza e aplicabilidade. A Figura 4.2 ilustra todas as componentes do perfil do utilizador a construir.

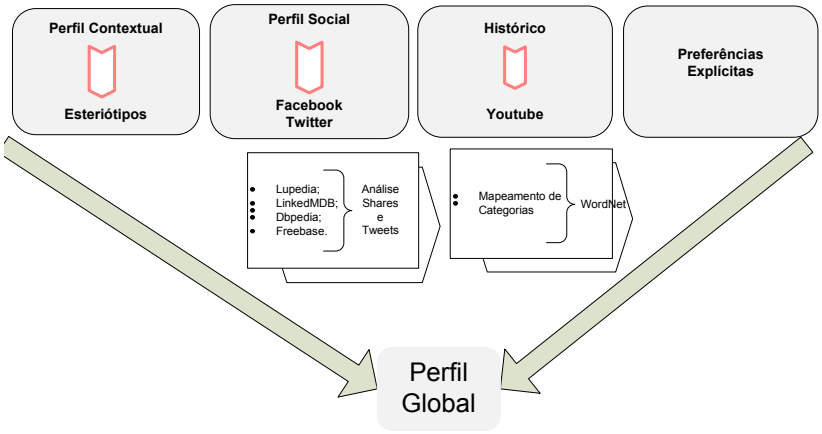

Figura 4.2: Perfil Utilizador.

O perfil social é construído recorrendo à plataforma Beancounter que, por sua vez, utiliza as redes sociais Facebook e Twitter. O historial do utilizador é constituído pelo historial de vídeos do do YouTube e de programas da BBC vistos pelo utilizador. A informação correspondente às preferências explícitas é proveniente dos "gostos" que o utilizador coloca nos vídeos do YouTube e pela classificação das recomendações construídas. O Perfil do Utilizador resulta da junção destas componentes. A cada componente do perfil será atribuído um peso consoante a sua relevância. A construção deste perfil será baseada nas categorias dos programas da BBC.

No próximo capítulo são descritas as tecnologias utilizadas no desenvolvimento deste projecto.

Capítulo 5

Tecnologias e Ambiente de Desenvolvimento

Neste capítulo são descritas todas as tecnologias utilizadas no desenvolvimento deste trabalho.

5.1 Linguagens de Programação

Todas as funcionalidades do Sistema de Recomendação projectado foram desenvolvidas em Java. Esta é uma linguagem de alto nível, orientada aos objectos e independente da plataforma. Para o desenvolvimento da interface *Web* com o utilizador adoptou-se a *framework* JQuery. Esta *framework* permite desenvolver *scripts* de JavaScript e estabelecer ligações assíncronas através de *Asynchronous JavaScript and XML* (AJAX) *scripts* de suporte à interacção do utilizador e à troca de dados assíncrona com os serviços desenvolvidos. A interface *Web* desenvolvida incorpora as funcionalidades da interface do Beancounter e as desenvolvidas no âmbito deste projecto.

5.2 Representação de Dados

Perante uma aplicação *Web*, em que a principal actividade é a troca de informação, a representação dos dados é bastante importante. Esta representação é feita através de linguagens que estruturam os dados para que estes possam ser acedidos e compreendidos facilmente. A XML, *Yet Another Multicolumn Layout* (YAML ou o JSON são exemplos des linguagens desse tipo. Neste projecto todas as bibliotecas de interface suportam o formato JSON.

5.2.1 JSON

O formato JSON é uma sintaxe para armazenamento e troca de dados muito parecido com a XML. Apresenta como vantagens ser menor, mais rápido e mais fácil de analisar. Na Figura 5.1 está a descrição do programa *A Girl By Any Other Name* retirado da API disponibilizada pela BBC.


```
- programme: {
      type: "episode",
      pid: "b03czdsr",
      position: 2,
      title: "A Girl By Any Other Name",
      short_synopsis: "The heroes are entangled in the perilous rites of Dionysus with a stranger called Medusa.",
      media_type: "audio_video",
      duration: 2700,
    - display_titles: {
          title: "Atlantis",
          subtitle: "A Girl By Any Other Name"
      },
      first_broadcast_date: "2013-10-05T20:25:00+01:00",
    - ownership: {
        - service: {
              type: "tv",
              id: "bbc_one",
              key: "bbcone",
              title: "BBC One"
          }
      },
```


Figura 5.1: Representação de dados em JSON.

O processamento de dados JSON é simples, podendo-se recorrer a bibliotecas e ferramentas específicas.

5.3 Ontologias

As ontologias são utilizadas para descrever formalmente conteúdos por meio de linguagens que possam ser processadas computacionalmente, permitindo a realização de inferências automáticas dentro de um domínio previamente determinado e restrito. Existem linguagens específicas como a DARPA *Agent Markup Language for Services* (DAML), a *Web Ontology Language for Services* (OWL-S) ou RDF-Schema que permitem a definição de ontologias. Neste projecto as ontologias criadas foram definidas em OWL.

5.4 NetBeans

O NetBeans[1] é um *Integrated Development Environment* (IDE) gratuito que suporta diversas linguagens de programação como Java, C, C++, *etc.* Permite o desenvolvimento de diversos tipos de aplicações desde aplicações *stand alone*, a bibliotecas ou aplicações *Web*. No IDE podem ser integrados diversos serviços como servidores de aplicações *Web* (Tomcat e Glassfish) ou servidores de base de dados (Derby, MySQL, *etc.*). Os serviços *Web* baseados em SOAP ou RESTful

[1] `https://netbeans.org/`

também são suportados, trazendo já embutidos alguns serviços *Web* RESTful, *e.g.*, Amazon, Flickr e Google, ou serviços SOAP, *e.g.*, o StrikeIron. O NetBeans ajuda no desenvolvimento, verificando a sintaxe do código automaticamente e efectuando sugestões de correcção de erros.

Neste trabalho o IDE escolhido foi o NetBeans. Adicionalmente, integrou-se o servidor de aplicações Tomcat e o servidor de base de dados MySQL. Nas aplicações *Web* são utilizados serviços RESTful.

5.4.1 Apache Maven

O Maven[2] é uma ferramenta que automatiza a preparação, compilação e execução de projectos. Utiliza um documento XML denominado POM onde são declaradas todas as dependências do projecto em causa como mostra a Figura 5.2.

```
<dependency>
    <groupId>com.google.code.gson</groupId>
    <artifactId>gson</artifactId>
    <version>${gson.version}</version>
</dependency>
<dependency>
    <groupId>commons-configuration</groupId>
    <artifactId>commons-configuration</artifactId>
    <version>${commons-configuration.version}</version>
</dependency>
<dependency>
    <groupId>org.codehaus.jackson</groupId>
    <artifactId>jackson-xc</artifactId>
    <version>${jackson-xc.version}</version>
</dependency>
```

Figura 5.2: Declaração de dependências no Maven.

Nestas dependências estão incluídos módulos, componentes externos, API, pastas ou *plug-ins* necessários. O Maven é semelhante ao Apache Ant mas mais poderoso porque faz o *download* automático das dependências declaradas para o repositório local. O NetBeans inclui o repositório Maven desde a versão 6.3, e já inclui o Maven 3 desde a versão 7.0, permitindo criar, compilar e executar um projecto Maven no IDE. A plataforma Beancounter utilizada neste projecto é um projecto Maven.

5.4.2 RESTful Services

A abordagem *Representational State Transfer* (REST) vê cada aplicação *Web* como um conjunto de recursos. Esta arquitectura é baseado em URI e é geralmente executada através do protocolo HTTP. Assim, um serviço RESTful é

[2] `http://maven.apache.org/`

uma implementação de um serviço *Web* utilizando o protocolo HTTP e os princípios REST, *i.e.*, adopta o modelo cliente-servidor, efectua pedidos HTTP GET e POST e obtém de respostas HTTP.

5.4.3 Apache Tomcat

O Apache Tomcat[3] é um servidor de aplicações *Web* do tipo *Servlet* e JSP*Java Server Pages*(JSP). É um servidor de aplicações *open source* desenvolvido pela Apache Software Foundation. Permite ainda alojar serviços *Web* através do motor de serviços *Web* Apache Axis2. O Tomcat efectua o *deploy* automático das aplicações através de ficheiros *Web Application Archive (WAR)*. O Tomcat é integrável com o NetBeans.

5.4.4 MySQL

O MySQL[4] é um servidor de bases de dados que suporta a *Structured Query Language* (SQL). É uma solução frequentemente adoptada para o armazenamento estruturado e persistente de informação. A versão MySQL Community Edition é livre e inclui um pacote completo com interface gráfica (o Workbench) e múltiplas bibliotecas de interface. O MySQL é integrável com o NetBeans.

5.5 Protégé

O Protégé[5] é uma plataforma *open source* de edição de ontologias. Implementa um vasto conjunto de funcionalidades que apoiam a criação, visualização e manipulação de ontologias em vários formatos de representação. A plataforma Protégé suporta duas diferentes formas para representação de ontologias:

- Protégé-Frames: Permite criar e preencher ontologias que são baseadas em frames, de acordo com o *Open Knowledge Base protocolo Connectivity* (OKBC). Neste modelo, uma ontologia consiste num conjunto de classes organizadas numa hierarquia de subordinação para representar conceitos salientes, conjuntos de *slots* associados a classes que descrevem as propriedades e relacionamentos, e um conjunto de instâncias dessas classes;
- Protégé-OWL: Permite construir ontologias para redes semânticas, particularmente para OWL. OWL é o mais recente desenvolvimento em linguagens de ontologia padrão, aprovado pelo World Wide Web Consortium (W3C) para promover a visão da Rede Semântica. Uma ontologia OWL pode incluir descrições de classes, propriedades e suas instâncias. A semântica

[3] `http://tomcat.apache.org/`
[4] `http://dev.mysql.com/`
[5] `http://protege.stanford.edu/`

formal OWL especifica como derivar suas consequências lógicas (factos não literalmente presentes na ontologia, mas inferidos pela semântica).

O Protégé é apoiado por uma forte comunidade de investigadores que o usam como solução para a representação do conhecimento em áreas tão diversas como a biomedicina, a recolha de informação, ou aplicações de aprendizagem automática [69]. O Protégé fornece uma série de *plugins* que ajudam o utilizador na definição e análise da ontologia.

5.6 Tecnologias Web

No âmbito do desenvolvimento da interface *Web* que permitirá ao utilizador interagir com o Sistema de Recomendação utilizou-se como tecnologias de suporte os *Servlets* e documentos *JavaServer Pages* (JSP).

5.6.1 Servlets e JavaServer Pages

Os *Servlets* são programas em Java que são executados do lado do servidor *Web*. Estes permitem a troca de informação entre o cliente e o servidor través de mensagens HTTP. Esta informação pode ser processada pelo uso conjunto de JSP e JavaBeans. Os JSP são modelos de documentos *HyperText Markup Language* (HTML) que podem incluir código Java e declarações JSP.

5.7 Tecnologias de Suporte do Beancounter

O Beancounter, que serve para construir o perfil social de cada utilizador, necessita de três tecnologias *open source* de suporte ao processamento e armazenamento da informação; o Elasticsearch, Redis e o Kestrel.

5.7.1 Elasticsearch

O Elasticsearch[6] é um motor de busca e análise textual baseado no Apache Lucene e desenvolvido em Java. É usado maioritariamente para indexação, armazenamento e recuperação de informação. Fornece uma interface de serviço REST e apenas efectua pesquisas em documentos com formato JSON. O Elasticsearch é, do ponto de vista funcional, um servidor HTTP. Numa implementação distribuída, os dados são divididos em fragmentos e atribuídos a diferentes nós. Embora todos os nós estejam ao mesmo nível de processamento, existe um nó mestre que distribui a informação pelos restantes nós. Cada nó apenas pode

[6] `http://www.elasticsearch.org/`

aceitar um pedido de um dado cliente. Sendo um serviço livre, não é necessária qualquer configuração prévia para a sua utilização. A compreensão do seu funcionamento resume-se à definição de três conceitos:

- *Index* - *Namespace* onde é definido um ou mais *Types*;
- *Type* - Local de armazenamento dos Documents, agrupando-os pelas características em comum;
- *Document* - Local de armazenamento das informações que vão ser indexadas e pesquisadas.

O Elasticsearch requer pelo menos a versão 6 do Java e é utilizado pelo Beancounter para indexação das actividades sociais. A versão utilizada é a 19.0.4. O serviço corre tanto em ambientes Windows como em Linux e para ser lançado basta executar o *script* disponibilizado dentro do directório *bin*. A informação é distribuída pelos diversos nós que o serviço cria automaticamente e é armazenada no directório *data*.

5.7.2 Redis

O Redis[7] é uma base de dados NoSQL baseada no modelo chave-valor, *i.e.*, os dados obtêm-se a partir de uma determinada chave. Os dados são armazenados em memória RAM, permitindo um acesso simples e rápido [70]. Estão disponíveis clientes para diversas plataformas de desenvolvimento, designadamente, o Jedis para a plataforma Java. A implementação do Redis pode ser efectuada tanto em sistemas Windows como Linux, sendo também de fácil instalação e configuração. O Beancounter usa o Redis 2.4. O Redis atribui uma chave a cada utilizador que se regista na plataforma, ficando associadas a essa chave todas as actividades do utilizador.

5.7.3 Kestrel

O Kestrel[8] é um serviço de fila de mensagens executado na *Java Virtual Machine* (JVM) com características de sistema de tempo real. Utiliza *Extensible Messaging and Presence Protocol* (XMPP) para notificar da chegada de novas mensagens à fila. O formato utilizado para representar os perfis é o JSON [71]. O Beancounter usa o Kestrel para colocar em fila todas as mensagens/actividades vindas do Facebook e do Twitter. Desta forma são armazenados em memória todos os perfis e actividades de cada utilizador. A versão utilizada na plataforma é a 2.4.1.

[7] `http://redis.io/`
[8] `http://robey.github.io/kestrel/`

5.8 Instalação e Configuração do Ambiente de Desenvolvimento

As aplicações desenvolvidas e utilizadas neste projecto foram instaladas, configuradas e executadas numa plataforma de 32 b com o sistema operativo Windows 7. Foi instalada a versão 1.7.40 do ambiente de desenvolvimento Java, incluindo *Java Runtime Environment* (JRE) e o *Java Development Kit* (JDK), servidor de aplicações (Tomcat 7.0), o servidor de base de dados MySQL 5.6, as tecnologias de suporte da plataforma Beancounter (Kestrel 2.4.1, Elasticsearch 0.19.4 e Redis 2.4) e o editor de ontologias Protégé 4.2. O servidor de aplicações e o servidor de base de dados foram configurados no IDE NetBeans 7.3.

5.8.1 Servidor de Aplicações

O Apache Tomcat 7.0.34 foi instalado e integrado no IDE NetBeans. Para isso foi necessário proceder à adição do respectivo servidor no separador *Services* do NetBeans e definir as variáveis *Catalina_Home* e *Catalina_Base* indicando o *path* local onde estão alocadas todas as bibliotecas e recursos do Apache Tomcat, como se mostra na Figura 5.3. Por omissão, os portos do serviço definidos são o 8080 e 8005.

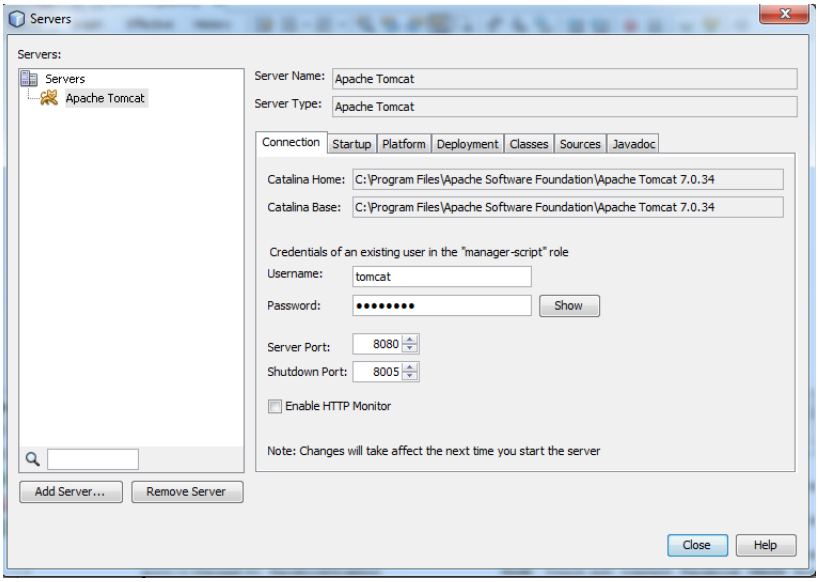

Figura 5.3: Configuração Apache Tomcat no NetBeans.

Para permitir a gestão das aplicações *Web* via navegador, o ficheiro `tomcat-users.xml` tem que ser alterado de forma a permitir a autenticação do utilizador em questão no serviço de aplicações. As alterações resultam da junção

da seguinte informação `<user name="xxxx"password="xxxx"roles="admin-gui, manager-gui,manager-script,admin"/>`. O nome do utilizador e correspondente palavra-chave deverão ser os mesmos do serviço no NetBeans. A Figura 5.4 mostra a página *Web* de gestão de aplicações.

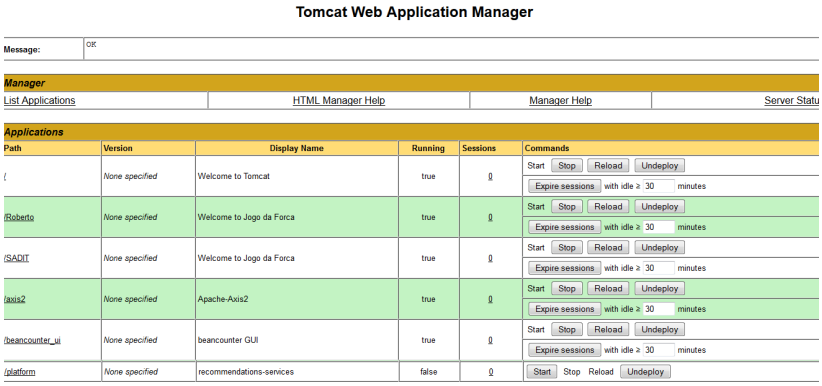

Figura 5.4: Gestão de Aplicações.

5.8.2 Armazenamento Persistente de Dados

Foi instalado um serviço de armazenamento persistente de dados MySQL *Community Server* 5.5 para guardar alguma informação necessária no âmbito da construção da recomendação. Apesar desta tecnologia incluir uma interface gráfica (MySQL Workbench), o serviço foi integrado no NetBeans após a instalação da biblioteca do respectivo connector J. Para proceder à configuração deste serviço no NetBeans, adiciona-se uma nova conexão no separador *Services* e efectua-se a configuração, segundo a Figura 5.5.

5.8.3 Plataforma Beancounter

A plataforma Beancounter utilizada para a recolha, armazenamento e construção do perfil social do utilizador necessita de três tecnologias de suporte o Kestrel, o Elasticsearch e o Redis. Para o lançamento destes três serviços foi definida uma variável de ambiente que identifica as pastas de instalação dos serviços. Desta forma, o lançamento consiste na execução dos três *scripts* representados no Excerto de Código 5.1, 5.2 e 5.3.

Uma vez lançados estes três serviços pode ser feito o arranque da plataforma Beancounter no NetBeans. Estes serviços asseguram o suporte ao armazenamento dos dados sociais dos utilizadores (autenticação e dados), assim como ao conjunto de aplicações que constituem a plataforma Beancounter. Estas aplicações/processos incluem:

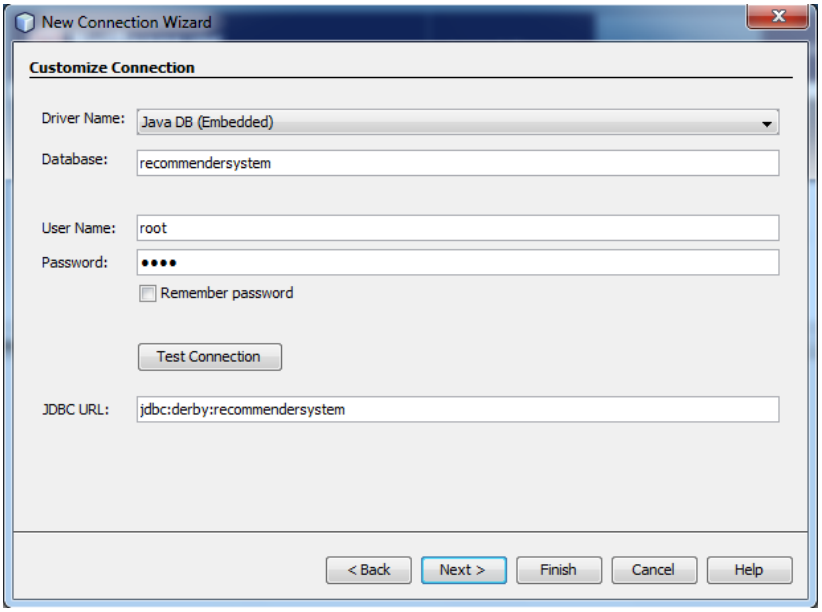

Figura 5.5: Configuração de uma base de dados MySQL no NetBeans.

Excerto de Código 5.1 *Script* de execução do serviço Kestrel.

```
1    @echo off set APP_NAME=kestrel set
2    ADMIN_PORT=2223 set VERSION=2.4.1 set SCALA_VERSION=2.9.2
3    set APP_HOME=%MY_BEANCOUNTER_PATH%\..\kestrel-2.4.1
4    set JAR_NAME=%APP_NAME%_%SCALA_VERSION%-%VERSION%.jar
5    set STAGE=production
6    set JAVA_OPTS=-server -Dstage=%STAGE%
7    echo launching kestrel service
8    java %JAVA_OPTS% -jar "%APP_HOME%\%JAR_NAME%"
```

Excerto de Código 5.2 *Script* de execução do serviço Elasticsearch.

```
1    @echo off
2    set VERSION=0.19.4
3    set EXEC=%MY_BEANCOUNTER_PATH%\..\elasticsearch-%VERSION%\bin\elasticsearch
4    set CONF=-Des.index.storage.type=memory echo starting elasticsearch
5    "%EXEC%" "%CONF%"
```

- `facebook-process`: Responsável pela actualização em tempo real das actividades do Facebook;
- `twitter`: Responsável pelo *update* em tempo real das actividades do Twitter;
- `resolver-process`: Responsável pela autenticação nas redes sociais. Após a autenticação, cria uma fila com todas as actividades de cada utilizador;
- `dispatcher`: Responsável pelo estabelecimento das ligações aos *endpoints*

Excerto de Código 5.3 *Script* de execução do serviço Redis.

```
1    @echo off
2    echo "starting redis"
3    cd %MY_BEANCOUNTER_PATH%\..\redis-2.4\bin
4    start redis-server.exe
5    cd %MY_BEANCOUNTER_PATH%\scripts
```

e encaminhamento dos dados para as respectivas filas;

- `realtime-profiler-process`: Responsável pela criação do perfil de cada utilizador agrupando as actividades existentes;
- `filter-process`: Responsável pela criação de filtros específicos para processar os perfis ou actividades do utilizador;
- `indexer`: Responsável pela interacção com o serviço de indexação Elasticsearch.

Estes processos, que são aplicações Java, são lançadas juntamente com os respectivos ficheiros de propriedades *guice* através da linha de comandos: `start java -jar xxxx-1.7.1-SNAPSHOT-jar-with-dependencies.jar -jndiProperties /guicejndi.properties`. Uma vez lançados todos os processos, os serviços RESTful disponibilizados pela plataforma ficam disponíveis. Logo após o *deploy* da plataforma efectuado pelo NetBeans, é apresentada uma página *Web* onde são fornecidas instruções sobre as funcionalidades do Beancounter (Figura 5.6).

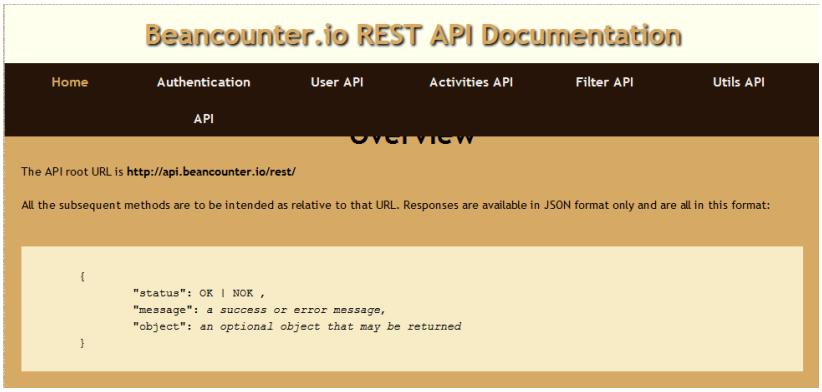

Figura 5.6: Página *Web* da documentação API Beancounter.

5.8.4 Editor de Ontologias

O *software* utilizado para a criação, edição e manipulação de ontologias foi o Protégé. Durante o desenvolvimento do trabalho foram usadas duas versões deste

editor de ontologias porque a versão mais recente não suporta o desenvolvimento de ontotogias no formato Protégé-Frames. Contudo, a ontologia acabou por ser desenvolvida no formato OWL na versão 4.2 do Protégé.

5.8.5 Resultado da Instalação

No final deste procedimento o ambiente no NetBeans deverá ficar como mostra a Figura 5.7.

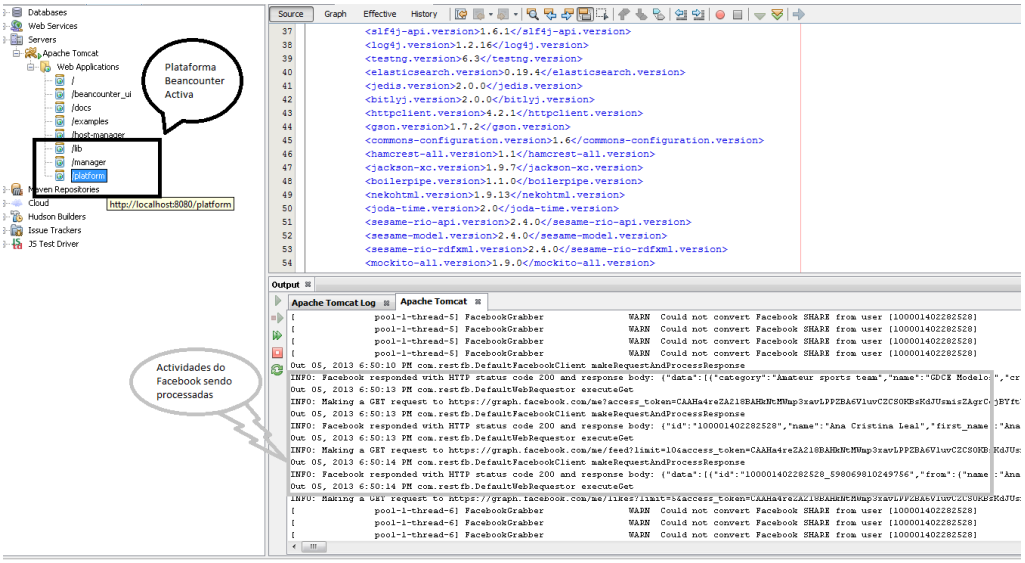

Figura 5.7: Ambiente de desenvolvimento instalado.

No final deste processo será efectuado o teste ao estado da API do Beancounter. Para isso faz-se uma chamada ao URL `http://localhost:8080/platform/rest/api/check` no navegador. A Figura 5.8 apresenta a resposta dada pela API em caso de sucesso, o que significa que todas as instalações e configurações, bem como os serviços correspondentes, foram efectuados com sucesso. Assim, todos os serviços do Beancounter estão prontos a ser consumidos.


```
{
    status: "OK",
    message: "system up and running at",
    object: "1381371806595"
}
```


Figura 5.8: Plataforma Beancounter activa.

5.9 Conclusão

Neste capítulo descreveram-se as tecnologias que vão ser utilizadas no desenvolvimento do Sistema de Recomendação. Adoptou-se maioritariamente serviços *Web* para a interacção com fontes de dados. Foram apresentadas todas as linguagens utilizadas para a programação (JAVA) ou para a representação dos dados (JSON). A representação do conhecimento será feita através de ontologias pelo que foi descrito o editor de ontologias utilizado, Prótégé. Todo o trabalho foi desenvolvido utilizando como base um IDE que foi configurado devidamente para suportar os diversos serviços. Por fim, foram ainda descritas as tecnologias que o Beancounter necessita. Neste capítulo fez-se ainda a descrição da instalação e configuração do ambiente de desenvolvimento. Todos os passos descritos, tanto na configuração do NetBeans como as tecnologias utilizadas pelo Beancounter, são necessários para que o sistema consiga ser compilado e executado com sucesso. Este ambiente de desenvolvimento apresenta como vantagens o uso de soluções *open source*, o ser independente da plataforma porque é baseada em ambientes, ferramentas e bibliotecas desenvolvidas em Java, e o assegurar um elevado nível de interoperabilidade através da adopção de interfaces do tipo serviço *Web*. Uma das dificuldades iniciais encontradas e que culmina numa desvantagem é o facto de ter sido adoptado para o desenvolvimento do sistema um ambiente Windows. Todo o Beancounter foi desenvolvido para Linux, pelo que foram desenvolvidos *scripts* alternativos para que a plataforma funcionasse em ambientes Windows. No capítulo seguinte serão descritos todos os algoritmos e interacções necessárias para a construção do Sistema de Recomendação.

Capítulo 6

Desenvolvimento do Sistema

Neste capítulo descreve-se o desenvolvimento do projecto, incluindo a arquitectura, a representação do conhecimento, os algoritmos, a interacção com as bibliotecas de interface utilizadas, os resultados obtidos e a interface gráfica.

6.1 Arquitectura

O sistema é constituído por diversos componentes que fornecem os seguintes serviços e funcionalidades: o serviço de construção do perfil social realizado pelo Beancounter, o serviço de criação de perfis globais, o serviço de recomendação de programas e a aplicação *Web* de interface com o utilizador. O Beancounter, cujas fontes de dados são o Facebook e o Twitter, é responsável por criar e manter actualizado o perfil social de cada utilizador.

O serviço de criação de perfis globais constrói o perfil global baseado nos conteúdos (pessoal e social) e no contexto associados a cada utilizador. Esta tarefa é suportada pelos dados do historial provenientes do YouTube e BBC, das classificações explícitas e do perfil social criado pelo Beancounter. A reutilização do perfil social criado pelo Beancounter recorre, no caso dos *shares* e *tweets*, aos repositórios da Freebase e IMDb para efectuar o necessário enriquecimento semântico. Este serviço recorre ainda ao mapeamento ontológico de conceitos efectuado para estabelecer a correspondências entre os diferentes sistemas de categorias utilizados pelo YouTube, Facebook, IMDb e a BBC.

O serviço de recomendação constrói uma grelha electrónica personalizada de programas com base no perfil global de cada utilizador, efectuando recomendações de programas da BBC utilizando a similaridade dos cossenos para determinar a semelhança entre os programas candidatos e o perfil do utilizador. Este ser-

viço armazena ainda as preferências explícitas efectuadas pelo utilizador após a recomendação.

Por fim, a aplicação de interface permite ao utilizador registar-se no serviço, solicitar recomendações e visualizar a grelha electrónica personalizada de programas construída pelo serviço de recomendação.

A Figura 6.1 apresenta a arquitectura global do sistema.

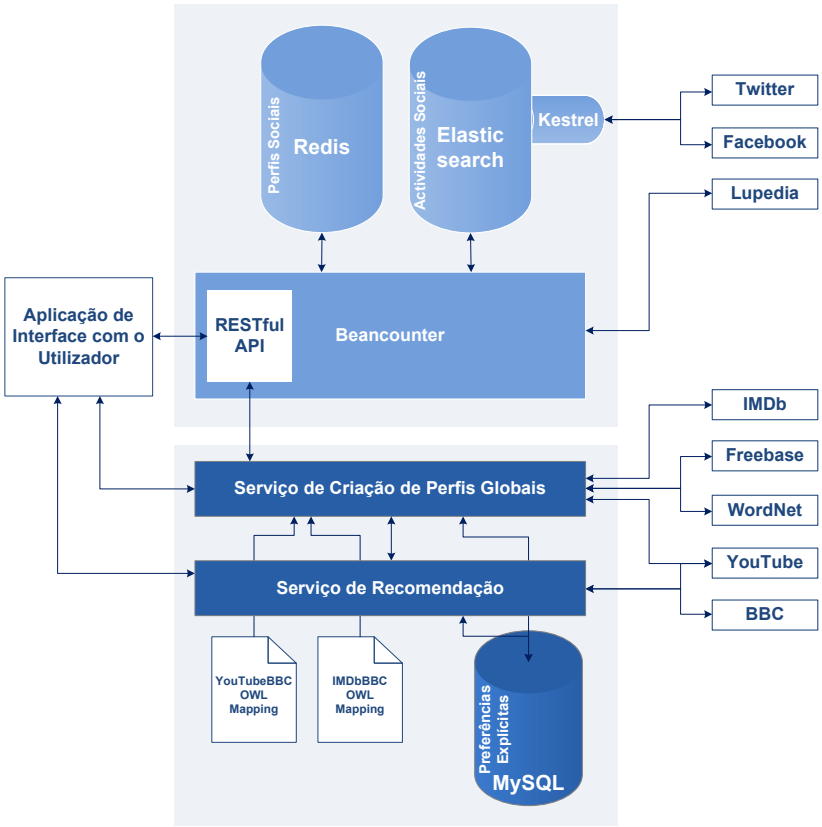

Figura 6.1: Arquitectura do Sistema.

6.2 Perfil Pessoal do Utilizador

O perfil pessoal do utilizador é baseado no seu comportamento que inclui o conjunto de vídeos vistos e as respectivas classificações atribuídas (YouTube) e o contexto espácio-temporal do utilizador (FreeGeoIP) (Figura 6.2). A Google disponibiliza uma biblioteca de interface, designada Google YouTube API, com diversos serviços incluindo o YouTube.

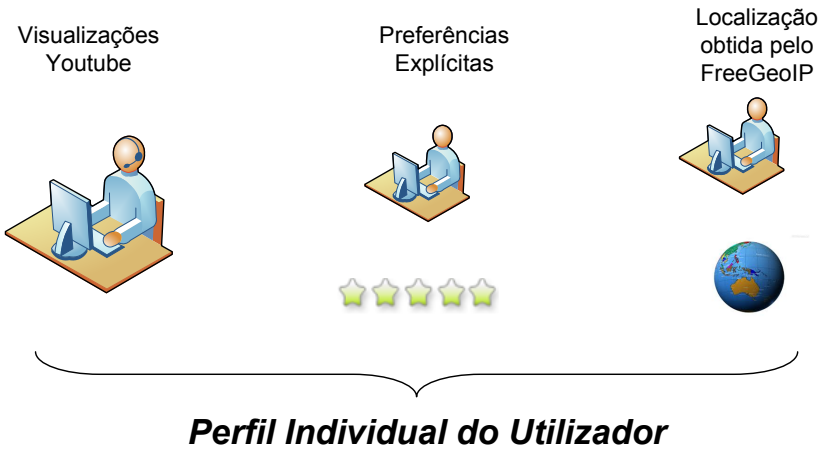

Figura 6.2: Componentes do Perfil Pessoal do Utilizador.

6.2.1 Interacção com o YouTube

A interacção com o YouTube implicou o registo e configuração da aplicação de interacção na API da Google (`https://developers.google.com/youtube/registering_an_application`). Esta biblioteca de interface fornece o acesso aos diversos serviços do YouTube. A aplicação registada foi denominada de aplicação de iteracção e foi configurado o acesso aos *YouTube Analytics API* e *YouTube Data API v3*, do YouTube. Seguiu-se a criação de um *client id*. Para que sejam criadas as chaves de acesso ao serviço esta configuração é realizada como mostra a Figura 6.3.

A API da Google fornece a possibilidade de *download* em formato JSON do ficheiro `clients_secrets.json` que contém os dados necessários para a interacção com o serviço e consequente troca de dados através do mecanismo de autenticação avançada `OAuth`. Este ficheiro é acedido pela classe Java desenvolvida (Excerto de código 6.1) para a interacção com esta API.

Excerto de Código 6.1 Leitura das chaves da API do ficheiro JSON.

```
1     GoogleClientSecrets clientSecrets = GoogleClientSecrets.load(JSON_FACTORY,
2               new InputStreamReader(YoutubeData.class.getResourceAsStream("client_secrets.json")));
```

Após este processo de autenticação e autorização, a aplicação pode interagir com o YouTube e aceder aos dados, utilizando para isso a fonte correspondente.

6.2.1.1 Biblioteca de Interface

O principal objectivo desta aplicação desenvolvida é a obtenção e a construção do historial de visualização do utilizador.

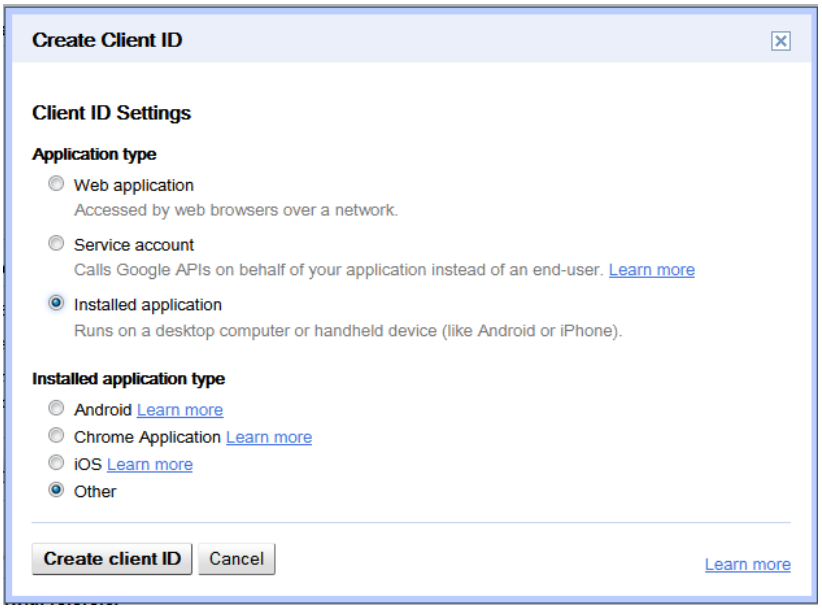

Figura 6.3: Configuração do *client id* na API da Google.

A biblioteca de interface utilizada disponibiliza operações que devolvem os seguintes dados:

- `Activities`- Informação sobre os canais e as actividades do utilizador no YouTube;
- `ChannelBanners` - URL do vídeo;
- `Channels` - Informação acerca de um determinado canal no YouTube;
- `GuideCategories` - Categorias que o YouTube atribui a um canal com base no conteúdo de outros indicadores, *e.g.*, a popularidade;
- `PlaylistItems` - Permite inserir, eliminar ou obter os vídeos de uma dada lista de reprodução;
- `Playlists` - Informação sobre uma lista de reprodução do YouTube, *i. e.*, a colecção de vídeos que podem ser vistos em sequência ou compartilhados com outros utilizadores;
- `Search` - Informação sobre o vídeo, canal, ou lista de reprodução especificado nos parâmetros de pesquisa de um pedido à API;
- `Subscriptions` - Informação sobre actualizações, quando um outro utilizador actua sobre um vídeo, *i. e.*, classificando-o ou apenas comentando, ou adição de vídeos;

- `Thumbnails` - Dimensão de uma miniatura (imagem);
- `VideoCategories` - Categorias associadas a um vídeo;
- `Videos` - Informação sobre um vídeo, incluindo o *rating* associado; Possibilita a inserção ou a eliminação de um vídeo;
- `Watermarks` -Imagem que é exibida durante a reprodução de vídeos de um canal específico [72].

Este projecto invoca as fontes/operações relativas aos *Channels*, *PlaylistItems* e *VideoCategories* para obter os dados necessários à construção do perfil do utilizador.

A informação devolvida pelo YouTube baseia-se na lista dos últimos 50 vídeos visualizados pelo utilizador. Assim, tem-se a informação do nome do vídeo e da categoria em que este se integra.

6.2.2 Interacção com o *Freegeoip*

A informação do contexto espacial do utilizador faz parte do perfil perfil pessoal do utilizador. Esta informação permite efectuar recomendações associadas à localização do utilizador. Por exemplo, no YouTube existem diversos canais associados a países evitando que sejam efectuadas recomendações de vídeos cuja visualização não é autorizada no país do utilizador.

Para obter esta informação é utilizado o serviço *Web* RESTful *freegeoip*. O pedido HTTP é efectuado a `http://www.freegeoip.net/json/` e como resposta obtém-se a localização da plataforma do utilizador. Este serviço obtém a geolocalização através dos endereços IP e nomes de *hosts*. As respostas são obtidas em formato JSON, como se pode verificar na Figura 6.4.


```
{
    ip: "188.80.69.229",
    country_code: "PT",
    country_name: "Portugal",
    region_code: "17",
    region_name: "Porto",
    city: "Paços De Ferreira",
    zipcode: "",
    latitude: 41.2766,
    longitude: -8.3762,
    metro_code: "",
    areacode: ""
}
```


Figura 6.4: Contexto espacial do utilizador.

Efectuado o processamento da resposta JSON devolvida, a informação da geolocalização do utilizador é incluída no Sistema de Recomendação.

6.2.3 Preferências Explícitas

As preferências explícitas do utilizador são contempladas no perfil pessoal do utilizador. No YouTube o utilizador pode manifestar a suas preferências colocando um "gosto". O *feedback* do utilizador face às recomendações efectuadas é essencial em qualquer Sistema de Recomendação. O utilizador efectua uma avaliação da recomendação proposta através de uma escala de 1 a 5. A Figura 6.5 apresenta um exemplo da interface utilizada para efectuar esta avaliação.

Figura 6.5: Preferências explícitas.

Esta funcionalidade foi implementada recorrendo a um *plugin* Jquery chamado de RateIt que permite criar as estrelas de avaliação. No Excerto de Código 6.2 está demonstrada a construção das estrelas de *rating*.

Excerto de Código 6.2 Construção da preferências explícitas.

```
1           <div productid="FamilyGuy" class="rateit"></div>
```

As preferências são armazenadas no serviço de base de dados instalado, sendo associado o título do programa à classificação efectuada pelo utilizador. Para efeitos de melhoramento da recomendação o sistema verifica se a base de dados contém preferências explícitas e procura os programas com maiores classificações verificando a categoria do programa.

#	idUser	rating	PreferencesName
1	FatimaLeal	5.0	The Choir - Sing While You Work: Series 2, Episode 1
2	FatimaLeal	4.0	Family Guy - Series 8, Dog Gone
3	FatimaLeal	4.5	Some Girls - Series 2, Episode 6
4	FatimaLeal	2.0	Citizen Khan - Series 2, Shazia's Gym Visit

Figura 6.6: Armazenamento de preferências explícitas de programas.

As categorias das preferências retiradas do YouTube e das preferências dos programas são somadas para posteriormente ser criado o vector de características.

6.3 Perfil Social do Utilizador

O perfil social do utilizador é construído com base nas actividades sociais do utilizador. Estas actividades são provenientes das redes sociais Facebook e Twitter e são processadas pela plataforma Beancounter (6.7). Para a troca de dados entre estas redes sociais e o Beancounter foi necessário o registo de uma aplicação em ambas as redes. O Beancounter disponibiliza uma biblioteca de interface que acede, não só, às actividades provenientes das redes sociais, mas também aos perfis criados.

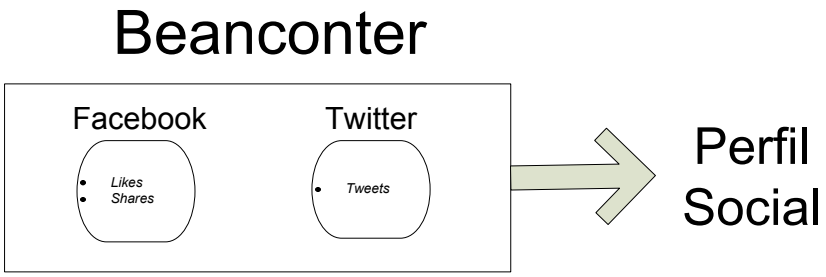

Figura 6.7: Componentes do Perfil Social.

6.3.1 Interacção com o Facebook

A aplicação Beancounter foi registada no Facebook permitindo aceder a todas as actividades, *i.e.*, *likes* e *shares*. Este registo e configuração é efectuado acedendo ao sítio do Facebook ou ao URL `https://developers.facebook.com/apps`. A Figura 6.8 ilustra esta operação. O *App ID* e *App Secret* são os dados de autenticação e validação.

Uma das funcionalidades do Facebook é permitir a troca de dados em tempo real, *i.e.*, sempre que um utilizador insere uma nova actividade é realizado de imediato o *update* dessa actividade nas aplicações registadas e configuradas. Esta requer a realização de uma configuração adicional que consiste na introdução de um endereço de *callback*. Este endereço será do *listener* das actividades do Facebook. Este endereço tem que ser forçosamente um endereço público, para que possa ser estabelecida a comunicação entre o Facebook e a aplicação que recebe as actualizações. A Figura 6.9 mostra a configuração desta funcionalidade para o Beancounter. Foi utilizada uma aplicação gratuita designada Free Dynamic DNS[1] que permite a configuração de um domínio para ultrapassar o problema da

[1] `http://www.noip.com/`

Figura 6.8: Configuração da aplicação Beancounter no Facebook.

mudança de IP que as operadoras de serviço de Internet efectuam frequentemente nos *routers* domésticos.

Figura 6.9: Configuração do *Real Time Update*.

6.3.1.1 Biblioteca de Interface

A Graph API do Facebook é uma biblioteca de interface que suporta pedidos HTTP GET e POST. Esta biblioteca dá acesso ao gráfico social do Facebook que é representado por objectos. Os objectos de seguida apresentados são os mais

utilizados para aceder às actividade do utilizador no Facebook:

- `Comment`;
- `Event`;
- `Message`;
- `Page`;
- `Photo`;
- `Video`;
- `User`.

A *Facebook Query Language* (FQL) é uma linguagem de interrogação do Facebook. Existem duas bibliotecas de interacção com o Facebook desenvolvidas em Java: a RestFB e a facebook-java-api. A biblioteca de interface utilizada pelo Beancounter é a RestFB[2] que contém todas as classes necessárias para a troca de informação, incluindo o envio de *queries* FQL e a recepção das respostas no formato JSON. As actividades como a publicação de mensagens, eventos ou fotos também são possíveis através desta biblioteca.

6.3.2 Interacção com o Twitter

Para a interacção com o Twitter foi necessário adicionar e configurar uma aplicação de interacção no Twitter. Este processo é efectuado no sítio `https://dev.twitter.com/`. A aplicação registada foi denominada de `BeanconterSystem` (6.10). A configuração da aplicação é bastante simples, sendo apenas necessária a introdução da designação da aplicação, uma pequena descrição, o URL do sítio para onde serão enviados os dados do Twitter e a selecção da opção *Read, Write and Access direct messages*. No final deste procedimento são gerados quatro tipos de chaves: (*i*) a *consumer key*; (*ii*) a *consumer secret*; (*iii*) o *access token*; e (*iv*) o *access token secret*. Estas chaves permitem efectuar a autenticação e validação da aplicação.

6.3.2.1 Biblioteca de Interface

O Twitter disponibiliza na sua REST API (versão 1.1) diversas fontes de dados que descrevem páginas, utilizadores ou *tweets*. Algumas das fontes disponibilizadas são:

- `Tweets`;

[2] `http://restfb.com/`

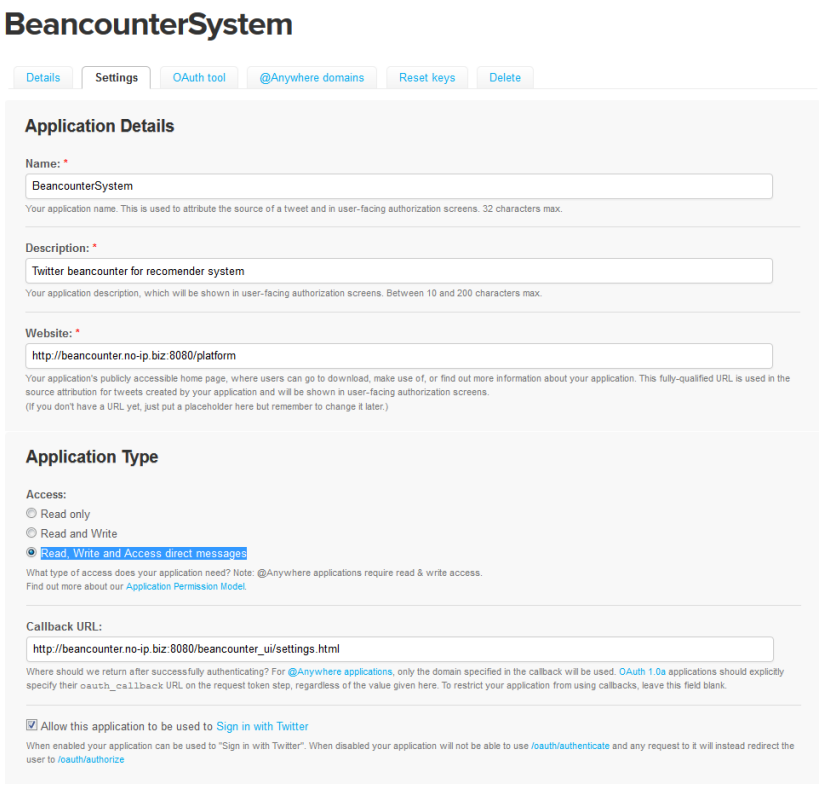

Figura 6.10: Configuração da aplicação Beancounter no Twitter.

- `Search`;
- `Streaming`;
- `Direct Messages`;
- `Friends & Followers`;
- `Users`;
- `Suggested Users`;

O Beancounter através da interacção com estas fontes extrai os *tweets* que o utilizador publica no seu perfil do Twitter.

Para o *update* em tempo real o Twitter disponibiliza a `Streaming API`. Esta API requer o estabelecimento de uma conexão HTTP persistente para que, sempre que o utilizador insira uma nova actividade, esta seja comunicada via HTTP para o cliente desenvolvido. A biblioteca de interface utilizada é o `twitter4J` desenvolvida em Java.

6.3.3 Interacção com Beancounter

O Beancounter foi integralmente desenvolvido em Java e está disponível no repositório de código *online* Github [73]. Este código disponível no Github estava adaptado para a estação de televisão italiana Rai.Tv, pelo que foi necessário proceder a diversas alterações, designadamente, a substituição do serviço de enriquecimento semântico da estação televisiva italiana pelo serviço de enriquecimento Lupedia. Esta mudança foi efectuada no módulo `realtime-profiler`, conforme se apresenta no Excerto de Código 6.3.

Excerto de Código 6.3 Alteração do serviço de enriquecimento de texto no Beancounter.

```
1    private NLPEngine initCogito() {
2    //return new CogitoNLPEngineImpl("http://test.expertsystem.it/IPTC_ITA/EssexWS.asmx/ESSEXIndexdata");
3          return new LUpediaNLPEngineImpl();
4       }
```

A equipa do Beancounter desenvolveu diferentes soluções suportadas por tecnologias alternativas. O código está disponível, cabendo ao utilizador escolher as tecnologias de suporte e proceder às respectivas alterações. Dado que o Beancounter agrega as actividades do Facebook e do Twitter, as aplicações de interface com estas redes sociais devem ser previamente criadas e configuradas. As chaves de autenticação destas aplicações encontram-se armazenadas no ficheiro `beancounter.properties` da plataforma Beancounter conforme a Figura 6.11. Adicionalmente, de suporte ao Beancounter, o Kestrel, o Elasticsearch e o Redis devem ser lançados para permitir a compilação da plataforma.

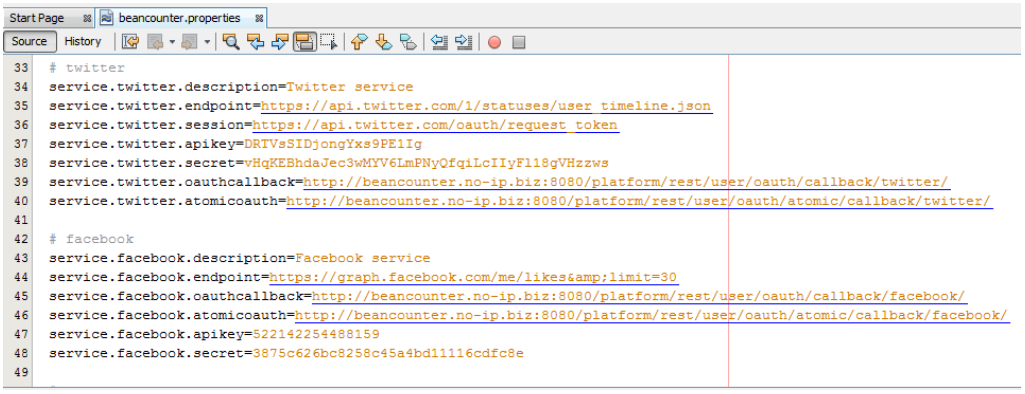

Figura 6.11: Configuração das propriedades do Beancounter.

Concluída a compilação da plataforma são lançadas as aplicações de interacção RESTful do Facebook e do Twitter. Utilizou-se a ferramenta de linha de comando cURL para transferir dados no protocolo HTTP, para definir o nome da

aplicação de interface, uma pequena descrição, endereço de *email* e um endereço de *callback* de acordo com o Excerto de Código 6.4 apresentado.

Excerto de Código 6.4 Pedido de registo da aplicação.

```
curl -d "name=Beancounterdescription=RecommenderSystem&email=
        fatimaleal2@gmail.com&oauthCallback=http://localhost:8080/platform"
        localhost:8080/platform/rest/application/register
```

A resposta contém a chave que deverá ser usada para aceder aos serviços RESTful. Esta chave é representada pelo *object* da Figura 6.12.

Figura 6.12: Resposta do registo da aplicação.

6.3.3.1 Biblioteca de Interface

O Beancounter é, por sua vez, construído por um conjunto de serviços *Web* RESTful, sendo a interacção efectuada através de pedidos HTTP (GET e POST) e as respostas devolvidas no formato JSON. Os serviços oferecidos estão enumerados e descritos na Tabela 6.1.

A invocação dos diferentes *endpoints* apresentados na Tabela 6.1 permitem o armazenamento, indexação e processamento dos dados sociais dos utilizadores. Os utilizadores registam-se no Beancounter através do serviço `ApplicationService` e associam as respectivas fontes sociais através do serviço `UserService`, após a autenticação através do `getOAuthToken()`. A consulta das actividades realizadas pelos utilizadores nas redes sociais associadas é efectuada através do serviço `ActivitiesService`, utilizando a operação `getallActivity()`. Este serviço (`ActivitiesService`) permite obter informações acerca de uma determinada actividade através do seu *id* assim como adicionar actividades nas redes sociais, *e.g.*, publicar no Twitter um *tweet* via Beancounter. O acesso ao perfil social do utilizador é realizado através do serviço `UserService` utilizando o método `getProfile()`. Os demais serviços do Beancounter são de filtragem (`FilterService`) e análise (`AnalysisService`) de dados sociais e teste (`AliveService`) da plataforma.

6.3.3.2 Respostas do Beancounter

Os resultados do Beancounter são fornecidos em formato JSON especificando, o tipo de actividade (*likes*, *shares* ou *tweets*) através do `verb`.

Na Figura 6.13 apresenta-se a informação em formato JSON relativa a um *share* efectuado pelo utilizador no Facebook. A descrição da actividade é com-

Tabela 6.1: Serviços RESTful do Beancounter.

Serviços RESTful	Operações	*Endpoint*
ActivitiesService	`addActivity();`	`/rest/activities/add/username`
	`getActivity();`	`/rest/activites/activityId`
	`getallActivity();`	`/rest/activities/all/username`
	`getUserActivity();`	`/rest/activities/username/activityId`
	`search();`	`/rest/activities/search`
	`searchWithToken();`	`/rest/activities/search/me`
	`setVisibility();`	`/rest/activities/activityId/visible/isVisible`
AliveService	`signUp();`	`/rest/api/check`
	`version();`	`/rest/api/version`
AnalysisService	`analysis();`	`/rest/analysis/list/all`
	`restart();`	`/rest/analysis/analysisId/restart`
	`result();`	`/rest/analysis/analysisId/result`
ApplicationService	`deregisterApplication();`	`/rest/application/apiKey`
	`register();`	`/rest/aplication/register`
FilterService	`delete();`	`/rest/filters/name`
	`filters();`	`/rest/filters/list/all`
	`get();`	`/rest/filters/name`
	`register();`	`/rest/filters/register/name`
	`start();`	`/rest/filters/name/start`
	`stop();`	`/rest/filters/name/stop`
UserService	`authenticate();`	`/rest/user/username/authenticate`
	`checkToken();`	`/rest/user/username/service/check`
	`deleteUser();`	`/rest/user/username`
	`getOAuthToken();`	`/rest/user/oauth/token/service/username`
	`getProfile();`	`/rest/user/username/profile`
	`getUserWithApiKey();`	`/rest/user/username`
	`getUserWithUserToken();`	`/rest/user/username/me`
	`handleOAuthCallback();`	`/rest/user/oauth/callback/service/username`
	`removeSource();`	`/rest/user/source/username/service`
	`signUp();`	`/rest/user/register`

posta pelas propriedades e contexto da partilha. As propriedades incluem o URL que direcciona para a partilha efectuada, o nome que identifica a partilha no Facebook e a descrição da actividade que contém a parte inicial do texto partilhado. O contexto da actividade é definido pela data, pelo serviço de origem da actividade e o *username* do utilizador nesse serviço. As fotos partilhadas, bem como as novas amizades do utilizador, são consideradas pelo Beancounter como partilhas (*shares*). O texto da descrição da partilha é submetido à Lupedia para a identificação e enriquecimento do conteúdo relevante para a construção do perfil do utilizador.


```
  userId: "ffd1aeee-1e71-4dd7-a0cb-a7528a8fab00",
- activity: {
    id: "cefdf8ef-2525-4b90-a23f-08f1a7b4beed",
    verb: "SHARE",
  - object: {
        type: "OBJECT",
        url: http://www.ionline.pt/iopiniao/dom-ser-dom,
        name: "O dom de ser dom",
        description: "A Filosofia congrega todos quantos procuram
        construir ou descobrir respostas às questões profundas da
        existência. Não é mais que um ponto de encontro e partilha, pois
        que cada homem deve decidir, de forma pessoal, o seu rumo e
        destino, fazer o caminho que há de depois pisar e... cumprir.
        Cu..."
    },
  - context: {
        date: 1382183230000,
        service: "facebook",
        username: "100000721681645"
    }
},
```


Figura 6.13: Representação do *Share* no Beancounter.

A informação correspondente aos *likes* difere ligeiramente da dos *shares* (6.14). Os *likes* devolvem as categorias das páginas de que o utilizador gostou. Estas categorias são extraídas para, posteriormente, serem processadas e incluídas no perfil do utilizador.

A estrutura da informação dos *tweets* não é muito diferente, incluindo as características e o contexto do *tweet*. A resposta JSON contém o URL do *tweet*, o conteúdo identificado através do parâmetro `text`, as *hashtags* que o *tweet* contém e, no caso de um *retweet*, inclui ainda o URL da fonte. Na Figura 6.15 encontra-se um exemplo de um *tweet*. O conteúdo ou texto do *tweet*, tal como no *share*, é submetido ao serviço de enriquecimento semântico Lupedia para identificação do conteúdo relevante.

O Beancounter constrói o perfil social do utilizador baseado nos resultados do processamento dos três tipos de actividades apresentadas. A Figura 6.16 apresenta um exemplo de um eventual perfil de utilizador.

O perfil social criado pelo Beancounter contém os seguintes dados: o URL do

```
- {
    userId: "b194318b-4164-4a5b-a5e4-42d4e635bc08",
  - activity: {
      id: "640eff7f-c52e-4e9c-ab46-366d1cda7b6b",
      verb: "LIKE",
    - object: {
        type: "FB-LIKE",
        url: http://www.facebook.com/29708230543,
        name: "Marco Frisina",
      - categories: [
          "band",
          "Musician"
        ]
      },
    - context: {
        date: 1380728711000,
        service: "facebook",
        username: "100000721681645"
      }
    },
```


Figura 6.14: Representação do *Like* no Beancounter.


```
  userId: "72045f5e-98b4-4f00-a6d7-ecc7ed42dab1",
- activity: {
    id: "259a7970-a55e-43d3-8549-80cc730b54ca",
    verb: "TWEET",
  - object: {
      type: "TWEET",
      url: "http://twitter.com/Fátima Leal/status/385863186827579392",
      text: "RT @JornalNoticias: Mulher encontrada viva entre vítimas
      mortais de naufrágio em Lampedusa http://t.co/xsBFaf9NsL Ler
      http://t.co/3JKqHPPd1f",
      hashTags: [ ],
    - urls: [
        http://dlvr.it/44c837,
        http://www.JN.pt
      ]
    },
  - context: {
      date: 1380831927000,
      service: "twitter",
      username: "1374347228"
    }
},
```


Figura 6.15: Representação do *Tweet* no Beancounter.

recurso da DBpedia que apresenta a descrição do interesse identificado, a *label* obtida pelo serviço de enriquecimento de texto, o peso do interesse, o número de actividades associadas e os respectivos *ids*.

Em termos de recomendação, os dados relevantes a extrair desta componente do perfil do utilizador são: as *labels*, que identificam os interesses, os pesos e as categorias provenientes dos *likes* do Facebook. Como os *likes* do Facebook já estão classificados em categorias não necessitam de ser processados por qualquer espécie de serviço de enriquecimento, sendo enviados directamente para a fase de

```
{
   status: "OK",
   message: "profile for user [fatimaleal] found",
 - object: {
      visibility: "PUBLIC",
      userId: "b194318b-4164-4a5b-a5e4-42d4e635bc08",
    - interests: [
       - {
            resource: http://dbpedia.org/resource/Nogueira_%28Braga%29,
            label: "Nogueira_(Braga)",
            weight: 5.0892161848066716e-8,
            visible: true,
            activitiesSize: 5,
          - activities: [
               "cb48948a-3cdf-4aa5-a43a-b45b15d4df14",
               "e72a74d1-64f1-41fc-bf05-d4fb5463e70d",
               "4c536d07-50c3-4391-a36b-fda778078563",
               "c336fda9-9e26-4cd2-a4f7-7fc2b4dc4b48",
               "cfd1740e-afb0-4ce9-a1fe-3a6ed639c2ea"
            ]
         },
       - {
            resource: http://dbpedia.org/resource/Pedro_%28film%29,
            label: "Pedro_(film)",
            weight: 5.0892161848066716e-8,
            visible: true,
            activitiesSize: 5,
          - activities: [
               "cb48948a-3cdf-4aa5-a43a-b45b15d4df14",
               "e72a74d1-64f1-41fc-bf05-d4fb5463e70d",
               "4c536d07-50c3-4391-a36b-fda778078563",
               "c336fda9-9e26-4cd2-a4f7-7fc2b4dc4b48",
               "cfd1740e-afb0-4ce9-a1fe-3a6ed639c2ea"
            ]
         },
       - {
```


Figura 6.16: Perfil social construído pelo Beancounter.

mapeamento das categorias. As *labels* são processadas pelos serviços de enriquecimento referidos para serem extraídas as respectivas categorias. O mapeamento entre sistemas de classificação é efectuado para converter todas as classificações para a taxonomia de categorias da BBC. O peso representa a relevância que o utilizador atribui a cada interesse.

6.4 Enriquecimento Semântico

O enriquecimento semântico é efectuado através da interacção com a Freebase, a IMDb e a WordNet para extrair informação semântica complementar. O conjunto de dados do perfil social a ser enriquecido é constituído pelas *labels* dos interesses e pelas categorias dos *likes*. O conjunto de dados do perfil individual que vai ser enriquecido é constituído pelas categorias dos vídeos visualizados no YouTube.

6.4.1 Interacção com a Freebase

A interacção com a Freebase assemelha-se à do YouTube dado que a interacção com ambos os serviços é suportada pela mesma biblioteca de interface da Google. A aplicação de interacção com a Freebase reutiliza o código de interacção com o YouTube e partilha a chave de segurança.

6.4.1.1 Biblioteca de Interface

A Freebase apresenta uma colecção de operações RESTful para leitura e escrita dos dados. A informação da Freebase está organizada em *Graphs*, *Topics*, *Types and Properties*, *Domains and IDs*, *Compound Value Types*, *Topic MIDs*, *Namespaces, Keys, and Topic IDs* e *More on Properties*.

A estrutura de armazenamento de dados é constituída por grafos (*Graphs*). Este esquema permite inserir novos temas sem ser necessário mudar a estrutura dos dados, criando apenas uma conexão arbitrária para um novo nó. Os nós, por sua vez, representam os tópicos (*Topics*). A Freebase possui mais de 39 milhões de tópicos relacionados com pessoas, lugares e coisas. Os tópicos são descritos através de tipos e propriedades (*Types and Properties*). Por exemplo, "Bob Dylan é um compositor, cantor, artista, escritor e actor" significa que a entidade "Bob Dylan" é um tópico descrito através das propriedades "compositor", "cantor", "artista", "escritor" e "actor".

A informação é também armazenada por domínios (*Domains and IDs*), correspondendo a cada domínio um identificador. A título ilustrativo listam-se alguns identificadores:

- `/business` é o identificador do domínio negócios;
- `/music` é o identificador do domínio música;
- `/film` é o identificador do domínio filmes;
- `/medicine` é o identificador do domínio medicina.

Os identificadores dos domínios facilitam o acesso à informação.

A Freebase permite a representação de dados compostos (*Compound Value Types*), *i.e.*, dados que são frequentemente alterados. Por exemplo, se a população de uma cidade for representada através deste tipo de dados, é possível obter a evolução da população da cidade ao longo de um período de tempo. Este acesso pode ser efectuado através do *Topic MIDs* que é constituído pelo selector `/m/` e um identificador que referencia o conceito a aceder.

A informação pode ser acedida através do conjunto *Namespaces, Keys, and Topic IDs*, *e.g.*, `/business/company/idTopic`. O *namespace* funciona como uma

pasta, *i.e.*, `/business` representa o *namespace* do domínio "negócios" e engloba toda a informação nele contida. A *Key* corresponde ao nome do ficheiro, *i.e.*, *company* é o nome de uma chave do *namespace* `/business`. O *Topic IDs* representa o identificador específico da informação, *i.e.*, *idTopic* é a referência do item de informação pretendido. Por exemplo, o pedido `business/company&key=AIzaSyCuyi5tqxJf6nRC67c2DsYSNg1_BrKYJuU` retorna a resposta em formato JSON da Figura 6.17. O par `key=AIzaSyCuyi5tqxJf6nRC67c2DsYSNg1_BrKYJuU` corresponde à chave da API do utilizador. O parâmetro `/search` do URL completo[3] devolve todos os resultados da `business/company`. Alternativamente,


```
{
   status: "200 OK",
 - result: [
    - {
         mid: "/m/09s1f",
         name: "Business",
       - notable: {
            name: "TV Genre",
            id: "/tv/tv_genre"
         },
         lang: "en",
         score: 389.702637
      },
    - {
         mid: "/m/0k989",
         id: "/en/commercial_law",
         name: "Commercial law",
       - notable: {
            name: "Field Of Study",
            id: "/education/field_of_study"
         },
         lang: "en",
         score: 313.105194
      },
    - {
         mid: "/m/012t_z",
         id: "/en/businessperson",
         name: "Businessperson",
       - notable: {
            name: "Profession",
            id: "/people/profession"
         },
         lang: "en",
         score: 255.759552
      },
```


Figura 6.17: Resposta da Freebase a uma *query* com *namespace* especificado.

pode-se obter a lista dos géneros dos programas televisivos através do pedido `tv/tv_genre&key=AIzaSyCuyi5tqxJf6nRC67c2DsYSNg1_BrKYJuU`[4] (Figura 6.18). As respostas da Freebase incluem o *MID* e o *ID* das instâncias retornadas, permitindo efectuar pesquisas mais específicas como se ilustra na Figura 6.17.

[3] O URL completo do pedido HTTP é `https://www.googleapis.com/freebase/v1/search?query=business/company&key=AIzaSyCuyi5tqxJf6nRC67c2DsYSNg1_BrKYJuU`

[4] O URL completo do pedido HTTP é `https://www.googleapis.com/freebase/v1/search?query=tv/tv_genre&key=AIzaSyCuyi5tqxJf6nRC67c2DsYSNg1_BrKYJuU`

```
- {
      mid: "/m/03d1v8",
      id: "/en/amc",
      name: "AMC",
    - notable: {
          name: "TV Network",
          id: "/tv/tv_network"
      },
      lang: "en",
      score: 136.218018
  },
- {
      mid: "/m/049tj1",
      id: "/en/the_incredible_hulk",
      name: "The Incredible Hulk",
    - notable: {
          name: "Science Fiction TV Program",
          id: "/m/06n90"
      },
      lang: "en",
      score: 98.499031
  },
- {
      mid: "/m/03c290",
      id: "/en/tv_land",
      name: "TV Land",
    - notable: {
          name: "TV Network",
          id: "/tv/tv_network"
      },
      lang: "en",
      score: 96.653671
  },
```


Figura 6.18: Resposta da Freebase a uma *query* com *ID* especificado.

Para a interligação com outros repositórios abertos e interligados da nuvem LOD, utilizam-se os parâmetros *topic* e *MID*. Por exemplo, no exemplo apresentado sobre *TV Genres* o *MID* é `m/09s1f`. Pode-se, então, efectuar um pedido específico do tipo `topic/m/09s1f?filter=/common/topic/topic_equivalent_webpage&limit=100`[5]. A resposta devolve todos os URI de interligação com outros objectos da nuvem LOD, permitindo o enriquecimento semântico.

Neste projecto a Freebase é utilizada para obter o URI IMDb dos filmes, séries ou celebridades que serão, posteriormente, enriquecidos através da biblioteca de interface utilizada para aceder aos dados da IMDb. Neste contexto, os dados que são processados provêm das *labels* do Beancounter que incluem o URI directo para a DBpedia. Recorre-se, então, à Freebase para aceder a informação complementar da IMDb e extrair os géneros das séries e filmes associadas às *labels* do Beancounter. Alternativamente, podia-se obter a mesma informação via DBpedia, recorrendo a *queries* federadas SPARQL.

[5]O URL completo do pedido HTTP é `https://www.googleapis.com/freebase/v1/topic/m/09s1f?filter=/common/topic/topic_equivalent_webpage&limit=100`

```
  id: "/m/09s1f",
- property: {
    - /common/topic/topic_equivalent_webpage: {
        valuetype: "uri",
      - values: [
          - {
              text: http://es.wikipedia.org/wiki/index.html?curid=235245,
              lang: "",
              value: http://es.wikipedia.org/wiki/index.html?curid=235245,
              creator: "/user/wikipedia_intl",
              timestamp: "2010-09-24T07:37:12.000Z"
            },
          - {
              text: http://ja.wikipedia.org/wiki/index.html?curid=24060,
              lang: "",
              value: http://ja.wikipedia.org/wiki/index.html?curid=24060,
              creator: "/user/wikipedia_intl",
              timestamp: "2010-09-25T22:29:11.000Z"
            },
          - {
              text: http://ru.wikipedia.org/wiki/index.html?curid=17715,
              lang: "",
              value: http://ru.wikipedia.org/wiki/index.html?curid=17715,
              creator: "/user/wikipedia_intl",
              timestamp: "2010-09-27T01:39:15.000Z"
            },
```


Figura 6.19: Resposta da Freebase a uma *query* com tópico especificado.

6.4.2 Interacção com a IMDb

A IMDb disponibiliza os seus dados através de ficheiros de texto. Os ficheiros disponibilizados estão divididos por temas e são de grande dimensão [74].

Dado que este tipo de suporte não é adequado para a interacção automática, vários investigadores desenvolveram interfaces do tipo serviços REST para os dados da IMDb. Alguns exemplos destas interfaces são a IMDB API de Dean Clatworthy (`http://deanclatworthy.com/imdb/`), a OMDB API de Brian Fritz (`http://www.omdbapi.com/`) e a My Movie API (`http://mymovieapi.com/`). Neste projecto utilizou-se a OMDB API que não necessita de registo prévio para a utilização. Os dados fornecidos consistem num aglomerado de diferentes bibliotecas de interface:

- Bing API: Utilizada para pesquisa de filmes;
- Rotten Tomatoes API: Utilizada para obter dados de um filme e respectivas classificações;
- Freebase API: Utilizada para obter dados de um filme;
- Wikipedia.org API: Utilizada para obter dados de um filme
- TheMovieDb.org API: Utilizada para obter dados de um filme

Os dados fornecidos pela OMDB API são fiáveis dada a lista de fontes utilizada. Foram ainda efectuados testes para verificar se os dados fornecidos pela OMDB API coincidiam com os da IMBb, tendo os dados, nos casos testados, coincidido integralmente. No entanto, caso ocorra uma alteração de políticas de disponibilização de dados de qualquer uma das empresas envolvidas, as bibliotecas de interface referidas podem ser, como cada empresa relembra, desligadas.

6.4.2.1 Biblioteca de Interface

O acesso aos dados é efectuado pela construção de um URL baseado nos parâmetros da Tabela 6.2. Com estes parâmetros é possível a construção do URL

Tabela 6.2: Descrição dos parâmetros suportados pela API OMDB.

Parâmetro	Valor	Descrição
i	String	Um *id* IMDb válido
t	String	Título do filme ou série a ser retornado
y	Year	Ano do filme ou série
r	JSON, XML	Formato dos dados (por omissão os dados são devolvidos no formato JSON)
plot	short, full	Descrição retorna curta, ou mais completa (por omissão a descrição é devolvida na forma curta)
callbalck	name	Nome da função JSONP
tomatoes	true	Adiciona dados do Rotten Tomatoes

e o acesso aos dados. Como exemplo, apresenta-se o pedido da descrição do filme que tem como título *The Mission* em que se pretende que os dados sejam devolvidos no formato XML. Então o resultado será `http://www.omdbapi.com/?t=TheMission&r=xml`, que apresenta como resposta a descrição do filme, como consta na Figura 6.20.


```
− <root response="True">
    <movie title="The Mission" year="1986" rated="PG" released="31 Oct 1986" runtime="2 h 5 min"
    genre="Adventure, Drama, History" director="Roland Joffé" writer="Robert Bolt" actors="Robert
    De Niro, Jeremy Irons, Ray McAnally, Aidan Quinn" plot="18th century Spanish Jesuits try to protect
    a remote South American Indian tribe in danger of falling under the rule of pro-slavery Portugal."
    poster="http://ia.media-imdb.com/images
    /M/MV5BMjA5MDgxOTI1NV5BMl5BanBnXkFtZTYwOTkxNjk5._V1_SX300.jpg"
    imdbRating="7.4" imdbVotes="34,284" imdbID="tt0091530" type="movie"/>
  </root>
```


Figura 6.20: Detalhes do filme “The Mission”.

Com este conhecimento é possível o enriquecimento de todos os filmes e séries que no mundo da multimédia tem um peso relevante. A utilização desta API para o enriquecimento dos dados que contêm o nome de actores ou realizadores pode

ser interessante, dado que a API tem ligação a outras fontes de enriquecimento. Sendo possível uma descrição longa do pedido efectuado, a interligação à API do Rotten Tomatoes para a pesquisa sobre actores ou realizadores é bastante completa. Neste trabalho a API é utilizada para obter os géneros de filmes e séries. Estes géneros são armazenados e processados para posterior mapeamento pelos géneros da BBC e a partir daí ser realizada a recomendação.

6.4.3 Interacção com a WordNet

A WordNet é uma base de dados lexical de Inglês onde substantivos, verbos, adjectivos e advérbios são agrupados em conjuntos de sinónimos cognitivos (*synsets*) relativos a conceitos distintos. Os *synsets* estão interligadas através de relações conceptuais, semânticas e lexicais. A rede de palavras e conceitos resultante pode ser acedida via navegador como exemplifica a Figura 6.21 [30].

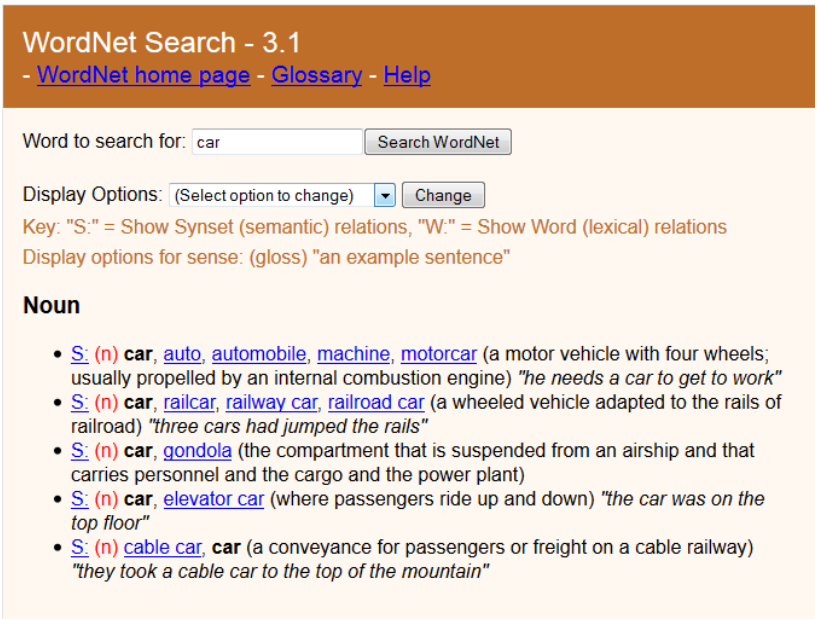

Figura 6.21: Relações semânticas da palavra “car” na WordNet.

Para se interagir com a WordNet procedeu-se à descarga do dicionário[6] e de dois pacotes adicionais, não havendo necessidade de se efectuar qualquer registo prévio.

A WordNet é utilizada neste trabalho para suportar o estabelecimento de relações semânticas entre conceitos de ontologias diferentes, *i.e.*, no âmbito do mapeamento de ontologias.

[6]Utilizou-se a versão 3.0 disponível em `http://wordnetcode.princeton.edu/3.0/WNdb-3.0.tar.gz`.

6.4.3.1 Biblioteca de Interface

A *Java API for WordNet Searching* (JAWS)[7] é uma biblioteca Java de interface com a WordNet. O Excerto de código 6.5 ilustra como se obtêm as relações semânticas de uma palavra através da JAWS.

Excerto de Código 6.5 Acesso ao dicionário WordNet.

```
1     System.setProperty("wordnet.database.dir", properties.getProperty("wordnet.database.dir"));
2     WordNetDatabase database = WordNetDatabase.getFileInstance();
3        Synset[] synsets = database.getSynsets(wordForm);
4        if (synsets.length > 0) {
5            for (int i = 0; i < synsets.length; i++) {
6                String[] wordForms = synsets[i].getWordForms();
7                for (int j = 0; j < wordForms.length; j++) {
8                    System.out.print((j > 0 ? ", " : "") + wordForms[j]);
9                }
10               System.out.println(": " + synsets[i].getDefinition());
11           }
```

Em primeiro lugar, configura-se o caminho para o dicionário; de seguida, através de um objecto do tipo `WordNetDatabase`, acede-se à base de dados; e, por último, especifica-se a palavra e a relação semântica pretendida.

A JAWS foi utilizada neste projecto para interagir com a WordNet no âmbito do mapeamento de categorias.

6.5 Mapeamento de Categorias

Os conteúdos provenientes das diferentes fontes do perfil do utilizador foram classificados através das categorias dos programas da BBC, que englobam os formatos e géneros dos programas da BBC. Dado que as categorias dos conteúdos da BBC e das fontes de enriquecimento utilizadas são distintas, foi necessário proceder a mapeamentos para integrar a informação heterogénea. As categorias utilizadas para descrever as actividades do utilizador são provenientes da IMDb, Facebook, Youtube e da BBC.

Os dados sociais obtidos pelo Beancounter foram filtrados por filmes, séries e celebridades, por isso as categorias utilizadas foram as do IMDb. O Facebook já caracteriza as suas páginas com dados que o Beancounter coloca no perfil social do utilizador. Estas categorias são provenientes apenas dos *likes*. A restante informação -*shares* e *tweets* do perfil social - é analisada e caracterizada pelas fontes de enriquecimento anteriormente referidas, para daí serem extraídos os filmes, as séries e as celebridades e então serem catalogados pelo IMDb. Assim, para o perfil social foram efectuados dois mapeamentos: do Facebook para a BBC e do IMDb para a BBC.

[7] `http://lyle.smu.edu/%7Etspell/jaws/jaws-bin.jar`

O YouTube, que constitui uma das fontes do historial do utilizador, atribui também categorias próprias aos seus vídeos, tendo, por isso, sido efectuado um mapeamento das categorias do Youtube para as categorias dos programas BBC.

Umas vez mapeadas todas as categorias para as categorias dos programas da BBC, é possível criar o vector do perfil global do utilizador.

Para estes mapeamentos recorreu-se, não só, às ontologias de representação de conhecimento definidas, mas também à WordNet para obter as relações semânticas. O mapeamento é estabelecido entre os conceitos de cada uma das quatro ontologias de representação de categorias do Facebook, IMDb e YouTube e os conceitos da ontologia de categorias dos programas da BBC. O algoritmo desenvolvido determina as relações semânticas entre as categorias de cada par de ontologias a mapear. As categorias que permaneçam não mapeadas, *i.e.*, não foi possível estabelecer uma relação semânticas entre aquele conceito e os conceitos da ontologia de destino, poderão vir a sê-lo através de um mapeamento ao nível de instâncias a realizar em tempo de execução. As ontologias construídas para efeito de mapeamento encontram-se nos Anexos B, C, D.

6.5.1 Conceitos

Na primeira fase, o algoritmo estabelece a relação semântica entre os conceitos de ambas as ontologias. As categorias dos programas da BBC são constituídas por formatos e géneros, estando os géneros estruturados em diversos níveis. Assim, os conceitos da ontologia a mapear são confrontados, em primeiro lugar, com os formatos e com o primeiro nível dos géneros. Caso não se identifique uma relação semântica, os conceitos da ontologia a mapear são confrontados com o segundo nível dos géneros da BBC como ilustra a Figura 6.22.

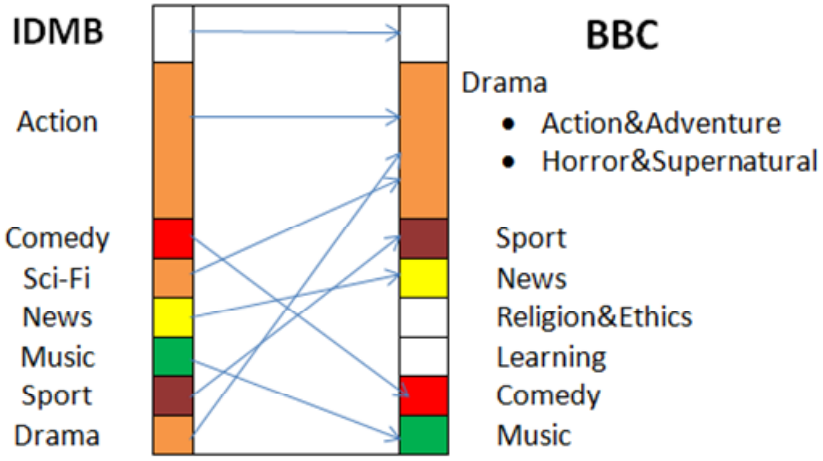

Figura 6.22: Exemplo de mapeamento da IMDb para BBC.

Este algoritmo foi aplicado aos três mapeamentos efectuados. A cobertura obtida no mapeamento do YouTube para a BBC ronda 75 % (Figura 6.24) e no mapeamento da IMDb para a BBC atinge 85 % (Figura 6.23). O resultado do mapeamento da IMDb e do YouTube encontra-se nos Anexos E e F, respectivamente. No caso do mapeamento do Facebook para a BBC, dado que as categorias das páginas do Facebook são muito diferentes das da BBC, obteve-se apenas 22 % de cobertura. Perante este resultado efectuou-se, neste caso, um mapeamento manual das categorias das páginas do Facebook para as categorias dos programas da BBC. A Tabela 6.3 apresenta alguns exemplos desse mapeamento. O Anexo G apresenta o mapeamento completo.

Tabela 6.3: Mapeamento de Categorias do Facebook para BBC.

Categoria Facebook	Categoria BBC
Church/Religion organization	Religion&Ethics
Dancer	Music
Athlete	Sport
Comedian	Comedy
Entertainer	Entertainment
Games/Toys	Children's
Health/Beauty	Factual
Media/News/Publishing	News

No final deste procedimento, restaram ainda algumas categorias do YouTube e da IMDb sem mapeamento. Para aumentar a percentagem de cobertura destes mapeamentos pode-se aplicar um mapeamento ao nível das instâncias em tempo de execução.

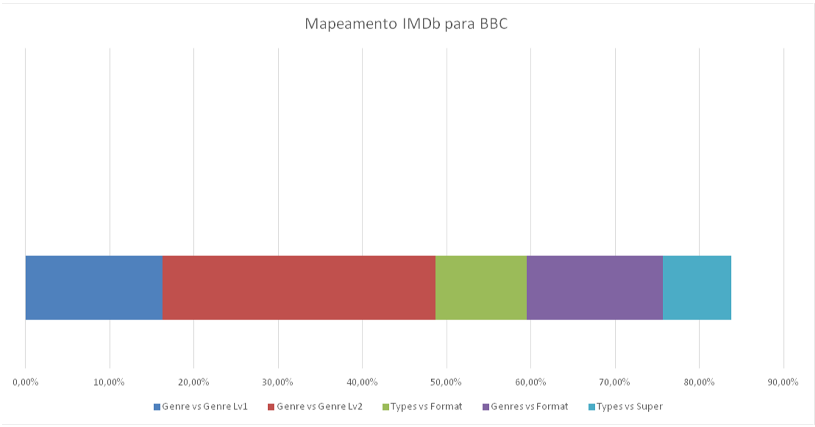

Figura 6.23: Cobertura do mapeamento IMDb para BBC.

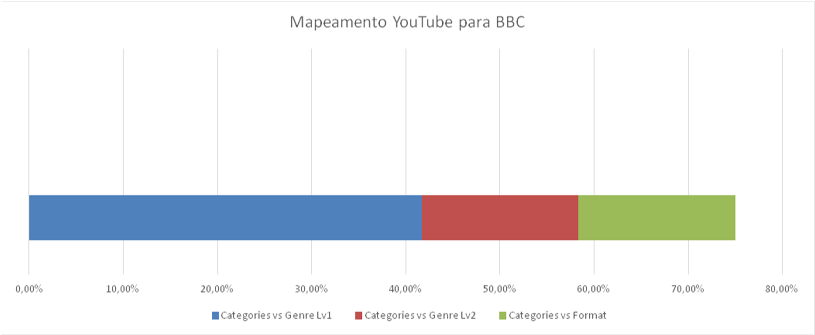

Figura 6.24: Cobertura do mapeamento Youtube para BBC.

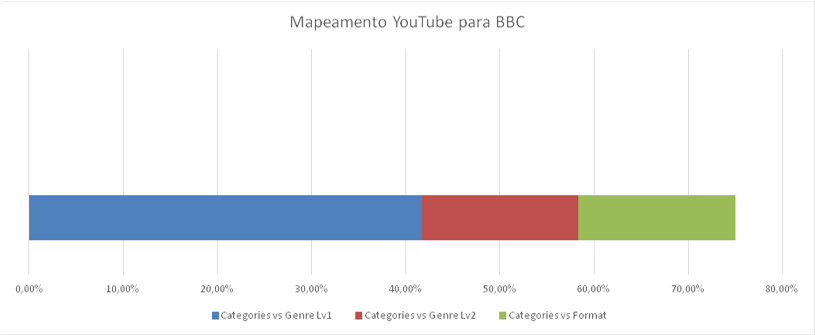

Figura 6.25: Cobertura do mapeamento Facebook para BBC.

6.5.2 Instâncias

O mapeamento de conceitos instanciados permitirá, numa segunda fase, mapear os conceitos não cobertos. Este algoritmo utiliza o peso das categorias não mapeadas para dirigir a sua procura no perfil do utilizador. Identifica as características com peso da mesma ordem de grandeza no perfil do utilizador e, recorre à WordNet, para determinar se a categoria não mapeada e a categoria do perfil do utilizador são sinónimos. A Figura 6.26 ilustra o mapeamento de conceitos instanciados.

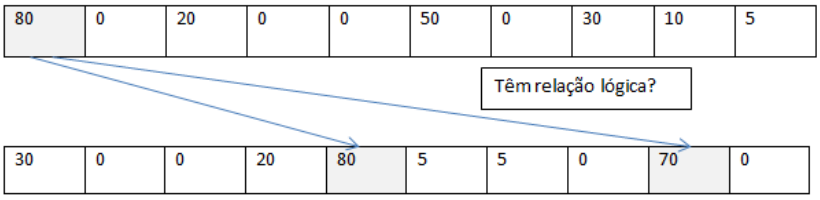

Figura 6.26: Mapeamento de conceitos instanciados.

Uma categoria não mapeada com peso 80 % tenta encontrar uma relação semântica com as categorias do peso do utilizador com pesos semelhantes. Caso a relação exista, o mapeamento é estabelecido, aumentando a percentagem de cobertura do mapeamento.

6.6 Construção da Recomendação

O sistema de recomendação desenvolvido é um Sistema de Recomendação baseado no conteúdo. Adoptou-se o *keyword vector space model* e a similaridade dos cossenos para processamento dos diferentes vectores de categorias. A Figura 6.27 ilustra o algoritmo de recomendação.

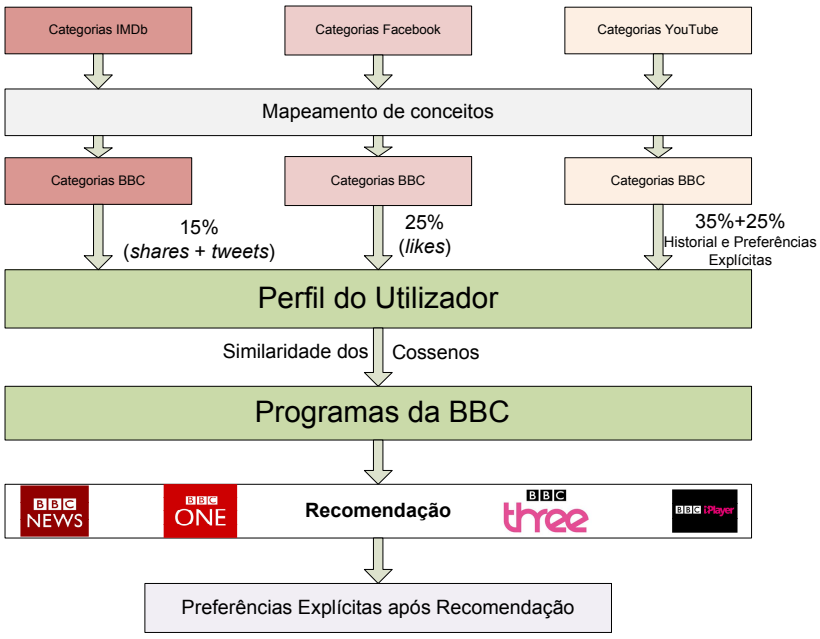

Figura 6.27: Construção da Recomendação.

6.6.1 Vectores Constituintes do Perfil Global

A ontologia do perfil do utilizador, que se encontra no Anexo A, representa as diferentes componentes do perfil através de quatro vectores (IMDb, Facebook, YouTube e Preferências Explícitas). Cada posição de um vector corresponde a uma categoria e tem um determinado peso, representado em percentagem.

O perfil social do utilizador proveniente do Beancounter apresenta uma lista de interesses e respectivos pesos que são utilizados para a construção dos vectores IMDb e Facebook.

A API do YouTube devolve os últimos 50 vídeos vistos pelo utilizador. O cálculo dos pesos do vector YouTube do historial contabiliza para as diferentes categorias o número de vídeos vistos pelo utilizador, normaliza e converte o resultado em percentagens. O mesmo procedimento é efectuado com as categorias do vector IMDb, do Facebook e das preferências explícitas.

Na Equação 6.1, Equação 6.2 e na Equação 6.3 são apresentados exemplos dos vectores criados no âmbito da construção de uma recomendação.

Equação 6.1: Categorias IMDb (*Posts* e *Tweets*).

$$IMDb = [0, 0, 0, 12, 25, 0, 0, 37, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 12, 14, 0, 0, 0, 0] \quad (6.1)$$

Equação 6.2: Categorias Facebook (*Likes*).

$$Facebook = [0, 0, 0, 0, 50, 0, 0, 0, 25, 0, 25, \ldots, 0] \quad (6.2)$$

Equação 6.3: Categorias YouTube (Historial).

$$YouTubeHist = [0, 0, 0, 0, 0, 0, 100, 0, 0, 0, 0, 0, 0, 0, 0] \quad (6.3)$$

Equação 6.4: Categorias YouTube (*Likes*).

$$YouTubePref = [0, 0, 0, 0, 0, 0, 100, 0, 0, 0, 0, 0, 0, 0, 0] \quad (6.4)$$

De seguida, estes vectores de categorias são mapeados para as categorias da BBC, passando a ser constituídos pelas 26 categorias da BBC, que incluem os géneros e os formatos dos programas.

O vector das preferências explícitas inclui os *likes* do YouTube e as preferências manifestadas pelo utilizador após a recomendação. As preferências dos vídeos do YouTube são, depois de mapeadas, somadas às preferências dos programas da BBC para criar o vector das preferências explícitas.

A contribuição de cada vector no perfil global foi definida de acordo com a importância da respectiva componente na recomendação final. Aos dados do historial do YouTube, que têm uma maior relevância, foi atribuído um peso de 35 %. Ao vector correspondente ao Facebook, que contém os *likes*, foi atribuído um peso de 25 %. Ao vector das Preferências Explícitas foi atribuído um peso de 25 %. Por fim, o vector que contém os restantes dados das redes sociais (*shares* e *tweets*) tem um peso de 15 %.

Cada vector, depois de mapeado, é convertido num `HashMap` para permitir o preenchimento da ontologia. Na Tabela 6.4 são apresentados os `HashMaps` construídos a partir dos quatro vectores mapeados.

Tabela 6.4: Hash Maps

Hash Map		Conteúdo
IMDb	=	[Weather=0, Learning=0, Phone-ins=0, PerformancesEvents=0, Factual=0, Animation=0, Talent-Shows=0, GameQuiz=0, Comedy=25, Bulletins=0, News=0, ReligionEthics=0, Documentaries=0, MagazineReviews=0, Children's=0, Music=0, Docugramas=0, Entertainment=0, Makeovers=0, Sport=12, DiscussionTalk=0, Appeals=0, Reality=0, Film=0, Drama=62]
Facebook	=	[Weather=0, Learning=0, Phone-ins=0, PerformancesEvents=0, Factual=0, Animation=0, Talent-Shows=0, GameQuiz=0, Comedy=0, Bulletins=0, News=0, ReligionEthics=0, Documentaries=0, MagazineReviews=0, Children's=0, Music=100, Docugramas=0, Entertainment=0, Makeovers=0, Sport=0, DiscussionTalk=0, Appeals=0, Reality=0, Film=0, Drama=0]
YouTube	=	[Weather=0, Learning=0, Phone-ins=0, PerformancesEvents=0, Factual=0, Animation=0, Talent-Shows=0, Religion=0, Game Quiz=0, Comedy=0, Bulletins=0, News=0, ReligionEthics=0, Documentaries=0, MagazineReviews=0, Children's=0, Music=100, Docugramas=0, Entertainment=0, Makeovers=0, Sport=0, DiscussionTalk=0, Appeals=0, Reality=0, Film=0, Drama=0]
PrefExplic	=	[Weather=0, Learning=0, Phone-ins=0, PerformancesEvents=0, Factual=0, Animation=0, Talent-Shows=0, Religion=0, Game Quiz=0, Comedy=10, Bulletins=0, News=0, ReligionEthics=0, Documentaries=0, MagazineReviews=0, Children's=0, Music=90, Docugramas=0, Entertainment=0, Makeovers=0, Sport=0, DiscussionTalk=0, Appeals=0, Reality=0, Film=0, Drama=0]

O Excerto de Código 6.6 ilustra a ponderação do vector IMDb.

Excerto de Código 6.6 Cálculo da ponderação do vector IMDb.

```
1      for (int j = 0; j < vectorIMDB.size(); ++j) {
2            int im = (int) ((vectorIMDB.get(j)) * 0.15);
3            vectorFinalI.add(im);
4        }
```

O vector global do perfil do utilizador é determinado através da média ponderada dos quatro Hash Maps. O Excerto de Código 6.7 apresenta o cálculo do vector global do perfil do utilizador.

Excerto de Código 6.7 Cálculo do vector global.

```
1     for (int s = 0; s < vectorFacebook.size(); ++s) {
2           int soma = (vectorFinalF.get(s) + vectorFinalI.get(s)+
3           + vectorFinalY.get(s) + vectorFinalP.get(s));
4           vectorFinal.add(soma);
5       }
```

Na Tabela 6.5 é apresentado o vector resultante. Neste caso, os resultados apresentam um desvio de 0.15 % devido aos arredondamentos tanto no mapeamento como no cálculo da média. As categorias recomendadas para este perfil são *Music*, *Drama*, *Comedy* e *Sport*.

Tabela 6.5: Hash Map final

Hash Map		Conteúdo
IMDb	=	[Weather=0, Learning=0, Phone-ins=0, PerformancesEvents=0, Factual=0, Animation=0, Talent-Shows=0, GameQuiz=0, Comedy=6.25, Bulletins=0, News=0, ReligionEthics=0, Documentaries=0, MagazineReviews=0, Children's=0, Music=82.5, Docugramas=0, Entertainment=0, Makeovers=0, Sport=1.8, DiscussionTalk=0, Appeals=0, Reality=0, Film=0, Drama=9.3]

Um vez determinadas as categorias a recomendar, interage-se com o serviço Programmes da BBC para se obter a lista dos programas com essas categorias. A lista de programas é filtrada pela data (contexto temporal) e os programas são ordenados em função da similaridade que possuem com o perfil do utilizador. A similaridade entre os programas candidatos e o utilizador é determinada através da similaridade dos cossenos.

6.6.2 Interacção com a BBC

A BBC disponibiliza o serviço RESTful Programmes para interacção com a sua grelha de programas. A interacção é efectuada através de pedidos HTTP e as respostas podem vir no formato JSON, RDF ou XML. Neste projecto interage-se com este serviço para obter os programas que se encaixam no perfil do utilizador.

6.6.2.1 Biblioteca de Interface

A biblioteca de interface é constituída pelo serviço Programmes que disponibiliza três operações: *schedules*, *genres* e *programmes*. A Tabela 6.6 apresenta os diferentes *endpoints* do serviço. O formato da resposta (JSON, RDF ou XML) é especificado no URL do *endpoint* pós-fixando a extensão pretendida (`.json`, `.rdf` ou `.xml`).

Tabela 6.6: Serviço RESTfull Programmes da BBC

Operações	*Endpoint*

Schedules	`/:service/programmes/schedules/:outlet`
	`/:service/programmes/schedules/:outlet/:year/:month/:day`
	`/:service/programmes/schedules/:outlet/:year/:month/:day/ataglance`
	`/:service/programmes/schedules/:outlet/yesterday`
	`/:service/programmes/schedules/:outlet/today`
	`/:service/programmes/schedules/:outlet/tomorrow`

Genres	`/:service/programmes/genres/:genre1/:genre2/:genre3/schedules`
	`/:service/programmes/genres/:genre1/:genre2/:genre3/schedules/:year/:month/:day`
	`/:service/programmes/genres/:genre1/:genre2/:genre3/schedules/:year/:month/:day/ataglance`
	`/:service/programmes/genres/:genre1/:genre2/:genre3/schedules/upcoming`

Formats	`/programmes/:groupPID/episodes/upcoming`
	`/programmes/:groupPID/episodes/upcoming/debut`
	`/programmes/:groupPID/episodes/player`

O URL de invocação do serviço é construído em função das categorias presentes no perfil do utilizador. Por exemplo, o URL `http://www.bbc.co.uk/tv/`

`programmes/genres/drama.json` retorna todos os programas do género *Drama*. A Figura 6.28 apresenta a informação de um dos programas dramáticos existentes na grelha de programas da BBC, incluindo o `pid`, título, duração, *etc.*


```
- programme: {
      type: "series",
      pid: "b0103y2x",
      title: "Series 6",
      position: 44,
      expected_child_count: 13,
      first_broadcast_date: "2011-04-23T18:00:00+01:00",
    - programme: {
          type: "brand",
          pid: "b006q2x0",
          title: "Doctor Who",
          position: null,
          expected_child_count: null,
          first_broadcast_date: "1963-11-23T17:15:00Z",
        - ownership: {
            - service: {
                  type: "tv",
                  id: "bbc_one",
                  key: "bbcone",
                  title: "BBC One"
              }
          }
      }
  },
```


Figura 6.28: Dados genéricos de um programa da BBC.

A informação do horário do programa, *i.e.*, a hora de início e fim, são dados importantes para a construção da recomendação e que podem obter-se através de um pedido HTTP de horário (*schedules*). Alternativamente, no caso de uma série, pode-se solicitar a informação do próximo episódio. A Figura 6.29 contém a resposta ao pedido `http://www.bbc.co.uk/programmes/b0103y2x/episodes/upcoming.json`. A referência `b0103y2x` consiste na identificação do programa (`pid`).

Desta forma, é possível recomendar programas da BBC compatíveis com o perfil do utilizador e que estão a ser ou vão ser transmitidos.

6.6.2.2 Vector de Categorias dos Programas

O serviço Programmes da BBC retorna todos os programas candidatos à recomendação, *i.e.*, com categorias iguais às categorias não nulas do perfil do utilizador.

Dado que na BBC cada programa é identificado por um formato e um género principal, interage-se de novo com a IMDb para classificar de forma mais detalhada os programas da BBC. Por exemplo, enquanto a série "American Dad" (Figura 6.30) possui apenas a categoria *Animation* na BBC, na IMDb apresenta

```
- {
    is_repeat: true,
    is_blanked: false,
    schedule_date: "2013-11-07",
    start: "2013-11-07T19:10:00Z",
    end: "2013-11-07T20:00:00Z",
    duration: 3000,
  - service: {
        type: "tv",
        id: "bbc_three",
        key: "bbcthree",
        title: "BBC Three"
    },
  - programme: {
        type: "episode",
        pid: "b011884d",
        position: 4,
        title: "The Doctor's Wife",
        short_synopsis: "By following a Time Lord distress signal, the Doctor
        endangers Amy, Rory and the TARDIS.",
        media_type: "audio_video",
        duration: 3000,
      - image: {
            pid: "p01h399k"
        },
      - display_titles: {
            title: "Doctor Who",
            subtitle: "Series 6, The Doctor's Wife"
        },
        first_broadcast_date: "2011-05-14T18:30:00+01:00",
      - ownership: {
          - service: {
                type: "tv",
                id: "bbc_one",
                key: "bbcone",
```


Figura 6.29: Dados detalhados de um programa da BBC.

as categorias *Animation* e *Comedy*. Aplica-se novamente o mapeamento das categorias do IMDb para BBC, constroem-se os vectores de características dos programas e, por fim, determina-se a similaridade entre os programas candidatos e o utilizador.

No caso da série "American Dad", o vector resultante encontra-se na Tabela 6.7. A similaridade entre esta série e o utilizador é determinada entre este vector e o vector do perfil do utilizador.

Tabela 6.7: `Hashmap` da série "American Dad"

Hash Map		Conteúdo
IMDb	=	[Weather=0, Learning=0, Phone-ins=0, PerformancesEvents=0, Factual=0, Animation=50, Talent-Shows=0, GameQuiz=0, Comedy=50, Bulletins=0, News=0, ReligionEthics=0, Documentaries=0, MagazineReviews=0, Children's=0, Music=0, Docugramas=0, Entertainment=0, Makeovers=0, Sport=0, DiscussionTalk=0, Appeals=0, Reality=0, Film=0, Drama=0]

6.6.3 Determinação da Similaridade

A determinação da similaridade entre os vectores de características do perfil do utilizador e dos programas é efectuada segundo a Equação 2.7 da similaridade

Figura 6.30: Representação do programa "American Dad" na BBC.

dos cossenos. O Excerto de Código 6.8 apresenta a implementação desta métrica.

Excerto de Código 6.8 Implementação da Similaridade dos Cossenos.

```
1        private double calcSim(int[] userModel, int[] programmeModel){
2        double sim = 0, numerator=0, denominatorum=0, denominator=0, denominatorpm=0;
3        if(userModel.length==programmeModel.length){
4            for(int i=0;i<userModel.length;i++){
5                numerator=numerator+(userModel[i]*programmeModel[i]);
6                denominatorum=denominatorum+Math.pow(userModel[i],2);
7                denominatorpm=denominatorpm+Math.pow(programmeModel[i], 2);
8            }
9            denominator=Math.sqrt(denominatorum)*Math.sqrt(denominatorpm);
10           sim=numerator/denominator;
11       }
12       return sim;
13   }
```

Quanto mais próximo da unidade for o resultado, mais o programa se identifica com o utilizador. Os programas candidatos são representados através da similaridade com o perfil do utilizador e de `Hashmaps` constituídos pelas categorias não nulas do perfil do utilizador.

6.6.4 Resultado da Recomendação

A recomendação é construída em tempo real e é apresentada de formas distintas. Em primeiro lugar, apresenta-se um calendário contendo os programas da BBC mais similares com o perfil do utilizador. Este calendário constitui a grelha de programação personalizada (EPG) semanal do utilizador e contém os programas recomendados e, no caso de séries, inclui as repetições no intervalo de uma semana. Em segundo lugar, apresenta-se um *link* que redirecciona o utilizador para as páginas dos programas da BBC das categorias de maior peso no perfil do utilizador, incluindo os programas do próprio dia ordenados por similaridade.

A Recomendação apresenta o título, uma pequena sinopse, a duração e o horário de transmissão do programa. Os programas recomendados são organizados por categoria (ordem decrescente do peso da categoria no perfil do utilizador) e por hora de exibição. Nesta etapa o utilizador pode classificar as recomendações do sistema, *i.e.*, manifestar explicitamente as suas preferências de uma forma colaborativa.

6.6.4.1 Interacção com o Calendário Google

O calendário de recomendações de programas é um calendário da Google. Reutilizou-se a API da Google, tendo apenas sido necessário a activação do serviço Google Calendar. No calendário são inseridos os programas da BBC recomendados. Antes da inserção das recomendações no calendário, o utilizador tem de permitir que o serviço interaja com o seu calendário (Figura 6.31) Google. A introdução dos eventos, *i.e.*, dos programas, é efectuada segundo o dia, hora de início e de fim e o canal da BBC de transmissão do programa. O Excerto de Código 6.9 exemplifica a inserção de um programa no calendário. A Figura 6.32 apresenta o resultado desta operação.

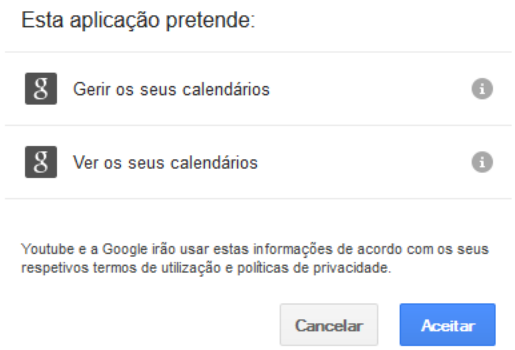

Figura 6.31: Aceitar a interacção com o calendário Google.

Excerto de Código 6.9 Introdução de eventos no calendário.

```
1        Event event = new Event();
2        Event eventTest = new Event();
3        event.setSummary(programme);
4        event.setLocation(channel);
5        DateTime start = DateTime.parseRfc3339(starthour + ".000-00:00");
6        DateTime end = DateTime.parseRfc3339(endhour + ".000-00:00");
7        event.setStart(new EventDateTime().setDateTime(start).setTimeZone("Europe/Lisbon"));
8        event.setEnd(new EventDateTime().setDateTime(end).setTimeZone("Europe/Lisbon"));
9        client.events().insert(calendarid, event).execute();
```

Figura 6.32: Informação do programa BBC no calendário.

6.7 Interacção com a Plataforma Desenvolvida

A interface gráfica reutiliza a estrutura e os ficheiros de estilo, *Cascading Style Sheets* (CSS), do Beancounter [75]. A tecnologia de desenvolvimento do lado do servidor é constituída por *Servlets* e, do lado do cliente, recorre à linguagem HTML e à biblioteca de JavaScript JQuery.

A interface com o utilizador permite efectuar o registo (Perfil Contextual), associar redes sociais, visualizar as actividades sociais (Perfil Social), aceder ao historial no YouTube (Perfil baseado no Conteúdo) e usufruir das recomendações criadas. Apresenta ainda a análise dos dados processados pelo Beancounter, sendo construído um gráfico circular (*pie chart*) com os interesses sociais identificados.

6.7.1 Registo do Utilizador

O registo do utilizador na plataforma é efectuado de forma simples, bastando inserir o nome do utilizador e a palavra-chave. A Figura 6.33 ilustra um exemplo de registo efectuado através de serviços RESTful que interagem com a base de dados NoSQL Redis.

6.7.2 Autenticação nas Fontes de dados

Após o registo na plataforma, o utilizador é direccionado para a página da Figura 6.34 para associar as fontes de dados (redes sociais e YouTube) usadas. A associação das contas das redes sociais do utilizador à plataforma consiste no redireccionamento para a página de autenticação da rede social e na introdução das credenciais do utilizador. Para aceder ao historial do YouTube, o utilizador tem de associar a sua conta Google. Terminada esta etapa, arranca a execução do Beancounter, sendo o utilizador novamente redireccionado para a página inicial.

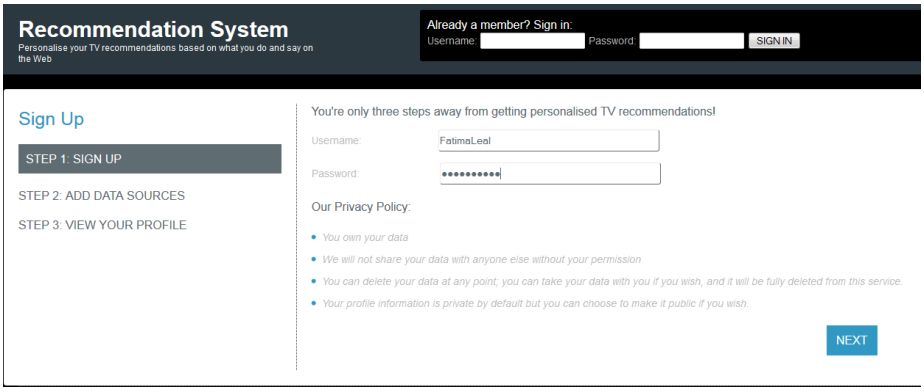

Figura 6.33: Registo do utilizador no Sistema de Recomendação.

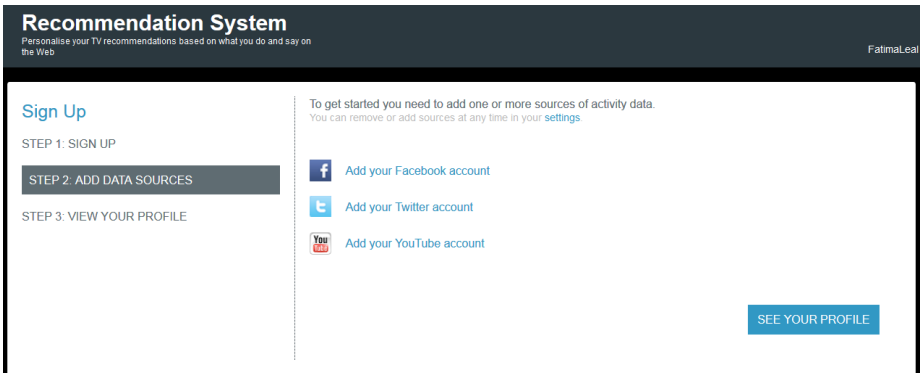

Figura 6.34: Associação das fontes de dados Facebook, Twitter e YouTube.

A interacção com as redes sociais é efectuada pelo Beancounter através de serviços RESTful chamando o método `handleOAuthCallback()` apresentado na Tabela 6.1.

Os dados do historial do YouTube pretendem simular a informação armazenada pela STB e não são armazenados localmente. Sempre que o utilizador solicite recomendações, a aplicação obtém os últimos 50 vídeos vistos pelo utilizador e constrói o perfil do historial do utilizador em tempo real. A interacção com os dados do utilizador é efectuada recorrendo à API da Google para o YouTube. O Excerto de Código 6.10 exemplifica como é efectuada a autenticação em linguagem Java.

Neste fase, a aplicação já apresenta as actividades sociais do utilizador, a *Pie Chart* de análise aos dados sociais e o historial do YouTube.

Excerto de Código 6.10 Autenticação no YouTube.

```
1     YouTubeService service = new YouTubeService(clientID, developer_key)
2     String requestUrl = AuthSubUtil.getRequestUrl("http://beancounter.no-ip.biz:8080/settings.html",
3     "http://gdata.youtube.com", false, true);
4     service.setAuthSubToken(sessionToken, null);
```

6.7.3 Interesses e Actividades Sociais

O Beancounter interage com as redes sociais, processa e armazena os dados. O método `getallActivity()` do Beancounter retorna os *likes*, *shares* e *tweets* do utilizador. A interface inclui uma hiperligação que redirecciona o utilizador para as actividades sociais que efectuou (Figura 6.35). O utilizador pode ainda consultar a data de realização das actividades sociais.

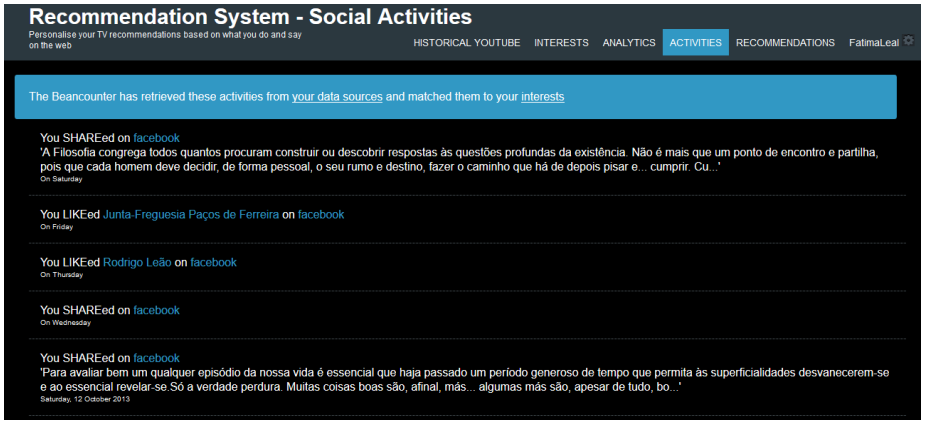

Figura 6.35: Actividades das redes sociais.

O texto dos *shares* e *tweets* é processado pelo serviço de enriquecimento semântico Lupedia. A Figura 6.36 apresenta a informação dos *likes* e dos interesses resultantes dos *shares* e *tweets*. Os *likes* são descritos através das categorias devolvidas pelo perfil social do Beancounter. Os interesses são descritos através da respectiva hiperligação para a DBpedia.

Facebook's Like	Likes's Name	Interests	Description	Number of Activities
Organization	Junta-Freguesia Paços de Ferreira	Springfield_(The_Simpsons)	http://dbpedia.org/resource/Springfield_(The_Simpsons)	1
Organization	Junta-Freguesia Paços de Ferreira	Hank_(Werkendam)	http://dbpedia.org/resource/Hank_(Werkendam)	1
band,Musician	Rodrigo Leão	Princesa_(2001_film)	http://dbpedia.org/resource/Princesa_(2001_film)	1
band,Musician	Marco Frisina	Castellaneta	http://dbpedia.org/resource/Castellaneta	1
band,Musician	Baroque Music	Isso_(Italy)	http://dbpedia.org/resource/Isso_(Italy)	1
band,Musician	The Gift	Jos	http://dbpedia.org/resource/Jos	1

Figura 6.36: Interesses retirados das redes sociais.

6.7.3.1 Gráfico de Interesses Sociais

O perfil social construído pelo Beancounter inclui o peso de cada interesse no perfil social do utilizador. Esse peso é utilizado na construção da *Pie Chart* de análise. No exemplo da Figura 6.37 verifica-se que os cinco principais interesses são *Isso*, *Springfield*, *Hank*, *Princesa* e *Castellaneta*. A informação contida entre parêntesis contextualiza o interesse. Estes interesses foram identificados pelo serviço de enriquecimento de texto Lupedia a partir dos *tweets* e *shares* do utilizador.

Como as contas do Facebook e do Twitter utilizadas continham informação em português e o serviço de enriquecimento utilizado apenas processa textos em inglês, alguns dos interesses encontrados podem não estar correctos. É o caso do interesse *Isso* que foi erradamente extraído, pois o serviço Lupedia associou a palavra "isso" a uma região italiana.

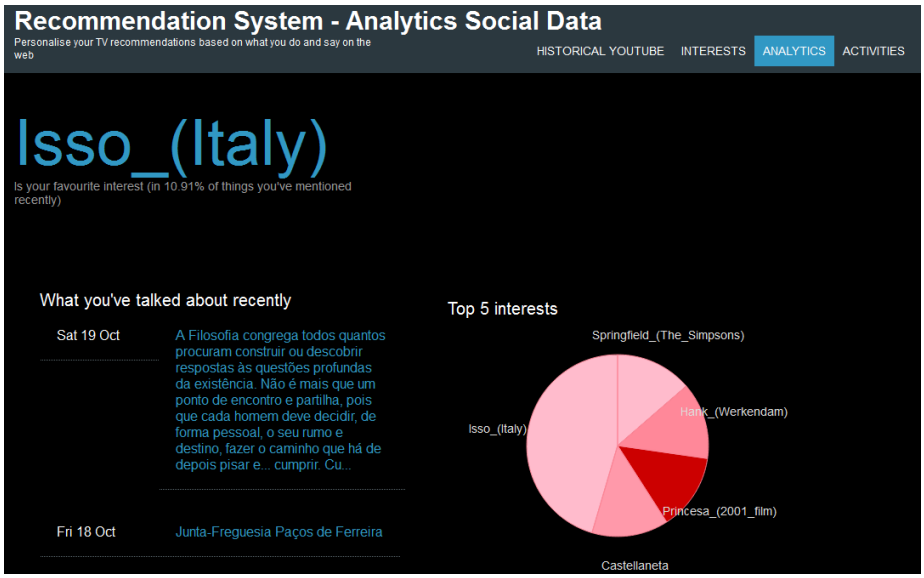

Figura 6.37: Análise das actividades sociais via *Pie Chart*.

É ainda possível obter uma breve descrição e a data de cada actividade. Os *likes* são processados separadamente dado que têm categorias próprias e possuem maior peso na recomendação final.

6.7.4 Historial YouTube

O historial do Youtube é constituído pelos último 50 vídeos vistos. A categoria YouTube dos vídeos pode ser obtida através do Excerto de Código 6.11.

A interface construída apresenta ao utilizador os últimos vídeos do seu historial. Como o YouTube apaga o historial semanalmente, se não houver historial,

Excerto de Código 6.11 Acesso às categorias do vídeos.

```
1           YouTube.VideoCategories.List as = youtube.videoCategories().list("snippet");
2           as.setId(sni.getCategoryId());
3           VideoCategoryListResponse asd = as.execute();
4           List<VideoCategory> v1 = asd.getItems();
```

a recomendação é efectuada tendo em conta apenas os dados sociais.

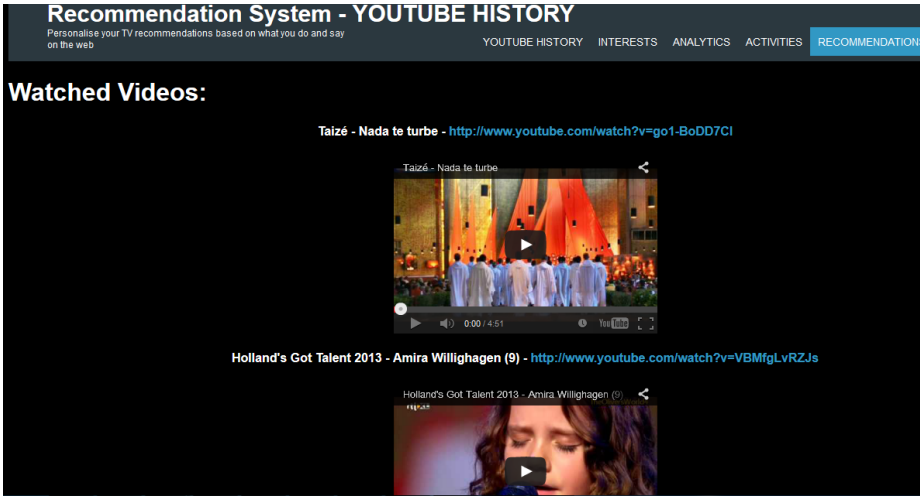

Figura 6.38: Historial do utilizador no YouTube.

6.7.5 Informação Contextual

O contexto do utilizador é constituído pelo seu nome (introduzido aquando o registo), pela data e pela localização. A informação da data é utilizada para gerar recomendações actuais, *i.e.*, do próprio dia, e a localização inclui ou exclui automaticamente alguns vídeos tanto no YouTube como na BBC. A Figura 6.39 apresenta a informação contextual de um utilizador no Sistema de Recomendação desenvolvido.

6.7.6 Recomendações

A recomendação é gerada a pedido do utilizador, podendo a interface gráfica ser utilizada apenas para consultar os dados sociais. Sempre que o utilizador acede à página de recomendação, o cliente automático de interacção com o Beancounter é activado começando a agregar os dados sociais o historial e preferências do YouTube para a criação dos vectores. O Excerto de Código 6.12 apresenta a autenticação do cliente automático no Beancounter sempre que o utilizador requerer recomendações.

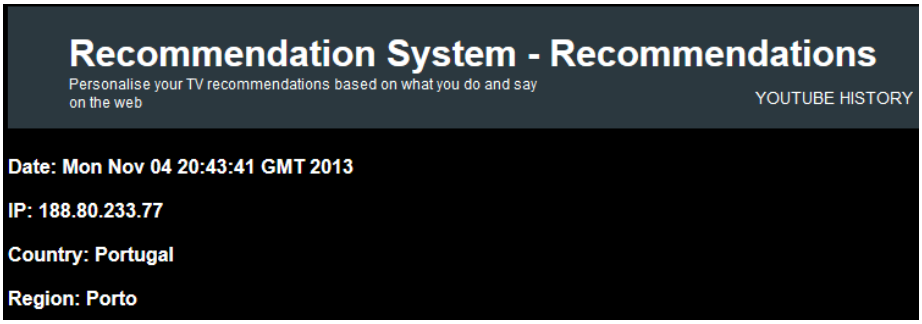

Figura 6.39: Informação contextual do utilizador.

Excerto de Código 6.12 Criação de cliente automático de acesso ao Beancounter.

```
1        PostMethod method = new
2        PostMethod("http://beancounter.no-ip.biz:8080/
3        platform/rest/user/"+username+"/authenticate?+
4        apikey=fd048738-3981-4d23-a6ea-6c8ce3f55c7b");
5        method.addParameter("password", password);
```

As recomendações baseiam-se nas categorias dos programas da BBC que englobam os géneros e os formatos. O utilizador obtém um EPG semanal com programas dos diversos canais da BBC como apresenta a Figura 6.40. São também apresentados os programas do próprio dia (Figura 6.41) que estão organizados pelas categorias preponderantes do perfil do utilizador e pelo horário de transmissão. A recomendação inclui uma sinopse do programa.

6.7.7 Preferências Explícitas

Feita a recomendação, o utilizador pode ainda classificar as recomendações, *i.e.*, explicitar as suas preferências face às recomendações geradas. Esta etapa permite aferir o grau de satisfação do utilizador, *i.e.*, o desempenho do Sistema de Recomendação. Essas preferências são representadas através de uma escala de *rating* de um a cinco que é apresentada através de cinco estrelas que o utilizador pode seleccionar(Figura 6.42).

6.8 Conclusão

Neste capítulo foram apresentados os algoritmos, as interacções e a interface gráfica do Sistema de Recomendação de conteúdos multimédia implementado. Todo o desenvolvimento foi efectuado através do NetBeans e utilizando as tecnologias descritas no capítulo 6. Embora no início da implementação tenha sido integrado um sistema base de dados MySQL, este apenas é utilizado no armazenamento

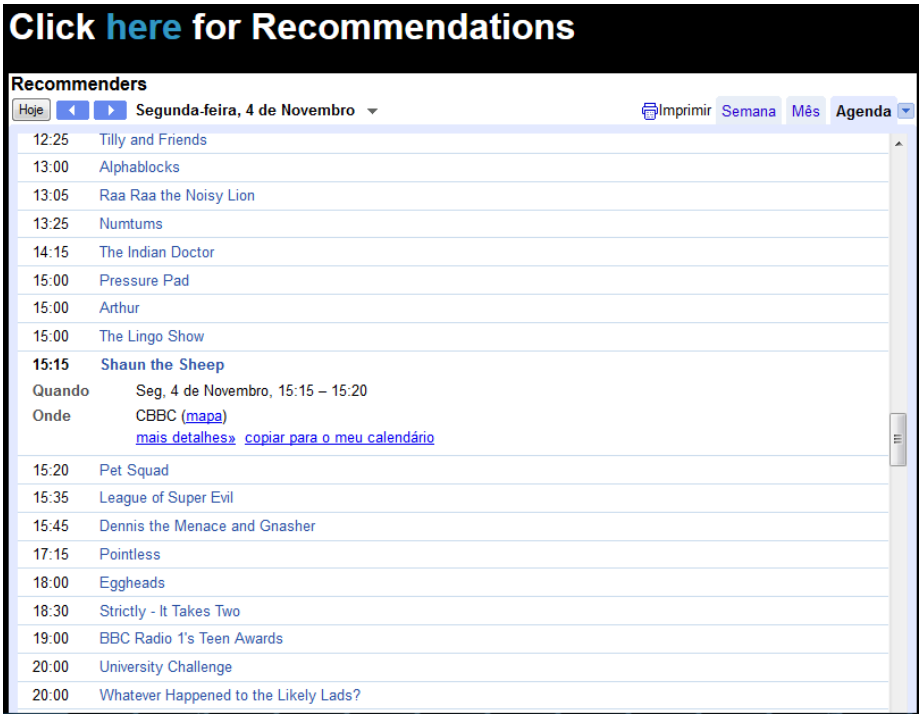

Figura 6.40: Calendário de recomendações.

das preferências explícitas dos programas dado que o Beancounter contém uma base de dados embutida que foi reutilizada para o armazenamento de todos os utilizadores que se registam na plataforma.

A interacção com todas as bibliotecas de interface utilizadas na construção do perfil do utilizador foi descrita. Ao longo do desenvolvimento surgiram dificuldades na interacção com as bibliotecas de interface apresentadas que foram provocadas pela limitação do número de pedidos por segundo, por hora ou por dia imposta pelos serviços.

O algoritmo de mapeamento entre categorias heterogéneas foi apresentado. Este mapeamento permite transformar as categorias de diferentes fontes de dados em categorias dos programas da BBC. A construção do vector do perfil global do utilizador foi detalhada. Por fim, foi apresentada a interface do utilizador que permite consultar os dados das diversas fontes assim como solicitar recomendações. Após a apresentação da recomendação o utilizador pode especificar as suas preferências.

O algoritmo de recomendação foi concebido para ser integrado num sistema de transacção de componentes multimédia suportado por agentes actualmente em desenvolvimento. Embora o desenvolvimento tenha sido efectuado de uma forma independente, todas as ontologias implementadas para representação do

Figura 6.41: Recomendação de programas da BBC.

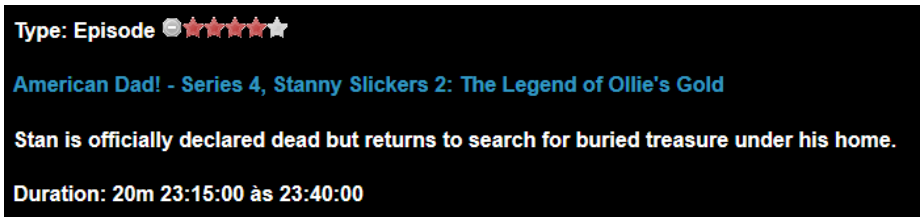

Figura 6.42: Classificação das recomendações.

conhecimento foram pensadas para serem integradas e preenchidas no sistema de transacção de componentes multimédia. Foram criadas cinco ontologias para representação do conhecimento. Para o mapeamento das categorias foram criadas quatro ontologias: Facebook, IMDb, YouTube e BBC. Para o perfil do utilizador foi criada uma ontologia de representação de todas as componentes do utilizador. Estas ontologias estão disponíveis nos Anexos A, B, C e D.

No próximo capítulo efectua-se o balanço do projecto realizado, analisando os resultados, identificando as limitações e propondo melhoramentos futuros.

Capítulo 7

Conclusões

Neste capítulo apresenta-se o balanço do trabalho desenvolvido, os problemas encontrados, os possíveis desenvolvimentos futuros e as conclusões finais.

7.1 Resultados Alcançados

O objectivo da recomendação de conteúdos multimédia personalizados foi alcançado através da construção de um Sistema de Recomendação baseado nas interacções implícitas e explícitas do utilizador. Este sistema é constituído pelo Serviço de Criação do Perfil do Utilizador e pelo Serviço de Recomendação. O perfil do utilizador é composto pelas componentes social, historial (baseada no conteúdo), contextual e pelas preferências explícitas. A componente social é construída pelo Beancounter a partir da actividade do utilizador nas redes sociais Facebook e Twitter. O historial ou a componente baseada no conteúdo consiste no historial de vídeos do YouTube e de programas da BBC vistos. A componente contextual consiste no nome, data e localização do utilizador. A componente das preferências explícitas do utilizador é composta pelos *likes* que atribuiu aos vídeos do YouTube vistos e pela classificação que atribuiu aos programas da BBC recomendados.

O Serviço de Criação do Perfil do Utilizador cria, para representar cada utilizador, uma instância da ontologia do perfil do utilizador desenvolvida. O perfil do utilizador é construído com a ajuda de fontes de enriquecimento de dados (Freebase e DBpedia) e de enriquecimento de texto (Lupedia e WordNet). A Lupedia é utilizada pelo Beancounter como ferramenta de enriquecimento do texto dos *shares* e *tweets* para identificar possíveis interesses. A DBpedia é utilizada pelo Beancounter como fonte enriquecimento de dados. As actividades identificadas pelo Beancounter como *likes* já estão classificadas, sendo as respectivas categorias directamente utilizadas na construção do perfil do utilizador. A Freebase é

utilizada para enriquecer os interesses identificados pela Lupedia que sejam relativos a filmes, séries e figuras públicas, catalogando-os segundo as categorias da IMDb. Deste conjunto de informação resultam quatro vectores que contêm as características YouTube, IMDb, Facebook e BBC (com as preferências explícitas do utilizador). Para se obter o perfil instantâneo do utilizador as diferentes categorias são mapeadas para as categorias dos programas da BBC, recorrendo-se ao mapeamento de categorias previamente definido. Este mapeamento entre as quatro ontologias que representam as categorias das quatro fontes heterogéneas de dados (IMDb, YouTube, Facebook e BBC) foi estabelecido para a ontologia das categorias dos programas da BBC. O perfil instantâneo do utilizador resulta da integração ponderada dos quatro vectores resultantes: o historial do utilizador contribui com um peso de 35 %, os *likes* do Facebook e as preferências explícitas contribuem com um peso de 25 % cada e a informação proveniente dos *shares* e *tweets* com um peso de 15 %.

O Serviço de Recomendação recomenda programas da grelha de programação da BBC com base no perfil instantâneo do utilizador. Uma vez determinada a similaridade entre os programas dos diversos canais da BBC em curso ou ainda a transmitir naquele dia e o perfil do utilizador, os programas são ordenados por ordem decrescente de peso no perfil do utilizador e similaridade.

A interface com o utilizador desenvolvida é uma aplicação *Web* que permite ao utilizador consultar as suas actividades sociais, listar os vídeos YouTube vistos e a recomendação construída. É apresentado um calendário com os programas que resultaram da recomendação. Adicionalmente, o utilizador pode aceder através de um *link* específico à lista completa dos programas da grelha compatíveis com o perfil do utilizador, *i.e.*, que correspondem a categorias com peso no perfil do utilizador, ordenados por ordem decrescente de peso e similaridade. O utilizador pode ainda classificar as recomendações apresentadas numa escala de 1 a 5 através do conjunto de cinco estrelas apresentado à frente de cada recomendação. A recomendação é efectuada em tempo real demorando a plataforma cerca de 30 segundos a processar toda a informação.

7.2 Problemas Ultrapassados

Este projecto implicou o estudo de metodologias e tecnologias desconhecidas e a interacção com fontes de dados heterogéneos e repositórios abertos de dados interligados, tendo sido encontradas dificuldades.

A reutilização do Beancounter implicou um estudo e esforço considerável de estudo da abordagem, fontes e tecnologias envolvidas. No sítio do projecto europeu NoTube[1] não havia documentação específica sobre o Beancounter, sendo

[1] `http://notube.tv`

apenas disponibilizado o código e enumeradas a tecnologias utilizadas. Os autores deste *blog* apresentam o Beancounter como uma ferramenta criação e gestão de perfis, mas o *link* encontra-se fora de serviço. Os responsáveis do projecto foram contactados, mas pouco acrescentaram sobre o seu funcionamento e confirmaram que o protótipo referido no *blog* se encontra fora de serviço. Desta forma, o estudo do funcionamento, tecnologias, configuração e depuração do Beancounter foi realizado de uma forma autónoma, tendo consumido uma grande parte do tempo disponibilizado para o desenvolvimento do projecto.

O primeiro problema ultrapassado foi o sistema operativo utilizado uma vez que o Beancounter foi desenvolvido para ambientes Linux. Inicialmente, recorreu-se a uma máquina virtual Linux onde se instalaram as tecnologias de suporte utilizadas (Elasticsearch, Redis e Kestrel), tendo-se posteriormente encontrado versões compatíveis com ambientes Windows, adaptado todos os *scripts* de execução destes serviços para Windows e migrado para uma plataforma Windows.

A compatibilidade das versões das bibliotecas e das suas dependências que o Beancounter utiliza assim como a versão do ambiente de desenvolvimento de Java produziu um novo conjunto de problemas. A solução deste problema passou pela adopção de diferentes versões da mesma biblioteca em diferentes módulos do Beancounter.

A interacção com as diferentes bibliotecas de interface apresentou novas dificuldades, nomeadamente, a interacção com a Freebase, que está limitada a 10 pedidos por segundo, e com a IMDb, que também limita o número de pedidos efectuados por hora.

7.3 Desenvolvimentos Futuros

O mapeamento entre instâncias, *i.e.*, vectores de características de ontologias diferentes não foi implementado. Esta abordagem permitiria em tempo de execução encontrar correspondências entre categorias não mapeadas através da busca de sinónimos na WordNet restringidos às categorias com pesos idênticos nos vectores de instâncias. Estes resultados permitiriam aumentar a percentagem de mapeamento entre as categorias das ontologias, conduzindo a uma recomendação mais fiável.

No que respeita ao perfil do utilizador, a construção de *tag clouds* pessoais a partir das *tags* adoptadas pelos utilizadores nas redes sociais para descrever itens, poderiam constituir uma nova fonte de informação a considerar. Faltou ainda estudar o efeito da calibração dos pesos das diferentes componentes do perfil e do seu impacto nas recomendações efectuadas aos utilizadores.

A influência do contexto no Sistema de Recomendação desenvolvido é reduzida, sendo apenas considerados o contexto temporal (a data) e espacial (o local).

A utilização desta informação pode ser reforçada através da definição e aplicação de estereótipos de grelhas de recomendação. Estes estereótipos podem ser criados em função do género e idade do espectador, do período do dia, do dia da semana ou da estação do ano. No entanto, o local, a cultura em que está inserido o utilizador podem ser igualmente relevantes no âmbito da geração de recomendações.

O agrupamento em *clusters* dos utilizadores assim como a adopção de um mecanismo de reutilização de resultados permitiria melhorar o desempenho do Sistema de Recomendação, libertando recursos e aumentando a rapidez de resposta. No primeiro caso, os utilizadores de um *cluster* partilhariam a mesma grelha personalizada e, no segundo caso, o *caching* de resultados, *e.g.*, o pedido dos programas diários de um dado género da BBC, permitiria reutilizar resultados que são idênticos para todos os utilizadores.

Dadas as limitações que algumas das fontes de dados utilizadas colocam ao número de pedidos por unidade de tempo (hora ou dia), seria importante efectuar a gestão parcimoniosa das respectivas interacções e, sempre que possível, promover a reutilização de resultados.

Por último, falta efectuar a análise da percepção e validação do sistema por parte dos utilizadores.

7.4 Conclusão

O protótipo desenvolvido gera recomendações personalizadas e os serviços implementados são integráveis com a plataforma B2B em curso. A quantidade, heterogeneidade e constante aumento da informação disponível, constituem um permanente desafio. Isto faz com que o protótipo desenvolvido posso ser melhorado assegurando que as recomendações englobam todas as características do utilizador.

O trabalho desenvolvido apresenta algumas limitações. O mapeamento das categorias, nomeadamente do Facebook, pode ser melhorado. Por outro lado, a ferramenta de enriquecimento de texto utilizada não processa texto em Português pelo que os utilizadores deverão utilizar apenas a língua inglesa. O melhoramento da recomendação deverá incluir também as preferências explícitas dos programas recomendados. Neste trabalho apenas estão a ser contempladas as preferências de alto valor. Por fim, a interacção com os repositórios abertos de dados interligados poderia ser uniformizada através de *queries* SPARQL federadas.

Contudo, o presente protótipo adoptou uma abordagem, ferramentas e tecnologias que permitem desenvolver novas funcionalidade e acrescentar novos dados ao perfil do utilizador, constituindo-se assim uma plataforma de teste de criação de perfis e de geração de recomendações personalizadas.

Anexos

Anexo A

Ontologia do Perfil do Utilizador

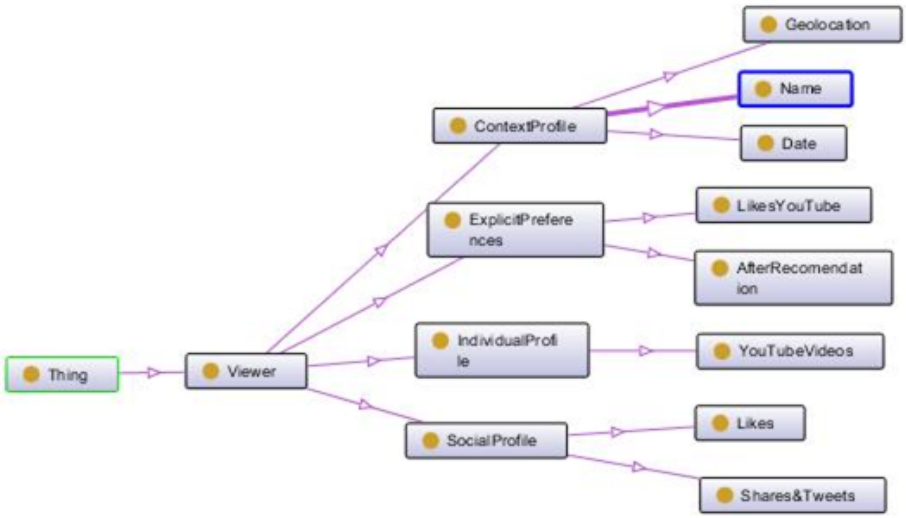

Figura A.1: Ontologia do Perfil do Utilizador.

Anexo B

Ontologia das Categorias da BBC

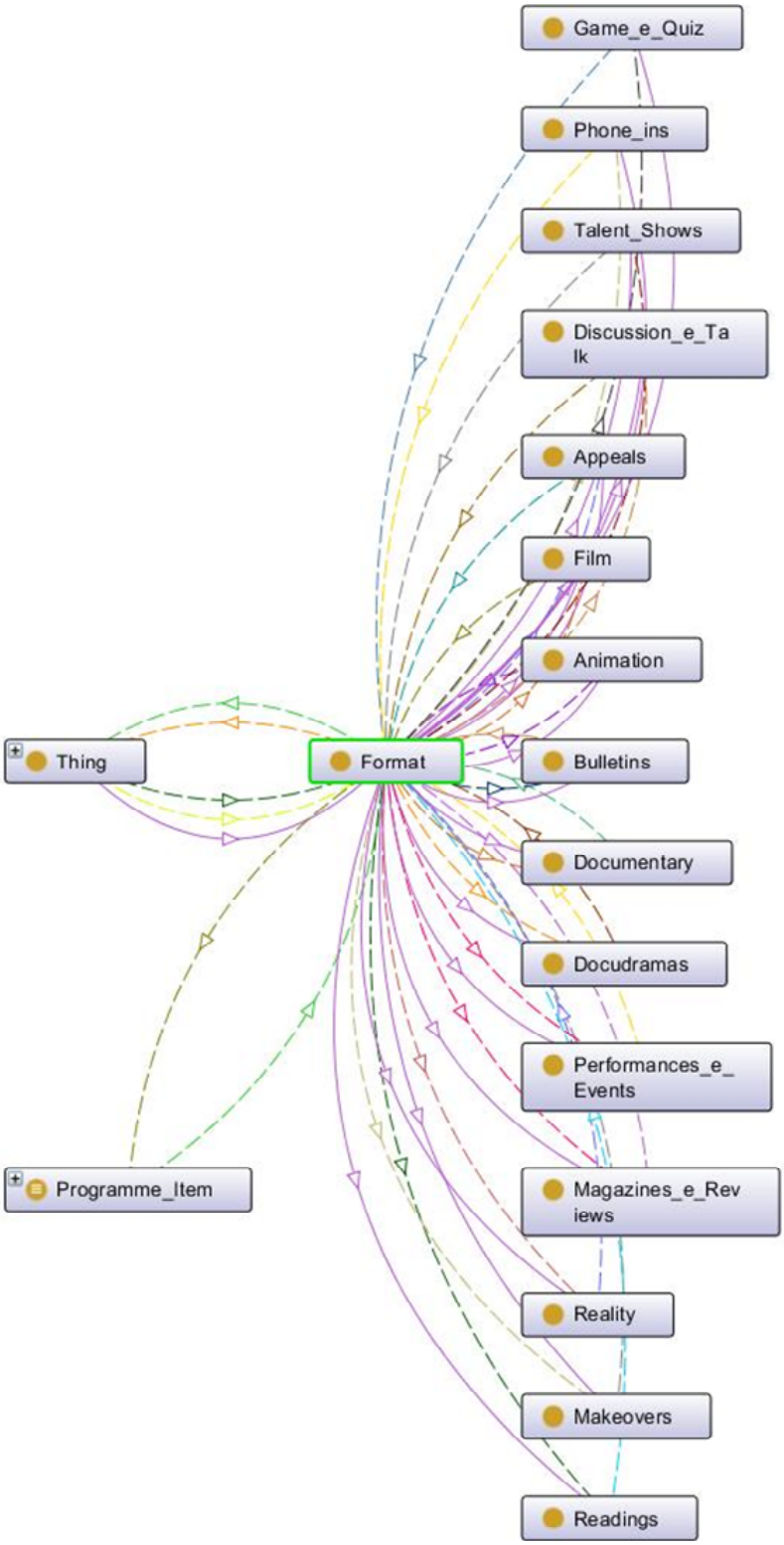

Figura B.1: Ontologia das Categorias da BBC: Formatos.

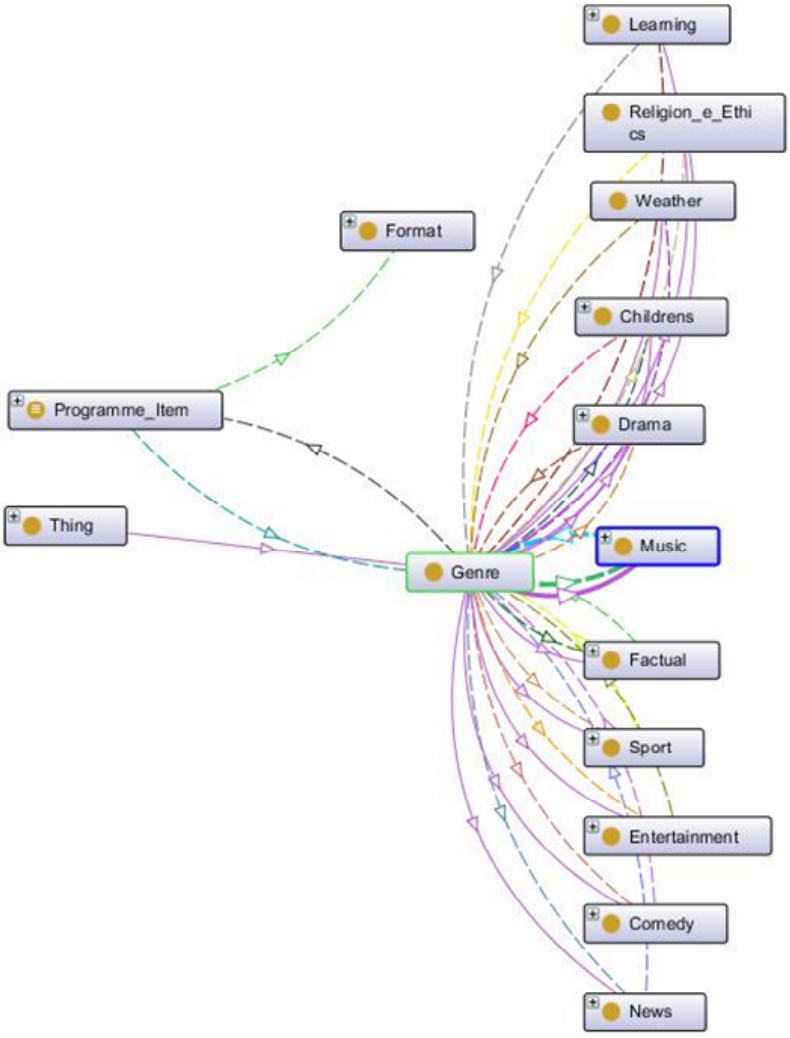

Figura B.2: Ontologia das Categorias da BBC: Géneros.

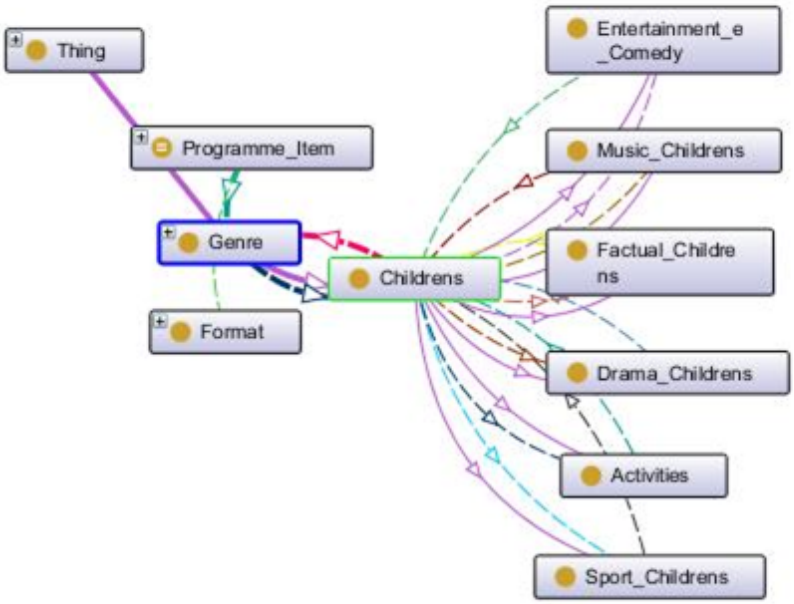

Figura B.3: Ontologia das Categorias da BBC: Subníveis do género "Childrens".

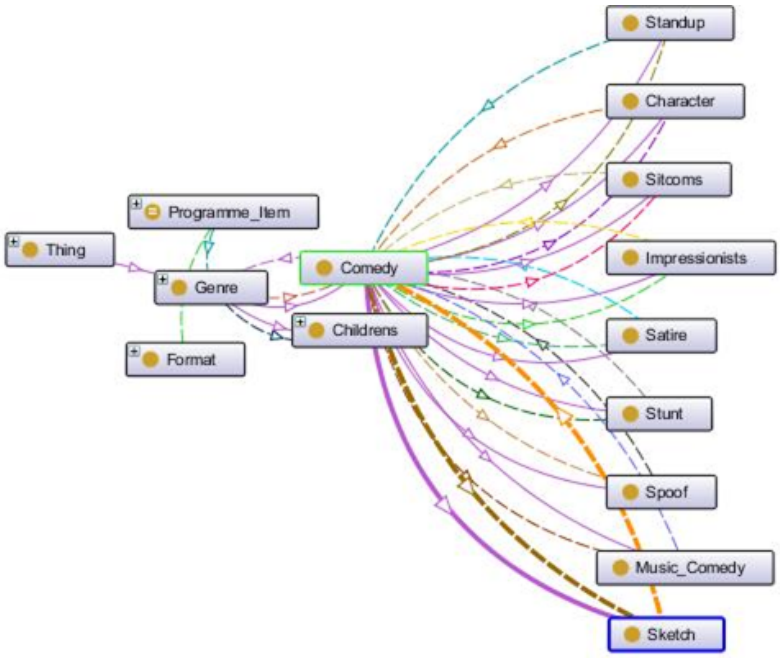

Figura B.4: Ontologia das Categorias da BBC: Subníveis do género "Comedy".

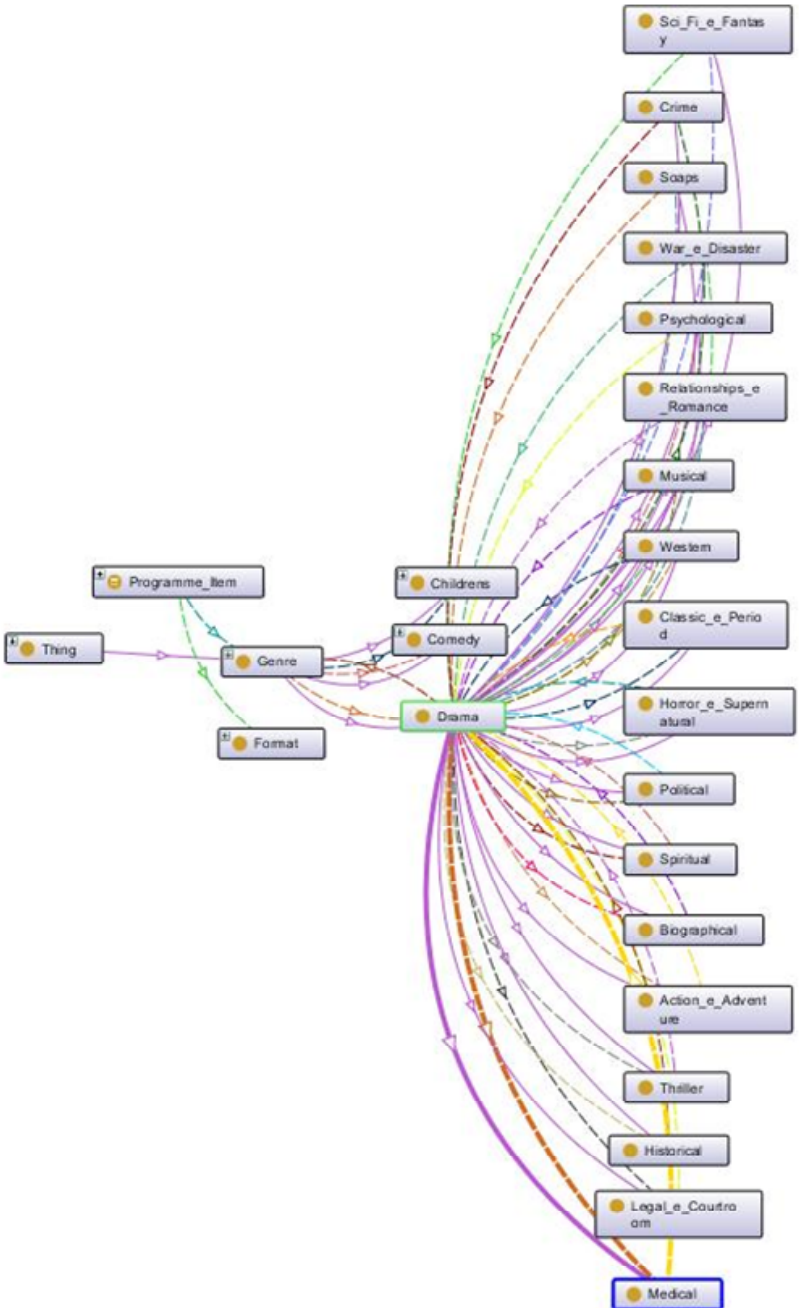

Figura B.5: Ontologia das Categorias da BBC: Subníveis do género "Drama".

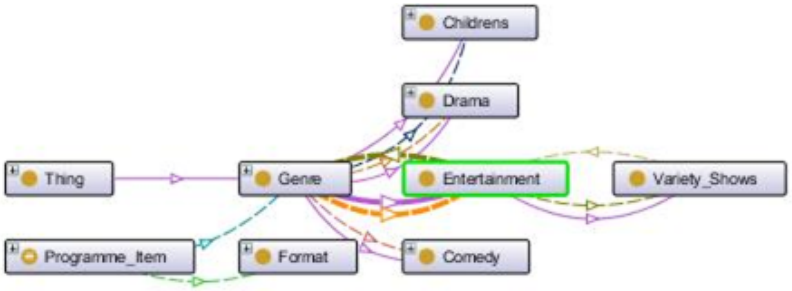

Figura B.6: Ontologia das Categorias da BBC: Subníveis do género "Entertainment".

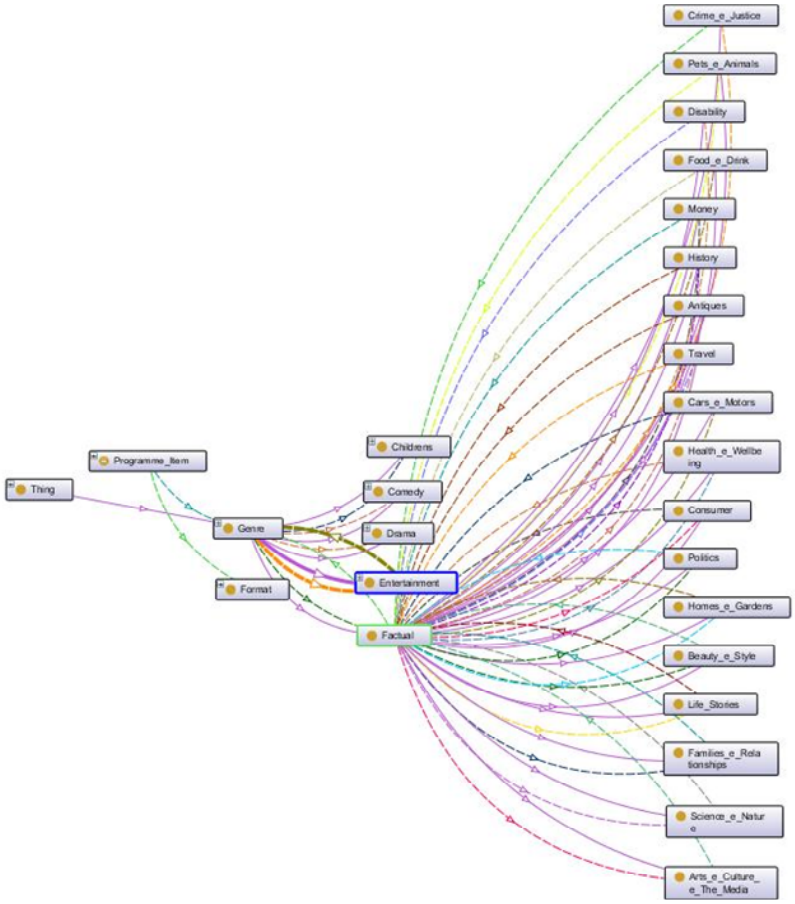

Figura B.7: Ontologia das Categorias da BBC: Subníveis do género "Factual".

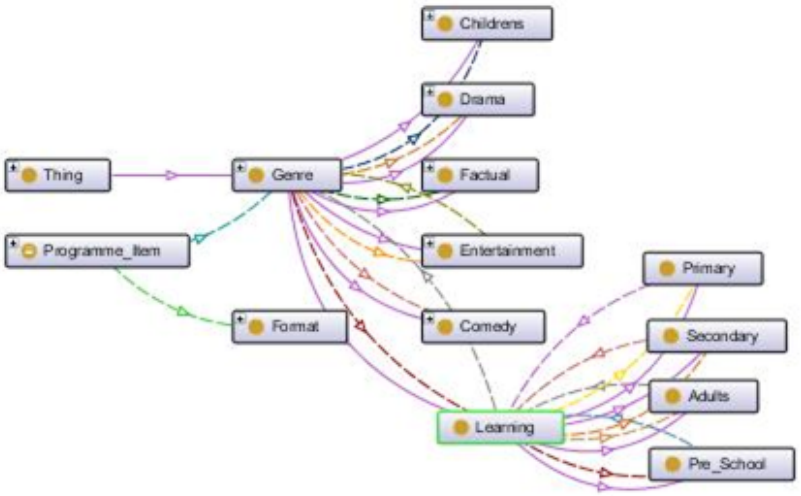

Figura B.8: Ontologia das Categorias da BBC: Subníveis do género “Learning”.

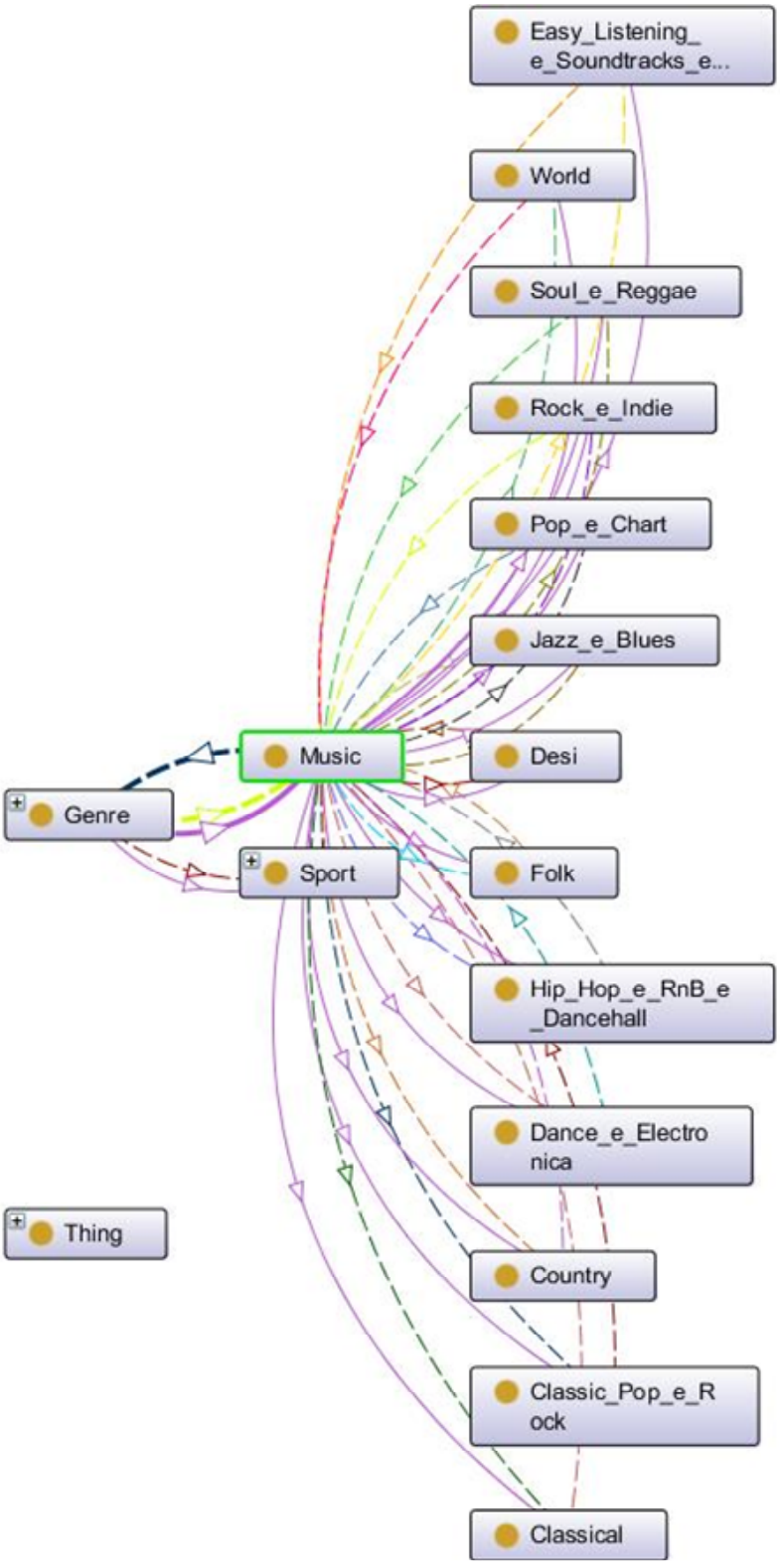

Figura B.9: Ontologia das Categorias da BBC: Subníveis do género "Music".

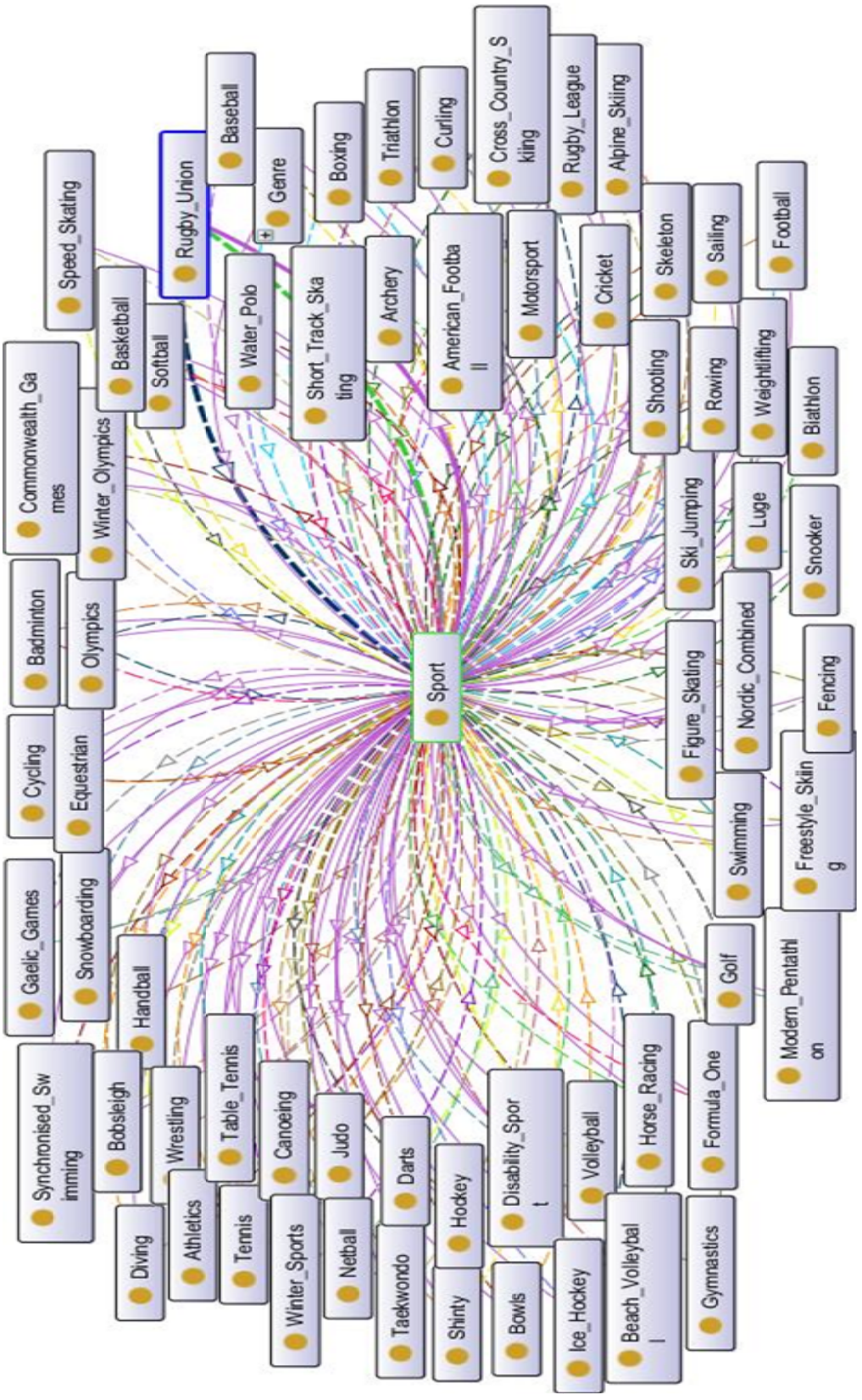

Figura B.10: Ontologia das Categorias da BBC: Subníveis do género "Sport".

Anexo C

Ontologia das Categorias do IMDb

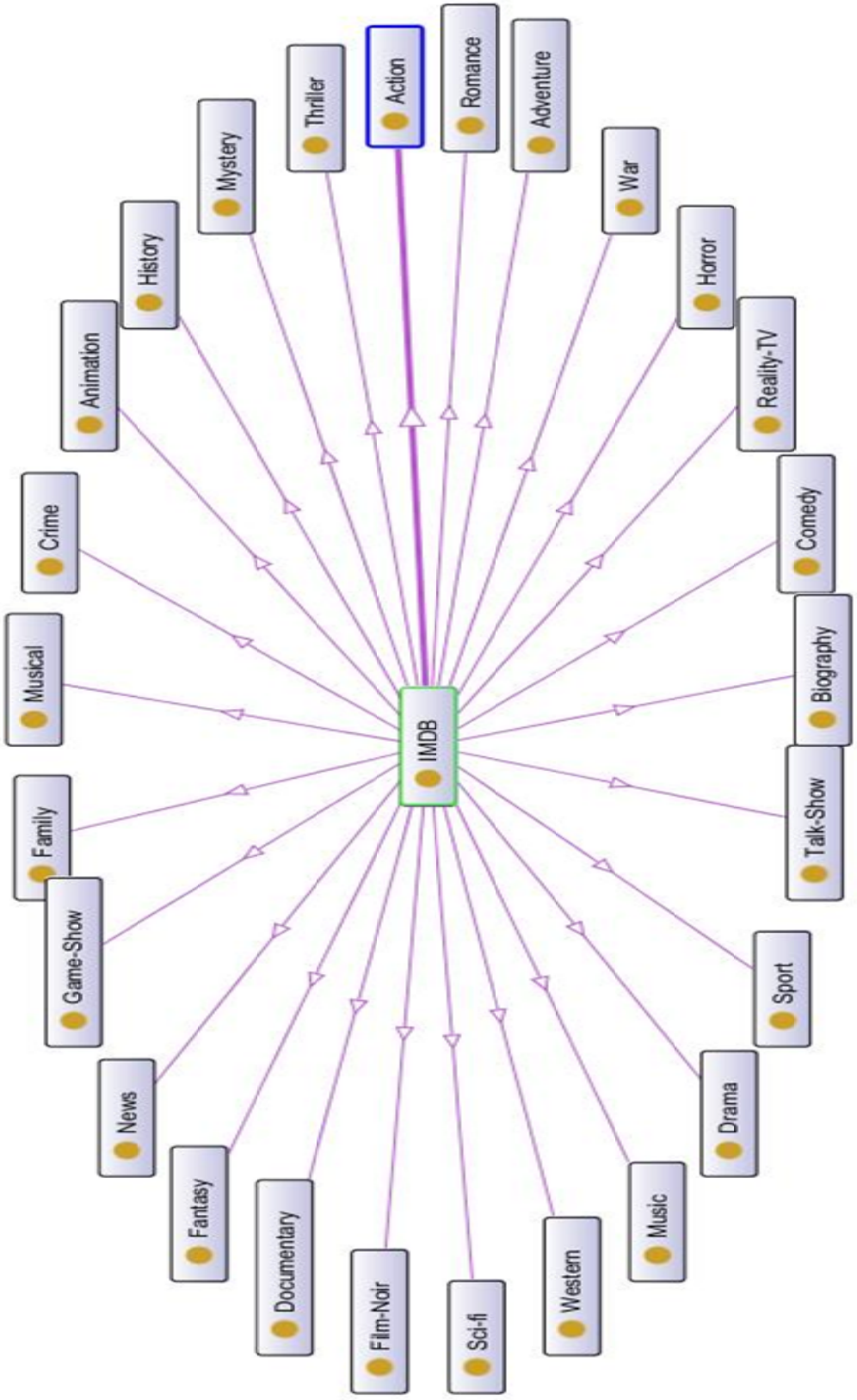

Figura C.1: Ontologia das Categorias do IMDb.

Anexo D

Ontologia das Categorias do YouTube

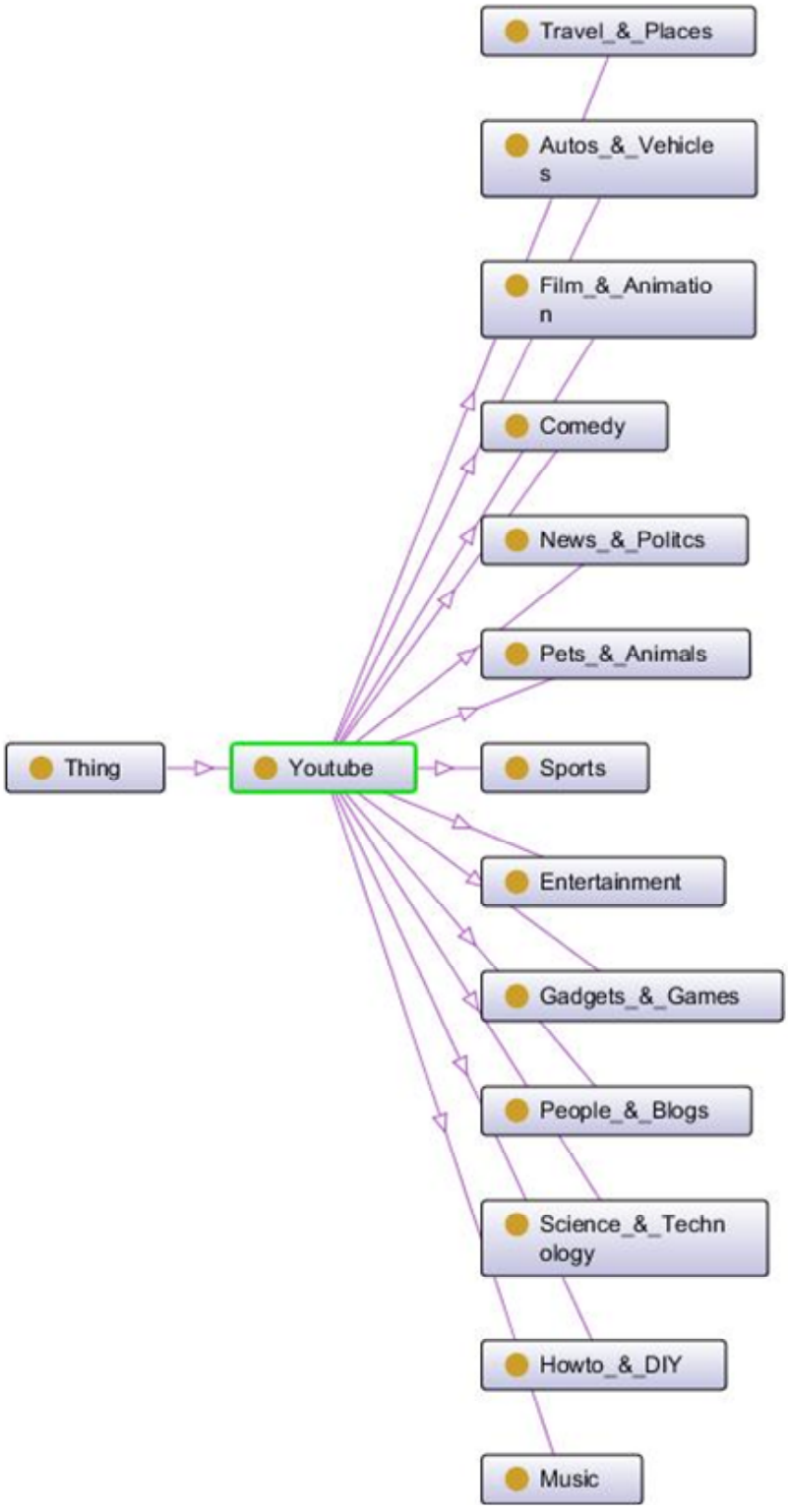

Figura D.1: Ontologia das Categorias do YouTube.

Anexo E

Mapeamento IMDb - BBC

Figura E.1: Mapeamento de Categorias IMDb para BBC utilizando a WordNet.

Anexo F

Mapeamento YouTube - BBC

Figura F.1: Mapeamento de Categorias YouTube para BBC utilizando a WordNet.

Anexo G

Mapeamento Facebook - BBC

Categoria Facebook	Categoria BBC
Actor/Director	Drama
Aerospace/Defense	Drama
Airport	Drama
Album	Music
Amateur Sports Team	Sport
App	Learning
Appliances	Learning
Artist	Music
Arts/Entertainment/Nightlife	Enterteinment
Athlete	Sport
Attractions/Things to Do	Factual
Author	Drama
Automobiles and Parts	Factual
Automotive	Factual
Baby Goods/Kids Goods	Children's
Bags/Luggage	Factual
Bank/Financial Institution	Factual

Continued on next page

Bank/Financial Services	Factual
Bar	Entertainment
Biotechnology	Learning
Book	Learning
Book Store	Learning
Building Materials	Factual
Business Person	Factual
Business Services	Factual
Camera/Photo	Entertainment
Cars	Factual
Cause	Factual
Chef	Factual
Chemicals	Factual
Church/Religious Organization	Religion
Clothing	Factual
Club	Sport
Coach	Sport
Comedian	Comedy
Commercial Equipment	Factual
Community Organization	Factual
Community/Government	Factual
Company	Factual
Computers	Learning
Computers/Technology	Learning
Concert Tour	Music
Concert Venue	Music
Consulting/Business Services	Factual
Dancer	Music
Doctor	Factual
Drugs	Factual

Continued on next page

Editor	Learning
Education	Learning
Electronics	Learning
Energy/Utility	Factual
Engineering/Construction	Factual
Entertainer	Entertainment
Event Planning/Event Services	Entertainment
Farming/Agriculture	Factual
Fictional Character	Factual
Food/Beverages	Factual
Food/Grocery	Factual
Furniture	Factual
Games/Toys	Children's
Government Official	Factual
Government Organization	Factual
Health/Beauty	Factual
Health/Medical/Pharmaceuticals	Factual
Health/Medical/Pharmacy	Factual
Home Decor	Factual
Home Improvement	Factual
Hospital/Clinic	Factual
Hotel	Factual
Household Supplies	Factual
Industrials	Factual
Insurance Company	Factual
Internet/Software	Learning
Jewelry/Watches	Factual
Journalist	News
Kitchen/Cooking	Factual

Continued on next page

Landmark	Factual
Lawyer	Factual
Legal/Law	Factual
Library	Learning
Local Business	Factual
Magazine	News
Media/News/Publishing	News
Mining/Materials	Factual
Monarch	Factual
Movie	Drama
Movie Theater	Drama
Movies/Music	Music
Museum/Art Gallery	Drama
Music Award	Music
Music Chart	Music
Music Video	Music
Musical Instrument	Music
Musician/Band	Music
News Personality	News
Non-Governmental Organization (NGO)	News
Non-Profit Organization	News
Office Supplies	Factual
Organization	Factual
Outdoor Gear/Sporting Goods	Sport
Patio/Garden	Factual
Pet Services	Factual
Pet Supplies	Factual
Playlist	Music
Political Organization	Factual
Political Party	Factual

Continued on next page

Politician	Factual
Producer	Factual
Product/Service	Factual
Professional Services	Factual
Professional Sports Team	Sport
Public Figure	Factual
Public Places	Factual
Radio Station	Music
Real Estate	Factual
Record Label	Factual
Restaurant/Cafe	Factual
Retail and Consumer Merchandise	Factual
School	Learning
School Sports Team	Sport
Shopping/Retail	Factual
Small Business	Factual
Software	Learning
Song	Music
Spas/Beauty/Personal Care	Factual
Sports League	Sport
Sports Venue	Sport
Sports/Recreation/Activities	Sport
Studio	Music
TV Channel	Entertainment
TV Network	Entertainment
TV Show	Entertainment
TV/Movie Award	Entertainment
Teacher	Learning
Telecommunication	Learning

Continued on next page

Tools/Equipment	Factual
Tours/Sightseeing	Factual
Transit Stop	Factual
Transport/Freight	Factual
Transportation	Factual
Travel/Leisure	Factual
University	Learning
Vitamins/Supplements	Factual
Website	Learning
Wine/Spirits	Factual
Writer	Learning

Bibliografia

[1] M. Rovira, J. Gonzalez, A. Lopez, J. Mas, A. Puig, J. Fabregat, and G. Fernandez, "Indextv: a mpeg-7 based personalized recommendation system for digital tv," in *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, vol. 2, 2004, pp. 823–826. [citado na p. vii, 27]

[2] T. Heath and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space*, 1st ed. Morgan & Claypool, 2011. [Online]. Disponível em: http://linkeddatabook.com/ [citado na p. vii, 43]

[3] R. Burke, "Hybrid recommender systems: Survey and experiments," *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331–370, Novembro 2002. [Online]. Disponível em: http://dx.doi.org/10.1023/A:1021240730564 [citado na p. xi, 5, 11]

[4] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, no. 6, pp. 734 – 749, Junho 2005. [citado na p. xi, 11, 12]

[5] BBC. (2013) Programmes genres. [Online]. Disponível em: http://www.bbc.co.uk/programmes/genres [Acedido em: Março 2013] [citado na p. xi, 35]

[6] BBC. (2013) Programmes formats. [Online]. Disponível em: http://www.bbc.co.uk/programmes/formats [Acedido em: Março 2013] [citado na p. xi, 35]

[7] IMDb. (2013) Genres. [Online]. Disponível em: http://www.imdb.com/genre/ [Acedido em: Março 2013] [citado na p. xi, 36]

[8] Desktop Vídeo. (2013) Youtube categories. [Online]. Disponível em: http://desktopvídeo.about.com/od/watchingonlinevídeo/ss/watchyoutube_4.htm [Acedido em: Março 2013] [citado na p. xi, 37]

[9] Open Knowledge Foundation. (2013) Datahub. [Online]. Disponível em: http://datahub.io/dataset [Acedido em: Outubro 2013] [citado na p. xi, 49]

[10] L. M. G. V. de Sousa, "Plataforma multiagente para transacção de componentes multimédia," Master's thesis, Instituto Superior de Engenharia do Porto, Instituto Politécnico do Porto, Novembro 2012. [citado na p. 1]

[11] B. M. D. Veloso, "Transacção de componentes multimédia suportada por agentes," Master's thesis, Instituto Superior de Engenharia do Porto, Instituto Politécnico do Porto, Novembro 2012. [citado na p. 1]

[12] R. D. Burke, A. Felfernig, and M. H. Göker, "Recommender systems: An overview," *AI Magazine*, vol. 32, no. 3, pp. 13–18, 2011. [citado na p. 5]

[13] F. J. Martin, J. Donaldson, A. Ashenfelter, M. Torrens, and R. Hangartner, "The big promise of recommender systems," *AI Magazine*, vol. 32, no. 3, pp. 19–27, 2011. [citado na p. 6, 7]

[14] M. J. Pazzani and D. Billsus, "The adaptive web," P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Berlin, Heidelberg: Springer-Verlag, 2007, ch. Content-based recommendation systems, pp. 325–341. [Online]. Disponível em: http://dl.acm.org/citation.cfm?id=1768197.1768209 [citado na p. 7, 53, 54]

[15] G. Adomavicius and A. Tuzhilin, "Context-aware recommender systems," in *Proceedings of the 2008 ACM conference on Recommender systems*, ser. RecSys '08. New York, NY, USA: ACM, 2008, pp. 335–336. [Online]. Disponível em: http://doi.acm.org/10.1145/1454008.1454068 [citado na p. 9]

[16] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th international conference on World Wide Web*, ser. WWW '01. New York, NY, USA: ACM, 2001, pp. 285–295. [Online]. Disponível em: http://doi.acm.org/10.1145/371920.372071 [citado na p. 9]

[17] V. M. Filho, "e-recommender: Sistema inteligente de recomendação para comércio eletrónico," Master's thesis, Escola Politécnica de Pernambuco, Novembro 2006. [citado na p. 12]

[18] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Novembro 1975. [Online]. Disponível em: http://doi.acm.org/10.1145/361219.361220 [citado na p. 13]

[19] N. N. Chan, W. Gaaloul, and S. Tata, "A web service recommender system using vector space model and latent semantic indexing," in *Advanced Information Networking and Applications (AINA), 2011 IEEE International Conference on*, mar 2011, pp. 602 –609. [citado na p. 13]

[20] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. [citado na p. 13]

[21] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "Measuring the privacy of user profiles in personalized information systems," *Future Generation Computer Systems*, no. 0, pp. –, 2013. [Online]. Disponível em: http://www.sciencedirect.com/science/article/pii/S0167739X1300006X [citado na p. 14]

[22] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *Proceedings of the 20th international joint conference on Artifical intelligence*, ser. IJCAI'07. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007, pp. 1606–1611. [Online]. Disponível em: http://dl.acm.org/citation.cfm?id=1625275.1625535 [citado na p. 15]

[23] P. Lops, M. Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Springer US, 2011, pp. 73–105. [Online]. Disponível em: http://dx.doi.org/10.1007/978-0-387-85820-3_3 [citado na p. 16, 53]

[24] A. E. U. Baldárrago, "Abordagem de recomendação baseada em conteúdo utilizando ontologia fuzzy de domínio e ontologia crisp de preferência do usuário." Universidade Federal de São Carlos, 2012. [Online]. Disponível em: http://www.bdtd.ufscar.br/htdocs/tedeSimplificado//tde_arquivos/3/TDE-2012-09-11T174610Z-4591/Publico/4477.pdf [citado na p. 16]

[25] M. Eirinaki, H. Lampos, M. Vazirgiannis, and I. Varlamis, "Sewep: Using site semantics and a taxonomy to enhance the web personalization process," 2003, pp. 99–108. [citado na p. 16]

[26] S. E. Middleton, N. R. Shadbolt, and D. C. De Roure, "Ontological user profiling in recommender systems," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 54–88, Janeiro 2004. [Online]. Disponível em: http://doi.acm.org/10.1145/963770.963773 [citado na p. 16]

[27] S. Aciar, D. Zhang, S. Simoff, and J. Debenham, "Informed recommender: Basing recommendations on consumer product reviews," *Intelligent Systems, IEEE*, vol. 22, no. 3, pp. 39 –47, Maio-Junho 2007. [citado na p. 16]

[28] I. Cantador, A. Bellogín, and P. Castells, "News@hand: A semantic web approach to recommending news," in *Adaptive Hypermedia and Adaptive Web-Based Systems*, ser. Lecture Notes in Computer Science,

W. Nejdl, J. Kay, P. Pu, and E. Herder, Eds. Springer Berlin Heidelberg, 2008, vol. 5149, pp. 279–283. [Online]. Disponível em: http://dx.doi.org/10.1007/978-3-540-70987-9_34 [citado na p. 16]

[29] Y. Blanco-Fernandez, J. Pazos-arias, A. Gil-Solla, M. Ramos-Cabrer, and M. Lopez-Nores, "Providing entertainment by content-based filtering and semantic reasoning in intelligent recommender systems," *Consumer Electronics, IEEE Transactions on*, vol. 54, no. 2, pp. 727 –735, Maio 2008. [citado na p. 16]

[30] Princeton University. (2013) Wordnet api reference. [Online]. Disponível em: http://wordnet.princeton.edu [Acedido em: Setembro 2013] [citado na p. 17, 92]

[31] Princeton University. (2013) Big huge thesaurus. [Online]. Disponível em: http://words.bighugelabs.com/ [Acedido em: Novembro 2013] [citado na p. 17]

[32] N. H. Sulaiman and D. Mohamad, "A jaccard-based similarity measure for soft sets," in *Humanities, Science and Engineering Research (SHUSER), 2012 IEEE Symposium on*, 2012, pp. 659–663. [citado na p. 17]

[33] E. E. H. Naw Naw, "Relevant words extraction method for recommendation system," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 111, pp. 680–684, 2013. [citado na p. 19]

[34] F. S. da Silva, "Personaltvware: Uma infraestrutura de suporte a sistemas de recomendação sensíveis ao contexto para tv digital personalizada," Master's thesis, Escola Politécnica de Pernambuco, Novembro 2011. [citado na p. 19]

[35] Red Bee Media. (2013) Reddiscover portfolio. [Online]. Disponível em: http://www.redbeemedia.com/sites/all/files/downloads/reddiscover_datasheet_red_bee_media.pdf [Acedido em: Novembro 2013] [citado na p. 25]

[36] Red Bee Media. (2013) Reddiscover. [Online]. Disponível em: www.redbeemedia.com/services/reddiscover [Acedido em: Novembro 2013] [citado na p. 25]

[37] A. Conconi, F. Cattaneo, R. D. Pero, L. Vignaroli, F. Negro, D. Brickley, L. Miller, V. Buser, M. Minno, D. Palmisano, G. Stoll, R. Zimmermann, M. Riethmayer, S. Dietze, N. Benn, and M. Yankova, "D6.1b notube system specifications and architectural design," 2010. [Online]. Disponível em: http://notube3.files.wordpress.com/2012/01/notube_d6-1-system-specifications-and-architectural-design_final.pdf [Acedido em: Maio de 2013] [citado na p. 26]

[38] R. D. Pero, L. Vignaroli, F. Cattaneo, F. Negro, P. Altendorf, D. Liu, P. Mihaylov, and D. Palmisano, "D6.4 notube integrated system 3rd prototype," 2011. [Online]. Disponível em: http://notube3.files.wordpress.com/2012/01/notube_d6-1-system-specifications-and-architectural-design_final.pdf [Acedido em: Maio de 2013] [citado na p. 26]

[39] D. Tsatsou, V. Mezaris, T. Kliegr, J. Kucha, M. Mancas, L. Nixon, R. Klein, and M. Kober, "User profile schema and profile capturing," 2012. [Online]. Disponível em: http://www.slideshare.net/linkedtv/linked-tv-d42user-profile-schema-and-profile-capturing [citado na p. 26]

[40] D. Tsatsou, M. Loli, V. Mezaris, T. Kliegr, J. Kucha, M. Mancas, J. Leroy, and L. Nixon, "Specification of user profiling and contextualisation," 2012. [Online]. Disponível em: http://www.slideshare.net/linkedtv/linked-tv-d41specificationofuserprofilingandcontextualisation [citado na p. 26]

[41] R. Klein, M. Kober, J. Xie, D. Tsatsou, S. Weinberg, and J. Thomsen, "Content and concept filter v1," 2012. [Online]. Disponível em: http://www.slideshare.net/linkedtv/linked-tv-d43content-and-concept-filter-v1 [citado na p. 26]

[42] S. Hsu, M.-H. Wen, H.-C. Lin, C.-C. Lee, and C.-H. Lee, "Aimed-a personalized tv recommendation system," in *Interactive TV: a Shared Experience*, ser. Lecture Notes in Computer Science, P. Cesar, K. Chorianopoulos, and J. F. Jensen, Eds. Springer Berlin Heidelberg, 2007, vol. 4471, pp. 166–174. [Online]. Disponível em: http://dx.doi.org/10.1007/978-3-540-72559-6_18 [citado na p. 28]

[43] M. L. Nores, J. J. P. Arias, Y. Blanco-Fernández, J. G. Duque, R. Tubio-Pardavila, and E. Casquero-Villacorta, "Mispot: dynamic product placement for digital tv through mpeg-4 processing and semantic reasoning," *Knowl. Inf. Syst.*, vol. 22, no. 1, pp. 101–128, Janeiro 2010. [Online]. Disponível em: http://dx.doi.org/10.1007/s10115-009-0200-8 [citado na p. 28]

[44] ImTV. (2013) Imtv on-demand immersive tv. [Online]. Disponível em: http://imtv.me/ [Acedido em: Maio 2013] [citado na p. 29]

[45] J. Magalhães, S. Strover, T. Chambel, P. Viana, T. Andrade, L. Francisco-Revilla, F. Martins, and N. Correia, "Imtv: Towards an immersive tv experience," in *Future TV Workshop, Adjunct Proceedings of EuroiTV'2012, the 10th European conference on Interactive tv and video*, 2012. [citado na p. 29]

[46] Y. Raimon, T. Gängler, F. Giasson, K. Jacobson, G. Fazekas, S. Reinhardt, and A. Passant. (2013) Music ontology. [Online]. Disponível em: http://www.musicontology.com/ [Acedido em: Outubro 2013] [citado na p. 34]

[47] T. Conlan. (2013) Bbc places next-generation iplayer at heart of digital strategy. [Online]. Disponível em: http://www.theguardian.com/media/2013/oct/06/bbc-next-generation-iplayer-digital [Acedido em: Outubro 2013] [citado na p. 34]

[48] Y. Hu, Z. Wang, W. Wu, J. Guo, and M. Zhang, "Recommendation for movies and stars using yago and imdb," in *Web Conference (APWEB), 2010 12th International Asia-Pacific*, 2010, pp. 123–129. [citado na p. 35]

[49] The eBusiness Knowledgebase. (2013) ebizmba. [Online]. Disponível em: http://www.ebizmba.com/ [Acedido em: Maio 2013] [citado na p. 39]

[50] M. D. anf Joanna Brenner. (2013) The demographics of social media users - 2012. [Online]. Disponível em: http://pewinternet.org/Reports/2013/Social-media-users.aspx [Acedido em: Março de 2013] [citado na p. 40]

[51] Wikimedia Foundation, Inc. (2013) Facebook statistics. [Online]. Disponível em: http://en.wikipedia.org/wiki/Facebook_statistics [Acedido em: Maio 2013] [citado na p. 40]

[52] Wikimedia Foundation, Inc. (2013) Twitter. [Online]. Disponível em: https://en.wikipedia.org/wiki/Twitter [Acedido em: Maio 2013] [citado na p. 40]

[53] T. Di Noia, R. Mirizzi, V. C. Ostuni, D. Romito, and M. Zanker, "Linked open data to support content-based recommender systems," in *Proceedings of the 8th International Conference on Semantic Systems*, ser. I-SEMANTICS '12. New York, NY, USA: ACM, 2012, pp. 1–8. [Online]. Disponível em: http://doi.acm.org/10.1145/2362499.2362501 [citado na p. 41]

[54] G. A. Miller, "Wordnet: a lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Novembro 1995. [Online]. Disponível em: http://doi.acm.org/10.1145/219717.219748 [citado na p. 41]

[55] Linked Data Community. (2013) Linked data - connect distributed data across the web. [Online]. Disponível em: http://linkeddata.org/ [Acedido em: Outubro 2013] [citado na p. 42]

[56] L. Yu, *A Developer's Guide to the Semantic Web.* Springer, 2011. [citado na p. 42]

[57] B. V. Keong and P. Anthony, "Meta search engine powered by dbpedia," in *Semantic Technology and Information Retrieval (STAIR), 2011 International Conference on*, 2011, pp. 89–93. [citado na p. 43]

[58] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human

knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, ser. SIGMOD '08. New York, NY, USA: ACM, 2008, pp. 1247–1250. [Online]. Disponível em: http://doi.acm.org/10.1145/1376616.1376746 [citado na p. 44]

[59] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th international conference on World Wide Web*, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 697–706. [Online]. Disponível em: http://doi.acm.org/10.1145/1242572.1242667 [citado na p. 44]

[60] M. Ghazanfar and A. Prugel-Bennett, "An improved switching hybrid recommender system using naive bayes classifier and collaborative filtering," in *The 2010 IAENG International Conference on Data Mining and Applications*, apr 2010, event Dates: 17-19 March, 2010. [Online]. Disponível em: http://eprints.soton.ac.uk/268483/ [citado na p. 53]

[61] M. Pazzani and D. Billsus, "Learning and revising user profiles: The identification ofinteresting web sites," *Mach. Learn.*, vol. 27, no. 3, pp. 313–331, Junho 1997. [Online]. Disponível em: http://dx.doi.org/10.1023/A:1007369909943 [citado na p. 53]

[62] L. Zhang, N. Ye, W. Zhou, and L. Jiao, "Support vectors pre-extracting for support vector machine based on k nearest neighbour method," in *Information and Automation, 2008. ICIA 2008. International Conference on*, 2008, pp. 1353–1358. [citado na p. 54]

[63] A. Abdelhalim and I. Traore, "A new method for learning decision trees from rules," in *Machine Learning and Applications, 2009. ICMLA '09. International Conference on*, 2009, pp. 693–698. [citado na p. 54]

[64] "Introduction to artificial neural networks," in *Electronic Technology Directions to the Year 2000, 1995. Proceedings.*, 1995, pp. 36–62. [citado na p. 54]

[65] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Agosto 2009. [Online]. Disponível em: http://dx.doi.org/10.1109/MC.2009.263 [citado na p. 55]

[66] P. R. G. Dias, "Recommending media content based on machine learning methods," Master's thesis, Faculdade de Ciência e Tecnologia da Universidade Nova de Lisboa, Novembro 2011. [citado na p. 55]

[67] D. Heckmann, T. Schwartz, B. Brandherm, M. Schmitz, and M. Wilamowitz-Moellendorff, "Gumo the general user model ontology," in *User Modeling*

2005, ser. Lecture Notes in Computer Science, L. Ardissono, P. Brna, and A. Mitrovic, Eds. Springer Berlin Heidelberg, vol. 3538, pp. 428–432. [Online]. Disponível em: http://dx.doi.org/10.1007/11527886_58 [citado na p. 55]

[68] LinkedTV. (2013) Linkedtv user model ontology. [Online]. Disponível em: http://data.linkedtv.eu/ontologies/lumo/ [Acedido em: Maio 2013] [citado na p. 55]

[69] Stanford Center for Biomedical Informatics Research. (2013) Protege. [Online]. Disponível em: http://protege.stanford.edu/ [Acedido em: Maio 2013] [citado na p. 63]

[70] R. M. Lerner, "At the forge: Redis," *Linux J.*, vol. 2010, no. 197, Setembro 2010. [Online]. Disponível em: http://dl.acm.org/citation.cfm?id=1883519.1883524 [citado na p. 64]

[71] R. Pointer. (2013) Kestrel. [Online]. Disponível em: http://robey.github.io/kestrel/readme.html [Acedido em: Maio 2013] [citado na p. 64]

[72] Google Developers. (2013) Youtube api reference. [Online]. Disponível em: https://developers.google.com/youtube/v3/docs/ [Acedido em: Agosto 2013] [citado na p. 75]

[73] NoTube. (2013) Notube beancounter. [Online]. Disponível em: https://github.com/dpalmisano/NoTube-Beancounter-2.0 [Acedido em: Maio 2013] [citado na p. 81]

[74] IMDb. (2013) Application programming interface. [Online]. Disponível em: http://www.imdb.com/interfaces [Acedido em: Agosto 2013] [citado na p. 90]

[75] NoTube. (2013) User interface for beancounter. [Online]. Disponível em: https://github.com/notube/beancounter_ui [Acedido em: Março 2013] [citado na p. 105]