



Instituto Politécnico de Viseu

Escola Superior de Tecnologia e Gestão de Viseu

Dedico este trabalho aos meus pais, Alda Ferreira e Miguel Ferreira, pois sem eles não seria possível estar aqui e também à mulher da minha vida, Odete Frazão, pela força de vontade e apoio que sempre me proporciona.

RESUMO

O acesso à informação é muito valorizado pelas pessoas, sobretudo quando este acesso preenche as necessidades de negócio ou quotidianas.

Atualmente os sistemas de informação na internet estão cada vez mais evoluídos: os motores de busca na internet são suficientemente inteligentes para apresentar informação de acordo com o que o utilizador pretende e com uma taxa de sucesso extremamente elevada; as redes sociais podem ser acedidas através de uma única interface; as notícias de várias fontes são integradas e estão disponíveis comodamente através de um único local; entre outros.

Quantas vezes não nos deparamos com necessidade de saber onde encontrar o melhor dentista; a clinica privada com a melhor reputação; a polícia para reportar um assalto quando estamos numa cidade desconhecida; ou até mesmo a farmácia mais próxima ou o contacto do hospital para saber sobre um ente querido com problemas de saúde?

Seria espetacular termos essa informação sempre atualizada e reunida num único sítio, podendo ser acedida em qualquer lugar de uma forma simplificada por qualquer dispositivo com acesso à internet.

Com este trabalho tem-se em vista um aprofundamento do conhecimento e competências na área de Sistemas de Informação. Numa fase inicial vai ser realizada uma pesquisa sobre quais as melhores fontes e formas para a obtenção de informação, prosseguindo-se com a criação de um sistema de informação de modo a dar resposta às necessidades do utilizador de uma forma comoda, simples e inteligente, reunindo a informação através de um mecanismo de recolha e seleção da informação de várias fontes melhorando autonomamente o processo de busca de acordo com vários critérios.

ABSTRACT

Information access is highly valued by people, especially when such access meets the needs of business or daily.

Currently, the information systems on the Internet are increasingly evolved: the internet search engines are smart enough to present information according to what the user wants and with a high success rate; social networks can be accessed through a single interface; multiple sources news are integrated and conveniently available through a single location; among others.

How often do we need to know where to find the best dentist, a private clinic with the best reputation; the police to report a robbery when we are in a strange city, or even the nearest pharmacy or call the hospital to know about a family member with health problems?

Would be fantastic that we can have always updated information gathered in one single place and we could access it anywhere in a simplified form using any device with network access.

This work has in its view a deeper knowledge and skill in the field of Information Systems. At an early stage, there will be a research of what are the best sources and ways to gather information. Next, will be developed an information system that will try to match the user needs in a comfortable, simple and smart fashion, gathering information through an autonomous mechanism that pick and select information from many sources, improving the search process according to several criteria.

PALAVRAS-CHAVE

Aplicações Moveis

Aplicações Web

Pesquisa de Empresas

Sistemas de informação

Sistemas de pesquisa

KEYWORDS

Company Search

Information Systems

Mobile Applications

Search Systems

Web Applications

AGRADECIMENTOS

Agradeço ao meu orientador Paulo Tomé, por todo o tempo que disponibilizou e por todos os conselhos que me ofereceu para melhorar a minha tese. Também agradeço no geral a todos os meus professores do mestrado pelo conhecimento transmitido que me ajudou a evoluir em várias subáreas dos Sistemas de Informação e que foi bastante útil na realização desta tese.

ÍNDICE GERAL

1. Introdução	1
1.1. Contextualização.....	1
1.2. Problema	2
1.3. Motivações.....	2
1.4. Objetivos.....	3
1.5. Metodologia de Investigação	3
1.6. Estrutura da Tese	3
2. Avaliação de Sistemas existentes.....	5
2.1. Introdução	5
2.2. Portal Nacional.....	5
2.3. Portugalio.....	6
2.4. Páginas Amarelas.....	7
2.5. Google Plus.....	7
2.6. Cylex.....	8
2.7. Conclusões	9
3. Análise e Desenvolvimento.....	10
3.1. Análise	10
3.1.1. <i>Estratégia de Pesquisa</i>	10
3.1.2. <i>Modelo de Dados</i>	11
3.2. Desenvolvimento	13
3.2.1. <i>Linguagens</i>	13
3.2.2. <i>Ferramentas</i>	14
3.2.3. <i>Arquitetura da aplicação</i>	14
3.2.4. <i>Obtenção de dados</i>	15

3.2.5.	<i>Implementação da Pesquisa</i>	17
3.2.6.	<i>Distribuição</i>	20
4.	Demonstração	22
4.1.	Funcionalidades Implementadas	22
4.2.	Funcionamento.....	26
4.3.	Vantagens.....	31
5.	Conclusão	33
6.	Referências	34

ÍNDICE DE TABELAS

Tabela 2-1 – Campos para pesquisa	5
Tabela 2-2 – Campos para pesquisa	7
Tabela 3-1 – Comparação entre MyISAM e InnoDB	19
Tabela 3-2 – Previsão do mercado das plataformas para <i>smartphones</i> 2013 – 2017	20
Tabela 4-1 – Comparação quanto à ordem da listagem inicial.....	27
Tabela 4-2 – Tempos de Resposta.....	31
Tabela 4-3 – Quantidade de resultados	31

ÍNDICE DE FIGURAS

Figura 1-1 – Estrutura da Tese	4
Figura 3-1 – Modelo de dados.....	11
Figura 3-2 – Exemplo de valores de uma linha na tabela empresas.....	12
Figura 3-3 – Linhas da tabela rel_emp_etiqueta com id_empresa = 1	12
Figura 3-4 – Linhas da tabela valores.....	13
Figura 3-5 – Valores e tipos de etiqueta.....	13
Figura 3-6 – Arquitetura exemplificada	15
Figura 3-7 – Paginação	16
Figura 3-8 – Página Inicial	17
Figura 3-9 – Procedimento “get_empresas”	18
Figura 4-1 – Processo de recolha de informação periodica.....	22
Figura 4-2 – Arquitetura do processo de pesquisa	23
Figura 4-3 – Classificação de um utilizador.....	23
Figura 4-4 – Processo de cálculo das médias de classificações	24
Figura 4-5 – Procedimento que decorre após o utilizador fazer uma pesquisa e visualizar uma entidade.....	25
Figura 4-6 – Aplicação: Início.....	26
Figura 4-7 – Pesquisa por Dentistas em Viseu.....	28
Figura 4-8 – Nova pesquisa para demonstração.....	29
Figura 4-9 – Demonstração de pesos.....	30

ABREVIATURAS E SIGLAS

ACID	Atomicity, Consistency, Isolation, Durability
DOM	Document Object Model
ESTGV	Escola Superior de Tecnologia e Gestão de Viseu
HP	Hewlett-Packard
HTML	HyperText Markup Language
IPV	Instituto Politécnico de Viseu
MVC	Model, View, Controller
OAuth	Open Authentication
OCR	Optical Character Recognition
PHP	Hypertext Preprocessor
SDK	Software Development Kit
SQL	Structured Query Language
UNLV	University of Nevada, Las Vegas
WAMP	Windows, Apache, MySQL, PHP
WEB	World Wide Web, Internet
XML	eXtensible Markup Language
XPath	XML Path Language

1. Introdução

Os sistemas de Pesquisa de Informação desempenham um papel importante no quotidiano de grande parte dos cidadãos. Frequentemente as pessoas recorrem a sistemas de pesquisa para assim conseguir resolver problemas do seu quotidiano.

Neste capítulo enquadra-se um trabalho de mestrado que tem como objetivo o desenvolvimento de um novo sistema de pesquisa de informação.

Na secção 1.1 faz-se uma primeira contextualização do domínio do trabalho. Na secção 1.2 apresenta-se o problema identificado e que levou ao desenvolvimento deste projeto. Na secção 1.3 apresentam-se as principais motivações para a realização desta tese. Na secção 1.4 referem-se quais objetivos que se pretendem alcançar. Na secção 1.5 é referida qual a metodologia de investigação adotada no processo de investigação. Por fim a secção 1.6 dá ênfase à estrutura deste documento.

1.1. Contextualização

Atualmente, os sistemas de pesquisa de informação têm um papel importante na vida dos cidadãos e organizações. Cada vez mais, as pessoas recorrem a sistemas de pesquisa para procurar sobre algo ou alguma coisa quando se deparam com um problema ou necessidade. Por outro lado, as organizações recebem várias vantagens com o uso de sistemas de pesquisa. O acesso rápido a informações, que de outro modo seriam de difícil acesso, permite às organizações aproveitar o tempo em inovação e empreendedorismo. Já como Benjamin Franklin (Franklin, 2001), um famoso cientista e escritor, dizia: “Tempo é dinheiro”, e neste caso não poderia estar mais certo.

O desenvolvimento de sistemas de pesquisa de informação não é uma tarefa fácil. De acordo com Roberts (Robertson, 2000) por um lado está pragmatismo comercial e por outro estão os aspetos tecnológicos. O pragmatismo comercial conduz a que os sistemas operem ou não em função de vários aspetos (satisfação dos clientes, competição com outros sistemas, etc). Por outro lado os aspetos tecnológicos também introduzem aspetos, como por exemplo tempo de resposta do sistema, que condicionam o sucesso.

1.2. Problema

Neste projeto, pretende-se contribuir no sentido de resolver o problema que muitos cidadãos comuns enfrentam quando têm uma necessidade e lhes falta conhecimento de qual a melhor entidade para dar resposta a essa necessidade. Por exemplo, uma pessoa precisa de uma cirurgia específica, que à partida sabe que é de risco, e tem necessidade em saber sobre uma boa clínica especializada nesse tipo de cirurgias.

A contribuição que se pretende dar centra-se na implementação de um sistema de pesquisa de informação.

1.3. Motivações

Assim como, atualmente, o Google é uma referência para pesquisa de informação genérica, este trabalho pretende contribuir para a elaboração de um sistema de integração de informação mais particular, neste caso de informação sobre entidades.

De futuro, o sistema derivado deste trabalho será uma ferramenta bastante útil à maioria das pessoas. Este sistema será útil principalmente pelos seus potenciais de integração e pesquisa de informação. A integração de dados da aplicação permitirá designadamente obter dados de várias fontes de uma forma automática. Por outro lado, a pesquisa de informação da aplicação irá potenciar que os seus utilizadores tenham acesso rápido e eficaz a informação que em determinado momento lhes vai permitir resolver um problema.

Outra das mais-valias desta aplicação está relacionada com os custos para os seus utilizadores e manutenção do *site* em termos financeiros. No que diz respeito à parte financeira, pretende-se associar publicidade não obstrutiva às páginas com os resultados, permitindo assim a manutenção e atualizações frequentes do *site*.

Por fim, considera-se que este trabalho possa vir a ser uma boa referência para quem quiser implementar trabalhos do género ou aprimorar os já existentes. Esta contribuição pretende ser alcançada através da identificação dos sistemas na área e da especificação de dados e ideias relevantes sobre o objeto de pesquisa.

1.4. Objetivos

O objetivo principal é a criação de um sistema inteligente de pesquisa e integração de informação que permita aos utilizadores encontrem o que procuram sem ter que pesquisar em vários *sites* diferentes, de uma maneira intuitiva e fluente.

Os outros objetivos fazem parte das próprias motivações, referidas na secção anterior.

1.5. Metodologia de Investigação

Nesta secção vai ser definida a metodologia de investigação seguida ao longo desta tese. Então, como metodologia de investigação foi adotada a pesquisa do género qualitativa, exploratória, bibliográfica e documental.

- Do ponto de vista da forma de abordagem do problema, a pesquisa será uma pesquisa qualitativa em que o processo e o seu significado vão ser os focos principais da abordagem.
- Do ponto de vista dos objetivos (Gil, 2008), a pesquisa será uma pesquisa exploratória onde se vai tentar uma familiarização mais próxima ao problema e tentar construir hipóteses das melhores maneiras de extrair e gerenciar os dados.
- Já do ponto de vista dos procedimentos técnicos (Gil, 2008), a pesquisa será uma pesquisa bibliográfica e documental, onde se vai recorrer a material já publicado e material que ainda não recebeu tratamento analítico.

A pesquisa vai ser direccionada aos principais métodos, algoritmos e maneiras de conceber o sistema. Será portanto direccionada a especialistas na área.

A captura dos dados vai ser feita através dos resultados experimentais durante o desenvolvimento.

1.6. Estrutura da Tese

Nesta secção vai ser descrita a estrutura adaptada para a realização desta dissertação.

Esta dissertação encontra-se dividida num conjunto de capítulos. Na Figura 1-1 – Estrutura da Tese é descrito o conteúdo de cada capítulo.

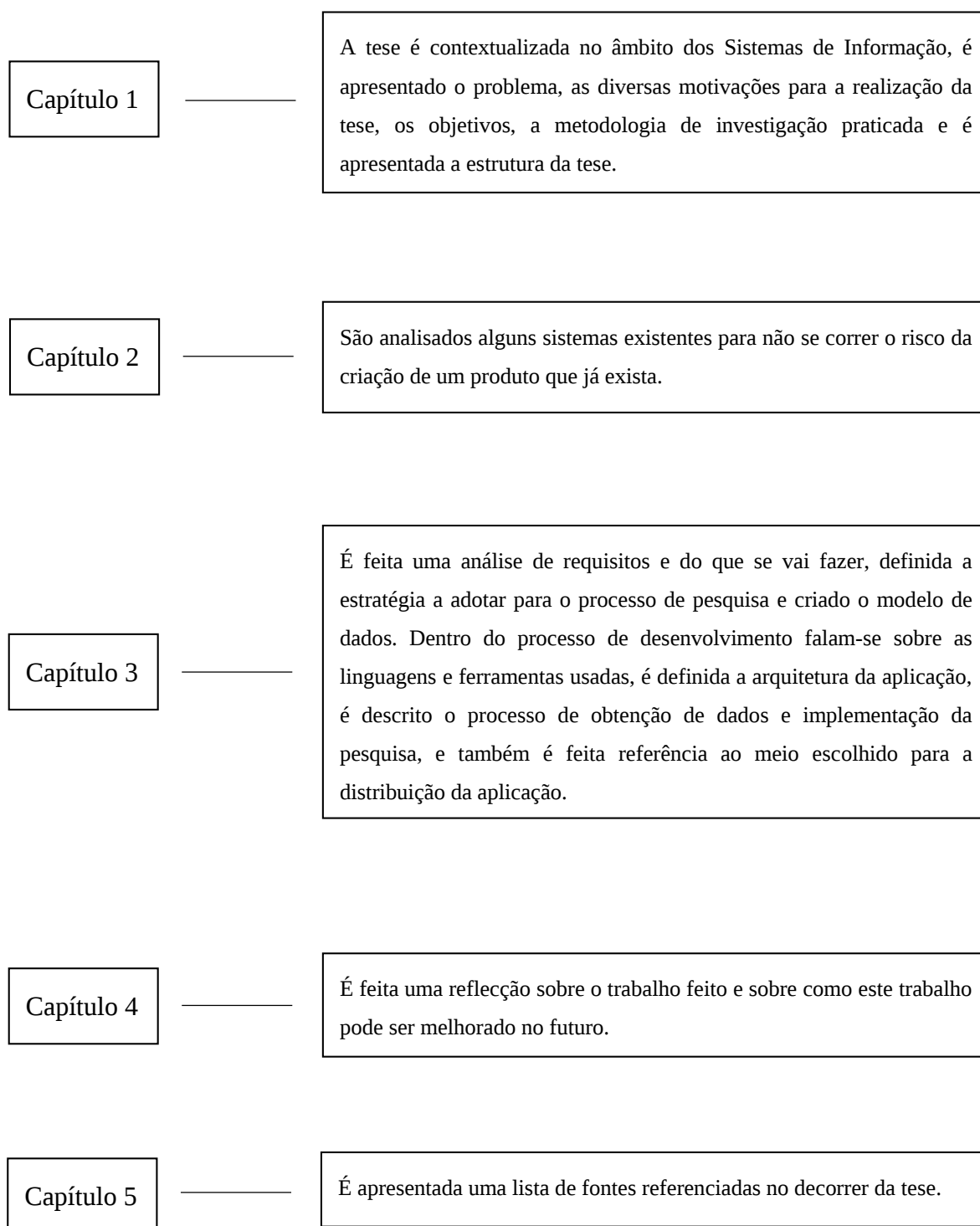


Figura 1-1 – Estrutura da Tese

2. Avaliação de Sistemas existentes

Um dos aspetos importantes no desenvolvimento de “produtos” é a análise de soluções similares ao pretendido. Esta tarefa permitirá, entre outros, reduzir o risco de se criar algo já anteriormente criado. Neste capítulo apresentam-se sistemas similares ao sistema que se pretende implementar neste projeto. Cada sistema é apresentado de acordo com um conjunto de características.

2.1. Introdução

Através de um trabalho de pesquisa, encontrou-se um conjunto de sistemas que têm como objetivo potenciar facilidades de pesquisa de informação aos seus utilizadores. Foram, designadamente seleccionados alguns *sites* interessantes em que é realizada uma pesquisa de entidades: o *Portal Nacional* (Webdados - Tecnologias de Informação, Lda., 2009), *Portugalia* (Portugalia, 2013), *Páginas Amarelas* (Páginas Amarelas SA, 1997), *Google Plus* (Google, 2011) e *Cylex* (CYLEX, 2013).

Todos os *sites* referidos no parágrafo anterior não possuem a componente de integração da informação, pois têm uma base de dados independente em que realizam a pesquisa. Se existe alguma outra forma de obtenção de dados diferente da própria inserção manual de empresas, essa forma é desconhecida.

2.2. Portal Nacional

Nesta secção vai ser feita uma pequena análise do *site* de pesquisa de empresas *Portal Nacional* (Webdados - Tecnologias de Informação, Lda., 2009).

Campo	Tipo
Distrito	Lista, Valores fixos
Concelho	
Área de atividade	
Categoria	
Texto Livre	Caixa de texto

Tabela 2-1 – Campos para pesquisa

De acordo com a informação analisada é possível concluir:

- Quantidade de campos para pesquisa: 5 como especificado na Tabela 2-1.
- Dados aos quais a pesquisa é direcionada: Nome e morada.
- Destaque: É efetuado manualmente, sendo que para obter destaque de uma empresa na listagem é necessário pagamento de uma quantia.
- Ordem dos resultados: Alfabética.
- Informação detalhada: Nome, morada, coordenadas, telefone, fax, *email*, *website*, *facebook*, *twitter*, área de atividade, descrição, produtos comercializados, serviços prestados e marcas comercializadas.
- Outras considerações: Sem tolerância a palavras mal escritas, provavelmente a usar o operando “%LIKE%” na consulta à base de dados, segundo os testes.

2.3. Portugalio

Nesta secção vai ser feita uma pequena análise do *site* de pesquisa de empresas *Portugalio* (Portugalio, 2013).

A análise da documentação permite concluir que:

- Quantidade de campos para pesquisa: 1 único campo de texto livre para pesquisa de empresas códigos postais e ruas.
- Dados aos quais a pesquisa é direcionada: Nome da empresa, código postal, localidade e ruas.
- Destaque: Existem empresas com destaque, empresas mais visitadas e classificação de empresas, sendo que para terem destaque, as empresas têm que pagar uma quantia.
- Ordem dos resultados: Não foi possível determinar o padrão de ordem. Quando é feita uma pesquisa por “viseu”, por exemplo, aparecem empresas nos primeiros resultados com Viseu no nome mas que não são situadas em Viseu. A listagem completa de empresas também não respeita a ordem de “destaques” ou a de “mais visitados” ou a ordem alfabética ou mesmo a classificação. Embora existam planos pagos para obter destaque na listagem e pesquisa, isso não é refletido na ordem dos resultados.
- Informação detalhada: Nome, morada, telefone, email e site.
- Outras considerações: Às vezes, mas nem sempre, são sugeridas recomendações quando se escreve uma palavra incorretamente. Com algumas pesquisas, aparecem resultados

sem lógica, por exemplo ao pesquisar “Engenhari” (falta apenas o “A”) aparece a entidade “Marques Vaz, Arquitecto”.

2.4. Páginas Amarelas

Nesta secção vai ser feita uma pequena análise do *site* de pesquisa de empresas *Páginas Amarelas* (Páginas Amarelas SA, 1997).

Campo	Tipo
O quê?	Caixa de texto
Onde	Caixa de texto

Tabela 2-2 – Campos para pesquisa

Da análise do *site* é possível concluir que:

- Quantidade de campos para pesquisa: 2 como especificado na Tabela 2-2.
- Dados aos quais a pesquisa é direcionada: Área de atividade, nome e telefone para o 1º campo, distrito, concelho, município e freguesia para o 2º campo.
- Destaque: É necessário um contrato com as páginas amarelas para a empresa obter mais destaque (mais campos nos detalhes). Também pode ser feita a classificação da empresa pelo utilizador.
- Ordem dos resultados: Relevância, ordem alfabética, distância ou top classificações.
- Informação detalhada: Página personalizada com múltiplas informações, textos sobre as atividades, etc.
- Outras considerações: O sistema de pesquisa mostra recomendações ao escrever. Não importa a ordem das palavras. Mesmo com erros, os resultados correspondem ao que se pretende encontrar.

2.5. Google Plus

Nesta secção vai ser feita uma pequena análise da componente de pesquisa de lugares que faz parte do *site Google+* (Google, 2011).

Relativamente ao Google+ é possível concluir que:

- Quantidade de campos para pesquisa: 2. Os mesmos referidos na Tabela 2-2 da secção anterior em que é analisado o sistema das páginas amarelas.
- Dados aos quais a pesquisa é direcionada: nome, área de atividade, categoria e telefone para o primeiro campo. País, morada, concelho, distrito e localidade para o segundo campo.
- Destaque: Para a empresa obter destaque, é necessário aderir ao *Google Adwords* (Google, 2005). Não foi encontrada relação com a pesquisa no Google+.
- Ordem dos resultados: Não foi possível determinar o padrão de ordem.
- Informação detalhada: Página personalizada com múltiplas informações, textos sobre as atividades, classificações, comentários e lugares semelhantes.
- Outras considerações: O serviço está integrado no *Google*, pelo que é possível obter direções, entre outros.

2.6. Cylex

Nesta secção vai ser feita uma pequena análise do *site* de pesquisa de empresas *Cylex* (CYLEX, 2013).

Da informação analisada, é possível concluir que:

- Quantidade de campos para pesquisa: 2. Os mesmos referidos na Tabela 2-2 da secção em que é analisado o sistema das páginas amarelas.
- Dados aos quais a pesquisa é direcionada: Nome, breve descrição da empresa e palavras-chave para o 1º campo e cidade para o segundo campo.
- Destaque: Este sistema não “vende” o destaque como os outros sistemas já analisados. Pelos dados aos quais a pesquisa é direcionada, a melhor opção para se obter destaque será o enriquecimento desses dados, logo a entidade aparecerá em mais pesquisas.
- Ordem dos resultados: Quando mais repetições da palavra pesquisada existirem (por exemplo no nome e descrição), mais acima aparece um resultado. Um ponto fraco é que são consideradas palavras de uma, duas e três letras quando efetuada a pesquisa de múltiplas palavras, então se procurar por exemplo “Ortodontia De Viseu em Viseu” o “Centro De Implantologia E Ortodontia De Viseu” não aparece em primeiro como pretendido.
- Informação detalhada: Nome, morada, código postal e localidade, telefone, fax, email, website, coordenadas e palavras chave ou breve descrição da empresa.

- Outras considerações: Não são apresentadas sugestões quando se escreve uma palavra mal formulada, assim como não são apresentados resultados. Deduz-se então que é efetuada uma pesquisa exata em uma tabela que possui uma lista de palavras com as respetivas ocorrências associadas. É interessante que na listagem existe um realce das palavras encontradas que correspondem à pesquisa.

2.7. Conclusões

Dos sistemas analisados foi possível chegar a algumas conclusões que vão ser a seguir especificadas:

1. Os campos de entrada nos sistemas analisados variam de 1 a 5, sendo que os de introdução de texto livre, são no máximo e em média 2 (o primeiro para o que se pretende procurar e o segundo onde se pretende procurar).
2. O *Google +* é o sistema em que são tidos em conta um maior número de campos quando realizada uma pesquisa: nome, área de atividade, categoria e telefone para o primeiro campo. País, morada, concelho, distrito e localidade para o segundo campo.
3. Em todos os sistemas, para uma entidade obter destaque, tem que ser paga uma quantia (exceto no *Cylex*). Não foi possível determinar exatamente a influência de destaques na listagem normal quando é efetuada uma pesquisa, mas ficou-se a saber que destaque não significa que a entidade apareça no topo da pesquisa.
4. Os padrões de ordem identificados nas várias pesquisas foram: relevância, ordem alfabética, distância e top classificações.
5. A informação mostrada nos detalhes de uma empresa não é linear, mas sim variada. O *Google +* apresenta, não só informação sobre a entidade em questão mas também sugere entidades relacionadas. Uma funcionalidade diferenciadora dos demais.
6. Tanto o *Páginas Amarelas* como o *Google +* apresentam um sistema de sugestões de palavras bem implementado e funcional, que mostra sugestões quando uma palavra na pesquisa não é reconhecida e reformula a pesquisa para a palavra mais próxima nesse caso.

3. Análise e Desenvolvimento

Neste capítulo vai ser detalhado o processo de pesquisa, vão ser explicados os caminhos percorridos para chegar aos objetivos propostos e apresentadas todas as especificações técnicas necessárias.

3.1. Análise

Para o sistema que se pretende desenvolver vai ser necessário um motor para acomodar uma base de dados. A base de dados deverá permitir que sejam guardadas as informações pesquisáveis e também algumas informações não pesquisáveis.

O ideal seria guardar apenas as informações pesquisáveis na base de dados, mas como se trata de ir buscar dados a outras fontes, existe o risco de essas fontes não estarem sempre disponíveis, então optou-se por guardar todas as informações, expeto informações que devido à sua natureza volátil sejam atualizadas num curto espaço de tempo (por exemplo os preços de combustíveis ou as farmácias de serviço).

Para gerar dinamicamente a informação, optou-se por uma linguagem de programação de servidor.

3.1.1. Estratégia de Pesquisa

Efetuada a análise dos sistemas no capítulo 2 e tendo em atenção as conclusões que daí derivaram, foram formuladas algumas ideias de como o sistema iria funcionar.

Ao ser efetuada uma pesquisa, o sistema necessita analisar essa pesquisa e redirecionar os resultados e ordem dos resultados de modo a obter maior relevância possível.

Para obter essa relevância foram tidos em consideração vários parâmetros de autoaprendizagem. Em primeiro lugar foi definido que a relação de uma palavra-chave com uma empresa teria um peso. Peso este que é incrementado quando um utilizador seleciona uma empresa no âmbito de uma determinada pesquisa. Este peso em conjunto com a relevância da pesquisa em relação às várias palavras-chave existentes vai influenciar a listagem de resultados, dando prioridade às relações com maior peso.

Em segundo lugar vão ser priorizadas, na listagem, as empresas mais bem classificadas e com maior número de visualizações.

As visualizações são incrementadas cada vez que alguém vê a página de detalhes de uma empresa e a média das classificações é recalculada regularmente através do conjunto de classificações atribuído pelos vários utilizadores.

3.1.2. Modelo de Dados

O Modelo de dados tem a finalidade de mostrar como os dados são fisicamente armazenados na base de dados e pode ser visualizado a seguir na Figura 3-1 – Modelo de dados.

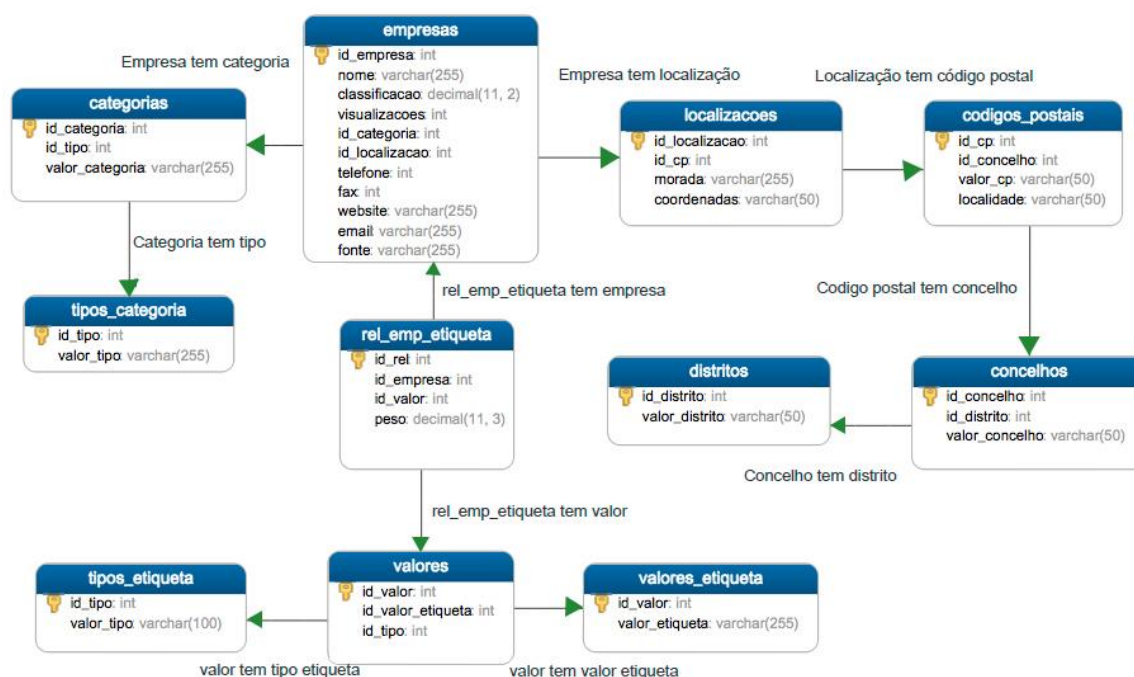


Figura 3-1 – Modelo de dados

Em primeiro lugar foi precisa a criação de uma tabela para armazenar os dados da entidade principal, neste caso uma Empresa.

Depois, para a pesquisa de informações foram criadas quatro tabelas:

1. A tabela principal de pesquisa denominada “empresas”, em que são armazenados os valores da etiqueta (por exemplo: Alimentação, Braga, etc).

2. A tabela “tipos_etiqueta”, em que são especificados os vários tipos que uma etiqueta pode ter (por exemplo: nome, área de atividade, concelho, categoria, distrito, etc).
3. A tabela “valores”, que contempla a relação concetual dos valores de etiqueta e tipos de etiqueta (muitos para muitos).
4. E por fim a tabela “rel_emp_etiqueta”, que é necessária para a relação concetual das empresas com a “tipos” (muitos para muitos).

Após a criação da entidade principal e as entidades e tabelas necessárias para a pesquisa, foram criadas outras tabelas apenas por uma questão de simplificar as consultas de dados sobre as empresas. Essas tabelas contêm as Categorias, Tipos de Categoria, Localizações, Códigos Postais, Concelhos e Distritos.

Por exemplo para a empresa “Afro - Gémeos - Importação, Exportação Produtos” temos uma linha com vários campos na tabela empresas como especificado na Figura 3-2:

id_empresa	nome	classificacao	visualizacoes	id_categoria	
1	Afro - Gémeos - Importação	3	0	1	
id_localizacao	telefone	fax	website	email	fonte
1	219813435	0			http://portalnacional.com.p

Figura 3-2 – Exemplo de valores de uma linha na tabela empresas

Para sabermos quais as linhas relacionadas da tabela “rel_emp_etiqueta” questionamos a base de dados com o comando “select * from rel_emp_etiqueta where id_empresa = 1” e obtemos as linhas especificadas na Figura 3-3:

id_rel	id_empresa	id_valor	peso
1	1	1	0
2	1	2	0
3	1	3	0
4	1	4	0
5	1	5	0

Figura 3-3 – Linhas da tabela rel_emp_etiqueta com id_empresa = 1

Por sua vez, a cada linha da tabela “rel_emp_etiqueta” corresponde uma linha na tabela valores, para obter essas linhas, questionamos a base de dados com o comando “select v.* from

rel_emp_etiqueta ree, valores v where id_empresa = 1 and ree.id_valor = v.id_valor;” e obtemos as linhas especificadas na Figura 3-4.

id_valor	id_valor_etiqueta	id_tipo
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5

Figura 3-4 – Linhas da tabela valores

Finalmente podemos então selecionar os valores e tipos das etiquetas questionando a base de dados com o comando: “select ve.*, te.* from rel_emp_etiqueta ree, valores v, valores_etiqueta ve, tipos_etiqueta te where id_empresa = 1 and ree.id_valor = v.id_valor and v.id_valor_etiqueta = ve.id_valor and v.id_tipo = te.id_tipo;” e então obtemos as linhas especificadas na Figura 3-5.

id_valor	valor_etiqueta	id_tipo	valor_tipo
1	Afro - Gémeos - Importaçã	1	nome
2	Alimentação	2	area_de_actividade
3	Odivelas	3	concelho
4	Comércio Grossista de Prodi	4	categoria
5	Lisboa	5	distrito

Figura 3-5 – Valores e tipos de etiqueta

3.2. Desenvolvimento

O desenvolvimento é uma fase crucial do projeto. É no desenvolvimento que se executam os testes, otimizam tempos, detetam-se *bugs* e aperfeiçoa-se a aplicação.

3.2.1. Linguagens

Foi utilizado o *PHP*¹ como linguagem de programação do lado do servidor e o *SQL*² como linguagem de consulta.

¹ Hypertext Preprocessor

² Structured Query Language

Segundo a informação contida no *site PHP.net* (The PHP Group, 2001), o *PHP* é uma linguagem de script *open source* de uso geral, muito utilizada e especialmente guarnecida para o desenvolvimento de aplicações Web.

Segundo a informação contida no *site 1keydata.com* (1keydata, 2013), o *SQL* é uma linguagem computacional virada para guardar, manipular e obter dados de bases de dados relacionais.

3.2.2. Ferramentas

Foram utilizadas várias ferramentas no desenvolvimento.

Em primeiro o *WAMP*³ que foi adquirido a partir do *site* (Brunel, Sadowski, Martin, Tellier, & Nichini, 2013) e posteriormente instalado numa máquina com Windows 8. O *WAMP* é um pacote de instalação múltipla e integrada de *software* que inclui tudo o necessário para começar a programar e analisar a execução da aplicação (servidor de base de dados, binários *PHP*, *Apache* e *phpmyadmin*).

Depois para editar o código *php*, qualquer editor de texto seria apto, mas foi optado por se usar o *Netbeans* em conjunto com o *Notepad++* e *Dreamweaver* para o efeito. Em resumo, o *Netbeans* é ótimo para detetar os erros durante a execução dos scripts *PHP*, o *Dreamweaver* é muito útil para escrever código em *html*⁴ porque nos diz erros relacionados com *tags* mal fechadas, recomendações e auto completação de código, por fim o *Notepad++* é o mais leve e portátil dos 3, apresentando várias cores que auxiliam a representação e formatação do código.

Finalmente foi utilizado o *Eclipse* e *Android SDK*⁵ (Google, 2013a) para integrar o sistema desenvolvido, nos dispositivos *android*.

3.2.3. Arquitetura da aplicação

Foi optado pela arquitetura *MVC*⁶ pela sua simplicidade e organização de código de acordo com cada uma das suas funções. A única particularidade do modelo utilizado em relação ao tradicional *MVC* é que embora os blocos de código dos controladores estejam separados dos blocos de códigos das vistas, existem vários ficheiros que agrupam um controlador com uma

³ Apache, PHP, MySQL em Windows

⁴ *HyperText Markup Language*

⁵ *Software Development Kit*

⁶ *Model, View, Controller*

vista correspondente. A ideia é obter maior simplicidade e um relacionamento mais forte entre as duas partes, sem sair de dentro do que é a arquitetura *MVC* em si.

Na Figura 3-6 – Arquitetura exemplificada, está exemplificado o que acontece quando um utilizador vê a página inicial de pesquisa.

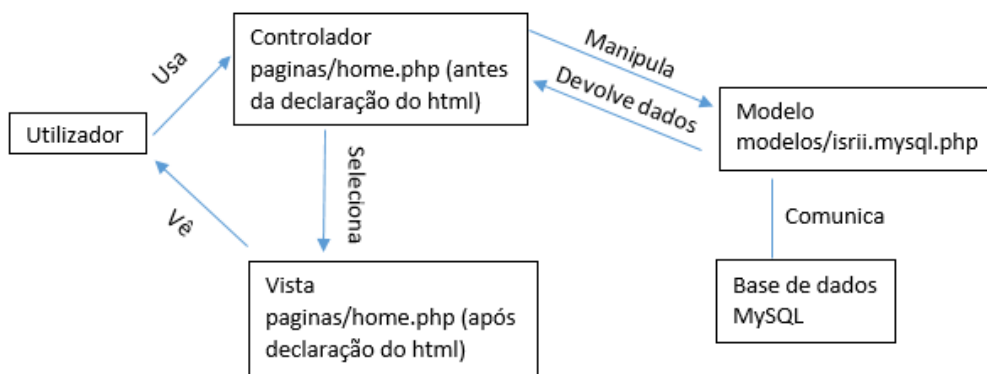


Figura 3-6 – Arquitetura exemplificada

3.2.4. Obtenção de dados

Para obtenção de dados iniciais foi codificada uma classe com o objetivo de extração de dados do *site* Portalnacional (Webdados - Tecnologias de Informação, Lda., 2009). A classe, tal como tudo o resto, foi codificada de raiz e permite a obtenção de todas as empresas e detalhes de empresas existentes no referido *site*. O funcionamento desta classe baseia-se na criação de um objeto *DOM*⁷ que é processado através de expressões *XPath*⁸ de modo a enquadrar os valores pretendidos e atribui-los aos atributos de objetos *PHP* e então usar um modelo para converter e gravar esses objetos na base de dados.

Primeiro é averiguado qual a última página que possui empresas através da paginação referente à listagem de empresas, como exemplificado na Figura 3-7 – Paginação.

⁷ Document Object Model

⁸ XML Path Language

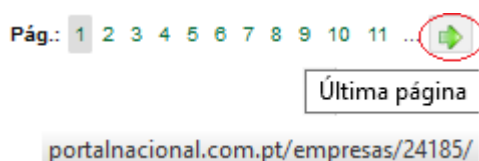


Figura 3-7 – Paginação

Depois, cada página (de 1 até à última) é carregada num novo objeto *DOM* e desse objeto são extraídos os *links* referentes a cada empresa.

Então, é repetido o processo de carregamento, agora para cada empresa, de onde são extraídos os vários dados e guardados na base de dados. As expressões *XPath* foram escritas de modo a serem o menos dependentes possível da estrutura da página, obtendo-se assim uma menor probabilidade de ter que alterar o código caso o *site* em questão seja estruturalmente alterado.

Algumas páginas com os dados das empresas possuem algumas lacunas de consistência, pelo que foi preciso identifica-las após vários testes iniciais e adaptar o código de modo obter os dados de todas as empresas.

Essas lacunas devem-se a algumas páginas apresentarem uma estrutura “antiga” (empresas que foram registadas à mais tempo) e outras páginas apresentarem uma estrutura “nova” (empresas que foram registadas depois de uma atualização do site). Por exemplo, a empresa “FisioLAR⁹” apresenta uma estrutura “antiga” enquanto que a empresa “Escola EB 2 e 3 Dr. Azeredo Perdigão¹⁰” apresenta uma estrutura “nova”.

Um outro desafio foi extrair os números de telefone e *email* dessas páginas, essa informação é representada na página de detalhes sobre a forma de imagens, então para resolver esse problema foi utilizado um sistema de reconhecimento de imagens cujo nome é “tesseract OCR¹¹”. Segundo Ray (Smith, 2007), este *software* foi inicialmente desenvolvido nos laboratórios *HP*¹² entre 1985 e 1995, e em 1995 esteve no TOP dos motores de reconhecimento de imagem mais exatos como evidenciado no teste de precisão *UNLV*¹³ (Rice, Jenkins, & Nartker, 1995). Desde 2006 que este sistema tem vindo a ser constantemente aperfeiçoado pelo *Google*.

⁹ <http://portalnacional.com.pt/empresa/fisiolar-fisioterapia-no-domicilio-279332/>

¹⁰ <http://portalnacional.com.pt/empresa/escola-do-ensino-basico-n-2-e-3-doutor-azeredo-perdigao-abraveses-94395/>

¹¹ *Optical Character Recognition*

¹² *Hewlett-Packard*

¹³ *University of Nevada, Las Vegas*

Foram extraídos com sucesso a maior parte dos números de telefone e *emails* dos detalhes das empresas, e obteve-se uma maior taxa de sucesso fazendo um zoom a cada imagem de 600%. Embora poucos, houveram alguns que não se obteve sucesso, pois as imagens aí presentes foram criadas utilizando um técnica de obscuração (os números não apresentam o mesmo *kernning* (espaçamento entre caracteres) dentro da mesma imagem de uma forma aleatória, levando, por vezes, o sistema de reconhecimento a interpretar dois números como um só.

3.2.5. Implementação da Pesquisa

Finalmente, após ter todos os dados (167760 registos de empresa) na base de dados, já se poderia apostar na implementação e otimização do processo de pesquisa.

Foi então criada a página inicial de pesquisa de entidades em Portugal como se pode visualizar na Figura 3-8 – Página Inicial.

Pesquisa de entidades em Portugal

Palavras-chave

[Cabeleireiro Bonanza](#)
Beleza, Cabeleireiros e Barbeiros
R. Zaire 34 - B, 1170-399 LISBOA
Lisboa, Lisboa ⭐⭐⭐

[FisioLAR - Fisioterapia no Domicílio](#)
Saúde, Fisioterapia, Massagens e Recuperação
Lisboa, 0000-000 Lisboa
Lisboa, Lisboa ⭐⭐⭐

[Farmácia Vriato - Unipessoal, Lda.](#)
Saúde, Farmácias
Avenida da Bélgica, LT 150, FR A, 3510-159 VISEU
Viseu, Viseu ⭐⭐⭐

[A Lisboa, Lda.](#)
Automóveis, Acessórios e Peças
R. Rodrigues Sampaio 63/5, 1150-279 LISBOA
Lisboa, Lisboa ⭐⭐⭐

[Bordados São Jacinto, Lda.](#)
Têxteis, Fiação e Tecelagem
Av. Jacinto Monteiro, 4815-015 CONDE
Braga, Guimarães ⭐⭐⭐

[Escola do Ensino Básico Nº 2 e 3 Doutor Azeredo Perdigão - Abraveses](#)
Ensino, Escolas de 1º, 2º e 3º Ciclos
Abraveses, 3515-126 VISEU
Viseu, Viseu ⭐⭐⭐

[Hotel Lisboa, Lda.](#)
Hotelaria, Hotéis, Pousadas, Estalagens e Albergarias
R. Barata Salgueiro 5, 1169-066 LISBOA
Lisboa, Lisboa ⭐⭐⭐

[Farmácia Portugal de Viseu, Lda.](#)
Saúde, Farmácias
Av. Alberto Sampaio 76, 3510-027 VISEU
Viseu, Viseu ⭐⭐⭐

[Borfil - Empresa de Bordados S.A.](#)
Têxteis, Rendas e Bordados
Pisca - Creixomil, 4810-000 GUIMARÃES
Braga, Guimarães ⭐⭐⭐

[Bordarte - Empresa Industrial e Técnica de Bordados, Lda.](#)
Têxteis, Rendas e Bordados
R. 1º Maio 211, 4800-124 GUIMARÃES
Braga, Guimarães ⭐⭐⭐

« 1 2 3 4 5 6 7 8 9 10 »

Figura 3-8 – Página Inicial

O processo de criação de *html* dinâmico assenta em três fases: obtenção do número total de empresas que corresponde aos critérios de pesquisa, obtenção de dados de um número definido de empresas (10) de acordo com os critérios de pesquisa e página corrente, e por fim cálculos para gerar a paginação utilizando o número de empresas e a página atual.

As consultas são feitas sob a forma de procedimentos armazenados na base de dados, ou seja, no código *php*, os modelos apenas necessitam de invocar os procedimentos com os respetivos parâmetros de entrada. Deste modo a estrutura da base de dados pode ser modificada sem que isso implique que se tenha que alterar o código da aplicação, em vez disso, é apenas necessário adaptar as funções associadas às tabelas cuja estrutura difere da anterior.

O procedimento principal onde se pode verificar a implementação da estratégia de pesquisa pode ser visualizado na Figura 3-9 – Procedimento “get_empresas”.

```

Begin
if(_keywords != '')
THEN
    select ree.id_empresa as id, (ree.peso + SUM(MATCH(ve.valor_etiqueta) AGAINST(_keywords))) as find
    from rel_emp_etiqueta ree, valores val, valores_etiqueta ve, empresas emp
    where ree.id_valor = val.id_valor
    AND val.id_valor_etiqueta = ve.id_valor
    AND emp.id_empresa = ree.id_empresa
    AND MATCH(ve.valor_etiqueta) AGAINST(_keywords)
    GROUP BY ree.id_empresa
    ORDER BY find DESC, emp.classificacao DESC, emp.visualizacoes DESC
    LIMIT _limit
    OFFSET _offset;
ELSE
    select emp.id_empresa as id
    FROM empresas emp
    ORDER BY emp.classificacao DESC, emp.visualizacoes DESC
    LIMIT _limit
    OFFSET _offset;
END IF;
END

```

Figura 3-9 – Procedimento “get_empresas”

As empresas são primeiro ordenadas pelo “*find*”, que é a soma da relevância da pesquisa em relação aos vários valores de etiqueta com os vários pesos das relações em que a relevância é maior que zero, depois pela classificação da empresa, que é a média pré calculada das várias classificações atribuídas pelos utilizadores, e por fim pelo número de visualizações. Depois são devolvidas apenas as empresas de acordo com o limite e *offset* especificado. Caso não hajam palavras como critério de pesquisa apenas ordena por classificação e visualizações.

Para obter um maior desempenho nas pesquisas utilizando o “*match ... against*” foi preciso alterar o motor das tabelas do *mysql* de *InnoDB* (Oracle Corporation and/or its affiliates, 2013a), para *MyISAM* (Oracle Corporation and/or its affiliates, 2013b). Antes da mudança, em média,

para pesquisar “lisboa” demorava 20 segundos, depois da mudança, em média demora 0,6 segundos para a mesma pesquisa, uma melhoria impressionante.

Vamos então efetuar uma análise entre os dois motores na Tabela 3-1 – Comparação entre MyISAM e InnoDB (Oracle Corporation and/or its affiliates, 2013a) (Oracle Corporation and/or its affiliates, 2013b).

MyISAM	InnoDB
Não compatível com ACID ¹⁴ e não transacional	Compatível com ACID e consequentemente transacional, com ROLLBACK, COMMIT e suporte para chaves estrangeiras
Ocupa menos espaço e o limite de informação que pode armazenar é de 256 TB	Ocupa mais espaço, os logs são mais complexos e também maiores. O limite de informação que pode armazenar é de 64 TB.
Requer reparação integral / reconstrução de índices / tabelas	Recuperação automática em caso de falha via repetição dos registos
Não existe ordem nos dados guardados	Os dados das linhas são guardados tendo em atenção a chave primária
Proteção por tabelas	Proteção por registos
Índices <i>FullText</i>	Índices <i>FullText</i> a partir da versão 5.6 do <i>mysql</i>

Tabela 3-1 – Comparação entre MyISAM e InnoDB

Através da comparação acima efetuada conclui-se que teoricamente para leituras de tabelas pouco modificadas o motor *MyISAM* é mais rápido pois não é transacional. Também concluímos que o motor *InnoDB* é mais apropriado para escritas concorrentes pois possui proteção por registos ao contrário do *MyISAM* que apresenta proteção por tabelas.

Na prática, foram efetuados testes na fase de recolha de informação para a base de dados com ambos os motores.

As características do ambiente de desenvolvimento foram:

- Windows 8 Pro com *Media Center*, versão 64 bits
- 8 GB de ram DDR3 1333MHz

¹⁴ *Atomicity, Consistency, Isolation, Durability*

- Disco de 7200 rpm com 300 GB livres, 465 Gb no total
- Processador Intel® Core™ i5-2410M CPU @ 2.3 GHz – 2.9 GHz com 2 cores físicas e 4 virtuais

Os resultados foram que com o motor *InnoDB*, o processo demorou 5 dias enquanto que o *MyISAM* demorou apenas 10 horas. Foi então concluído que o *InnoDB* é o motor mais apto para a aplicação.

3.2.6. Distribuição

Nesta secção vai ser apresentada a opção escolhida para distribuição desta aplicação ao público.

Atualmente as tecnologias do presente e futuro são dispositivos móveis como *tablets* e *smartphones*. De acordo com a IDC (Restivo, Lamas, & Shirer, 2013), no âmbito do mercado dos *smartphones*, o android é a plataforma que lidera e se prevê que continuará a liderar nos próximos 4 anos, com uma grande vantagem sobre os seus principais concorrentes (Tabela 3-2).

Smartphone OS	Partilha de Mercado 2013	Partilha de Mercado 2017
Android	75.3%	68.3%
IOS	16.9%	17.9%
Windows Phone	3.9%	10.2%
BlackBerry OS	2.7%	1.7%
Outros	1.2%	1.9%
Total	100%	100%

Tabela 3-2 – Previsão do mercado das plataformas para *smartphones* 2013 – 2017¹⁵

Todos os dispositivos móveis suportam navegação na internet a partir de vários *browsers*, incluído o que vem de fábrica, pelo que a aplicação *web* pode ser acedida dessa mesma forma nesses dispositivos.

Embora a aplicação possa ser acedida como um simples *site*, os utilizadores preferem ter uma aplicação instalada que evite todo o processo de abrir o browser e escrever ou seleccionar o site.

¹⁵ IDC Worldwide Mobile Phone Tracker, 4 de Setembro de 2013

Uma aplicação instalada tem várias vantagens, além do seu fácil acesso e integração com a plataforma móvel, esta é integrada num sistema de distribuição específico, encontrando-se pesquisável e disponível para instalar a partir do mesmo.

Na plataforma *android* o sistema de distribuição é o “*Google Play Store*”, anteriormente designado por “*Android Market*”. Este sistema pode ser acedido via *site* com auxílio de qualquer dispositivo com possibilidade de acesso à internet além da própria aplicação que vêm de fábrica com os vários dispositivos *android*.

Para poder utilizar o *Google Play* para distribuir esta aplicação, foi preciso converte-la numa aplicação *android*.

A aplicação *android* tem como base uma *webview* (Google, 2013b) que mostra o conteúdo da aplicação *web* criada anteriormente. Os eventos de navegação pela interface são intercetados para que toda a navegação seja feita dentro da aplicação e ao clicar numa ligação não seja aberta uma aplicação externa assim como ao clicar no na tecla de *software* para voltar, não feche a aplicação e em vez disso volte para a página anterior.

4. Demonstração

Neste capítulo vai ser feita uma demonstração passo-a-passo do funcionamento da aplicação, com principal relevância às vantagens do sistema de pesquisa implementado.

Um sistema de pesquisa é um sistema para obtenção de informação, designadamente, para ajudar a encontrar informação presente num sistema computacional. Os sistemas de pesquisa ajudam a diminuir o tempo necessário para encontrar informação e reduzir a quantidade de informação que necessita de ser consultada.

4.1. Funcionalidades Implementadas

Os sistemas de pesquisa oferecem uma *interface* que permite aos utilizadores especificar os critérios de pesquisa acerca de uma área de interesse e encontram os itens correspondentes. Os critérios de pesquisa, normalmente são referidos como “pergunta de pesquisa”. No caso de sistemas de pesquisa de texto, a “pergunta de pesquisa” é tipicamente expressa em palavras que identificam o conceito que um ou mais documentos podem conter.

Um sistema de pesquisa tem que obter os dados pesquisáveis de alguma forma. Os sistemas de pesquisa mais populares indexam o conteúdo pesquisável em bases de dados. O sistema criado, periodicamente recolhe informação de fontes externas, essa informação é processada e então guardada numa base de dados assim como esquematizado na Figura 4-1 – Processo de recolha de informação periodica.

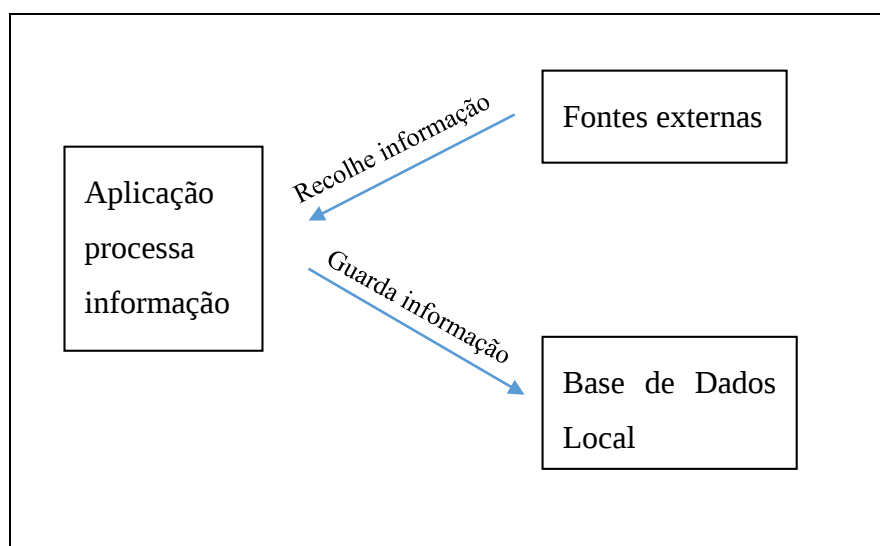


Figura 4-1 – Processo de recolha de informação periodica

Quando um utilizador realiza uma pesquisa, quer seja de um *brower*, quer seja de um dispositivo móvel, ele está a ligar ao servidor em que se encontra o ambiente *php*. No servidor *php*, encontra-se a aplicação cuja arquitetura foi explicada na secção 3.2.3. A aplicação vai então estabelecer ligação com a base de dados, e, através de procedimentos armazenados, vai ser gerada uma resposta com os dados correspondentes à pesquisa que o utilizador fez (Figura 4-2 – Arquitetura do processo de pesquisa).

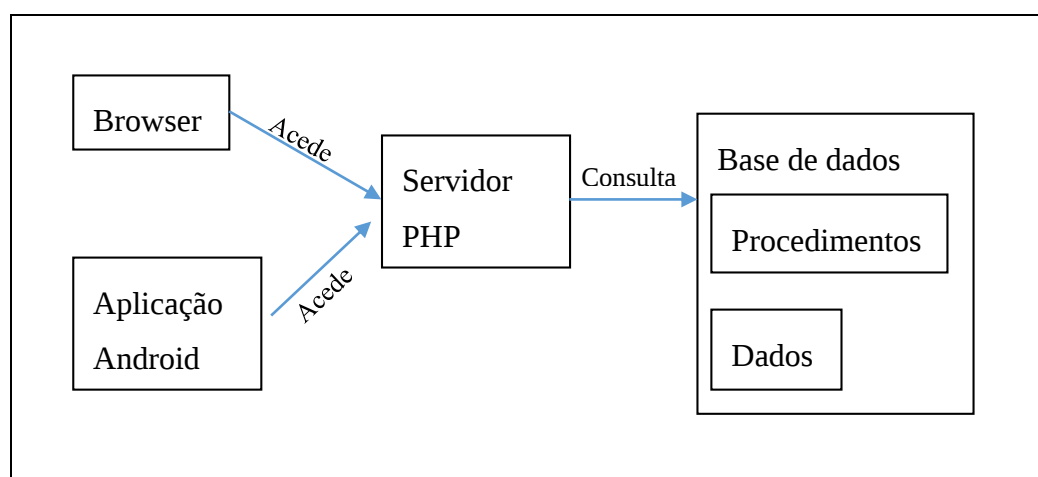


Figura 4-2 – Arquitetura do processo de pesquisa

A lista de itens que vão de encontro aos critérios de pesquisa é tipicamente ordenada. Um dos parâmetros pelos quais o sistema implementado ordena as entidades no ato de pesquisa é a classificação dada pelo utilizador. Quando um utilizador classifica uma entidade, essa informação é guardada na tabela classificações (Figura 4-3 – Classificação de um utilizador).

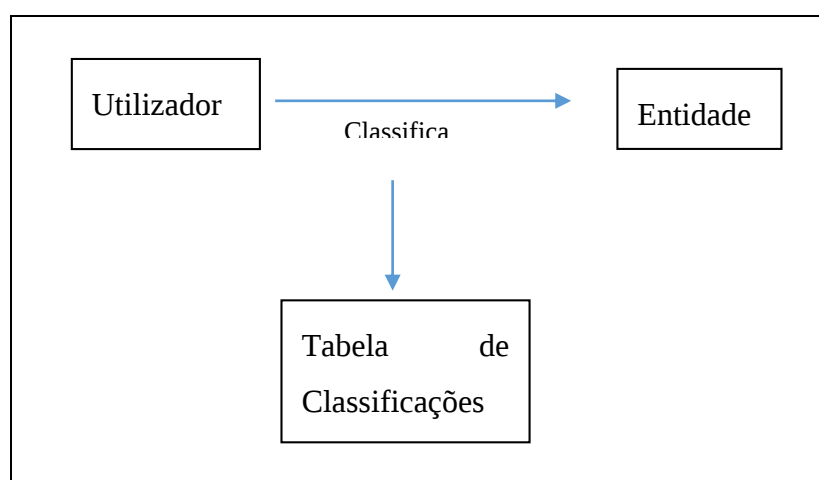


Figura 4-3 – Classificação de um utilizador

Como especificado na secção 3.2.5, o motor de base de dados utilizado para as tabelas, o *Myisam*, funciona com bloqueio a nível de tabela. Isto faz com que, em relação a outros, como o *InnoDB*, teoricamente haja uma perda de desempenho quando ocorrem múltiplas escritas concorrentes, então optou-se por aplicar o motor *InnoDB* na tabela de classificações, contornando assim o problema da melhor forma possível. A média das classificações de cada empresa é calculada e atualizada periodicamente, e não quando um utilizador atribui uma classificação a uma empresa. Para isso, existe um procedimento na base de dados, responsável pelo acontecimento, que consulta as classificações dadas pelos utilizadores, e para cada empresa classificada, calcula a média que é atualizada no registo respetivo da entidade (Figura 4-4 – Processo de cálculo das médias de classificações).

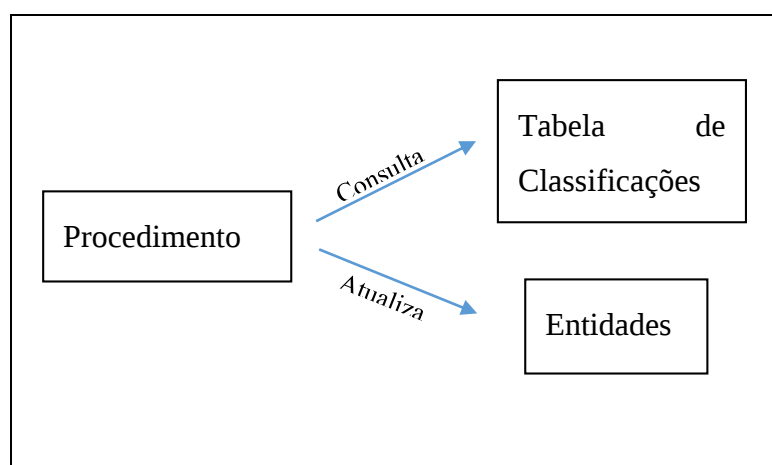


Figura 4-4 – Processo de cálculo das médias de classificações

É necessário identificar o utilizador para guardar o seu voto relativo a uma entidade. Para a sua identificação, o mais correto é que o utilizador proceda à sua autenticação. Estando os votos associados ao utilizador, não se corre o risco do mesmo utilizador classificar várias vezes a mesma empresa. Estando o voto associado ao utilizador, ao voltar a classificar a mesma empresa, esse voto será apenas atualizado.

A funcionalidade de autenticação foi implementada com base a *framework* de autenticação *Oauth*¹⁶ (Parecki, 2013). A *framework* permite que o utilizador use credenciais externas à aplicação, evitando assim o processo de registo. Foi selecionado o Google, como provedor único por ser o mais utilizado para o efeito, mas também poderia ser utilizador o *facebook* entre outros, ou um conjunto de provedores. A autenticação *Oauth* consiste em três passos: a aplicação cliente obtém uma chave de autenticação, o utilizador autoriza a aplicação cliente na

¹⁶ Open Authentication

aplicação servidora e então a aplicação cliente troca a chave de autenticação pela chave de acesso (Mafra, 2013).

Num sistema de pesquisa, por vezes não obtemos os melhores resultados do ponto de vista do utilizador. O sistema precisa de aprender quais as preferências em comum à maioria dos utilizadores para que sejam valorizados os melhores resultados e desvalorizados os outros. A classificação por parte do utilizador, nem sempre é a melhor solução sozinha, por exemplo, uma entidade com um voto de 5 valores e outra entidade com 20 votos, mas com a média de 4 valores, ficamos em dúvida sobre qual a entidade que deveria aparecer em primeiro nos resultados de pesquisa.

Então, para agilizar o processo de pesquisa, foi implementada uma função que é desencadeada sempre que um utilizador visualiza os detalhes de uma entidade no âmbito de uma pesquisa. Sempre que ocorre o evento, é chamado um procedimento que atualiza o numero de visualizações da entidade e também o peso das várias relações etiqueta-entidade (Figura 4-5). Estes parâmetros são então todos usados na ordenação dos resultados de pesquisa.

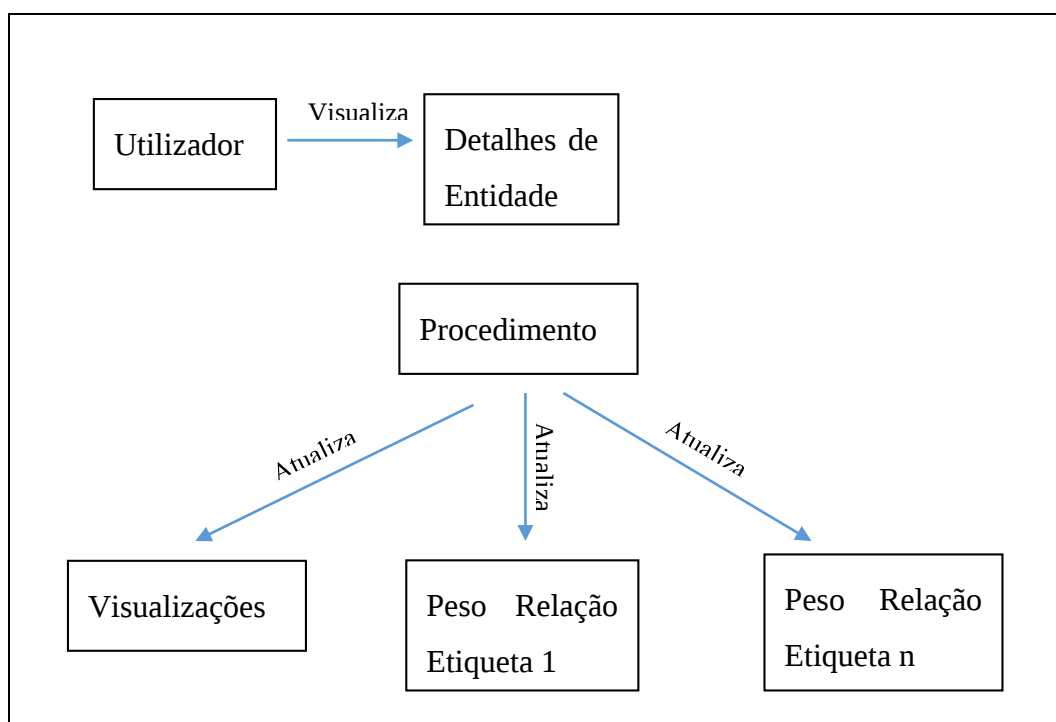


Figura 4-5 – Procedimento que decorre após o utilizador fazer uma pesquisa e visualizar uma entidade

4.2. Funcionamento

Inicialmente, quando acedemos à aplicação encontramos uma interface simples e familiar que nos oferece imediatamente uma listagem das empresas mais bem classificadas e populares como pode ser percebido através da visualização da Figura 4-6 – Aplicação: Início.

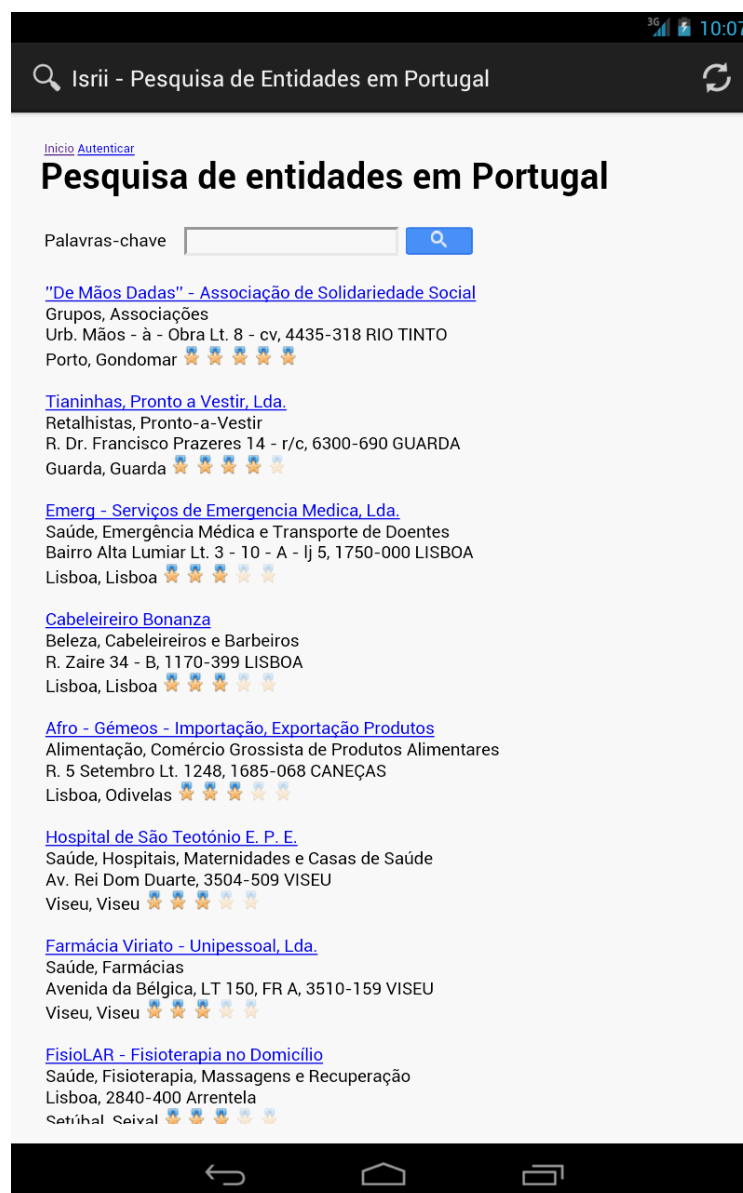


Figura 4-6 – Aplicação: Início

As empresas que aparecem em primeiro quando a aplicação é aberta pela primeira vez por um utilizador, são aquelas que têm uma maior classificação, e dentro das que têm maior classificação, estas são ordenadas pelas que têm maior número de visualizações.

Em relação aos sistemas analisados, esta é uma abordagem mais justa, pois não dá relevância a empresas que pagaram uma certa quantia para estarem nos resultados iniciais (*Cylex*, *Portugalia*) e não se limita às empresas perto (*Google+*). Segue uma pequena comparação na Tabela 4-1.

Sistema desenvolvido	Portal Nacional	Portugalia	Páginas Amarelas	Google +	Cylex
Ordem por classificação, visualizações, últimas Adicionadas	Alfabética	Ultimas atualizadas ou em destaque ou mais visitadas	Top pesquisas	Locais recomendados perto da localização atual	Empresas com desconto

Tabela 4-1 – Comparação quanto à ordem da listagem inicial

Um utilizador só precisa introduzir o que deseja pesquisar: se por exemplo quiser procurar dentistas em Viseu, apenas precisa de escrever “dentistas em Viseu” e então surge uma lista com as entidades com maior relevância na pesquisa, por sua vez ordenados por classificação e visualizações.

A pesquisa efetuada, como demonstrado na Figura 4-7 – Pesquisa por Dentistas em Viseu, devolve uma lista dos dentistas em Viseu, mas como ainda não existem classificações, nem visualizações apenas ordena pelas últimas adições.

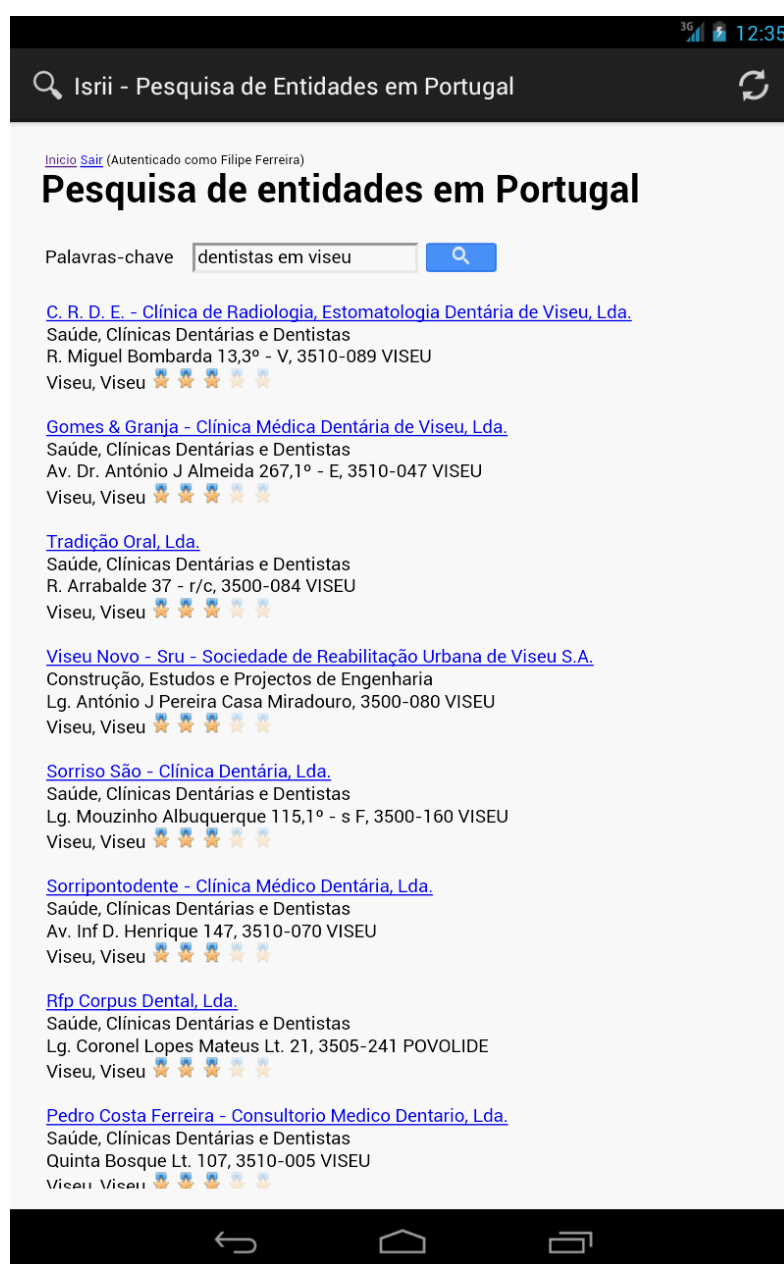


Figura 4-7 – Pesquisa por Dentistas em Viseu

Para demonstração, vai ser dada uma classificação de 4 estrelas à entidade “Sorripontodente” e vai-se verificar o que acontece.

Como se pode ver na Figura 4-8 – Nova pesquisa para demonstração, a “Sorripontodente” passou para a 3ª posição, e não para a 1ª. Isto deve-se ao facto de que os dois primeiros resultados oferecem uma maior relevância na pesquisa efetuada pois contêm a palavra Viseu no próprio nome.

Das muitas dicas encontradas na internet de como dar “relevância nas pesquisas à sua empresa”, esta é uma delas e como se pode ver funciona no sistema desenvolvido até certo ponto.

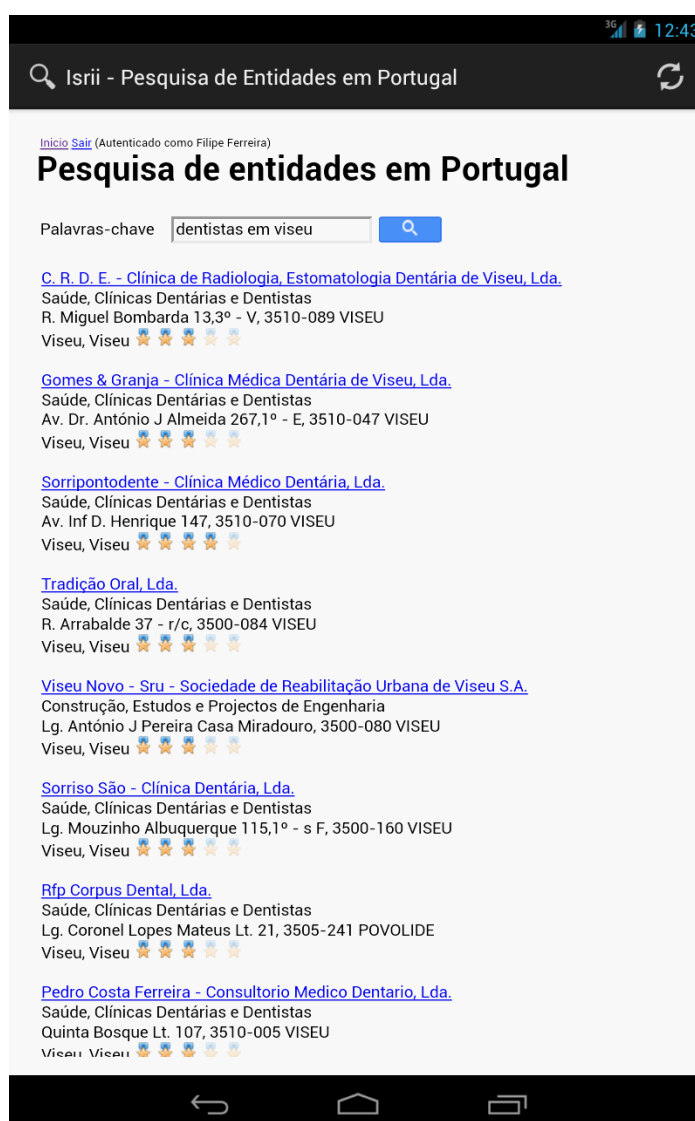


Figura 4-8 – Nova pesquisa para demonstração

Quando as relações que as etiquetas têm com uma entidade passam a ser maiores, por outras palavras, cada vez que uma palavra numa pesquisa se assemelha ao valor de uma etiqueta e depois de efetuada essa pesquisa, o utilizador fica interessado numa empresa e vê os detalhes, essa entidade passa a ter mais relevância para as etiquetas que possuem esses valores.

Por exemplo se quatro pessoas realizarem a pesquisa anterior e virem os detalhes da entidade “Tradição Oral”, a relação daquela empresa com a pesquisa passa a ter peso suficiente para que seja um resultado mais relevante que os outros e então apareça em 1º nos resultados de pesquisa como se pode ver na Figura 4-9 – Demonstração de pesos.

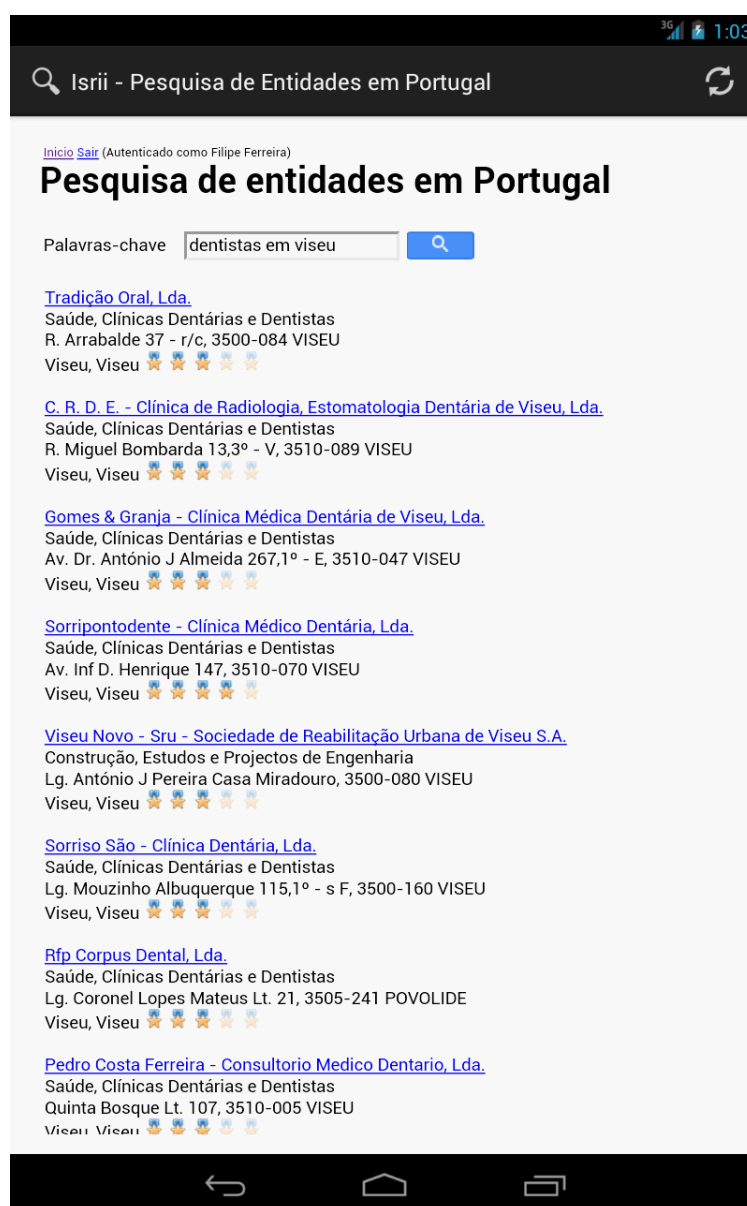


Figura 4-9 – Demonstração de pesos

4.3. Vantagens

Nesta secção vão ser analisadas as principais vantagens deste sistema em relação a outros sistemas.

Para começar, foi feita uma análise dos tempos que uma pesquisa demora a decorrer em alguns sistemas e o implementado. Para tal, foi utilizado o *firebug* (Hewitt, 2013).

O *firebug* é uma poderosa extensão para *firefox* (Mozilla, 2013) que coloca à disposição um conjunto de ferramentas de desenvolvimento para editar, depurar e monitorizar *CSS*, *HTML* e *Javascript*, em tempo real e em qualquer página Web¹⁷.

Vão então ser apresentados os tempos para a pesquisa “dentistas em viseu” na Tabela 4-2 – Tempos de Resposta.

Sistema desenvolvido	Portal Nacional	Portugalio	Páginas Amarelas	Google +	Cylex
165 ms	233 ms	289 ms	823 ms	790 ms	450 ms

Tabela 4-2 – Tempos de Resposta

Como se pode observar pelos resultados da tabela anterior, em termos de tempos de pesquisa, a aplicação desenvolvida supera qualquer um dos sistemas analisados.

Agora, para complementar a análise, vai ser analisada a quantidade de resultados obtidos em cada sistema.

Sistema desenvolvido	Portal Nacional	Portugalio	Páginas Amarelas	Google +	Cylex
7070	0	209	49	49	20

Tabela 4-3 – Quantidade de resultados

¹⁷ World Wide Web

Ao analisar a quantidade de resultados, verificou-se que uma busca no sistema atual, encontra muito mais entidades que uma busca em qualquer um dos outros sistemas. Felizmente, os primeiros resultados são aqueles que são bastante satisfatórios e relevantes. Quanto mais distantes da primeira página, menos relevantes são os resultados, mas estes continuam a conter uma relação com a pesquisa embora menor.

5. Conclusão

Neste capítulo vão ser referidas as principais conclusões que derivaram do trabalho realizado e vai ser feita uma breve referência quanto ao futuro desenvolvimento da aplicação que resultou deste trabalho de pesquisa.

A utilização de palavras-chave no sistema de pesquisa, garante que em qualquer momento uma entidade ou conjunto de entidades pode-se tornar mais relevante nas pesquisas, bastando para isso apenas adicionar novas palavras-chave e associar essas mesmas palavras-chave às entidades relacionadas. Outros métodos viáveis para destacar as entidades mais relevantes no contexto da listagem da pesquisa também foram implementados com sucesso, esses métodos preenchem qualquer lacuna que derive da relevância programática inicial atribuída a cada entidade quando efetuada uma pesquisa e são os parâmetros: número de visualizações, média das classificações por parte dos utilizadores e pesos das relações das entidades com as várias etiquetas que possuem relevância numa pesquisa.

Mesmo que uma palavra seja mal formulada, o sistema de pesquisa encontra o que se quer graças ao uso das pesquisas de texto inteiro proporcionadas pelo *Myisam*. Por vezes, muito raramente, quando se procura alguma coisa em algum lugar, por exemplo “dentistas em Leiria”, aparece um resultado pelo meio, que não se enquadra com um dos termos de pesquisa. Esse pequeno acontecimento é normal porque o sistema ainda não foi treinado, o que será resolvido com o uso, em que os pesos das relações dos resultados relevantes sobem e conseqüentemente os resultados que não se enquadram tanto com ambos os termos de pesquisa vão descendo de posição.

O sistema ainda se encontra em desenvolvimento, mas já apresenta um mecanismo de pesquisa elaborado, funcional, usável e inteligente. Estão planeados certos detalhes para o futuro, dos quais, a análise dos preços de combustível, que, por exemplo quando se procura por uma gasolinera, na página dos detalhes irão aparecer sugestões de outras gasolineras em que os preços estejam de momento mais baratos. Isto irá ser efetuado em tempo real com a integração de um dos vários *sites* que possuem os preços das gasolineras. Outra funcionalidade que ainda não está implementada é mostrar nos detalhes das farmácias se a farmácia está de serviço ou não e sugestões de alternativas mais perto que estejam de serviço. Entre outras funcionalidades que vão sendo sugeridas e que a seu tempo irão ser implementadas também será relevante implementar a adição e alteração manual dos dados.

6. Referências

- 1keydata. (2013). SQL Tutorial - SQL Query Reference and Programming Examples. *1keydata*. Obtido 27 de Setembro de 2013, de <http://www.1keydata.com/sql/sql.html>
- Brunel, D., Sadowski, C., Martin, M., Tellier, X., & Nichini, Q. (2013). WampServer. *WampServer*. Obtido 27 de Setembro de 2013, de <http://www.wampserver.com/en/>
- CYLEX. (2013). Guia de empresas Cylex Portugal. Obtido 16 de Setembro de 2013, de <http://www.cylex.pt/>
- Franklin, B. (2001). *The autobiography and other writings*. New York: Signet Classic.
- Gil, A. C. (2008). *Como elaborar projetos de pesquisa*. São Paulo: Atlas.
- Google. (2005). Google AdWords: Publicidade On-Line do Google. Obtido 19 de Setembro de 2013, de <https://adwords.google.pt>
- Google. (2011). Google+. Obtido 16 de Setembro de 2013, de <https://plus.google.com/u/0/local>
- Google. (2013a). Android SDK | Android Developers. Obtido 3 de Outubro de 2013, de <http://developer.android.com/sdk/index.html>
- Google. (2013b). Building Web Apps in WebView | Android Developers. Obtido 3 de Outubro de 2013, de <http://developer.android.com/guide/webapps/webview.html>
- Hewitt, J. (2013). Firebug. Obtido 10 de Outubro de 2013, de <https://addons.mozilla.org/pt-pt/firefox/addon/firebug/>
- Mafra, D. (2013). Como funciona a autenticação OAuth. Obtido 11 de Outubro de 2013, de <http://www.diogomafra.com.br/2010/09/como-funciona-autenticacao-oauth.html>
- Mozilla. (2013). Navegador web Mozilla Firefox. Obtido 10 de Outubro de 2013, de <http://www.mozilla.org/pt-PT/firefox/>
- Oracle Corporation and/or its affiliates. (2013a). MySQL :: MySQL 5.7 Reference Manual :: 14.2 The InnoDB Storage Engine. Obtido 29 de Setembro de 2013, de <http://dev.mysql.com/doc/refman/5.7/en/innodb-storage-engine.html>

- Oracle Corporation and/or its affiliates. (2013b). MySQL :: MySQL 5.7 Reference Manual :: 14.3 The MyISAM Storage Engine. Obtido 29 de Setembro de 2013, de <http://dev.mysql.com/doc/refman/5.7/en/myisam-storage-engine.html>
- Páginas Amarelas SA. (1997). Páginas Amarelas: Telefones, Moradas e Contactos de Empresas Portuguesas. Obtido 16 de Setembro de 2013, de <http://www.pai.pt/>
- Parecki, A. (2013). OAuth Community Site. Obtido 9 de Outubro de 2013, de <http://oauth.net/>
- Portugalio. (2013). Portugalio - Directório de Empresas, Código Postal e Ruas de Portugal. Obtido 16 de Setembro de 2013, de <http://www.portugalio.com/>
- Restivo, K., Lamas, R., & Shirer, M. (2013). IDC: The premier global market intelligence firm. *Worldwide Mobile Phone Market Forecast to Grow 7.3% in 2013 Driven by 1 Billion Smartphone Shipments, According to IDC - prUS24302813*. Obtido 3 de Outubro de 2013, de <http://www.idc.com/getdoc.jsp?containerId=prUS24302813>
- Rice, S., Jenkins, F., & Nartker, T. (1995). The Fourth Annual Test of OCR Accuracy.
- Robertson, S. (2000). Salton Award Lecture on theoretical argument in information retrieval. *ACM SIGIR Forum*, 34(1), 1–10. doi:10.1145/373593.373597
- Smith, R. (2007). tesseract-ocr. Obtido 29 de Setembro de 2013, de <http://code.google.com/p/tesseract-ocr/>
- The PHP Group. (2001). PHP: O que é PHP? - Manual. Obtido 27 de Setembro de 2013, de http://php.net/manual/pt_BR/intro-what-is.php
- Webdados - Tecnologias de Informação, Lda. (2009). Portal Nacional. Obtido 16 de Setembro de 2013, de <http://portalnacional.com.pt/>