Gestão de Sistemas e Computação

Modelação de Dados na Web 2.0
Aplicada ao Domínio Hoteleiro

Trabalho Final – Laboratório de GSC Aplicado

Elaborado por Eduardo João Nixa Santos

Discente n.º 20071030

Orientador: Prof. Doutor Marcirio Silveira Chaves

Barcarena – Oeiras, Novembro de 2011

Universidade Atlântica

Gestão de Sistemas e Computação

Modelação de Dados na Web 2.0
Aplicada ao Domínio Hoteleiro

Trabalho Final – Laboratório de GSC Aplicado

Elaborado por Eduardo João Nixa Santos

Discente n.º 20071030

Orientador: Prof. Doutor Marcirio Silveira Chaves

Barcarena – Oeiras, Novembro de 2011

O autor é o único responsável pelas ideias expressas neste relatório.

Agradecimentos

Muitos foram aqueles que contribuíram de alguma forma para concluir com sucesso mais esta importantíssima etapa da minha vida. A conclusão duma licenciatura é para mim, essencialmente, um marco de realização pessoal, que poderá marcar igualmente a evolução da minha carreira profissional, de há vinte e dois anos a esta parte, desenvolvida na área dos Sistemas de Informação.

Quero neste momento deixar expressos os meus agradecimentos a todos quantos de alguma forma contribuíram para a chegada deste momento, e mais em particular:

- Ao meu Pai, Fernando Manuel da Costa Santos, pelo exemplo de vida, de perseverança e honestidade, e por todos os demais valores que sempre tentou passar-me, fazendo de mim a pessoa que hoje sou;
- A todos os meus Amigos, sendo que a Amizade é para mim o valor mais nobre de todos;
- À minha vasta família, que adoro;
- A todos os meus Professores, sem excepção, desde a primeira, Gertrudes Rebelo, ao Professor que me orientou neste Projecto, e que muito me ajudou à sua conclusão, Marcírio Silveira Chaves;
- À minha companheira, Ana Isabel Ramos, pelo amor, apoio e compreensão;
- Aos nossos gatos, Nenúfar e Cereja, pela companhia;
- Ao meu actual empregador, BCM-Bricolage S.A. (Leroy Merlin Portugal), por todo o apoio e ajuda facultadas durante a minha Licenciatura, na pessoa do meu responsável hierárquico, e amigo pessoal, Fernando Gonçalves;
- À Turma de Gestão de Sistemas e Computação de 2007 da Universidade Atlântica, pela amizade, união e entreajuda sempre patenteadas;
- À Universidade Atlântica.

A todos um grande bem-hajam, e o meu muito obrigado.

Abstract

Internet hosts a large and ever growing amount of information, yet not meaning its availability on a structured basis, allowing users its easy access or use.

Reviews written by users on Web 2.0 platforms, describing their experiences on products or services' use, relies a rich base of information that can become essential when making a decision is necessary. However, it seems there are no free available tools to process that information.

This work presents a model for building a knowledge base that allows accessing that information, filtered by the main subjects and sub-subjects of interest, aiming to obtain results able to support decision making.

Key-Words: internet, web 2.0, reviews, meta-model, modeling.

Resumo

A Internet alberga uma vasta quantidade de informação, em constante crescimento, não significando que ela se encontra acessível de forma estruturada, permitindo aos utilizadores fácil acesso e utilização.

O registo de avaliações por parte dos utilizadores em plataformas na Web 2.0, descrevendo por escrito as suas experiências decorrentes da utilização de bens, produtos ou serviços, constitui uma base rica em informação, que se pode tornar essencial para a tomada de decisão. No entanto, no melhor do nosso conhecimento, não existe nenhuma ferramenta gratuita disponível para processar essa informação.

Este trabalho apresenta um modelo para a constituição de uma base de conhecimento que permita consultar essa informação, filtrada por temas e subtemas do seu interesse, com vista à obtenção de resultados que facilitem e sustentem a sua tomada de decisão.

Palavras-Chave: internet, web 2.0, avaliações online, meta-modelo, modelação.

Índice

1. Introdução

Uma das vertentes da internet que mais cresce nos tempos actuais é sem dúvida a disponibilização online de opiniões sobre da aquisição de serviços e/ou produtos. Como resultado, vários utilizadores proferem comentários descritivos das suas experiências, avaliando os produtos e classificando os serviços em função das suas características. No entanto, a análise dessa informação, numa perspectiva duma ferramenta de apoio à decisão, é manifestamente complexa e apresenta várias condicionantes importantes, incluindo:

- A análise pode recair sobre centenas, ou mesmo milhares de avaliações expressas, causando potencial de dispersão;
- A diversidade global dos idiomas utilizados;
- Vasta diversidade temática nas características avaliadas;
- Impossibilidade de consultar informação de forma selectiva, sob vários pontos de vista (tipo de cliente, faixa etária, género, nacionalidade, etc.);
- Impossibilidade de analisar estritamente aquelas variáveis de maior interesse para o cliente, que procura ajuda para a sua tomada de decisão.

Este trabalho apresenta um modelo para utilização de uma ferramenta informática que tem por objectivo o armazenamento de informação recolhida a partir da Web 2.0, para criação de uma base de conhecimento que possa, ela própria, constituir-se como base para uma ferramenta de apoio à decisão para clientes e potenciais clientes.

Os principais objectivos subjacentes a este trabalho são:

- Criar um modelo conceptual que seja possível aplicar de forma generalizada à problemática das avaliações online;
- Validar o modelo conceptual, aplicando-o especificamente a um sector das avaliações online, o sector do alojamento no domínio hoteleiro, e realizando a respectiva prova de conceito.

Neste trabalho, a modelação de dados é arquitectada de forma que o registo e tratamento das avaliações seja amplamente multidisciplinar. Isto é, a solução proposta deverá evoluir para que se constitua numa ferramenta de apoio à decisão no que toca à aquisição duma panóplia de bens e serviços, do mais variado tipo de produtos, que podem ir desde os livros ao software, do vestuário desportivo aos bens alimentares, entre outros.

Numa fase embrionária, o presente projecto trata um ramo específico das avaliações online, as avaliações registadas sobre o domínio hoteleiro.

A estrutura deste trabalho é a seguinte:

- No Capítulo II encontra-se uma análise da literatura subjacente aos temas abordados no trabalho;
- No Capítulo III é apresentado o Meta-Modelo e a sua aplicação concreta ao caso do objecto "Alojamento", com os respectivos Diagramas de Classes, subdivididos nos seus principais domínios;
- No Capítulo IV são apresentadas as consultas levadas a cabo com a finalidade de realizar uma Prova de Conceito da solução implementada;
- No Capítulo V são apresentadas as conclusões do presente trabalho, juntamente com as limitações e trabalhos futuros.

2. Análise da Literatura

A partir da literatura consultada durante a realização deste trabalho, fez-se uma análise que permite dissertar sobre alguns pontos de maior interesse para a problemática abordada neste trabalho.

2.1. As Avaliações Online no Processo de Tomada de Decisão

Existem hoje centenas de portais online, onde os utilizadores podem deixar expressa a sua opinião, uma crítica normalmente com polaridade positiva, mista ou negativa, decorrente de uma sua experiência de aquisição ou utilização de um bem ou de um serviço. Daí resultam milhões de registos de avaliações, informação que pode e deve ser utilizada para sustentar uma tomada de decisão com vista às opções por parte de futuros utilizadores.

No entanto, parece não estar ainda disponível aos utilizadores uma ferramenta que lhes permita consultar essa informação de forma estruturada, possibilitando-lhes uma filtragem por determinados temas ou subtemas específicos, que resulte numa consulta mais granular da informação, ou seja, não tão quantitativa, mas sim mais qualitativa.

Quanto ao gestor do sistema, este terá que ter sempre em mente tudo aquilo que possa assistir uma tomada de decisão por parte do utilizador, devendo ser sempre esse o seu principal vector de focalização: qual a melhor forma de facilitar a tomada de decisão aos utilizadores do sistema?

2.2. Ferramentas Comerciais que Processam Conteúdo da Web 2.0

Foram analisadas algumas ferramentas comerciais existentes no mercado, na área da análise e interpretação de texto. Constatou-se que todas as ferramentas analisadas possuem custos dispendiosos de licenciamento, apenas ao alcance de grandes empresas. Mesmo a nível empresarial, dando como exemplo o actual panorama português, em que mais de 90% do tecido empresarial é composto por micro, ou pequenas e médias empresas (PME), o investimento financeiro numa destas soluções muito dificilmente obteria um retorno que o justificasse.

Algumas das ferramentas mais conhecidas nesta área são:

- Clarabridge (www.clarabridge.com);
- Sentiment Metrics (www.sentimentmetrics.com);
- Attensity (www.attensity.com);
- Sentimetrix (www.sentimetrix.com);
- Radian6 (www.radian6.com).

Todas estas ferramentas permitem a análise selecção e tratamento de dados "amontoados", dispersos e não estruturados na Web 2.0, com o intuito de serem retirados preciosos indicadores de gestão, que suportem acções e decisões futuras ao nível da gestão. Pecam, na sua larga maioria, por não estarem financeiramente ao alcance dos utilizadores mais comuns, ou seja, não empresariais.

Estas ferramentas não só se encontram inacessíveis financeiramente na maioria dos casos, como na sua generalidade não possuem um GUI (*graphical user interface*) de fácil utilização, mas antes complexas consolas de gestão, cuja utilização requer um outro nível de conhecimentos que, muito naturalmente, o utilizador comum poderá não deter.

Para se ter uma ideia mais concreta da dimensão dos clientes-tipo destas ferramentas comerciais, encontrámos as seguintes empresas listadas como clientes, nos *websites* acima referidos: Airbus, AOL, Fujitsu, Nissan, Siemens, Sony.

Optou-se pela não realização de uma comparação directa entre as cinco ferramentas atrás referidas, pois a informação disponível sobre cada uma delas é bastante dispersa e pouco conclusiva. Todas elas remetem a obtenção de mais informação para um pedido de informações, ou a realização de uma consulta comercial que possa resultar na apresentação de uma proposta comercial formal.

No entanto, relativamente a algumas das problemáticas abordadas no presente trabalho, foi possível efectuar algumas comparações. Por exemplo, no que toca à análise da informação com carácter multilíngue, todos estes produtos dão uma resposta pouco flexível, contemplando normalmente apenas entre duas e quatro línguas diferentes (exemplo *Clarabridge*: inglês, francês, castelhano e português).

Já no que toca às fontes de informação que estes produtos podem utilizar, a panóplia de opções contempladas é muito ampla, podendo ir desde todas inúmeras redes sociais hoje existentes (LinkedIn, Facebook, Twitter, etc.) até aos mais variados portais e blogues.

2.3. Consultas Estruturadas Sobre Textos na Web

É enorme, e em constante crescimento, a quantidade de informação disponível na internet, a grande maioria da qual sob a forma de texto não estruturado. Localizar parte desta informação, não oferece hoje em dia muitas dificuldades, recorrendo, por exemplo, à utilização dos actuais e bastante potentes motores de busca (exemplo www.google.com). No entanto, a localização dessa informação, não implica de forma alguma o seu correcto tratamento ou assimilação.

Existem três modelos estruturados de consulta sobre textos na *web*: *Schema Extraction*, *Text Query* e *Extraction Database* (Cafarella *et al.*, 2006). Explicando os três modelos:

- O *Schema Extraction Model* consiste numa análise de texto proveniente da internet, sendo construída uma matriz que explica a informação presente no texto;

- No *Text Query Model* verifica-se numa interacção directa com o utilizador, em que este fornece parte da informação que pretende localizar no texto;
- No *Extraction Database Model* são utilizadas técnicas de extracção de informação para detectar correspondências no texto.

Sendo que, no modelo de solução proposto neste trabalho, o utilizador deve indicar parte relevante da estrutura da consulta, consideramos que o mesmo se insere na abordagem denominada de "*Text Query Model*". Neste modelo não há extracção antecipada da informação, há sim lugar a uma indexação de texto *web* de forma que os utilizadores possam realizar as suas próprias consultas de extracção.

2.4. Meta-modelos e Modelos

A abordagem do problema utilizando a meta-modelação de dados, permite um nível de abstracção que vai possibilitar posteriormente a sua implementação em modelos mais específicos, ajustados caso a caso de acordo com as especificidades e os requisitos de cada objecto.

A expressão "Meta" provém do grego, significa "mais além", e é utilizada em inglês para indicar um conceito que é uma abstracção de outro conceito, utilizado para acrescentar ou completar este último (Wikipedia, n.d.).

Ou seja, no caso da modelação de dados efectuada durante este trabalho, o uso de um meta-modelo possibilita a sua implementação em vários modelos de dados ligeiramente acrescentados ou completados, especificamente de acordo com as necessidades do objecto em análise.

A noção de semântica é importante para o meta-modelo (Infogrid, 2009). Semântica, na sua definição mais lata, é o estudo do significado. No contexto do meta-modelo, pode-se afirmar que a semântica é a necessidade da atribuição de significado ao que se modela.

3. Modelação Conceptual

Neste Capítulo serão explicados o desenho e modelação de dados criados para sustentar uma Base de Conhecimento. A utilização de um meta-modelo de dados possibilita genericamente a sua aplicação a qualquer tipo de objecto em análise. Em seguida, é explicada a sua aplicação na criação de um modelo de dados concreto, o modelo criado para a Base de Conhecimento dos objectos do tipo "Alojamento".

São apresentados os diagramas de classes, cujo desenho e modelação permitiram a implementação da base de dados, bem como os diagramas UML (*Unified Modeling Language*) descritivos do modelo, divididos nos principais domínios da solução. No final deste capítulo é apresentado o diagrama de classes completo, aplicado ao sector do Alojamento.

3.1. Meta-modelo Conceptual, Multidomínio e Multilíngue

A informação disponível na Web 2.0 é bastante heterogénea e varia muito ao nível do detalhe (granularidade) em cada domínio do conhecimento. Para modelar tal conteúdo, é necessária a criação de meta-modelo de dados que possa ser estendido e aplicado a domínios específicos.

O meta-modelo apresentado na Fig. 1 foi idealizado de forma a poder ser aplicado aos mais variados tipos de objectos em análise. A modelação foi desenvolvida em torno da classe *Review*, que é a classe que regista informação relativa às avaliações sobre cada objecto, por exemplo, a identificação do próprio objecto avaliado, a tipificação do avaliador, o texto da avaliação, a data em que foi registada, entre outros.

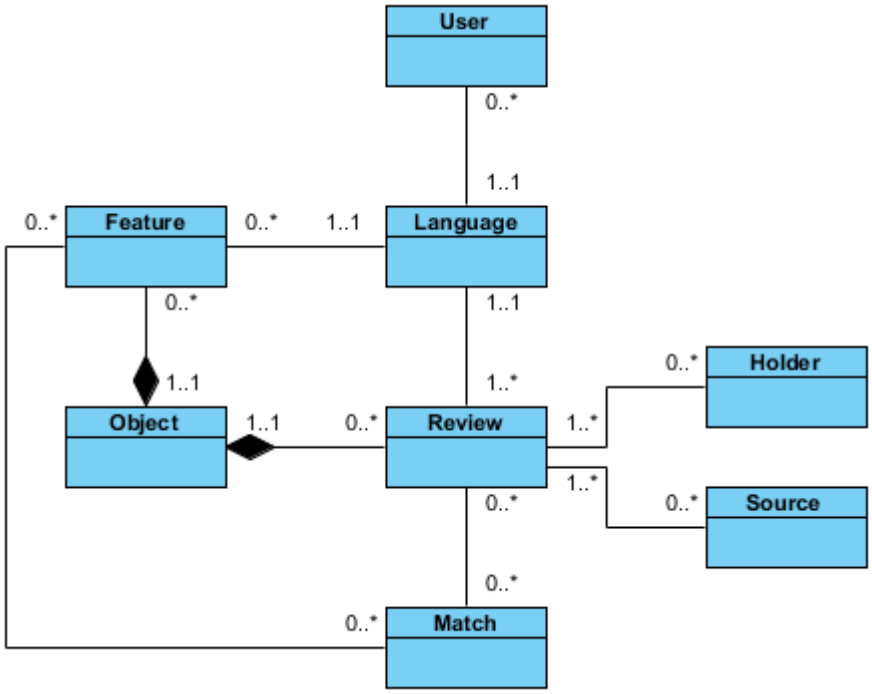

Fig. 1 - Meta-Modelo para receber conhecimento gerado por avaliadores na Web 2.0.

A classe *Review* possui uma associação do tipo composição à classe *Object*. Esta ligação permite conhecer o objecto associado a cada avaliação, e associá-lo a todos os detalhes do próprio objecto único relacionado. Por exemplo, para um *Object* "Livro", a classe *Object* pode ser estendida em três níveis, Editora, Autor e Livro, permitindo o detalhe em cada nível desta hierarquia relacionada.

A classe *Object* também possui uma associação do tipo composição à classe *Feature*, onde se encontram todos os factores de busca pelos quais a classe *Review* poderá ser pesquisada para um determinado objecto. Para o *Object* "Alojamento", algumas *features* a serem utilizadas são: "Limpeza", "Conforto" ou "Localização".

A classe *Match* está directamente associada, quer à classe *Review*, quer à classe *Feature*, e foi introduzida no sistema com o intuito de permitir resultados instantâneos, isto é, que determinada pesquisa em função de determinada *Feature*, tenha de imediato o seu resultado conhecido. Desta forma, o desempenho do sistema e o tempo de resposta às consultas solicitadas pelos utilizadores não vai depender de varrimento

completo da classe *Review*. De referir que, para o exemplo do domínio hoteleiro, se estima que a classe *Review* possa conter muitos milhares, ou mesmo milhões de registos.

A classe *Review* está também associada à classe *Holder*, onde se registam informações sobre os *holders*, isto é, os revisores ou avaliadores, autores de cada *Review*, como por exemplo a sua nacionalidade, o tipo de *holder* (exemplo: "Casal jovem"), entre muitas outras.

A classe *Review* está também associada à classe *Source*, que permitirá ao sistema registar e conhecer a proveniência de cada avaliação (exemplo: www.booking.com).

3.1.1. Aspecto Multidomínio

A abordagem repartida por sub-domínios permite um maior enfoque, mais específico e detalhado, sobre cada uma das áreas do domínio global.

Foram identificados os sub-domínios chave presentes no meta-modelo, o que permitiu uma explicação detalhada de cada um deles, compreender o porquê da sua existência, as suas relações com classes vizinhas e as suas funcionalidades.

O presente meta-modelo foi desta forma idealizado visando ser possível a sua aplicação a qualquer tipo de objecto. O único requisito para tal, é que o objecto em causa tenha comentários ou avaliações registadas, e alojadas na Web 2.0. Exemplos de instâncias que o modelo suporta em outros sectores, ou para outros objectos: Livro, Filme, etc.

O nível de abstracção intrínseco à utilização do meta-modelo permite instanciar as classes e seus respectivos atributos de acordo com o objecto em questão.

3.1.2. Aspecto Multilíngue

Na Fig. 1, a classe *Feature* está associada à classe *Language*, permitindo ao sistema um carácter multilingue no que toca às diversas *features* de pesquisa (exemplos: limpeza; PT ou cleanness; EN).

A classe *Review* está associada à classe *Language*. O sistema ditará ele próprio, e de forma automática, a língua em que cada avaliação se encontra escrita. Esta análise/decisão será tomada pelo sistema, e registada na própria classe *Review*, levando em linha de conta a quantidade de *features* encontradas em cada avaliação, de determinada língua. Este automatismo ocorrerá aquando da execução do algoritmo de alimentação, ou de actualização, da classe *Match*.

A classe *User* está associada à classe *Language*. Esta ligação serve para indicar a língua preferencial de cada utilizador, no âmbito do carácter multilingue do sistema. Na classe *User* estão registados todos os utilizadores com acesso ao sistema. Um dos seus atributos estipula o nível de direitos e permissões de cada utilizador perante o sistema.

Nas secções que se seguem, é explicada a aplicação do meta-modelo ao sector Alojamento.

3.2. Domínio *Object*

Analisadas as particularidades do objecto Alojamento, decidiu-se por uma extensão de três níveis hierárquicos de detalhe, relativamente à classe *Object*, como patente na Fig. 2:

- A classe *ObjCountry*, onde figura uma lista completa de todos os países, possibilitando a tradução do nome do país em qualquer outra língua, através da extensão na classe *CountryTrans*;
- A classe *ObjCity*, com a lista de todas as cidades com unidades de hotelaria, possibilitando a tradução do nome da cidade em qualquer outra língua, através da extensão na classe *CityTrans*;

- A classe *ObjectHost*, onde constará todo o detalhe sobre cada hotel, motel, hostal, pensão, albergaria, etc., ou seja, o detalhe de todas as unidades de hotelaria objecto de avaliações ou comentários.

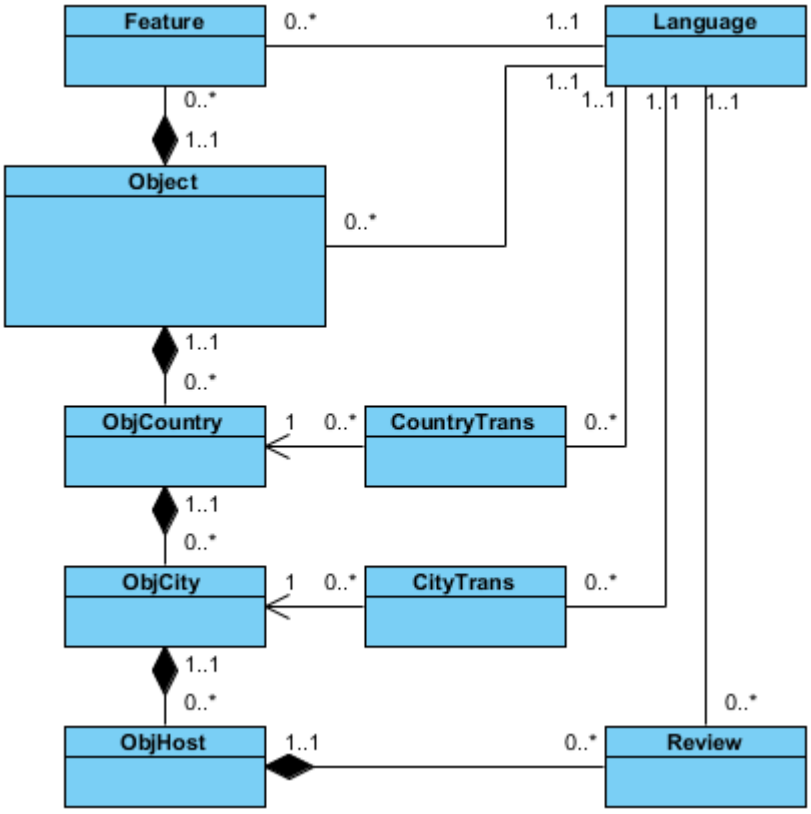

Fig. 2 - Diagrama de Classes do Domínio "*Object*".

Na classe *Object*, a ligação à classe *Language* assegura o carácter multilíngue do sistema. Por exemplo, é possível a existência de entradas significando um mesmo objecto, mas em línguas diferentes, como é o caso de "Alojamento" para português, "Hébergement" para o francês ou "Accommodation" para o inglês.

Para assegurar a mesma disponibilidade multilíngue do sistema, relativamente às classes *ObjCountry* e *ObjCity*, foram introduzidas, respectivamente, as classes *CountryTrans* e *CityTrans*. Como exemplo, para uma entrada "França", em língua portuguesa, na classe *ObjCountry*, haverá uma entrada correspondente "Frankreich", para a língua alemã, na classe "*CountryTrans*". Já para uma entrada "Lisboa", em

língua portuguesa, na classe *ObjCity*, haverá uma entrada correspondente "Lissabon", para a língua dinamarquesa, na classe *CityTrans*.

3.3. Domínio *Review*

Na classe *Review* encontrar-se-ão registadas todas as avaliações sobre cada objecto específico, no caso particular, todas as avaliações sobre cada uma das unidades hoteleiras registadas no sistema, na classe *ObjHost*.

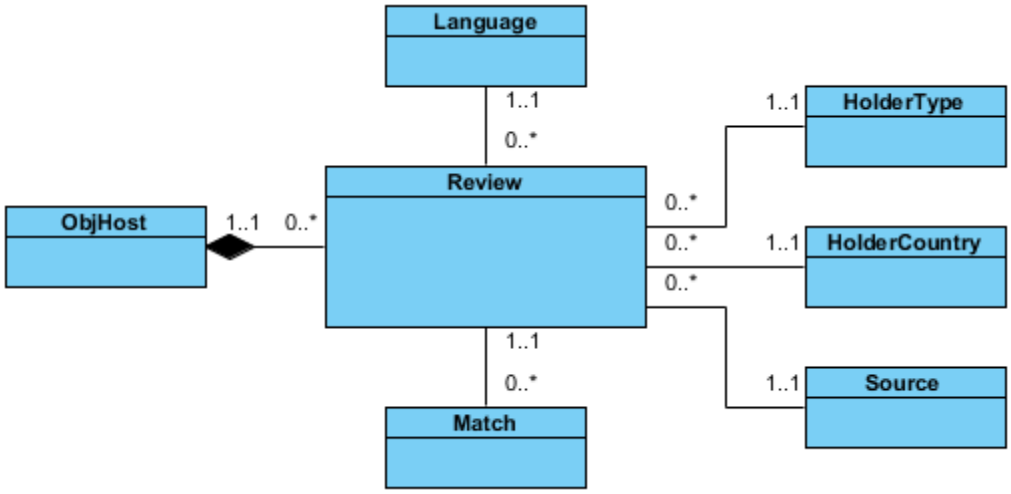

Fig. 3 – Digrama de Classes do domínio "*Review*".

Como podemos observar na Fig. 3, as avaliações carregadas na classe *Review* dependem directamente da classe *ObjHost*, que indica de forma inequívoca a que unidade hoteleira de alojamento diz respeito determinada avaliação.

A ligação existente entre a classe *Language* e a classe *Review*, como anteriormente explicado no meta-modelo de dados, permite ao sistema conhecer a língua em que cada avaliação foi escrita. Na eventualidade desse dado não ser conhecido, o sistema poderá determinar a língua em que se encontra escrita determinada avaliação. Essa detecção será automática, baseada no número de entradas correspondentes encontradas na classe *SubFeature*, detalhada na secção 3.2.3.

A classe *Match* conterá única e exclusivamente índices (apontadores) para registos correspondentes entre a classe *Review*, e as classes *Feature* e, por consequência, *SubFeature*.

As classes *HolderType*, *HolderCountry* e *Source* vão permitir ao sistema associar determinados atributos presentes na classe *Review*, como por exemplo a tipologia em que se insere o autor do comentário (exemplo para a classe *HolderType*, registo "Casal Jovem com filhos"), ou a nacionalidade do mesmo (exemplo para a classe *HolderCountry*, registo "Noruega"), ou ainda o registo da fonte de proveniência do comentário (exemplo para a classe *Source*, registo "www.tripadvisor.com").

3.4. Domínio *Feature*

No modelo de dados para o *Object* "Alojamento", as *Feature* representam o grande grupo temático pelo qual o utilizador poderá pesquisar avaliações. Alguns exemplos muito comuns, no que toca à pesquisa de comentários sobre unidades hoteleiras de alojamento: "limpeza", "conforto", "localização", "quarto", "pequeno-almoço", etc.

Como se pode observar na Fig. 4, optou-se por considerar a classe *Feature* como o registo desses "grandes temas", estendendo-a numa segunda classe *SubFeature*, permitindo associar todas as palavras ou expressões relacionadas.

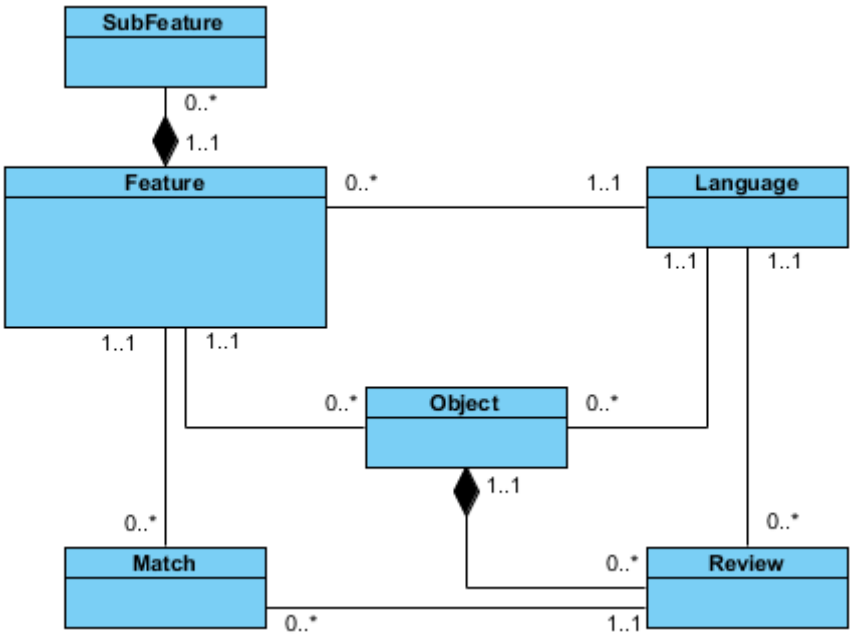

Figura 4 - Diagrama de Classes do Domínio "*Feature*".

A extensão da classe *Feature* para a classe *SubFeature* permite ao sistema registar infinitas terminologias associadas a um mesmo grande tema. Como exemplo, ao grande tema "limpeza", registado como termo de língua portuguesa na classe *Feature*, corresponderão registos associados na classe *SubFeature* como: "limpo", "sujo", "suja", "sujidade", "sujeira", "pó", "poeira", "asseio", "asseado", "asseada", "lixo", "imundo", etc.

Um dos atributos da classe *Subfeature* permitirá ao administrador do sistema registar todos os termos relacionados que não serão mais do que parte de expressões, e não a expressão ou a palavra completas. Continuando o exemplo da *Feature* "Limpeza", uma provável entrada como *SubFeature* relacionada é "%limp%". Com esta entrada única na classe *SubFeature*, e a utilização conjunta do *wildcard* "%", serão detectadas todas as correspondências contento a expressão parcial "limp", ou seja, serão detectadas a presença de palavras como "limpo" ou "limpeza". Fica desta forma igualmente assegurada a detecção de expressões com variação de género (masculino, feminino) ou de número (singular, plural). Para o exemplo anteriormente dado: "limpo" ou "limpa", "limpos" ou "limpas". Esta problemática é explicada com maior detalhe na secção 3.2.4.

A manutenção das *SubFeatures* é uma incumbência do gestor do sistema. Para o efeito, recorreu-se a várias fontes possíveis, para se proceder ao carregamento e manutenção desses dados na respectiva classe:

- Palavras-chave, decorrentes do próprio conhecimento linguístico do gestor;
- Consulta de dicionários, bem como de dicionários de sinónimos, disponíveis online (exemplo para a língua portuguesa: www.priberam.pt), ou em ferramentas aplicacionais (exemplo mais comum, o próprio Microsoft Word);
- Recurso a ferramentas de tradução linguística, também com informação sobre sinónimos (exemplo: ferramenta "Tradutor" em www.google.com).

A interposição da classe *Match*, entre a classe *Feature* e a classe *Review*, permite ao sistema, em qualquer instante ou, se quisermos, em "tempo real", ter presente informação relacionada sobre a presença de todas as *features* detectadas em todas as avaliações. Esta classe foi introduzida no sistema do modelo de dados com o principal objectivo de evitar a demora que poderia decorrer do constante varrimento completo da classe *Review*, aquando de cada pesquisa efectuada pelo utilizador sobre uma determinada *feature*, uma vez que se prevê que, num avançado estado de maturidade do sistema, este possa contar com um número bastante elevado de registos na classe Review. Por estimativa, a classe *Review* pode vir a contar muitos milhares de registos, ou mesmo milhões, e a tipologia do seu atributo mais importante (texto) tornaria o sistema seguramente bastante pesado e muito pouco célere ou performante.

3.5. Modelo Completo para o Sector do Alojamento

Concretizando a arquitectura anteriormente apresentada, foi aplicado o meta-modelo na criação dos diagramas de classe para o exemplo do sector Alojamento. O diagrama de classes apresentado na Fig. 5, representa a estrutura inter-relacional do modelo de dados completo que sustenta a respectiva base de conhecimento.

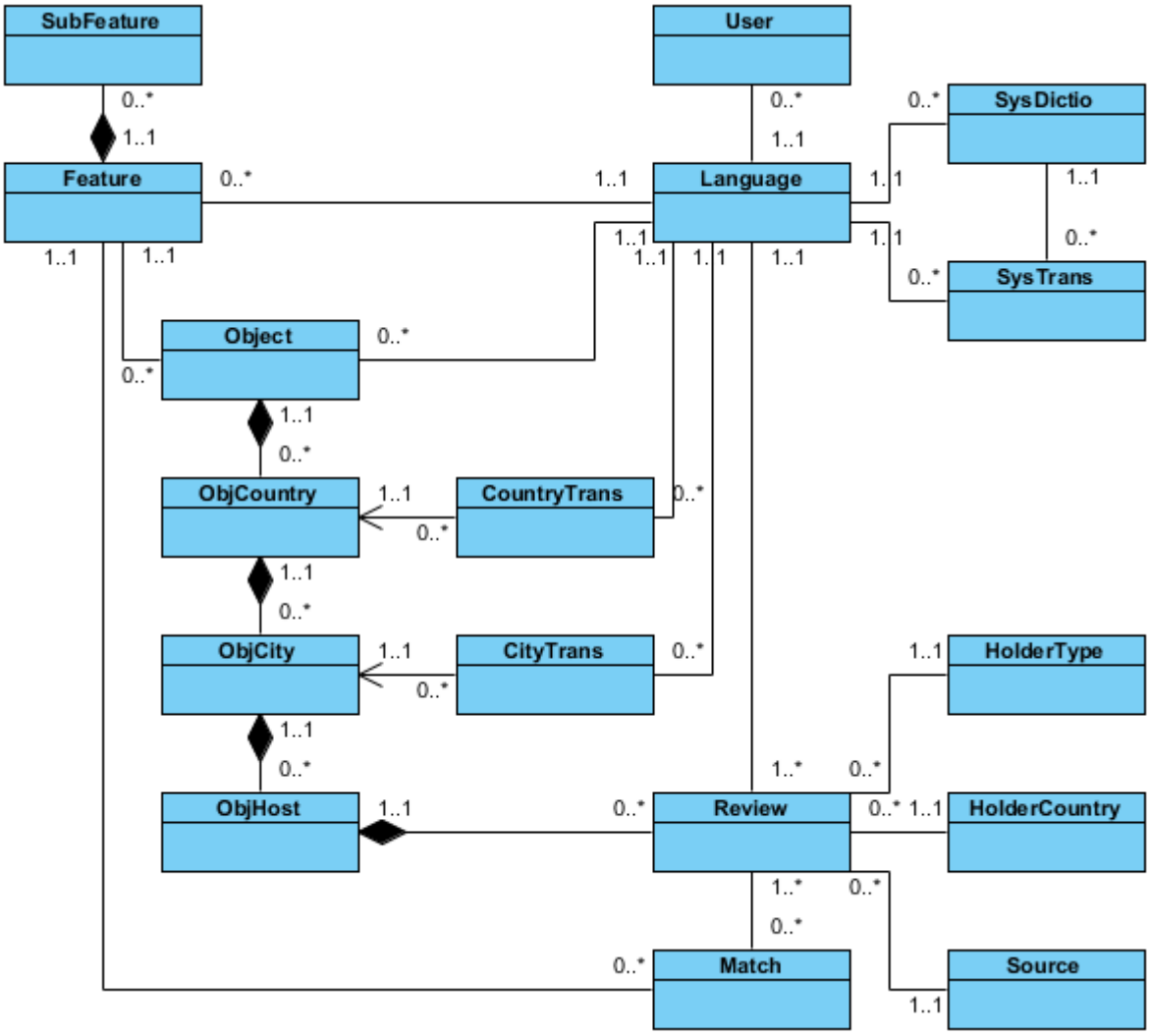

Fig. 5 - Diagrama de Classes completo para o sector Alojamento.

No diagrama de classes representado na Fig. 5 podemos observar a presença dos principais domínios presentes no meta-modelo, e anteriormente escalpelizados (*Object, Review e Feature*). O grau de abstracção inerente à criação de um meta-modelo permitiu a extensão de algumas classes consideradas úteis para este modelo, como se pode observar com o aparecimento das classes *SysDictio* e *SysTrans*, cujo objectivo foi prever o carácter multilíngue também para o futuro interface gráfico para o utilizador.

3.6. Tecnologias Associadas

É importante referir que hoje em dia várias plataformas de bases de dados possuem *add-ons* (funcionalidades modulares que podem ser instaladas ou acrescentadas a uma determinada configuração de base) proprietários, específicos para o tratamento de texto, incluindo inúmeras funcionalidades de *text parsing*, como é o caso da componente *Oracle Text*, fornecida com versões de *Oracle* a partir da versão *Oracle Enterprise 11g*. Este produto específico, para além de muitas funções nele embutidas, permitindo um vasto rol de operações sobre componentes de texto, possui também dicionários de sinónimos, que poderiam constituir uma ajuda preciosa para a implementação deste Projecto.

No entanto, este Projecto, e o seu respectivo modelo de dados, foi idealizado de uma forma completamente agnóstica, no que à base de dados diz respeito. Nesse sentido, uma das decisões tomadas incidiu sobre a escolha da utilização de *windcards*, tendo-se optado pela utilização dos *wildcards* genéricos de SQL (*Structured Query Language*), como são o "%" (percentagem), o "_" (*underscore*) ou "[ ]" (parentesis rectos) para a inquirição da base de dados através de consultas (*queries*), com recurso ao operador "LIKE".

Explicando a utilização destes *wildcards:*

- O *wildcard* "%" é geralmente utilizado no início ou no final duma porção de texto, e ignora tudo o que se encontrar, respectivamente, à esquerda ou à direita do texto;
- O *wildcard* "_" representa um qualquer caracter, na mesma posição onde é evocado;
- O *wildcard* "[ ]" permite indicar um intervalo de caracteres, por exemplo, a sua utilização no formato "[2-5]" permitiria detectar numa determinada posição os caracteres entre o "2" e o "5", inclusive.

Deste modo, o sistema não fará depender a sua implementação de uma ou de outra ferramenta de bases de dados, podendo funcionar perfeitamente em qualquer uma das bases de dados mais comummente utilizadas, desde as bases de dados de licenciamento *freeware*, como o Microsoft SQL Express, o postgreSQL ou o MySQL, até às que carecem de licenciamento pago, como por exemplo Microsoft SQL Server ou Oracle.

4. Validação do Modelo através da Realização de Prova de Conceito

O sucesso da implementação de uma Base de Conhecimento como aquela que se decidiu implementar com a realização deste Projecto depende, em larga medida, de três fases distintas: a recolha, a limpeza e o carregamento dos dados, formalmente conhecidos como processo de ETL (*extraction, transformation and loading*) (Rahm e Do, 2000).

Este capítulo descreve estas três fases e, em seguida, apresenta a prova de conceito realizada para validar o modelo proposto no capítulo anterior.

4.1. Processo de Recolha da Informação

Caberá ao gestor do sistema recolher e compilar toda a informação relevante, nomeadamente aquela relativa às avaliações.

Durante a fase inicial deste Projecto, foram estabelecidos contactos directos com algumas entidades detentoras de portais de maior renome nesta área, nomeadamente o booking.com e tripadvisor.com. Estes contactos visaram a tentativa de obter consentimento expresso para, através da utilização de API's ou Webservices disponibilizados por estes portais a terceiros, se automatizar o acesso e/ou recolha de informação da Web 2.0 relevante para este Projecto. O booking.com dispõe de um sistema de parcerias (*affiliated*), que permite aos utilizadores registados como tal, a recolha directa e automatizada de alguma informação das suas bases de dados. É inclusivamente disponibilizada uma ferramenta denominada *URL Contructor*, que permite a construção de URL's (*Uniform Resource Locator*) directos, de acesso a páginas com conteúdos específicos. Infelizmente, nada possuem no que toca às avaliações registadas.

Já o tripadvisor.com, nos contactos estabelecidos com os seus escritórios em Londres, Reino Unido, escusou-se, tendo comunicado nada poderem disponibilizar, por indicações expressas da sua sede em Boston, Estados Unidos da América.

Frustradas que foram estas demandas iniciais, da tentativa de tornar automática a recolha de informação da Web 2.0, outras possibilidades foram exploradas.

A complexidade da recolha e compilação da informação a ser angariada, para posterior carregamento de dados no sistema, pode variar desde os processos mais básicos, como a recolha por meio de meros e "artesanais" *copy-paste*, avaliação a avaliação, até processos bem mais complexos, como por exemplo o desenvolvimento ou a implementação de *web-crawlers*. Segundo a Wikipedia, um "*Web crawler* é um programa de computador que navega pela *World Wide Web* de uma forma metódica e automatizada", podendo também ser denominados de indexadores automáticos, *bots*, *web spiders*, *web robots*, ou ainda *web scutters.*

4.1.1. Caracterização da Informação Recolhida

Para validação do modelo proposto neste projecto, a fase de recolha da informação foi concretizada utilizando uma base de dados já existente, em formato Microsoft Excel, utilizado em outro projecto (Gomes et al., 2011).

A utilização dessa base de dados, aplicando-lhe as duas fases seguintes do processo de ETL, *transformation* e *loading*, permitiu a este projecto, essencialmente, ganhar tempo que decorreria da recolha manual ou semi-automática dessa informação. Alguns dados estatísticos sobre a informação recolhida são:

- 1503 registos de avaliações;
- 50 unidades hoteleiras;
- Avaliações expressas em 3 línguas diferentes (português, inglês e castelhano);
- 10 tipologias diferentes de avaliadores;
- 72 nacionalidades diferentes de avaliadores.

4.2. Processo de Limpeza de Dados

O processo de limpeza e normalização dos dados é fundamental, pois o sucesso do funcionamento do sistema pode depender dele. Este processo trata da deteção e remoção de inconsistências e erros, com o intuito da melhoria da qualidade dos dados.

Para a realização da prova de conceito da solução, foram realizadas algumas operações de limpeza dos dados adquiridos, dados esses que se encontravam num formato cru (*raw*), antes do seu respectivo carregamento na base de dados.

Exemplos de operações de limpeza de dados aplicadas sobre a colecção de dados, são os seguintes:

- Substituição de todos os pontos-e-vírgula por vírgulas, para obviar a eventuais problemas com a integração do ficheiro CSV, em que o separador de campos utilizado foi o caracter ponto-e-vírgula;
- Eliminação das mudanças de linha presentes no texto, que têm correspondência na combinação de teclas Alt+Enter, ou o caracter com o número de ordem 10 na tabela ASCII (*American Standard Code for Information Interchange*);
- Ordenação dos dados por forma a espelharem o seu destino, no caso específico a tabela *Review*;
- Substituição de alguma informação presente por extenso, pelos seus respectivos índices relacionais (exemplos: país do avaliador, "Portugal" passou a "1", ou no caso da tipologia de avaliador, "Casal Maduro" passou a "3").

A informação recolhida encontra-se registada numa base de dados Excel. A opção recaiu sobre a transformação dessa base de dados em formato CSV, com uma ordenação de dados em colunas que espelha os atributos da respectiva classe (tabela) de destino, a classe *Review*.

Como se pode observar através do seguinte exemplo de um registo do ficheiro CSV final:

```
3;2;4;61;1;;02-11-2010;Pos: Staff were very helpful and the only problem
was rectified immediately. Hotel was very clean. Location was very good,
just out of main tourist area but easy walking distance and plenty of well
priced ammenities around. Good breakfast despite being the same each day.
Neg: Walls were a bit thin so it could be a bit noisy at times. ;+;4;5
```

Tabela 1 - Registo em ficheiro CSV.

O ficheiro CSV final, com vista à realização da Prova de Conceito, conta com 1503 registos, informação proveniente do www.booking.com e do www.tripadvisor.com. Esses registos foram importados automaticamente para a tabela *Review* da base de dados, com recurso às funcionalidades e automatismos de importação directa, disponibilizados pela plataforma de base de dados utilizada, Microsoft SQL Server 2008.

4.3. Processo de Carregamento de Dados

Também o carregamento de dados poderá ser efectuado das mais diversas formas, começando desde logo pela mais básica, o processo de carregamento manual de dados, o qual possui duas enormes desvantagens, a morosidade, que poderá repercutir-se no custo global do projecto, e uma muito provável falta de normalização dos dados. Devido ao volume de dados que se prevê que a base de dados possa vir a atingir, estes processos manuais são altamente desaconselháveis.

Uma das tecnologias mais utilizadas hoje em dia para este efeito é a integração da informação estruturada em ficheiros XML (*Extensible Markup Language*), que são depois lidos ou importados para as respectivas classes.

Outros ficheiros estruturados de texto poderão vir a ser utilizados, como por exemplo os ficheiros CSV (*Comma Separated Value*).

O modelo poderá receber dados provenientes das mais variadas fontes, e tanto mais fácil será a sua alimentação com os dados relevantes, quanto maior for o número de interfaces disponíveis para a efectuar os carregamentos de dados de forma previamente validada e automatizada.

Para o exemplo do objecto "Alojamento", uma vez alimentadas as informações referentes aos países, cidades e suas respectivas unidades de hotelaria, as actualizações a essa informação ficariam praticamente restritas apenas ao surgimento de novas unidades hoteleiras. Já no que diz respeito às avaliações registadas online pelos utilizadores, essas poderão surgir a cada segundo.

Por último, os utilizadores do sistema, para além de efectuarem consultas, que são o core do sistema, podem registar os seus próprios comentários, sendo esta uma fonte adicional de entrada de dados na classe *Review*.

Para validação do modelo proposto, o carregamento de dados será efectuado através das seguintes formas:

- Lançamento manual de dados, pelo administrador do sistema;
- Integração automática de ficheiros estruturados de texto (CSV, XML, etc.), por parte do administrador do sistema, através da criação e execução de pequenos programas (*scripts*) que automatizem o processo, ou com recurso às ferramentas disponibilizadas pela própria solução de base de dados a utilizar;
- Utilização directa das ferramentas de importação existentes na ferramenta de base de dados utilizada.

Visando a alimentação da base de dados, para a realização da prova de conceito, foram compiladas num ficheiro Microsoft Excel cerca de 1500 avaliações. O objectivo foi transformar a informação recolhida num ficheiro cuja integração na base de dados pudesse ser automatizada. Para o efeito, optou-se pelo formato CSV, anteriormente referido. No entanto, procedeu-se a um trabalho de normalização da informação recolhida, para que a mesma pudesse ser perfeitamente compatível com os seus formatos de destino, primeiro CSV, e no final a própria base de dados.

4.4. Cenários para Realização da Prova de Conceito

Os cenários para realização da prova de conceito do modelo estão divididos em duas perspectivas: a perspectiva do utilizador e a perspectiva do gestor do sistema. Todas as consultas a seguir apresentadas foram realizadas em SQL *standard*, sem optimização com recurso a outra linguagem de programação.

Foram efectuadas três consultas (*queries*) visando a validação dos resultados obtidos, na óptica do utilizador. O objectivo foi proceder a uma simulação, o mais próxima possível da realidade, relativamente a eventuais pesquisas efectuadas pelos utilizadores.

No que diz respeito às necessidades do gestor do sistema, foram também efectuadas três consultas para validação dos resultados obtidos, tendo por objectivo a simulação de ferramentas de apoio à gestão do sistema.

4.4.1. Perspectiva do Utilizador

- **Cenário A**: O utilizador procura avaliações em língua portuguesa, que mencionem aspectos relacionados com o tema “Limpeza”.

Sintaxe de código SQL aplicado:

```
SELECT *
FROM Review
WHERE ID_Language = '1' AND (
ReviewText LIKE '%limp%' OR ReviewText LIKE '% p%' OR
ReviewText LIKE '%assead%' OR ReviewText LIKE '%asseio%' OR
ReviewText LIKE '%assp%' OR ReviewText LIKE '%desinfe%' OR
ReviewText LIKE '%encard%' OR ReviewText LIKE '%imund%' OR
ReviewText LIKE '%lavad%' OR ReviewText LIKE '%limp%' OR
ReviewText LIKE '%lixeir%' OR ReviewText LIKE '%lixo%' OR
ReviewText LIKE '%poeir%' OR ReviewText LIKE '%porca%' OR
ReviewText LIKE '%porco%' OR ReviewText LIKE '%suja%' OR
ReviewText LIKE '%sujeir%' OR ReviewText LIKE '%suji%' OR
ReviewText LIKE '%sujo%')
```

Tabela 2 - Código SQL Cenário A.

Dos 1503 registos de avaliações presentes na classe *Review*, resultaram seleccionados 88 registos correspondendo à pesquisa efectuada. Na Tabela 3 é apresentado o conteúdo dos atributos *ID_Review* e *ReviewText* de 2 desses 88 registos:

66:	*Limpeza e asseio excelentes, atmosfera familiar e funcionários extremamente corteses e eficientes. Trata-se de um hotel moderno dentro de um prédio antigo, o que lhe dá bastante charme. Os quartos são pequenos mas confortáveis e bem equipados. Ótimo banho*
271:	*O banheiro estava sujo, o ralo do banheiro não funcionava, a agua também não esquentava. Pessimo Hotel!*

Tabela 3 - Resultados obtidos no Cenário A.

- **Cenário B**: O utilizador procura avaliações em língua inglesa, que mencionem aspectos relacionados com o tema "Location".

Sintaxe de código SQL aplicado:

```
SELECT ID_Review, ReviewText
FROM Review
WHERE ID_Language = '2' AND (
ReviewText LIKE '%location%' OR ReviewText LIKE '%local%' OR
ReviewText LIKE '%centr%' OR ReviewText LIKE '%far%' OR
ReviewText LIKE '%near%' OR ReviewText LIKE '%situat%' OR
ReviewText LIKE '%center%' OR ReviewText LIKE '%away%' OR
ReviewText LIKE '%close%' OR ReviewText LIKE '%position%' OR
ReviewText LIKE '%placem%')
```

Tabela 4 - Código SQL Cenário B.

A aplicação sobre a base de dados do comando SQL acima apresentado, resultou na selecção de 556 registos, sendo apresentados na Tabela 5 quatro exemplos do conteúdo dos atributos *ID_Review* e *ReviewText* na tabela *Review*:

904:	The hotel is very well located. Near Marques de Pombal and Av. Liberdade. It is possible to reach Rossio, Baixa and Chiado in 15/20 minutes walking through the streets. . . And besides this, it is silent and calm at night.
981:	The air conditioning during the summer is the great thing. The personal was friendly, breakfast is simple, wifi works not so good but still works, not too far from subway station. Sometimes you can hear the noise from other rooms and from the street.
1025:	The pensao is really close (20 min walk) to the city center of lisbon and has a metro station in front of the house. Neighborhood not so good. Room small, without wradrobes or cupboards, smelled very smokey although non-smoker room, very cheap, clean but
1290:	This was our second stay and we plan to go back again. The location is in an interesting neighborhood and the views are spectacular. The staff is friendly and very helpful. If you go be sure to visit the bar in the evening as the city views are great.

Tabela 5 - Exemplo de resultados obtidos no Cenário B.

- **Cenário C**: O utilizador procura avaliações em língua inglesa, registadas por casais maduros, que se pronunciem sobre o tema "Cleanness".

Sintaxe de código SQL aplicado:

```
SELECT ID_Review,ID_Language,ID_HolderType, ReviewText
FROM Review
WHERE ID_Language = '2' AND ID_HolderType = '3' AND (
ReviewText LIKE '%neat%' OR ReviewText LIKE '%tidy%' OR
ReviewText LIKE '%sanitar%' OR ReviewText LIKE '%dirty%' OR
ReviewText LIKE '%filthy%' OR ReviewText LIKE '%unclean%' OR
ReviewText LIKE '%messy%' OR ReviewText LIKE '%grimy%' OR
ReviewText LIKE '%nasty%' OR ReviewText LIKE '%dingy%' OR
ReviewText LIKE '%stained%' OR ReviewText LIKE '%grubby%' OR
ReviewText LIKE '%unwashe%' OR ReviewText LIKE '%greasy%' OR
ReviewText LIKE '%untidy%' OR ReviewText LIKE '%squalid%' OR
ReviewText LIKE '%mucky%' OR ReviewText LIKE '%crummy%' OR
ReviewText LIKE '%smudgy%' OR ReviewText LIKE '%smutty%' OR
ReviewText LIKE '%lousy%' OR ReviewText LIKE '%splotchy%' OR
ReviewText LIKE '%sloven%' OR ReviewText LIKE '%boarish%' OR
ReviewText LIKE '%frowzy%' OR ReviewText LIKE '%sloven%' OR
ReviewText LIKE '%swinish%' OR ReviewText LIKE '%grouty%' OR
ReviewText LIKE '%scurril%' OR ReviewText LIKE '%miry%' OR
ReviewText LIKE '%fecule%' OR ReviewText LIKE '%ebon%' OR
ReviewText LIKE '%pig%' OR ReviewText LIKE '%obscene%' OR
ReviewText LIKE '%soiled%' OR ReviewText LIKE '%mucky%' OR
ReviewText LIKE '%muddy%' OR ReviewText LIKE '%sloppy%' OR
ReviewText LIKE '%dingy%')
```

Tabela 6 - Código SQL Cenário C.

A aplicação sobre a base de dados do comando SQL acima apresentado, resultou na selecção de 4 registos, sendo apresentado na Tabela 7 o respectivo conteúdo dos atributos *ID_Review* e *ReviewText* da tabela *Review*:

18:	*Pos: Its central town location near beach & restaurants Neg: It was a bit grubby. could do with a good cleaning, painting and minor repairs*
222:	*The hotel is clean and tidy with basic facilities. It is a family run hotel and we found everyone very helpful and friendly. The price was relative to the accommodation. It was difficult to find because of one way systems and narrow streets.*
507:	*pos : location only Neg: staff totally demotivated - rumors that the hotel is bankrupt and about to close down (this would explain the attitude of the staff) - the ongoing renovation works noisy and dirty - the broken glass on our balcony - the shower of*
835:	*The only positives were the breakfast, the view, and the staff was helpful. The pictures on booking. com are nothing like the room we had. The room needed updating, worn brown/orange carpet, dingy gold bedspread, curtains. We've booked 12 hotels for Italy*

Tabela 7 - Exemplo de resultados obtidos no Cenário C.

4.4.2. Perspectiva do Gestor

- **Cenário D**: O gestor do sistema pretende gerir as *SubFeatures* registadas, relacionadas com a *Feature* "Limpeza".

Sintaxe do comando SQL aplicado:

```
SELECT SubFeature
FROM SubFeature
WHERE ID_Feature = '1'
ORDER BY SubFeature
```

Tabela 8 - Código SQL Cenário D.

Foram seleccionados 18 registos, apresentados na Tabela 9 ordenados alfabeticamente:

1. %assead%
2. %asseio%
3. %assép%
4. %desinfe%
5. %encard%
6. %imund%
7. %lavad%
8. %limp%
9. %lixeir%
10. %lixo%
11. % pó%
12. %poeir%
13. %porca%
14. %porco%
15. %suja%
16. %sujeir%
17. %suji%
18. %sujo%

Tabela 9 - Resultados obtidos no Cenário D.

- **Cenário E:** Para efeitos de gestão do sistema, o administrador pretende efectuar uma consulta conjunta de avaliações registadas em língua inglesa que contenham referências à *feature* "Cleanness", bem como de avaliações registadas em língua portuguesa que contenham referências à *feature* "Limpeza"**.**

Sintaxe do comando SQL aplicado:

```
SELECT ID_Review,ID_Language,ID_HolderType, ReviewText
FROM Review
WHERE (ID_Language = '2' AND (
ReviewText LIKE '%neat%' OR ReviewText LIKE '%tidy%' OR
ReviewText LIKE '%sanitar%' OR ReviewText LIKE '%dirty%' OR
ReviewText LIKE '%filthy%' OR ReviewText LIKE '%unclean%' OR
ReviewText LIKE '%messy%' OR ReviewText LIKE '%grimy%' OR
ReviewText LIKE '%nasty%' OR ReviewText LIKE '%dingy%' OR
ReviewText LIKE '%stained%' OR ReviewText LIKE '%grubby%' OR
ReviewText LIKE '%unwashe%' OR ReviewText LIKE '%greasy%' OR
ReviewText LIKE '%untidy%' OR ReviewText LIKE '%squalid%' OR
ReviewText LIKE '%mucky%' OR ReviewText LIKE '%crummy%' OR
ReviewText LIKE '%smudgy%' OR ReviewText LIKE '%smutty%' OR
ReviewText LIKE '%lousy%' OR ReviewText LIKE '%splotchy%' OR
ReviewText LIKE '%sloven%' OR ReviewText LIKE '%boarish%' OR
ReviewText LIKE '%frowzy%' OR ReviewText LIKE '%sloven%' OR
ReviewText LIKE '%swinish%' OR ReviewText LIKE '%grouty%' OR
ReviewText LIKE '%scurril%' OR ReviewText LIKE '%miry%' OR
ReviewText LIKE '%fecule%' OR ReviewText LIKE '%ebon%' OR
ReviewText LIKE '%pig%' OR ReviewText LIKE '%obscene%' OR
ReviewText LIKE '%soiled%' OR ReviewText LIKE '%mucky%' OR
ReviewText LIKE '%muddy%' OR ReviewText LIKE '%sloppy%' OR
ReviewText LIKE '%dingy%'))
OR (ID_Language = '1' AND (
ReviewText LIKE '%limp%' OR ReviewText LIKE '% p%' OR
ReviewText LIKE '%assead%' OR ReviewText LIKE '%asseio%' OR
ReviewText LIKE '%assp%' OR ReviewText LIKE '%desinfe%' OR
ReviewText LIKE '%encard%' OR ReviewText LIKE '%imund%' OR
ReviewText LIKE '%lavad%' OR ReviewText LIKE '%limp%' OR
ReviewText LIKE '%lixeir%' OR ReviewText LIKE '%lixo%' OR
ReviewText LIKE '%poeir%' OR ReviewText LIKE '%porca%' OR
ReviewText LIKE '%porco%' OR ReviewText LIKE '%suja%' OR
ReviewText LIKE '%sujeir%' OR ReviewText LIKE '%suji%' OR
ReviewText LIKE '%sujo%'))
```

Tabela 10 - Código SQL Cenário E.

Da aplicação desta consulta sobre a base de dados, mais concretamente sobre a tabela *Review*, resultou a selecção de 470 registos. Isto permite ao gestor do sistema concluir desde logo que é elevada a probabilidade de se encontrarem referências ao tema "Limpeza" num registo de avaliação. Esta pesquisa foi efectuada em avaliações em português e inglês, e resultou na selecção de 31% do total dos registos de avaliações. Na Tabela 11, apresenta-se o conteúdo dos atributos *ID_Review* e *ReviewText*4 de 4 desses registos da classe *Review*:

616: Hotel simples com acomodações modestas e limpas. Está bem localizado próximo a estação de metro e farta opção de transporte. O Hotel necessita de reformas e troca do mobiliário, bastante antigo.
835: The only positives were the breakfast, the view, and the staff was helpful. The pictures on booking. com are nothing like the room we had. The room needed updating, worn brown/orange carpet, dingy gold bedspread, curtains.
1046: Apesar da Residencial se localizar numa zona considerada de risco na cidade, e de exteriormente não apresentar grande aspecto, a verdade é que no seu interior, as instalações surpreendem pela positiva. Instalações com boas condições, limpas, acolhedoras.
1476: This place is a hellhole with the smell of raw sewage all over the hotel. the bathrooms are mouldy and dirty. The rooms facing the square are in no way sound proofed, leaving it difficult to get any sleep

Tabela 11 - Exemplo de resultados obtidos no Cenário E.

- **Cenário F:** Para prosseguir o seu trabalho de gestão do sistema, e poder decidir em que novas línguas apostar, o gestor do sistema analisou as nacionalidades dos avaliadores. Nesse sentido, elaborou consulta para apurar o número de avaliações por cada nacionalidade, listando os 10 países com mais avaliadores.

  Sintaxe do comando SQL aplicado:

```
SELECT TOP(10) dbo.Review.ID_HolderCountry, dbo.ObjCountry.Country
FROM Review,ObjCountry
WHERE dbo.Review.ID_HolderCountry = dbo.ObjCountry.ID_ObjCountry
```

Tabela 12 - Código SQL Cenário F.

O resultado permite ao gestor do sistema decidir quais as línguas prioritárias pelas quais deve prosseguir o carregamento de dados. Para a realização da prova de conceito, foram carregadas avaliações nas línguas portuguesa, inglesa e castelhana. Com os resultados apurados por esta consulta, será provável que o gestor se decida a prosseguir esse trabalho pelas línguas francesa, holandesa e russa.

ID	Country	Count	Língua
1	Portugal	385	Português
61	Reino Unido	229	Inglês
11	Brasil	151	Português
28	EUA	115	Inglês
26	Espanha	99	Castelhano
39	Irlanda	65	Inglês
13	Canada	30	Inglês/Francês
35	Holanda	28	Holandês
32	França	25	Francês
64	Russia	25	Russo

Tabela 13 - Exemplo de resultados obtidos no Cenário F.

4.5. Análise dos Resultados

Os resultados encontrados mediante aplicação de seis consultas de SQL *standard* permitem-nos validar o funcionamento do modelo idealizado e implementado, nas diferentes perspectivas do utilizador e do gestor do sistema.

Foi possível realizar pesquisas temáticas sobre o texto das avaliações, com carácter multilíngue, obtendo sempre os resultados esperados. Na perspectiva do utilizador, foram realizadas consultas temáticas que retornaram a informação pretendida, potenciando eventuais tomadas de decisão. Na perspectiva do gestor, foram realizadas consultas mais técnicas que retornaram informação útil para a gestão do próprio sistema.

Não seria possível obter nenhum destes resultados directamente nos portais de origem da informação (www.booking.com e www.tripadvisor.com). A consulta de avaliações nesses portais não disfruta de qualquer tipo de dinamismo, sendo o utilizador obrigando a consultar (ler) centenas, ou mesmo milhares de avaliações sobre um determinado objecto, não lhe sendo conferida a possibilidade de consultar directa e estritamente aquilo que lhe interessa.

5. Considerações Finais

Neste capítulo são descritas as conclusões retiradas da realização deste trabalho. São ainda referidas algumas limitações do modelo, bem como possíveis trabalhos futuros.

5.1. Conclusões

A quantidade de informação disponível na internet tem vindo a aumentar exponencialmente. Mais concretamente, estima-se que a informação em formato de texto registada em aplicações dentro da web 2.0 cresce, a nível mundial, na ordem dos vários *TeraBytes* por segundo.

Existem diversas ferramentas que possibilitam tratar essa vasta quantidade de informação, mas nenhuma delas acessível ao comum dos utilizadores, para que a possa utilizar como ferramenta de suporte à tomada de decisão (*DSS - Decision Suport System*). Todas as ferramentas comerciais disponíveis carecem de licenciamento muito dispendioso, e são normalmente utilizadas por grandes empresas ou instituições.

A realização deste trabalho resultou na validação de um modelo que permite aos utilizadores pesquisar, numa base de conhecimento, grandes quantidades de texto proveniente da Web 2.0, obtendo resultados estruturados e filtrados pelas temáticas que realmente lhes interessam, com vista à sua tomada de decisão.

Também na perspectiva do gestor, foi possível validar, através da realização de prova de conceito, a extracção de informação que lhe possibilite a tomada de decisões estratégicas sobre a evolução do próprio modelo, e da alimentação futura da base de conhecimento.

5.2. Limitações

Uma ferramenta de análise contextual de texto depara-se com várias problemáticas, sendo complexo dar resposta a todas elas. O carácter multilíngue da ferramenta desenhada com este Projecto confere ainda complexidade adicional.

Foram identificadas as seguintes dificuldades, ou limitações, no decurso do desenvolvimento deste Projecto, podendo todas elas ser alvo duma análise específica que possa resultar na resolução, ou pelo menos no atenuar dessas lacunas:

- Dificuldade em lidar com acentuações (exemplo: utilização da expressão "asséptico" *versus* "asseptico");
- Dificuldade em lidar com a implementação de alterações à língua, ou de períodos paralelos, como o que decorre presentemente devido à entrada em vigor do Acordo Ortográfico (exemplo: utilização da expressão "asséptico" *versus* "assético");
- Dificuldade em lidar com erros ortográficos (exemplo: presença da expressão "chero" em vez de "cheiro");
- Dificuldade em lidar com práticas de escrita com diminutivos ou abreviações, frequentemente adoptada por uma determinada faixa de utilizadores da Internet, decorrentes de linguagem própria muito utilizada em ferramentas de conversação instantânea (*chat*), bem como em mensagens escritas em telemóveis (exemplos: "qq" *versus* "qualquer", "kem" *versus* "quem").

5.3. Trabalhos Futuros

Durante o desenvolvimento deste Projecto, várias ideias foram surgindo, sempre numa perspectiva de melhoria contínua da solução. Essas ideias ficam expressas nesta secção, para posterior análise e decisão sobre o seu desenvolvimento:

- Desenho e implementação do interface Web para uma disponibilização global da plataforma através da internet;
- Existem diversos dicionários disponibilizados online, bem como dicionários de sinónimos. A criação de interfaces que possibilitem uma interligação automatizada da solução a essas bases de conhecimento poderá constituir uma forte mais-valia para a mesma;
- Com recurso ao desenvolvimento, ou à utilização de *webservices* previamente desenvolvidos, a solução pode evoluir para a recolha, tratamento e carregamento de dados de forma completamente automatizada;
- Com a emergência das Redes Sociais e a sua crescente utilização, a solução poderá potenciar a recolha de dados directamente destas vastas fontes de informação (exemplo: www.facebook.com);
- Integração com correctores ortográficos, que permitam suporte gramatical e ortográfico aos utilizadores quando registam as suas avaliações directamente no sistema;

Bibliografia

Attensity, n.d. *Attensity*. [Online] Available at: www.attensity.com [Accessed 7 Novembro 2011].

Cafarella, M.J., Etzioni, O. & Suciu, D., 2006. *Structured Queries Over Web Text*. Washington, E.U.A.: University of Washington.

Chaves, Marcirio Silveira, Silva, Mário J. & Martins, Bruno, Outubro 2005. A Geographic Knowledge Base for Semantic Web Applications. In *20th Brazilian Symposium on Databases - SBBD*. Uberlândia, Minas Gerais, Brazil, Outubro 2005.

Clarabridge, n.d. *Clarabridge*. [Online] Available at: www.clarabridge.com [Accessed 7 Novembro 2011].

Cranefield, S., Pan, J. & Purvis, M., 2003. *A UML ontology and derived content language for a travel booking scenario*. Dunedin, Nova Zelândia: Universidade of Otago.

Fernandez, I., 2009. *Beginning Oracle Database 11g Administration - From Novice to Professional*. Boston: Apress.

Gomes, R., Chaves, M.S. & Pedron, C., 2011. Impacto da Web 2.0 nos Pequenos e Médios Hotéis em Portugal. In *Anais da 11ª Conferência da Associação Portuguesa de Sistemas de Informação (CAPSI 2011)*. Lisboa, 2011.

Haruechaiyasak, C., Kongthon, A., Palingoon, P. & Sangkeettrakarn, C., 2010. Constructing Thai Opinion Mining Resource: A Case Study on Hotel Reviews. In Processing, A.F.f.N.L., ed. *Proceedings of the 8th Workshop on Asian Language Resources*. Beijing, China, 2010.

Infogrid, T.W.G.D., 2009. *What is metamodeling, and what is it good for?* [Online] Available at: http://infogrid.org/wiki/Reference/WhatIsMetaModeling [Accessed 14 September 2011].

Metrics, S., n.d. *Sentiment Metrics*. [Online] Available at: www.sentimentmetrics.com [Accessed 7 Novembro 2011].

Niemann, M., Mochol, M. & Tolksdorf, R., 2006. *Improving Online Hotel Search - What Do We Need Semantics For?* Berlim, Alemanha: Universität Berlin.

Niemann, M., Mochol, M. & Tolksdorf, R., 2008. *Enhancing Hotel Search with Semantic Web Technologies*. Talca, Chile: Universidad de Talca.

Radian6, n.d. *Radian6*. [Online] Available at: www.radian6.com [Accessed 7 Novembro 2011].

Rahm, E. & Do, H.H., 2000. *Data Cleaning: Problems and Current Approaches*. Leipzig, Alemanha: Universidade de Leipzig.

Ramos, P.N., 2007. *Desenhar Bases de Dados com UML*. 2nd ed. Lisboa: Edições Sílabo.

Sentimetrix, n.d. *Sentimetrix*. [Online] Available at: www.sentimetrix.com [Accessed 7 Novembro 2011].

Wikipedia, n.d. *Decision Support System*. [Online] Available at: http://en.wikipedia.org/wiki/Decision_support_system [Accessed 7 Novembro 2011].

Wikipedia, n.d. *ETL - Extract, Transform, Load*. [Online] Available at: http://pt.wikipedia.org/wiki/Extract,_transform,_load [Accessed 3 Novembro 2011].

Wikipedia, n.d. *Meta*. [Online] Available at: http://en.wikipedia.org/wiki/Meta [Accessed 13 Novembro 2011].

Wikipedia, n.d. *Meta-modeling*. [Online] Available at: http://en.wikipedia.org/wiki/Metamodeling [Accessed 12 Novembro 2011].

Wikipedia, n.d. *Raw data*. [Online] Available at: http://en.wikipedia.org/wiki/Raw_data [Accessed 30 Outubro 2011].

Wikipedia, n.d. *SQL*. [Online] Available at: http://pt.wikipedia.org/wiki/SQL [Accessed 23 Outubro 2011].

Wikipedia, n.d. *URL*. [Online] Available at: http://pt.wikipedia.org/wiki/URL [Accessed 12 Setembro 2011].

Wikipedia, n.d. *Web 2.0*. [Online] Available at: http://en.wikipedia.org/wiki/Web_2.0 [Accessed 7 Outubro 2011].

Wikipedia, n.d. *Web crawler*. [Online] Available at: http://pt.wikipedia.org/wiki/Web_crawler [Accessed 12 Setembro 2011].

Wikipedia, n.d. *XML*. [Online] Available at: http://pt.wikipedia.org/wiki/XML [Accessed 14 Setembro 2011].

Yoo, D., n.d. *Semantic Hotel Search*. [Online] Available at: http://donghee.info/research/SHSS/index.html [Accessed 14 Outubro 2011].