INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




User Profiling and Knowledge Extraction For Tourism

DIOGO MIGUEL RODRIGUES SOARES DE ALMEIDA
Outubro de 2015

POLITÉCNICO
DO PORTO

Instituto Superior de Engenharia do Porto

User Profiling and Knowledge Extraction for Tourism

Diogo Miguel Rodrigues Soares de Almeida

Dissertação para obtenção do Grau de Mestre em Engenharia Informática, Área de Especialização em Arquitetura, Sistemas e Redes

Orientador: Ana Maria Neves de Almeida Baptista Figueiredo

Co-orientador: Nuno Luz

Júri:

Presidente:

[Nome do Presidente, Categoria, Escola]

Vogais:

[Nome do Vogal1, Categoria, Escola]

[Nome do Vogal2, Categoria, Escola] (até 4 vogais)

Porto, outubro 2015

Dedicatória

Dedico a todas a pessoas que, direta ou indiretamente, proporcionaram a realização deste projeto e a persecução dos objetivos traçados.

Todas elas tiveram um papel importante e determinante ao longo do seu ciclo.

Resumo

O sector do turismo é uma área francamente em crescimento em Portugal e que tem desenvolvido a sua divulgação e estratégia de marketing.

Contudo, apenas se prende com indicadores de desempenho e de oferta instalada (número de quartos, hotéis, voos, estadias), deixando os indicadores estatísticos em segundo plano.

De acordo com o " Travel & tourism Competitiveness Report 2013", do World Economic Forum, classifica Portugal em 72º lugar no que respeita à qualidade e cobertura da informação estatística, disponível para o sector do Turismo. Refira-se que Espanha ocupa o 3º lugar.

Uma estratégia de mercado, sem base analítica, que sustente um quadro de orientações específico e objetivo, com relevante conhecimento dos mercados alvo, dificilmente é compreensível ou até mesmo materializável.

A implementação de uma estrutura de *Business Intelligence* que permita a realização de um levantamento e tratamento de dados que possibilite relacionar e sustentar os resultados obtidos no sector do turismo revela-se fundamental e crucial, para que sejam criadas estratégias de mercado. Essas estratégias são realizadas a partir da informação dos turistas que nos visitam, e dos potenciais turistas, para que possam ser cativados no futuro.

A análise das características e dos padrões comportamentais dos turistas permite definir perfis distintos e assim detetar as tendências de mercado, de forma a promover a oferta dos produtos e serviços mais adequados.

O conhecimento obtido permite, por um lado criar e disponibilizar os produtos mais atrativos para oferecer aos turistas e por outro informá-los, de uma forma direcionada, da existência desses produtos. Assim, a associação de uma recomendação personalizada que, com base no conhecimento de perfis do turista proceda ao aconselhamento dos melhores produtos, revela-se como uma ferramenta essencial na captação e expansão de mercado.

Palavras-chave: Turismo, Data Mining, POI, clustering, sistemas de recomendação, similaridade.

Abstract

The tourism sector is a frankly growth area in Portugal and has developed its outreach and marketing strategy. However, it only relates to performance indicators and installed supplies (number of rooms, hotels, flights, accommodation), leaving statistical indicators in the background. According to "Travel & tourism Competitiveness Report 2013," of the World Economic Forum, ranks Portugal in 72nd place in quality and coverage of statistical information available to the tourism sector. It should be noted that Spain occupies the 3rd place. A market strategy without analytical basis, to sustain a specific and objective framework of guidance, with relevant knowledge of the target markets, is difficult to understand or even achievable. Implementing a *Business Intelligence* structure that allows the realization of a survey and data processing that enables to relate and sustain the results achieved in tourism, proves to be essential and crucial to create market strategies. These strategies are performed from the information of tourists who visit us, and potential tourists, so they can be captivated in the future.

Analysis of characteristics and behavioural patterns of tourists allows a definition of distinct profiles and thus detecting market trends, in order to promote appropriate offering products and services.

The knowledge obtained allows, on the one hand, creating and delivering the most attractive products to offer tourists and on the other, to inform them in a targeted manner, the existence of these products. Thus, the combination of a personalized recommendation which is based on the knowledge of the tourist profiles proceed to advice the best products, proves to be an essential attracting tool and in market expansion.

Keywords: Tourism, Data Mining, POI, clustering, recommender systems, similarity.

Agradecimentos

Gostaria de agradecer á Professora Ana pela sua disponibilidade no esclarecimento de dúvidas, pelo seu contributo, pela orientação constante e oportunidade de desenvolver este projeto.

Também um agradecimento especial ao Professor Nuno Luz pela grande importância que teve no decorrer do projeto, dando sempre uma visão coerente e análise construtiva na abordagem das várias fases do projeto, bem como a disponibilidade garantida ao longo do projeto.

Por último gostaria de agradecer á minha família pelo apoio e pela oportunidade que me proporcionaram, sem eles não seria possível.

Índice

Lista de Figuras

Lista de Tabelas

Acrónimos e Símbolos

Lista de Acrónimos

BI	Business Intelligence
CF	Collaborative Filtering
CRM	Customer Relationship Management
CSV	Comma-separated values
ERP	Enterprise Resource Planning
PCC	Pearson correlation coefficient
POI	*Point of Interest*
SP	Stored Procedure
tf-idf	term frequency- inverse document frequency

Lista de Símbolos

∩	Interseção

1 Introdução

Este documento apresenta o projeto desenvolvido no âmbito da unidade curricular Tese / Dissertação / Estágio, do Mestrado em Engenharia Informática, do Departamento de Engenharia Informática do ISEP, cujo objetivo principal consiste no desenvolvimento de um sistema de análise, definição e classificação para a área do Turismo.

Neste primeiro capítulo será explicado o contexto em que decorreu o projeto, a sua motivação, os objetivos e a metodologia adotada para os atingir.

1.1 Enquadramento e Motivação

Sendo Portugal um país cada vez mais atrativo para o setor do turismo, estando inclusive no 15º lugar como um dos países mais competitivos a nível mundial para receber turistas (segundo a fonte Global Travel & Tourism Competitiveness Index de 2015, elaborado pelo Fórum Económico Mundial) revela-se crucial a captação de novos turistas, bem como a preservação dos que já se sentem motivados a voltar.

Para isso, é necessário ir ao encontro das necessidades e expetativas de cada um deles, sendo a forma mais adequada para o fazer consiste na obtenção de um conhecimento mais aprofundado dos seus desejos e motivações. Para tal considera-se que a definição de métricas de classificação que permitam a criação de padrões e índices de qualidade na tomada de decisão possa ser uma primeira abordagem a esta questão.

O projeto desenvolvido nesta tese constitui assim uma resposta a essa necessidade, sendo o objetivo o desenvolvimento de um sistema que permita analisar, definir e implementar métodos de extração de conhecimento, com base na recolha de dados interativos do utilizador/turista. Este processo foi aplicado á região do Porto podendo, no entanto, ser aplicado e/ou expandido para outras regiões.

De uma forma generalizada, a abordagem a utilizar para este problema será baseada no processo de classificação já existente da rede social *Twitter*, no que respeita às grandes fases de desenvolvimento (Lee et al. 2011).

Esta será descrita com maior detalhe no seção 2.3, no qual será definida em que consiste bem como a caracterização da arquitetura do sistema de classificação.

1.2 Objetivos

Com base nos pontos anteriores, é assim possível enumerar um conjunto de objetivos para esta Tese.

Será realizada a análise, definição e implementação dos métodos de extração de conhecimento, a partir de dados partilhados pelos utilizadores/turistas, recolhidas das mais diversas fontes tecnológicas de apoio (por exemplo portais web, aplicações móveis e quiosques).

Este processo terá como auxilio uma base de dados existente (Toursplan), que constitui uma base de apoio nos dados até agora já recolhidos. Esta base de dados contém atualmente incluídos módulos para a recolha de dados interativos do turista, recomendação de produtos de turismo, planeamento de trajetos e apresentação de "*dashboards*" com base nos dados recolhidos.

Para suporte também será adicionada informação, de forma a sustentar e a reforçar o perfil de um utilizador, bem como as suas atividades turísticas, locais e pontos de interesse, sem nunca os identificar pessoalmente.

Após a extração e tratamento da informação será possível identificar um perfil de utilizador, a partir das diferentes iterações realizadas, mostrando a similaridade entre eles.

A análise destas relações, juntamente com as ferramentas já existentes, irá disponibilizar uma base analítica para a obtenção de conhecimento sobre o turista, face aos respetivos mercados alvo.

Descrita a motivação na implementação do sistema, este terá como fases de desenvolvimento as seguintes:

1- Análise prévia de um utilizador e Data Collection

   Ou seja é feita uma análise prévia de cada utilizador de modo a realizar-se a extração de informação de objetos turísticos com os quais o utilizador interagiu (exemplo: POI). Para este processo de monitorização e análise será utilizada a metodologia da rede social Twitter (Twitter, 2015) .

2- Data Modeling

Este processo denominado de *Text-based Modeling*, permitirá modelar os dados de cada documento associado a cada utilizador, permitindo avaliar no final a importância das palavras utilizadas.

3- User Similarity

Para este processo, é realizado o processo de similaridade entre utilizadores, ou seja para cada documento, a partir do cálculo do cosseno do ângulo entre dois vetores. Este cálculo é descrito como a similaridade do cosseno.

4- Clustering

Este processo será realizado de acordo com os dados do ponto anterior.

O resultado serão *clusters*/grupos de perfis de utilizadores semelhantes. Se estes perfis forem de aplicações diferentes, estes resultados poderão ser muitos importantes para se tentar perceber se os perfis são da mesma pessoa ou não.

1.3 Metodologia

A metodologia utilizada neste trabalho está dividida em dois grandes segmentos. Numa primeira fase foi efetuado um levantamento das principais aplicações existentes no mercado, bem como as tecnologias utilizadas. Numa segunda fase procedeu-se à aplicação do método incremental partindo na análise de requisitos, passando pelo projeto, implementação, teste e integração e finalmente pela manutenção.

As etapas definidas foram baseadas no modelo de *software* baseado em cascata, onde se pode verificar na figura 1 que demonstra as principais etapas utilizadas na elaboração do sistema, identificando o conjunto de passos necessários para a sua construção.

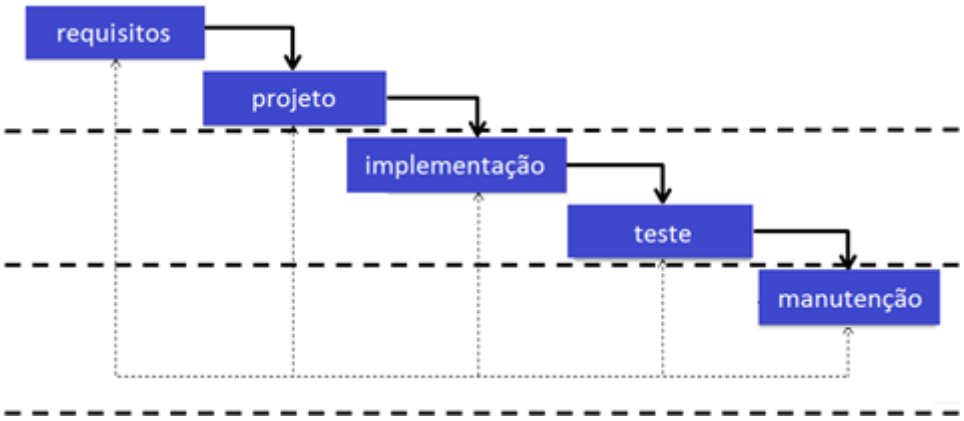


Figura 1- Principais etapas do modelo em cascata

Esta metodologia permite identificar as principais etapas do ciclo de vida do sistema para a realização da arquitetura do sistema.

Tem como objetivo a passagem por uma sequência de fases em que os resultados de cada fase se tornam na entrada para a próxima, ou seja cada fase se completa num período de tempo específico, fazendo com que esta se mova para a fase seguinte (Balaji 2012).

Numa primeira fase é feita a recolha de dados, identificando as principais características de um utilizador para posterior análise (data de nascimento, sexo, estado civil, género, país).

A estes dados vão-se juntar outras variáveis de análise para reforçar e sustentar na fase seguinte os requisitos necessários para a criação de um perfil de utilizador (POI, action timestamp). No caso do campo action timestamp, este caracteriza-se por determinar a data em que uma pesquisa foi realizada por um utilizador.

De seguida é necessário desenhar o sistema, encontrando respostas tecnológicas para as necessidades que vão surgindo á sua elaboração.

Depois da implementação do sistema este será testado e validado, de modo a colmatar algumas lacunas e aperfeiçoando-o até que se encontre com o que é pretendido.

Assim sendo, a metodologia adotada é caracterizada pelos seguintes pontos:

1. Levantamento do estado de arte relativamente ao sector de turismo: Na primeira fase, foi feita uma verificação de algumas tecnologias existentes que permitem recolher e analisar dados relativos a cada utilizador que partilhem informação que possa também se relacionar com a área do turismo.

2. Criação do desenho do sistema: Nesta fase procurou-se perceber quais as fases que o sistema deveria ter para responder ao problema. Esta inicia-se

desde a fase de extração e análise dos dados, com base nas características e métricas consideradas mais importantes que identifiquem cada utilizador, levando á fase da modelação dos dados onde permitirá definir e visualizar perfis de utilizador relacionados, até á fase final de capacitar a criação de agrupamentos com base no grau de semelhança dos utilizadores, onde permitirá assim caracterizar a dimensão do perfil do utilizador.

3. Implementação do sistema de classificação: Esta fase constitui o início da implementação do sistema, procurando satisfazer os requisitos especificados com as tecnologias escolhidas anteriormente.

4. Teste do sistema: Teste e validação do trabalho desenvolvido, corrigindo alguns erros que surjam no decorrer da validação do sistema.

5. Manutenção do sistema de classificação: Nesta fase será definido o trabalho futuro, contendo melhorias para o sistema bem como adição de futuras características e funcionalidades.

1.4 Estrutura do Documento

Para além do presente capítulo, onde se apresenta o enquadramento geral do trabalho e as motivações que levaram ao desenvolvimento da investigação, assim como a metodologia que foi adotada para a realização do projeto, este documento encontra-se estruturado da forma que se descreve de seguida.

No capítulo 2 é identificado a metodologia utilizada para o desenvolvimento do sistema, procurando transmitir ao leitor o que é o sistema de recomendação, abordando o conceito de *User Profiling*, os principais tipos de sistema de recomendação bem como exemplos de aplicação na área do turismo.

No capítulo 3 contextualiza-se um dos conceitos relacionados com o âmbito do projeto, nomeadamente o conceito de *Business Intelligence*, demonstrando seus componentes e funcionalidades. Também é abordado o conceito de *Data Mining* e mais concretamente o conceito de *Text Mining*.

No capítulo 4 são ilustradas e mencionadas todas as fases que constituem a arquitetura da rede social Twitter, sendo esta usada como referência no desenvolvimento do sistema.

No capítulo 5 é efetuada a proposta de arquitetura da solução, identificando e explicando cada uma das fases necessárias para o desenvolvimento do sistema. Também se identificam os algoritmos que compõem o mesmo.

No capítulo 6 são apresentadas algumas conclusões relativamente ao desenvolvimento do projeto, mencionando os objetivos alcançados, as limitações que ocorreram no decorrer do projeto e trabalho futuro sugerido neste âmbito.

Em anexo encontram-se a descrição das tecnologias utilizadas, o algoritmo principal do sistema, as principais características de cada classe implementada no sistema, e ainda o modelo de dados relativo ao Toursplan.

2 Sistemas de Recomendação

Hoje em dia recomendar produtos, serviços, informação ou até mesmo perfis de utilizador torna-se numa tarefa primordial mas, por outro lado, desafiante devido ao grande volume de informação existente na web.

Muitas vezes uma pessoa possui pouca ou quase nenhuma experiência para realizar escolhas de acordo com as várias alternativas que lhe são apresentadas. Saber se deve escolher determinado produto ou até mesmo se aquela informação é ou não viável pode revelar-se num entrave á sua seleção.

Neste ponto, os tipos de recomendações que podem auxiliar uma pessoa são a de forma direta, ou seja a opinião de alguém que convence essa pessoa, pelas suas razões pessoais como sendo fidedigna e correta, e por outro lado de forma indireta, por via de especialistas em determinadas áreas nos quais nos potenciam pelo poder de discurso e de justificação, de que essa mesma opção é a mais assertiva e a mais favorável para a nossa escolha.

Sabendo que uma recomendação pode levar a que um utilizador a aceite, e desta forma capta a sua atenção, ou por outro lado que se perca, é fundamental que sejam praticados os instrumentos de captação e de motivação para atrair público-alvo.

Dessa forma, os sistemas de recomendação contêm ferramentas personalizadas que permitem a que esse mecanismo de ação seja mais adaptado e focado nas principais áreas em que opera, tais como de carácter social, como as redes sociais existentes onde se sugere pessoas ou grupos nos quais determinado conjunto de pessoas do nosso círculo de amizades fazem parte, como exemplo o Facebook ou o Twitter. Também se pode verificar no comércio eletrónico onde se sugere determinado conjunto de produtos, de acordo com as suas compras ou pesquisas realizadas, tais como a Amazon ou o eBay.

As pessoas fornecem recomendações como entradas que o sistema agrega e direciona para os indivíduos considerados potenciais interessados neste tipo de recomendação. Um dos grandes desafios deste tipo de sistema é realizar a combinação adequada entre as expectativas dos utilizadores e dos produtos, serviços e pessoas a serem recomendados aos mesmos, ou seja, definir e descobrir este relacionamento de interesses é o grande problema (Cazella et al. 2010). O problema da recomendação foi estudado de forma intensiva e foram categorizados três tipos de sistemas de recomendação: sistema de recomendação baseado no conteúdo, na recomendação colaborativa e a abordagem híbrida.

O número inicial de dados quando o sistema de recomendação inicia é um grande problema. Dada as recomendações serem geralmente baseadas em dados existentes (por exemplo nos perfis de utilizador e no histórico de escolhas), que os sistemas necessitam de atacar este problema, de modo a que não sofram de *cold start problem* (Luz et al. 2015) .

Cold start problem caracteriza-se como sendo a adição de utilizadores no sistema que não se encontravam previamente nele, fazendo com que a identificação de utilizadores similares seja um problema, em que as similaridades resultantes não serão de boa qualidade. Este problema encontra-se nos sistemas de recomendação baseado em conteúdo e colaborativo, sendo que o último também tem problemas ao nível de *gray sheep*.

Gray sheep problem é traduzido no número de utilizadores cujos dados não têm correspondência, de modo a se encaixar num grupo de perfis de utilizador. Como resultado poderão existir utilizadores individuais, sem correspondência para com outros grupos de perfis de utilizador.

Neste subcapítulo será referenciado as principais técnicas de classificação de um sistema de recomendação, destacando o sistema baseado nos conteúdos (content-based profile) e o sistema colaborativo, dado que estes permitirão a construção de um sistema hibrido utilizado como sistema de classificação desenvolvido.

Esta técnica foi a escolhida, pelo facto do sistema ser baseado no sistema de recomendação da rede social Twitter, em que esta realiza a mineração de utilizadores existentes, fazendo posteriormente um pré-processamento textual para a construção do vetor de pesos.

Para o cálculo de pesos é implementado o algoritmo tf-idf, permitindo posteriormente que com o vetor de pesos, resultado do cálculo anterior, seja realizada a distância entre o perfil de utilizador e os restantes, feito a partir da medida de similaridade entre o cosseno do ângulo de dois utilizadores. Esta servirá para determinar a similaridade entre dois documentos, característica principal de um sistema colaborativo.

2.1 Perfil do utilizador

Para identificar o tipo de perfil do utilizador existem dois tipos de classificações:

- **Extração implícita**- A criação do perfil do utilizador na forma implícita ocorre de uma maneira simples. A informação do utilizador é adquirida pela utilização do sistema, sendo esta adquirida de forma transparente tais como, informações pessoais, histórico de compras, itens procurados, menus navegados entre outros (Gazzana & Silveira 2009) .

  Depois de adquiridas, essas informações são armazenadas numa base de dados para posterior utilização.

- **Extração explícita**- Neste método, é necessário que o próprio utilizador informe quais os seus interesses e preferências, em que esses interesses podem ser extraídos através de questionários, na avaliação de produtos de um determinado conjunto de categorias, pela interação com objetos de uma determinada área, ou de outra maneira que seja possível obter as informações pertinentes para a composição do perfil do utilizador. Depois desses dados serem adquiridos, cria-se o primeiro perfil no qual pode ser alterado pelo próprio utilizador ou através de outras perguntas que o sistema poderá fazer (Gazzana & Silveira 2009) .

2.2 Recomendação baseada em conteúdo

O perfil baseado no conteúdo é um sistema caracterizado por recomendar itens idênticos aos itens que um conjunto de utilizadores do sistema pesquisaram no passado (Sun et al. 2012) .

A descrição das recomendações para este tipo de recomendação pode ser obtida através de um conjunto de documentos que o próprio utilizador disponibiliza ou pelas suas ações como por exemplo, pelas suas visualizações, compras ou seleções, permitindo a construção de um modelo ou perfis de utilizadores baseado nos itens que cada utilizador avalia.

Um perfil de utilizador é uma representação estruturada dos interesses do utilizador, adotado para recomendar novos itens interessantes. O processo de recomendação basicamente consiste em corresponder os atributos de um perfil de utilizador, contra os atributos de um conteúdo de um objeto, resultando num julgamento entre o nível de interesse do utilizador naquele objeto (Tintarev & Masthoff 2011).

Sendo uma recomendação baseada no conteúdo uma recomendação que permite sugerir a utilizadores um conjunto de itens com base em itens pesquisados no passado, estas são feitas pela comparação positiva que essas pesquisas foram avaliadas e as mais similares serão recomendadas para os utilizadores que fazem parte do mesmo sistema.

Uma das técnicas utilizadas para classificar a informação de um documento, bem como a importância e dos itens pesquisados e sugeridos de um perfil de utilizador, é a técnica tf-idf. Esta técnica baseia-se em *Information Retrieval*, que caracteriza-se na forma de obter os recursos de uma informação relevante, a partir de uma coleção de recursos de informação existente.

Quanto maior for o valor desse índice, mais importante é o termo para o documento em que ele ocorre.

Para que seja estabelecida a similaridade entre itens como roupas e brinquedos, por exemplo, seria necessária a identificação dos atributos nos itens a serem comparados (peso, preço, marca, etc.). No caso de os itens serem artigos (ou documentos), este processo de comparação pode ser facilitado, pois documentos podem ser considerados similares se compartilharem termos em comum. Sendo assim, a filtragem baseada em conteúdo é mais indicada para a recomendação de itens textuais, onde o conteúdo é geralmente descrito com *keywords* (Cazella et al. 2010) .

A tabela 1 demonstra as vantagens e desvantagens na utilização da recomendação baseada no conteúdo.

Tabela 1 Características na utilização da técnica baseada no conteúdo

Vantagens	Desvantagens
Não necessita de avaliação por parte dos utilizadores para recomendar itens.	Avaliação de conteúdo textual.
Consegue recomendar novos objetos introduzidos no sistema, sem ter sido ainda pontuados.	Não lida com questões complexas do sistema.
Utilizador independente, ou seja explora apenas os utilizadores ativos para a construção do perfil.	Superespecialização, ou seja utiliza métodos que avaliam os itens com os valores mais altos, não havendo espontaneidade.
Simples para informação textual.	Não é adequado para informação multimédia.

2.3 Recomendação colaborativa

Um sistema de recomendação colaborativa caracteriza-se como sendo uma recomendação em que o utilizador será recomendado com itens de pessoas com gostos ou preferências similares no presente (Sun et al. 2012), ou seja com base em pesquisas feitas e preferencialmente aceites por utilizadores que também participam no mesmo sistema. Por outras palavras "Utilizadores similares que têm preferências similares".

Pode-se concluir assim que num sistema de recomendação colaborativo, este funciona a partir da informação obtida, fruto da interação entre os utilizadores que possuem experiências similares. Estes métodos permitem prever uma avaliação que um utilizador faz a um item, baseada na similaridade entre a avaliação que outros utilizadores fizeram a outros itens.

Existem dois tipos de classificação relacionada com a recomendação colaborativa: baseada na similaridade entre dois itens (item–based) ou na similaridade entre dois utilizadores (user- based).

Na recomendação baseada no item, esta tem como objetivo identificar a avaliação entre dois itens, a partir da avaliação que os utilizadores fazem sobre cada um deles.

Por outro lado na recomendação baseada no utilizador, esta é feita a partir da avaliação dos itens que ambos os utilizadores realizaram.

Em cada uma das classificações o princípio é o mesmo, existindo vários conceitos para o cálculo de similaridade, sendo algumas nomeadas de seguida.

As medidas de similaridade apresentadas são classificadas como *cosine similarity*, *Jaccard coeficient* e *correlation distance*.

Cosine similarity, ou similaridade do cosseno, é uma medida de similaridade que mede o cosseno do ângulo entre dois documentos. Nesta medida, os dois utilizadores são tratados como sendo dois vetores em espaços m-dimensionais. Esta medida é muito utilizada em documentos, para *information retrieval* e também para a técnica de *clustering*. Uma das características importantes nessa medida é a independência no tamanho dos documentos (Huang 2008). Por outras palavras, documentos com o mesmo número de atributos, mas com tamanhos diferentes, são tratados de forma semelhantes incluindo os que deles possam derivar.

Para o cálculo da similaridade entre o cosseno do ângulo entre dois utilizadores pode ser descrita a partir da expressão (1) (Udagawa 2013).

$$sim(u,v) = \frac{\sum_{i=1}^{N} w_{i,j} * w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} * \sqrt{\sum_{i=1}^{N} w_{i,q}^2}} \quad (1)$$

Para a técnica *Jaccard coefficient*, esta representa a medida de similaridade como uma representação dividida entre a união dos objetos.

Para o cálculo de documentos textuais, *Jaccard coefficient* compara a soma dos pesos dos termos partilhados com a soma dos pesos dos termos que estão presentes nos dois documentos, mas que não são partilhados (Huang 2008) .

O cálculo da técnica *Jaccard coefficient* pode ser descrita a partir da expressão (2) (Huang 2008) , em que o valor de similaridade varia entre 0 e 1.

$$SIM_J(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a} \cdot \overrightarrow{t_b}}{\left|\overrightarrow{t_a}\right|^2 + \left|\overrightarrow{t_b}\right|^2 - \overrightarrow{t_a} \cdot \overrightarrow{t_b}}. \quad (2)$$

No caso da técnica de *correlation distance*, uma das técnicas existentes é denominada de *Pearson correlation coefficient*. Esta mede a relação entre dois vetores arbitrários, que não necessariamente devem ter a mesma dimensão, permitindo representar linearmente dois *data sets*.

Para o cálculo da técnica de *Pearson correlation coefficient*, esta pode ser realizada a partir da expressão referida em (3) (Huang 2008) .

$$SIM_P(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{m\sum_{t=1}^{m} w_{t,a} \times w_{t,b} - TF_a \times TF_b}{\sqrt{[m\sum_{t=1}^{m} w_{t,a}^2 - TF_a^2][m\sum_{t=1}^{m} w_{t,b}^2 - TF_b^2]}} \quad (3)$$

Ao contrário das outras técnicas descritas, o valor de similaridade varia entre -1 e 1. Quando $\overrightarrow{t_a} = \overrightarrow{t_b}$ o valor é 1, significando que são iguais, e quando o valor é zero, significa que são diferentes.

A construção de um sistema fiável e prático a partir de medidas similares é fundamental em tópicos de pesquisa no campo dos sistemas colaborativos (Sun et al. 2012) .

A filtragem baseada na recomendação colaborativa difere da recomendação anterior, por esta não necessitar de obter compreensão da informação existente nos itens.

A tabela 2 demonstra as características na utilização da recomendação colaborativa.

Tabela 2 Características na utilização da técnica colaborativa

Vantagens	Desvantagens
Simula recomendações reais entre utilizadores.	Requer um grande volume de dados para funcionar corretamente.

Não necessita de compreensão da informação existente.	Problema de novo utilizador.
Utiliza conteúdo social existente na Web.	Dispersão de dados.

2.4 Recomendação híbrida

No caso de um sistema hibrido, este une o sistema de recomendação de perfil baseado no conteúdo com o sistema de recomendação colaborativa, permitindo juntar o que de melhor existe em cada uma das recomendações indicadas.

2.5 Sistemas de recomendação no turismo

Os sistemas de recomendação para o turismo, também denominado de *e-tourism*, resultaram numa necessidade de promoção das agências e empresas ligadas ao sector dos seus serviços e nos pacotes turísticos, sugerindo destinos, pontos de interesse e eventos.

Este tipo de recomendação pode resultar de duas formas: implícita, em que os sistemas adquirem informação relativamente á interação e que o utilizador teve *online* e de forma explicita, em que o utilizador divulga a sua própria informação.

De uma forma ou de outra, ambos os tipos de recomendação permitem responder às necessidades e preferências que cada utilizador tem.

Existem vários sistemas de recomendação ligados ao turismo, a seguir referem-se alguns dos que se consideraram mais importantes. Procede-se ainda a uma breve comparação entre os objetivos da abordagem proposta e cada um dos sistemas referidos.

2.5.1 TripAdvisor

- É um site de turismo que avisa a cada utilizador informação relacionada com viagens, atividades e recomendações de hotéis personalizadas, com base nas suas preferências individuais e histórico de pesquisas realizadas anteriormente no site. Também tem uma componente social que permite a partilha de opiniões e avaliar um conjunto de elementos, tais como hotéis, restaurantes, atrações e reserva de um fim-de-semana, para a assistência no processo de tomada de decisão da empresa.
  A figura 2 demonstra a página inicial do site TripAdvisor (TripAdvisor, 2015) .

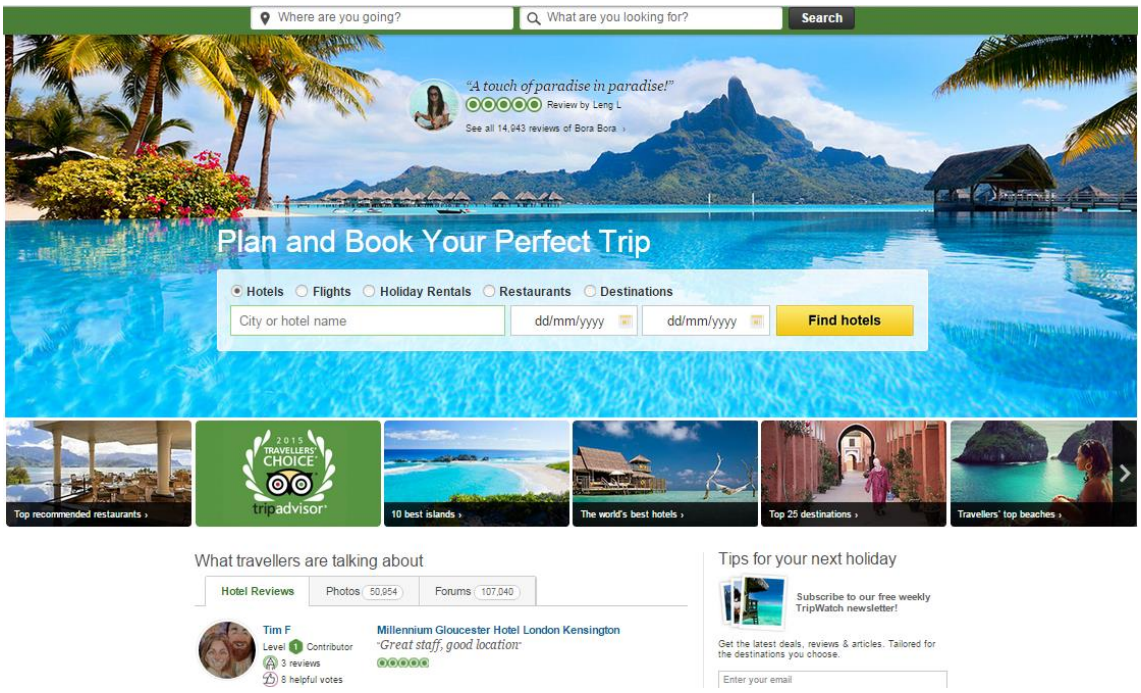


Figura 2 Site TripAdvisor (TripAdvisor, 2015)

2.5.2 Heracles

- Este sistema realiza uma filtragem baseada no conteúdo da informação dos turistas através de várias fontes de informação e de motores de busca ao longo da web.

  Dada a informação ser usada instantaneamente, em vez de ser acumulada para a geração de conhecimento, o sistema apresenta também um método com supervisão automática que, através de um maior número de parâmetros de entrada, refina-os permitindo um melhoramento nos resultados encontrados.

  A figura 3 ilustra o sistema de recomendação descrito anteriormente, onde a área de planeamento de uma viagem, desde o início até ao fim da viagem no qual, segundo o sistema Heracles, reúne o melhor das fontes de viagem da web num único ambiente integrado (Heracles, 2015).

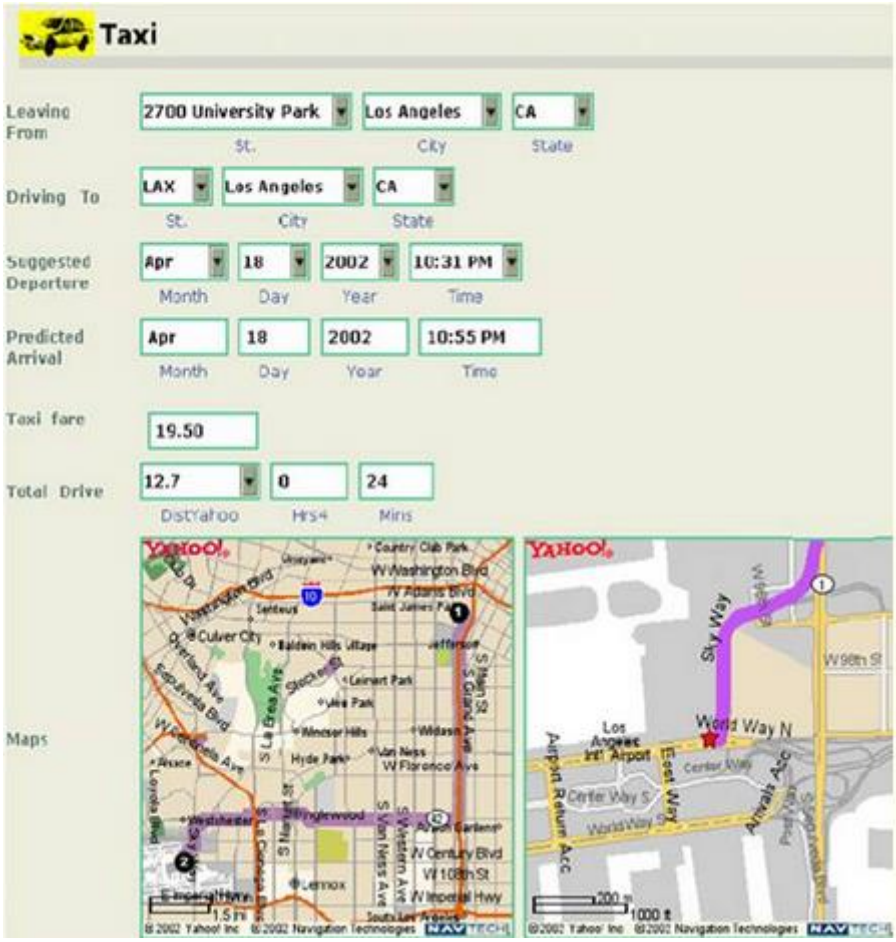


Figura 3 Sistema de recomendação Heracles (Heracles, 2015)

Depois de selecionar um local, este extrai as datas para esse local, verifica a sua localização, o tempo meteorológico e até realiza uma recomendação sobre de que forma se deve viajar, seja de avião, conduzir um carro ou ir de autocarro para o local (Heracles, 2015).

2.5.3 Travel-Buddy

- Este sistema usa a abordagem hibrida, combinando a recomendação baseada no conteúdo com a de filtragem colaborativa, demográfica e semântica.

  Na filtragem colaborativa, esta foca na utilização do utilizador em *user link*, porque a abordagem baseada no conteúdo é mais eficaz em itens dinâmicos. Os dados demográficos do utilizador são utilizados por defeito para a possibilidade do utilizador em utilizar o *user link*. Á medida que o perfil do utilizador constrói o seu perfil colaborativo, este vai permitir o utilizador na utilização de *user links*, com base na avaliação dada a alguns itens (Sharef 2013) .

A abordagem colaborativa é depois combinada com a recomendação semântica, onde a recomendação semântica armazena os *links* dos conceitos, numa base de dados subjacente. Ou seja numa vista em que o resultado de uma seleção de dados pode ser representada numa tabela (view).

A figura 4 demonstra a página inicial do site Travel-Buddy (Travel-Buddy, 2015) .

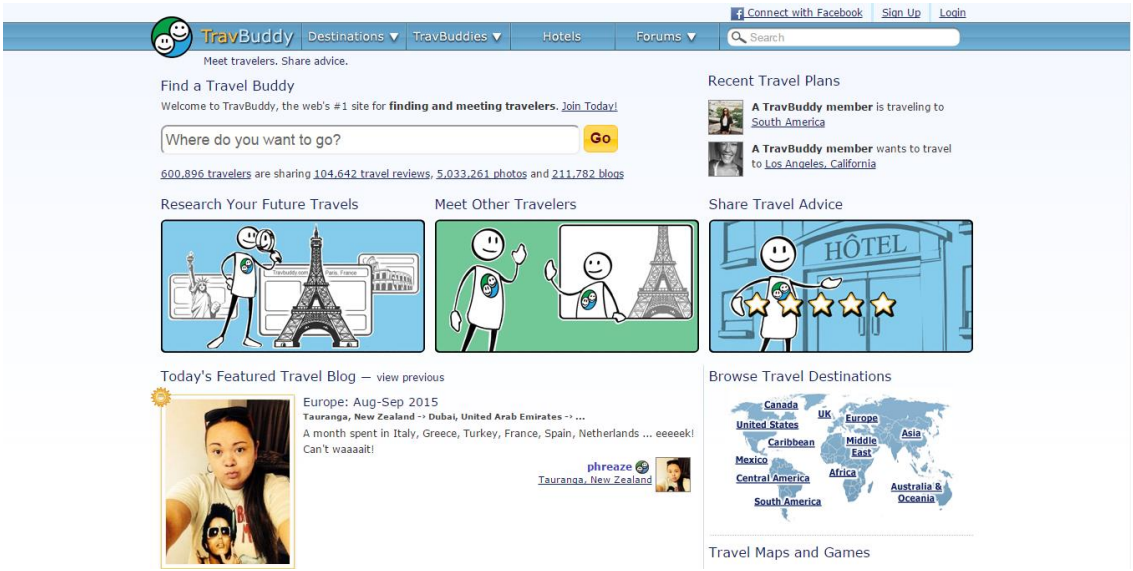


Figura 4 Site Travel-Buddy (Travel-Buddy, 2015)

2.5.4 Dietorecs

- Classificado como sendo um sistema de recomendação hibrido, Dietorecs é um sistema que suporta múltiplas decisões e disponibiliza recomendações personalizadas para o utilizador. Assim, o utilizador deve escolher que tipo de pesquisa deseja realizar quando entra no sistema.

  O utilizador pode aceder ao sistema a partir de três portas funcionais, que encaixa com grupos complementares de estilos de decisão (Fesenmaier & Ricci 2003) .

  A primeira porta contém informação relacionada com destinos de viagem e categorias de estadia.

  Para a segunda porta é baseada na personalização de uma viagem, permitindo juntar um pacote de itens disponíveis no catálogo.

No caso da porta três, esta é mais orientada á recomendação do utilizador, sendo este processo conduzido pelo sistema, com o objetivo de guiar o utilizador em opções de recomendação. O utilizador pode rotular a opção como interessante ou no caso de não ser do seu agrado, como não interessante.

A figura 5 demonstra a página inicial na primeira porta do sistema Dietorecs (Fesenmaier & Ricci 2003) .

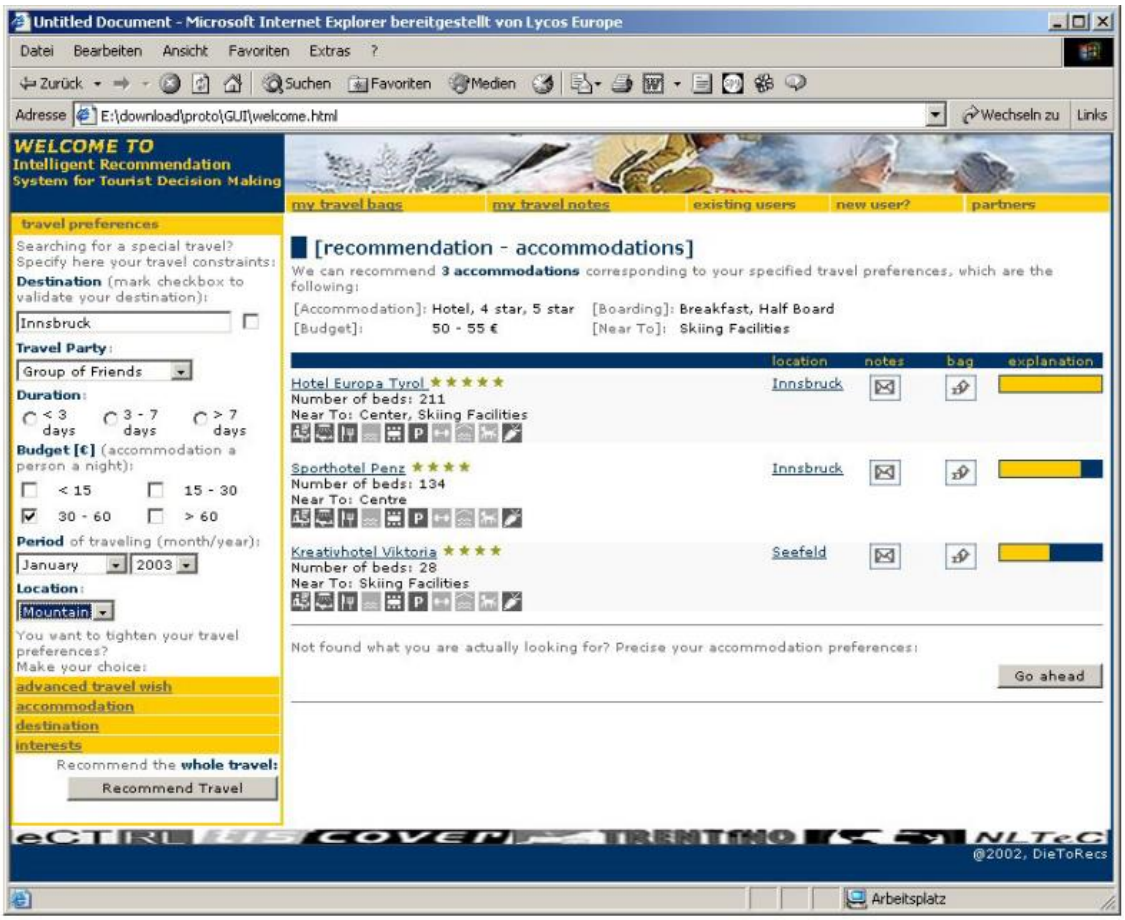


Figura 5 Sistema Dietorecs na primeira porta (Fesenmaier & Ricci 2003)

2.5.5 Breve comparação com a abordagem proposta

Relativamente ao sistema desenvolvido, este diferencia-se dos mencionados anteriormente, pelo facto de ser um sistema cujo principal objetivo é determinar a importância de cada perfil de utilizador analisado, permitindo agrupar aqueles que têm o mesmo valor similaridade, a partir do valor do cosseno de similaridade entre dois documentos.

O sistema TripAdvisor, para além da recomendação de atividades, viagens, hotéis personalizados ou ate informação que possa ser do interesse do utilizador, de acordo com as

suas preferências, permite também a avaliação e partilha de informação sobre as suas experiências turísticas ao longo da web. O que em comparação com o sistema desenvolvido, tem particularidades distintas já que este extrai informação proveniente de uma base de dados e não efetua avaliação das suas experiências.

Relativamente ao sistema Heracles, este é apresentado como sendo um sistema baseado no conteúdo, tendo um método com supervisão automática que, através de um maior número de parâmetros de entrada, refina-os, melhorando os resultados encontrados para o planeamento de uma viagem. No caso do sistema desenvolvido, este não tem por base um motor de busca nem de planeamento de uma viagem, mas sim em determinar similaridades entre perfis de utilizador.

O sistema Travel-Buddy baseia-se numa abordagem hibrida, combinando a recomendação, para além de baseada no conteúdo e a de filtragem colaborativa, também a recomendação demográfica e semântica. O sistema desenvolvido apenas se baseia numa recomendação baseada no conteúdo e colaborativa, para além de que para a construção do perfil do utilizador, este não se baseia na avaliação prévia de itens.

Finalmente o sistema Dietorecs que se baseia num sistema hibrido de múltiplas decisões, disponibilizando recomendações personalizadas para o utilizador. Esta diverge do sistema desenvolvido na recomendação personalizada, estando orientada aos gostos e aptidões do utilizador bem como na avaliação dos destinos turísticos e hotéis.

3 Business Intelligence

Business Intelligence pode definir-se como um conjunto de arquiteturas, bases de dados, ferramentas de análise, aplicações e metodologias, cujo principal objetivo é permitir fácil acesso aos dados (e modelos) para facilitar aos gestores a realização de análises aos indicadores de negócio (informação em conhecimento).

O BI é uma ferramenta/tecnologia importante para o desenvolvimento de uma organização, pois possibilita o tratamento e sumarização dos dados que são produzidos pelas várias aplicações colaborativas da organização tais como ERP, CRM, entre outros.

ERP, caracterizado como *Enterprise Resource Planning* é uma solução que integra todos os processos organizacionais numa só plataforma, cobrindo todas as áreas funcionais de uma empresa. Esta tem como principal objetivo uma gestão mais eficiente do negócio da empresa, automação da informação e integração entre os departamentos e sectores de uma empresa.

No caso do CRM, ou *Customer Relation Management*, esta solução tem como objetivo o planeamento das ações dirigida ao cliente. Ou seja, tenta potenciar as ações de vendas e em como a empresa pode chegar de forma eficiente a um cliente, numa altura em que o mercado se encontra cada vez mais competitivo e concorrente. Em conjunto com o BI, permite uma melhor análise do cliente, dando métricas e o planeamento necessário para que a qualidade da comunicação por parte da empresa possa resultar nas exigências do cliente.

Aplicações de BI com *Cloud Computing*, são aplicações de trabalho que permitem o acesso leve e ágil a aplicações BI, permitindo a tomada de decisões assentes em conhecimento e na criação de oportunidades de negócio. Agir de acordo com estas oportunidades, possibilita a criação de estratégias de ação de forma a atingir vantagem competitiva (Goel & Aggarwal, 2013). Esta ferramenta, para além das vantagens referidas anteriormente, disponibiliza ao utilizador um acesso rápido, escalabilidade da aplicação, no caso de esta aumentar em número de utilizadores, e também na sua disponibilidade e acessibilidade.

Do ponto de vista das organizações virtuais, o BI é visto como uma ferramenta de prospeção. É também um termo que abrange uma grande gama de software capaz de analisar, tratar, consolidar e disponibilizar informação ao utilizador, deixando-o assim mais preparado para tomar certas decisões dentro da organização. O BI engloba também software de extração, transformação e carregamento de dados denominado por ETL, data mining, processamento de dados multidimensional (OLAP) e *reporting* (Gangadharan & Swami, 2004).

3.1 Componentes e funcionalidades

As funcionalidades fundamentais das ferramentas de BI podem ser: sumarização de informação, estruturando e integrando dados; criação de relatórios e extração de informação. Geralmente, as ferramentas de BI oferecem um conjunto de funcionalidades, tecnologias e software que integram dados heterogéneos de diferentes fontes de forma a produzir conhecimento para o utilizador final (Olszak & Ziemba, 2007). As principais funcionalidades e componentes são:

- OLAP (Online analytical processing): OLAP é um componente capaz de fazer análises em tempo real sobre um DW ou um só Data Mart – Subconjunto de dados de um DW. O OLAP produz views multidimensionais dos dados, sumarizados e preparados para serem rapidamente analisados pelo utilizador comum. Basicamente são feitas consultas que procuram tendências ou fatores críticos para um objetivo previamente definido (Ranjan, 2009).

- Data Mining: Permite explorar enormes quantidades de dados com a finalidade de descobrir padrões que representem informação, procurando padrões ocultos e extraindo conhecimento de DW organizacionais. O Data Mining não substitui outras técnicas de análise de dados como o anteriormente falado OLAP ou análises estatísticas, simplesmente é um complemento (Fayyad et al. 1996).

O Data Mining pode envolver várias técnicas como redes neuronais, estatística ou árvores de decisão para lidar com os dados (Rygielski, Wang, & Yen, 2002). Técnicas como associação, *clustering* ou classificação servem para dividir e organizar os dados permitindo assim a descoberta de padrões mais facilmente (Chien & Chen, 2008).

Esta fase será referida posteriormente em maior detalhe na seção 2.4, dado ser um dos componentes utilizados para a realização do sistema proposto.

- Data Warehouse (DW): Como já foi possível entender durante este capítulo, este componente essencial é um único, completo e consistente armazém de dados. Está otimizado para distribuir e manter disponíveis os dados perante os utilizadores finais de uma forma compreensível (Ranjan, 2009).

- Data Marts: Conjunto mais pequeno de dados que normalmente dizem respeito a um conjunto de processos ou áreas organizacionais. Desenhados para facilitar analises feitas pelos utilizadores, estes dados são baseados em necessidades previamente definidas pelos departamentos da organização virtual (Tuncay & Belgin, 2010).

- ETL (Extraction-Transformation-Load): Conjunto de ferramentas responsáveis por tratar os dados vindos de várias fontes, torna-los homogéneos entre si e integra-los no DW. Processo bastante importante de um sistema de BI pois é quando os dados são preparados para serem interpretados pelas diversas aplicações de BI (Tuncay & Belgin, 2010).

O sistema ETL adiciona valor significativo á informação:

- Deteta e corrige problemas relacionados com a qualidade dos dados;
- Disponibiliza medidas documentais de confiança na informação;
- Ajusta a informação de múltiplas fontes para serem usadas em conjunto;
- Estrutura a informação para ser usada por ferramentas relativas a utilizadores finais.

Os passos tradicionais de um processo ETL foram expandidos para quatro passos (Kimball & Wiley,2004):

- Extrair (Extract) a informação de fontes existentes no sistema;
- Limpar (Clean) para aumentar a qualidade dos dados e consistência;
- Consolidar (Conform) os dados para que fontes separadas possam ser utilizadas em conjunto;
- Entregar (Delivery) a informação num formato disponível para apresentação.

A figura 6 identifica os passos realizados num processo ETL (Kimball & Wiley, 2004).

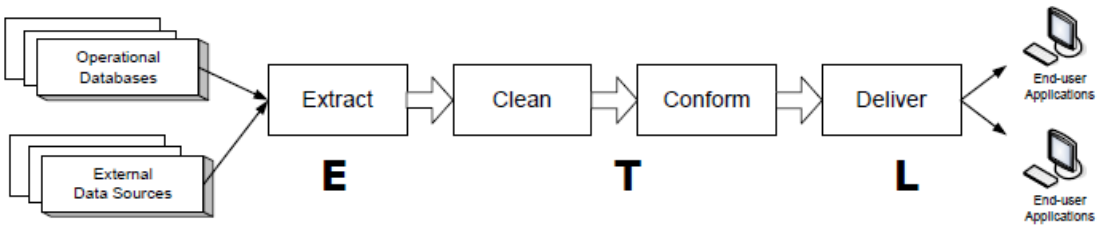


Figura 6 Processo ETL (Kimball & Wiley, 2004)

Esta fase é utilizada para a extração de informação da base de dados Toursplan, fazendo a respetiva limpeza dos dados extraídos, a partir da remoção de caracteres não necessários para a análise da informação, consolidando-a quando esta for armazenada num documento

estruturado para cada utilizador, verificando a respetiva similaridade entre cada utilizador. No final será realizada a fase de agrupamento dos utilizadores, de acordo com a sua similaridade.

A figura 7 demonstra uma arquitetura com os componentes e funcionalidades descritos anteriormente, adaptado de (Renko, 2011).

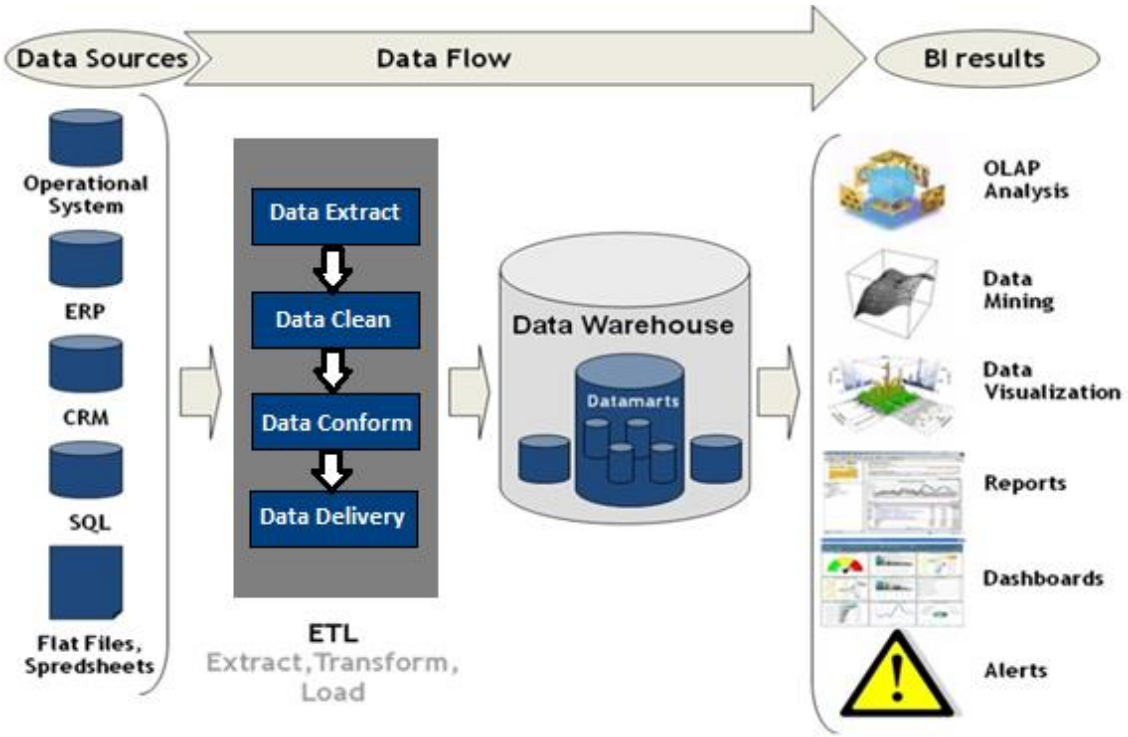


Figura 7 Etapas realizadas num processo de BI (Renko, 2011)

O termo *business intelligence tool* representa o software que permite aos utilizadores verem e utilizarem grandes conjuntos de dados complexos (Schiff, 2010).

3.2 Descoberta do conhecimento

A era da informática na sociedade traduziu-se num aumento exponencial de dados provenientes nas capacidades de geração e coleção de dados, oriundas de várias fontes.

Isto traduziu-se num problema ao nível da qualidade e importância dos dados armazenados em que, segundos estudos recentes, cerca de 40% dos dados adquiridos são sujos, de uma forma ou de outra (Maimon & Rokach 2010).

Para solucionar este problema seria necessário, que as pessoas despendessem parte significativa do seu tempo para corrigir erros e limpar dados desnecessários. Ora, no mundo real este processo manual traz consequências a nível temporal e na respetiva eficácia, já que o erro humano é propício e não garante que o processo se encontre completamente de acordo com o pretendido. Á medida que os dados gerados iam crescendo, mais se verificava que era necessário utilizar técnicas e ferramentas auxiliares que permitissem transformar esse grande volume de dados, em informação importante e em conhecimento.

Em conjunto com o Data Mining, o processo de descoberta de conhecimento permite automatizar um processo e adquirir, de um grande volume de dados, informação relevante e conhecimento útil que até então não era possível, tendo custos eficazes quando se trata em obter um nível aceitável na qualidade de dados.

O processo de descoberta de conhecimento em bases de dados é explicado na figura 8, consistindo nas seguintes etapas (Maimon & Rokach 2010):

1. Os dados de limpeza para remover ruídos e inconsistências.
2. A integração de dados para obter dados de várias fontes.
3. Etapa de seleção de dados, onde os dados relevantes para a tarefa são recuperados.
4. Os dados da etapa de transformação, onde os dados são transformados numa forma adequada para a análise de dados.
5. Análise de Dados, onde consultas complexas são executadas para uma análise em profundidade.

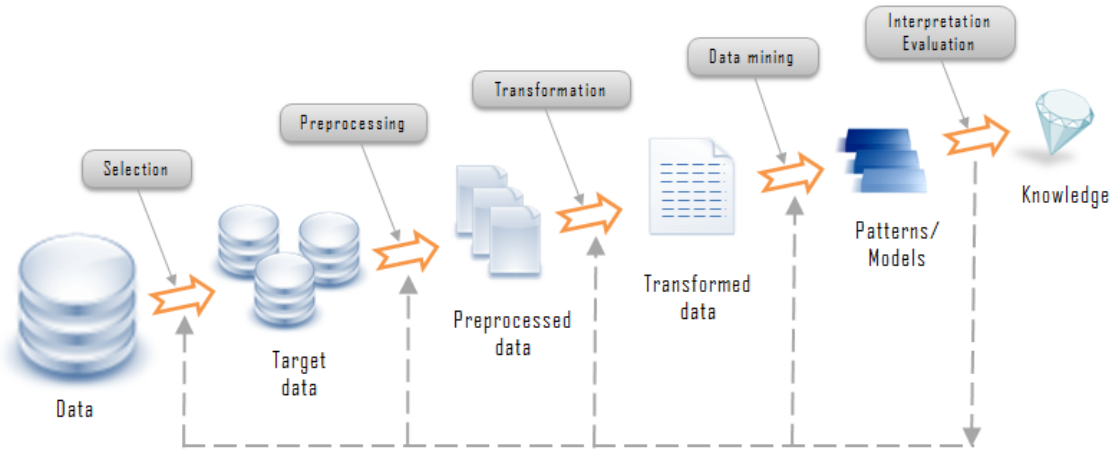


Figura 8 Processo de extração de conhecimento (Rithme, 2015)

3.3 Data Mining

Data Mining é o processo de proposição de várias consultas e extração de informações úteis, padrões e tendências frequentemente desconhecidos, a partir de grandes quantidades de dados armazenados em bases de dados (Thuraisingham 1998).

Para que possa ser realizada, é previamente necessário um pré-processamento de dados, bem como a posterior interpretação dos dados obtidos. Utiliza técnicas de inteligência artificial para detetar relações de similaridade ou discordância entre dados, tais como a descoberta de Classes, Associação, *Clusters* ou Padrões.

De entre as técnicas utilizadas existem dois tipos: a aprendizagem não supervisionada (*unsupervised learning*), caracterizada por não necessitar de conhecimento prévio para a aprendizagem, ou seja a partir de observações e descobertas, e a aprendizagem supervisionada (*supervised learning*), que caracteriza-se pela classificação de dados a partir de exemplos de dados rotulados, para o treino do *data set*.

Apesar de existir mais técnicas para o processo de extração e descoberta de conhecimento, tais como o Clustering, Classificação, Regressão, Associação, Sequenciação e Análise de Desvios, será descrita pormenorizadamente apenas a técnica de *clustering*, tendo esta sida utilizada para o desenvolvimento do sistema.

3.3.1 Clustering

Um algoritmo de *clustering*, é um algoritmo classificado como não supervisionado que cria automaticamente grupos, com base na partição dos dados em análise, de forma a identificar instâncias dentro de um *cluster* que sejam similares, de outras que não o sejam.

Esta técnica é usada para encontrar dentro de um determinado conjunto de grupos, aqueles que partilham o mesmo conjunto de dados.

Após a sua realização é possível a:

- Sintetização dos resultados de cada segmento base para determinar apenas as características mais comuns em cada grupo;
- Utilização dos dados para a aplicação de outras técnicas de *data mining*.

Dependendo do método utilizado para a utilização do algoritmo de *clustering*, esta pode ser classificada nos seguintes tipos:

- **Algoritmo partitivo** - Classificado como sendo não hierárquico, este tipo de agrupamento realiza partições de um conjunto de dados na geração de agrupamentos, dividindo n dados existentes em k partições.

  O valor de partições tem de ser conhecido previamente, cujo método de partição utilizado são o de heurísticas iterativas.

  Permite dessa forma encontrar alguma partição de dados que possa maximizar algum critério relacionado com a similaridade entre as observações dentro do agrupamento com as não similaridades observadas entre diferentes observações no agrupamento. Quando comparado com o método hierárquico, o método por particionamento é mais rápido porque não é necessário calcular e armazenar, durante o processamento, a matriz de similaridade (Doni 2004).

  A figura 9 demonstra o agrupamento partitivo, em que os valores indicam o número de iterações e para cada iteração, são identificados os *clusters* resultantes dessa mesma partição.

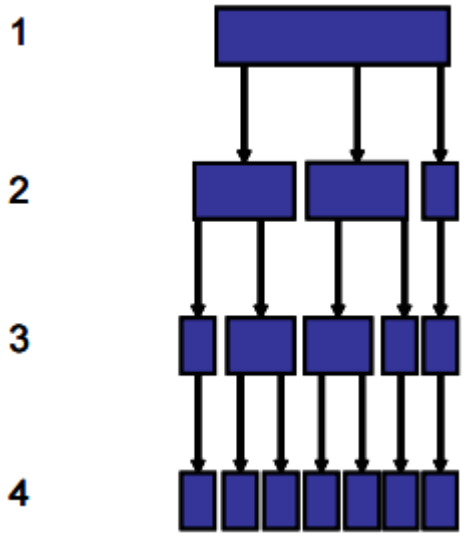


Figura 9 Algoritmo partitivo

- **Algoritmo hierárquico** - Neste tipo de agrupamento é classificado de dois tipos: o agrupamento hierárquico aglomerativo e por divisão.

No agrupamento hierárquico aglomerativo este realiza, para cada nível, a geração dos *clusters* pela junção dos *clusters* anteriores, sendo que este processo começa pelo nível inicial, em que apenas existe um único objeto para cada cluster.

Por outro lado, no agrupamento hierárquico por divisão, este é realizado a partir da divisão do *cluster* inicial até chegar ao número de *clusters* necessários.

O resultado é uma árvore de decisão, denominada dendograma, que traduz graficamente para cada nível, a relação entre os *clusters*.

Um corte no dendrograma a qualquer nível de aglomeração produz uma classificação em k subgrupos ($1 \leq k \leq n$), sendo n o número de indivíduos.

A figura 10 demonstra um exemplo de uma árvore de decisão dendograma, com os dois tipos de agrupamento hierárquico descrito anteriormente (Metz & Monard 2005) .

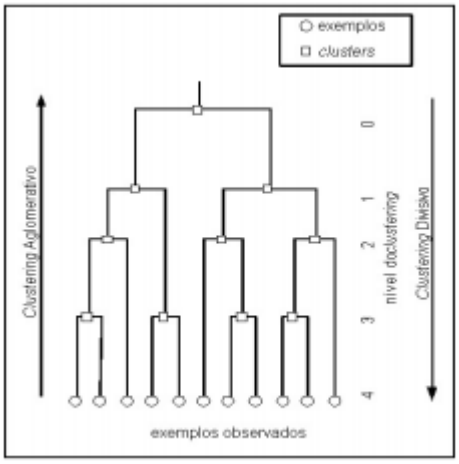


Figura 10 Dendograma resultante de uma classificação hierárquica (Metz & Monard 2005)

- **Algoritmo baseado em grelha** - Este agrupamento tem como objetivo quantificar o espaço de dados num número finito de células para que os dados sejam as operações de *clustering* realizadas sobre o espaço. Este agrupamento lida com grandes volumes de dados.

- **Algoritmo baseado em densidade** - Este agrupamento tem como principal característica continuar o crescimento de um dado *cluster*, na medida em que a

sua densidade permita obter uma aproximação dos objetos vizinhos. Desta forma, o *cluster* ao verificar um objeto na sua vizinhança que não pertença a um cluster, integra-o.

Também é caracterizado pelo agrupamento de *clusters* de forma arbitrária, ao contrário de agrupamentos cujo principal objetivo é o cálculo da distância entre objetos.

Existem várias algoritmos de *clustering*, alguns deles descritos anteriormente, sendo que o algoritmo escolhido para a fase de agrupamento denomina-se como algoritmo partitivo, denominado de método *k-Means*.

O método de *k-Means* é um algoritmo simples, classificado como algoritmo partitivo e caracteriza-se por utilizar uma partição inicial aleatória, onde vai atribuindo aos *clusters* novos padrões, com base na similaridade entre o padrão e o *cluster*, até atingir o ponto de convergência. Esta é feita pela análise e comparação de valores numéricos de dados.

A sua simplicidade deve-se ao facto de este se traduzir na complexidade temporal O(tkn), em que t são as iterações, *k* são os objetos, e *n* o número de *clusters* e *t* as iterações . Geralmente a medida usada para calcular o valor médio entre os objetos é a partir da distância entre o objeto e o centroide, denominada de distância Euclidiana.

O algoritmo fornece automaticamente uma classificação automática sem a necessidade de nenhuma supervisão humana, ou seja, sem a necessidade de nenhuma pré-classificação. Por causa desta característica, o *k-Means* é considerado como sendo um algoritmo de mineração de dados não supervisionado.

O objetivo de algoritmo *k-Means* é um conjunto de *clusters* que minimize o critério de erro quadrado.

A figura 11 exemplifica a utilização da técnica de *clustering k-Means* (MATLAB, 2012) .

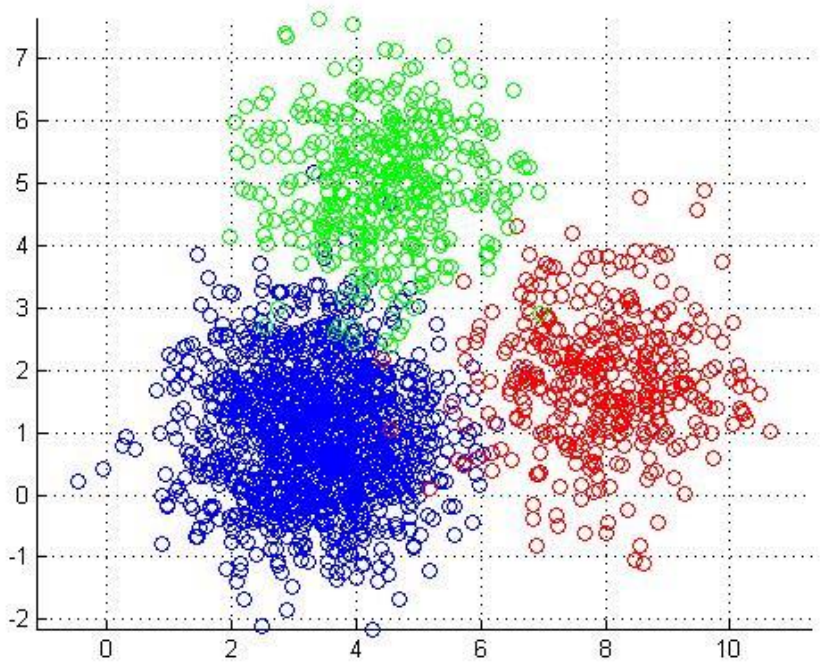


Figura 11 *Clustering k-Means* (MATLAB, 2012)

Os passos que são realizados neste algoritmo de *clustering* são os seguintes (Gualtar 2011) :

1. Escolher k centroides de *clusters* coincidentes com k padrões escolhidos aleatoriamente dentro do conjunto de padrões;
2. Atribuir cada padrão ao centro do *cluster* mais próximo (baseando nos valores médios já existentes em cada objeto do *cluster*);
3. Recalcular o centro do *cluster* utilizando os membros correntes, ou seja o novo centroide;
4. Se o critério de convergência não é atingido, vai para o passo 2, até haver estabilidade dos objetos nos *clusters*.

Este tipo de algoritmo permite obter bons resultados em *clusters* isolados e compactos.

3.3.2 Classificação

Representa a classificação de um item e atribui a outros tipos de classes predefinidas (Fayyad et al. 1996).

3.3.3 Regressão

Permite o ajustamento de dados para duas variáveis e a descoberta de relações funcionais entre elas (Fayyad et al. 1996), desde que uma possa ser predita por outra.

3.3.4 Associação

Permite definir que dados estão relacionados, ou seja identificar regras de associação que identifiquem aqueles que ocorrem juntos num conjunto de dados.

Ou seja permite encontrar uma descrição compacta para um subconjunto de dados (Fayyad et al. 1996). Aplica-se nos casos em que se deseja estudar preferências e afinidades.

3.3.5 Sequenciação

Esta técnica permite utilizar um tipo de padrão nos dados para determinar que tipos de sequências podem ser determinadas (Thuraisingham 1998).

3.3.6 Análise de desvios

Descoberta das alterações mais significativas nos dados a partir de dados previamente medidos ou de valores normativos (Fayyad et al. 1996).

3.4 Text Mining

As técnicas de *text mining* caracterizam-se como sendo o processo de extração e geração de conhecimento de informação considerada importante, ao nível textual de um documento. Torna-se assim evidente que um dos principais objetivos desta técnica é o processamento de documentos não estruturados, ou seja documentos em que ainda não foi realizada nenhuma modelação, de forma a extrair índices numéricos que sejam considerados importantes.

Esta técnica permitirá assim que o resultado do processo de *text mining* seja utilizado em técnicas de *data mining*, tais como análises estatísticas ou aprendizagem da máquina.

A sua utilidade específica terá em consideração no âmbito do projeto, contudo a técnica pode ser aplicada para (Figueiredo 2010):

- Resumir longos textos em versões mais compactas;
- Filtrar informação de um conjunto extenso de documentos;
- Extrair palavras-chave importantes de corpos textuais, entre outras utilidades.

O algoritmo de *text mining,* funciona através da aplicação de um filtro linguístico e de um filtro estatístico, aos quais é depois adicionado um importante filtro de domínio (Figueiredo 2010).

- O filtro linguístico trata de filtrar o texto de entrada para que fiquem apenas os termos que possam ser mais interessantes;
- O filtro de domínio dá mais valor a certas palavras filtradas anteriormente, por forma ao resultado final ser ponderado;
- O filtro estatístico trata de ordenar e classificar todas as palavras-chave que tiverem sido assumidas como valiosas nas fases anteriores do algoritmo.

4 Twitter Trending Topic Classification

Com a explosão de informação em sites de *microbloging* e apesar do Twitter permitir identificar uma lista de tópicos populares de *tweets*, conhecidas como *Trending Topics*, em tempo real é difícil de entender o que significam. Esta classificação é realizada com o objetivo de melhor identificar alguns tópicos, permitindo uma melhoria na eficácia e precisão no retorno da informação.

São utilizadas duas abordagens para a classificação de tópicos de tendências: a classificação baseada em texto (Text-based Modeling) e a classificação baseada em rede (Network-based Modeling).

a) Classificação baseada em texto é feita a partir do modelo *Naive Bayes Multinomial*. Este consiste num método simples e eficaz de classificação que será importante para a classificação de texto, dado que se caracteriza como sendo um método de aprendizagem probabilístico que determina a probabilidade de um documento ocorrer numa classe (Manning et al. 2009).
b) Classificação baseada em rede é realizada a partir do algoritmo *user-similarity*, com base nos dados recolhidos pelo algoritmo *Weighted Page Rank*. Este algoritmo tem em conta a importância da existência de *inlinks* e *outlinks* nas páginas web e distribui os resultados do ranking com base na popularidade dessas mesmas páginas (Tyagi & Sharma 2012).

A figura 12 mostra a arquitetura do sistema utilizado para a classificação dos tópicos do Twitter.

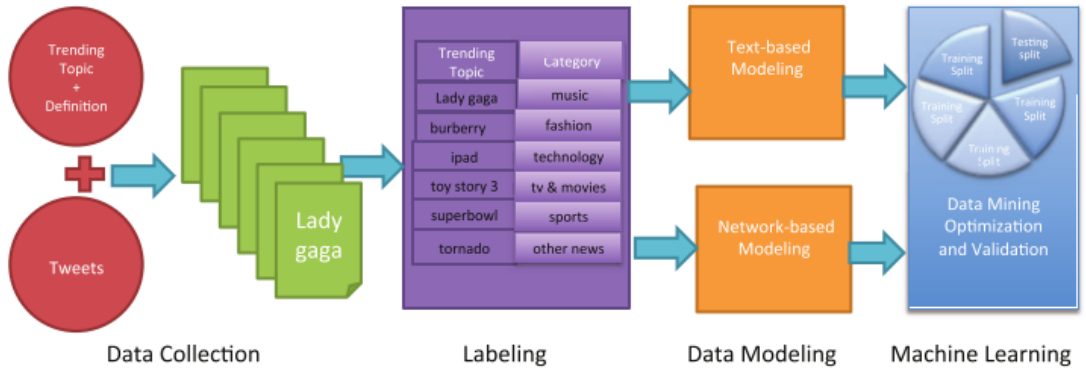


Figura 12 Arquitetura do sistema para classificação de tópicos Twitter (Lee et al. 2011)

4.1 Data Collection

Caracterizada pela coleta de informação a partir do site "What The Trend" é criada, para cada tópico, uma tendência de tópico.
É tido em conta a definição de tendência, ou seja a indicação da importância daquele termo para o utilizador.

Para cada tendência de tópico é gerado um documento para que, enquanto os tópicos estiverem na fase de processo de agrupamento de tendências, os restantes tópicos com o mesmo nome sejam armazenados num documento com o mesmo nome do tópico.

Tendo o tweet mais do que um tópico de interesse, é armazenado o mesmo número de tópicos em documentos.

4.2 Labeling

Nesta fase os dados são rotulados pelas definições de tendência do tópico e por alguns tweets. O processo de rotulagem é feito a partir de dois anotadores que, em caso de discordância, um terceiro anotador intervém. Foram rotuladas cerca de 1000 amostras de tópicos selecionadas, tendo sido estreitadas para cerca de 768, dada a ausência em algumas amostras de definição de tendência de tópico ou devido ao terceiro anotador não finalizar a *label*.

O processo de *labeling* foi feito manualmente em cerca de 3005 tópicos, porque alguns temas são comuns em mais do que um tópico.

As figuras 13 e 14 demonstram a distribuição dos 768 tópicos pelas 18 classes escolhidas e a tendência de tópicos para a classe Tecnologia.

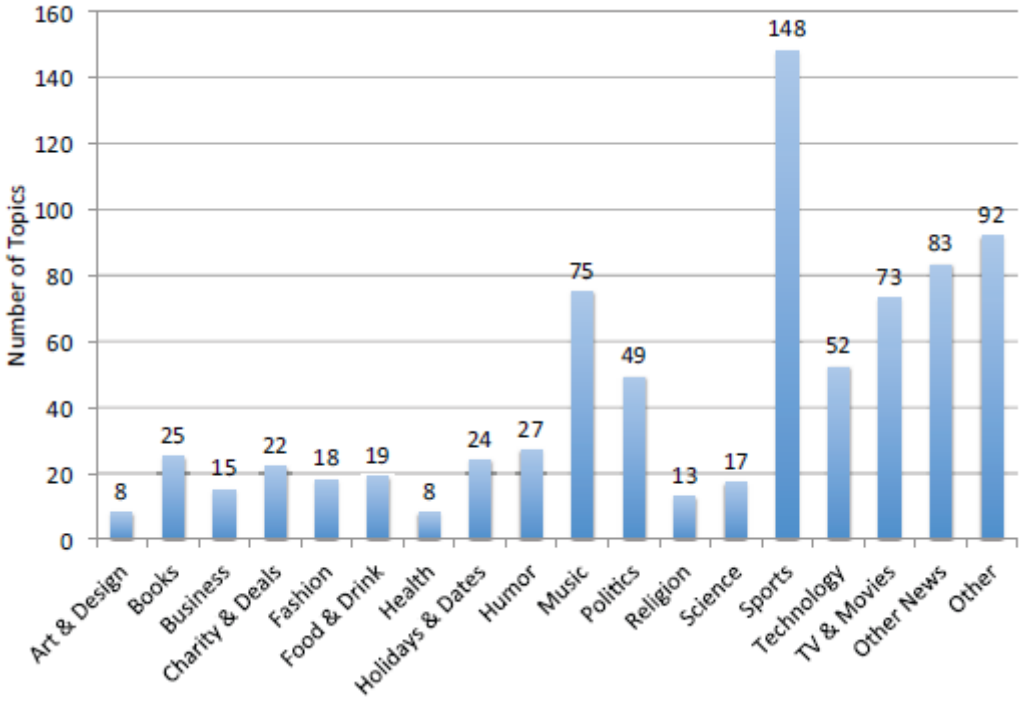


Figura 13 Distribuição de 768 tópicos pelas 18 classes escolhidas (Lee et al. 2011)

Figura 14 Tópicos de tendências existentes na classe Tecnologia (Lee et al. 2011)

4.3 Data Modeling

a) Text-based Data Modeling, são utilizadas duas fases neste tipo de modelação.
Na primeira fase, é feito a partir da definição de tendências e pelo número variável de tweets. Ao documento será assim atribuído um rótulo correspondente ao tópico.
Na segunda fase, o documento é executado a partir de um vector "string-to-word", que consiste em dois componentes:

O primeiro componente é um tokenizer que remove caracteres delimitados e inibe as palavras de darem palavras ao documento.

O segundo componente transforma os tokens em pesos tf-idf.

Para cada um dos 18 rótulos, as palavras mais usadas com os seus pesos tf-idf são usadas para construir o conjunto de dados para a aprendizagem da máquina.

b) Network-based Data Modeling, utiliza informação específica da rede social Twitter. É utilizado o modelo de *user-similarity* para encontrar os cinco temas mais semelhantes para a similaridade de tópicos.

Utilizadores comuns influentes nos tópicos são identificados a partir do algoritmo de aprendizagem *Weighted Page Rank,* bem como informações da rede social, tais como a hora do tweet, o número de tweets, a relação que a pessoa tem com a pessoa que segue. Dessa forma, usando o número de utilizadores comuns influentes entre dois tópicos, a maioria dos tópicos semelhantes são calculados a partir do modelo de similaridade entre utilizadores.

O modelo de *user-similarity* assume que, se houver uma sobreposição significativa entre os utilizadores na geração de tweets sobre dois tópicos, então implica que existe uma estreita relação entre os tópicos.

A fórmula de cálculo para a similaridade entre utilizadores é descrita em (4) :

$$user\_similarity(t_i, t_j) = \frac{|U^s_{influencer_{t_i}} \cap U^s_{influencer_{t_j}}|}{s} \quad (4)$$

em que $U^s_{influencer_{t_i}}$ é o conjunto topo de influências s do tópico ti.

Tabela 3 Lista dos 5 tópicos mais semelhantes do tópico “macbook” da classe Tecnologia (Lee et al. 2011)

Similar Topic Y	Class of Topic Y	No. of Common Influential Users between Topic X and Topic Y
iwork	technology	11
magic trackpad	technology	11
#landsend	charity & deals	11
apple ipad	technology	11
mobileme	technology	10

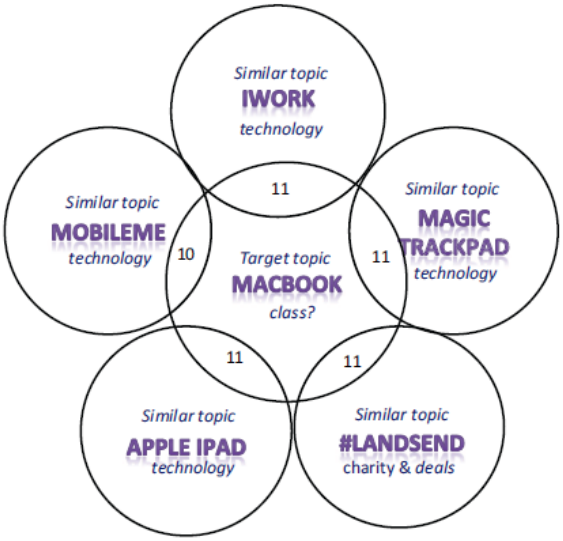


Figura 15 Tendência de tópico “macbook” e os seus 5 tópicos similares e relacionados (Lee et al. 2011)

A tabela 3 e a figura 15 mostram um exemplo do tema "macbook", os seus cinco temas mais similares, e o número de utilizadores comuns mais influentes entre o tópico "macbook" e os seus temas semelhantes. A tendência de tópico "macbook" é classificada como tecnologia por etiquetagem manual e os seus cinco tópicos mais semelhantes ("iWork", "magic trackpad", "#landsend", "apple, ipad" e "mobileme") estão marcados manualmente como tecnology, tecnology, charity & deals, tecnology, tecnology. Os números na Figura 6 indicam o número de utilizadores comuns influentes que fizeram tweet sobre o mesmo "macbook" e o seu tópico semelhante.

No caso do tópico #LANDSEND, este não faz parte da tendência de tópico MACBOOK, por pertencer á classe charity & deals.

Os dados resultantes para a aprendizagem da máquina, neste caso, consiste em 768 linhas e 19 colunas. Cada linha representa um tópico de interesse. As 18 colunas representam as 18

classes e a última coluna representa a classe *label*. Desde que o tópico "macbook" tem quatro tópicos similares em tecnologia, a soma dos quatro valores dos utilizadores comuns influentes, correspondentes aos seus tópicos semelhantes em tecnology (11 + 11 + 11 + 10 = 43), torna-se no valor para a linha "macbook" e para a coluna tecnology na tabela. E o valor correspondente ao seu tópico semelhante #landsend torna-se no valor para a linha "macbook" e para a coluna de charity & deals.

4.4 Machine Learning

Nesta fase são recebidos os dados dos dois tipos de modelação de tópicos e que são utilizados para a aprendizagem da máquina, usando modelos preditivos, a partir de técnicas de classificação de modelos e seleciona-se os que têm maior precisão.

5 Arquitetura proposta

Neste capítulo descreve-se a análise da solução proposta, com foco nos processos e funcionalidades necessárias a incluir no sistema, para que seja possível atingir os pressupostos e requisitos estipulados.

Cada uma das fases que constituem a arquitetura do sistema serão posteriormente detalhadas, mencionando as suas características e funcionamento.

A figura 16 representa a arquitetura do sistema de classificação utilizado, exemplificando todas as fases desde o início da recolha de dados, passando pela fase de modelação de dados baseados em texto, identificando a semelhança entre utilizadores, até chegar à fase de *clustering*. No final é obtido o perfil de utilizador melhorado.

Figura 16- Arquitetura do sistema de classificação

5.1 Enquadramento da arquitetura proposta no âmbito do Toursplan

Nesta seção é apresentada a proposta de arquitetura orientada a serviços para enquadrar os módulos de recomendação de pontos de interesse e de planeamento de viagens turísticas do projeto Toursplan no contexto do portal de turismo do município do Porto.

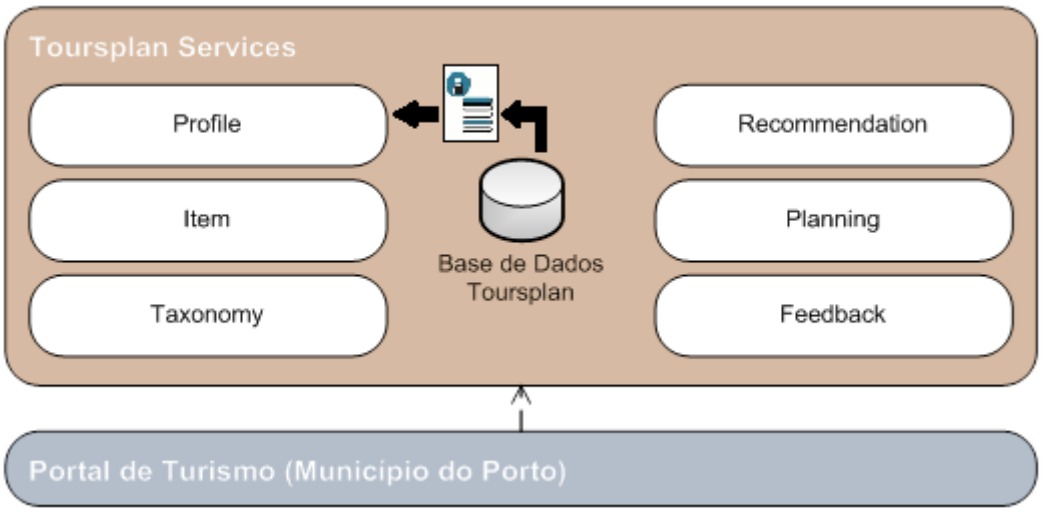


Figura 17 Diagrama de componentes do serviço Toursplan

O componente Toursplan Services representa um conjunto de serviços desenvolvido pelo GECAD-ISEP (ver figura 17), que irá conter os módulos de planeamento e recomendação. Este componente inclui uma base de dados local, de forma a guardar os valores que associam o perfil do utilizador, *itens* e taxonomias aos diversos elementos do sistema de recomendação e planeamento. Sendo assim, uma versão adaptada e simplificada da base de dados original é utilizada.

O foco deste trabalho será no perfil de utilizador, como indicado na figura 17. Este terá como objetivo a seleção e tratamento da informação relativa a um utilizador, sendo que para isso é necessário identificar previamente que informação é considerada relevante para a construção de um perfil. Pode-se considerar como sendo uma pré-recomendação do utilizador, ou seja o processo que antecede uma recomendação, estando a informação tratável e limpa, permitindo uma análise mais fidedigna de um perfil de utilizador e que posteriormente se possa realizar recomendações com maior fiabilidade e eficácia. No final, o perfil de utilizador melhorado será integrado no serviço de perfil do Toursplan.

As operações dos serviços Toursplan são descritas na tabela 4. Todos os serviços encontram-se implementados utilizando Web Services REST (REpresentational State Transfer).

Tabela 4 Serviços envolvidos na arquitetura

Serviço	Operações	Categoria
• Recommendation	GET (Obter)	Recomendação
• Planning	POST (Requisitar)	Planeamento
• Feedback	POST (Registar)	Inicialização e Manutenção
• Profile	POST (Registar) e PUT (Alterar)	
• Item	POST (Registar), PUT (apenas para o Item) e DELETE	
• Taxonomy	POST (Registar) e DELETE	

Para todos os serviços Toursplan, os resultados poderão ser apresentados em três diferentes formatos. A escolha dos mesmos depende do valor enviado no campo *content-type* do cabeçalho HTTP. Os possíveis valores são:

- XML – application/xml
- JSON – application/json
- HTML – text/html

Na devolução de uma página HTML, será apresentada a descrição do serviço e um formulário de testes e de invocação do serviço.

Os detalhes técnicos atualizados, bem como uma versão de demonstração do componente Toursplan Services estão disponíveis no seguinte endereço:

https://www.gecad.isep.ipp.pt/ToursplanServices/

Na seção Anexos, Anexo 6, encontra-se disponível a base de dados Toursplan.

5.2 Análise prévia de um utilizador e Data Collection

O *twitter* é uma rede social que é utilizada mensalmente por cerca de 300 milhões de utilizadores ativos, onde se estima que sejam feitos cerca de 50 milhões de *tweets* em todo o mundo (Twitter, 2015) .

Cada *tweet* contém um *timestamp* explicito que identifica a hora exata em que este foi gerado e, no caso dos utilizadores estes têm um perfil bem definido, contendo as suas informações pessoais (nome, localização, esboço biográfico) (Mathioudakis & Koudas 2010). Cada um deles representa um documento que contém informação bastante rica para ser explorada e analisada com melhor detalhe, sendo desafiante a sua análise e extração. Neste sentido, a tecnologia já implementada na rede social *Twitter* revela-se numa mais-valia e que servirá como base de apoio para o desenvolvimento do sistema de classificação.

Esta fase tem como objetivo verificar previamente a informação pessoal e demográfica de cada utilizador (data de nascimento, sexo, estado civil, género, país), adicionando características relacionadas com a sua pesquisa ou partilha de informação, tais como POI e *action timestamp*.

A figura 18 demonstra um excerto da base de dados Toursplan, contendo apenas as tabelas utilizadas para a extração da informação referente a cada utilizador.

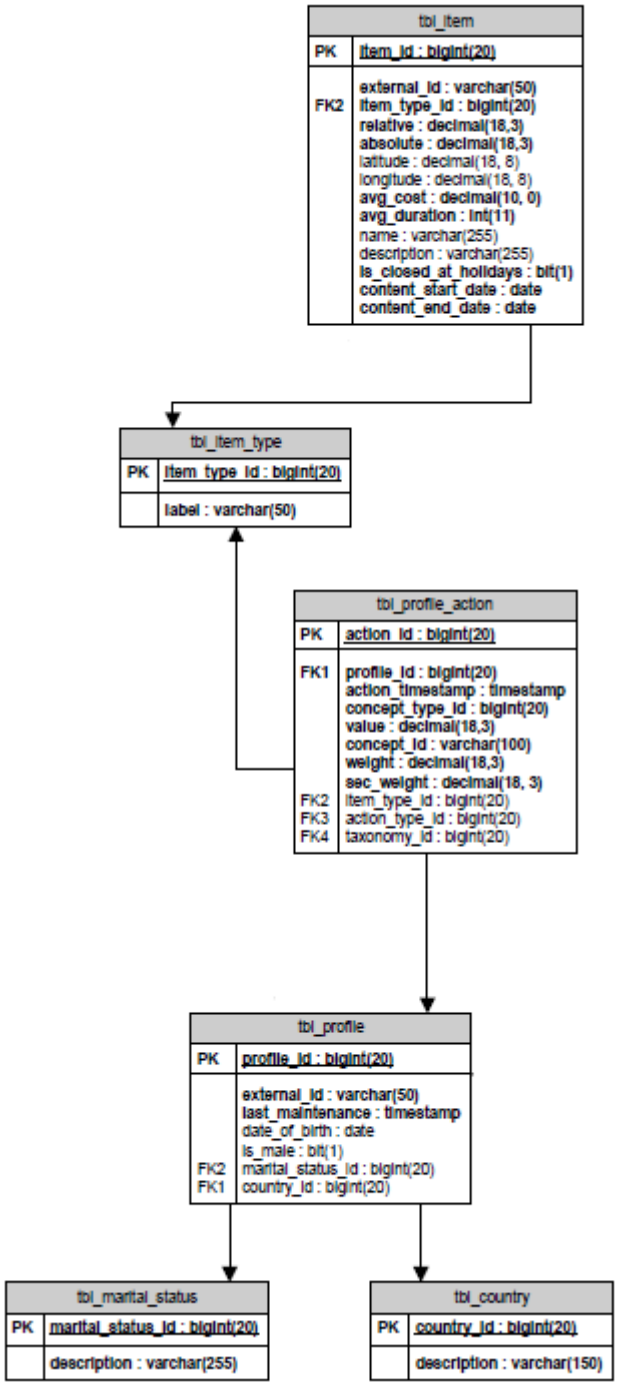


Figura 18 Tabelas utilizadas para extração de informação de um utilizador

Porém, para que esta informação fosse extraída da base de dados Toursplan foi implementado um procedimento que permitisse, não só a extração da informação necessária para a construção do perfil de cada utilizador, como também a criação de um documento único em formato csv para cada um deles.

Permite assim definir diferentes utilizadores, sendo estes identificados pelo seu identificador numérico, presente em cada nome de documento gerado. Por outro lado, dado que a informação de cada utilizador provém de fontes diferentes, este valor não garante que o identificador numérico seja único para cada utilizador, traduzindo na possibilidade de diferentes utilizadores terem o mesmo identificador.

Uma das formas possíveis de tratar este problema, e isto devido á impossibilidade em identificar de forma única cada utilizador, é o armazenamento de todos os registos referentes á identificação de um perfil de utilizador por documento, cujo nome do documento é o identificador do utilizador.

Para a criação da instrução SELECT, criaram-se dois agrupamentos de informação: um para identificar a informação referente a um utilizador, e outro para a informação de um POI.

Com estes agrupamentos de dados, torna-se mais fácil a realização da instrução *Left Join* para que seja possível a correspondência entre duas tabelas, retornando a tabela referente aos dados do cliente com os registos equivalentes da tabela POI, ou seja que coincidam nas duas tabelas. No caso dos registos que não tenham equivalência, é retornado o valor NULL nos seus campos.

A figura 19 ilustra todos os passos realizados na seleção dos dados descritos anteriormente.

```
SELECT user.profile_id as ID_USER, user.external_id as EXTERNAL_ID, poi.name as POI,
user.date_of_birth as DATE_OF_BIRTH,(case when user.is_male=1 then 'Male' else 'Female' end) as GENDER,
poi.action_timestamp as ACTION_TIMESTAMP,
user.marital_status_description as MARITAL_STATUS, user.country_description as COUNTRY
(SELECT * FROM tbl_profile_action, tbl_item WHERE tbl_profile_action.concept_id = tbl_item.item_id
AND tbl_item.item_type_id = tbl_profile_action.item_type_id
AND tbl_profile_action.concept_type_id = 2) as poi left join

(SELECT tbl_profile.*, tbl_marital_status.description as marital_status_description,
tbl_country.description as country_description
FROM tbl_profile, tbl_marital_status, tbl_country
WHERE tbl_profile.marital_status_id = tbl_marital_status.marital_status_id
AND tbl_profile.country_id = tbl_country.country_id AND tbl_profile.profile_id = ",theUser,") as user
ON poi.profile_id = user.profile_id");
```

Figura 19 Excerto da query utilizada

Todo o processo realizado na implementação de um procedimento é descrito com maior detalhe no Anexo 4.

Com esta informação é possível criar um documento não estruturado, ou seja um documento que contém apenas os dados extraídos de cada utilizador, que servirá de base para a modelação de dados descrita na fase seguinte.

5.3 Data Modeling

Depois de realizada a extração da informação referente a cada utilizador, será utilizado o processo de modelação de dados baseado em texto (Text-based Data Modeling). Este processo vai permitir a descoberta de características a partir de dados não marcados, de forma a criar um perfil de utilizador.

Para utilizar os modelos de documentos baseados em texto, para cada documento serão processados os dados que permitam caracterizar e definir um perfil de utilizador, tais como a sua informação pessoal e as características relacionadas com a sua pesquisa, mencionados anteriormente.

O documento é executado passando por duas etapas:

O primeiro componente usado é um *tokenizer* que remove caracteres delimitados e extrai todos os tokens relevantes encontrados no documento.

O segundo componente transforma as palavras existentes em pesos tf-idf (Manning et al. 2009).

A fórmula (5) demonstra a expressão que permite calcular o tf-idf:

$$w_{i,j} = tf_{i,j} \cdot idf_i \quad (5)$$

Em que *w* determina o peso de um termo *i*, referente ao documento *j*, tendo como resultado o produto entre o número de vezes que um termo ocorre num documento (tf) e a medida geral da importância de um termo (idf) (Udagawa 2013) .

Para o cálculo do idf, esta é feita a partir da fórmula (6):

$$idf_i = log_2 \left[\frac{M}{df_i}\right] \quad (6)$$

Em que *M* define o número de documentos totais existentes e df_i caracteriza o número de vezes que um termo *i* aparece num documento *d*.

Algoritmo1 Construção dos documentos existentes (buildAllDocuments)

Por cada iterar $it \in documentosemTfIdf$, **repetir**
    $palavra \leftarrow it$
    calcularTfIdf(tfIdf)
Fim

Algoritmo2 Cálculo da ocorrência de palavras para cada documento (addWordOccurence)

Se $(todasPalavras(nome) == nulo)$
valorpalavra= 1.0
Adiciona a todasPalavras(nome, valorpalavra)
Fim
Senao
$valorpalavra \leftarrow obter\ todasPalavras(palavra)$
$valorpalavra[0] \leftarrow valorpalavra[0] + 1$
Adicionar a todasPalavras(palavra,valorpalavra)
Fim
Retorno de valorpalavra

Algoritmo3 Atualização do corpus da palavra , ou seja o idf referente ao cálculo de pesos TfIdf (updateCorpus)

Por cada iterar $it \in TodasPalavras$, **repetir**
$palavra \leftarrow it$
$valorcorpopalavra \leftarrow buscaTodasPalavras(palavra)$
$valorcorpopalavra[1] = log(totaldocumentos \div valorpalavra[0])$
$tfidf = valorpalavra[1] * valorcorpopalavra[1]$
$valorpalavra[2] = tfidf$
$vec = vector + tfidf * tfidf$
Adiciona a TodasPalavras(palavra,valorpalavra)
Fim

Algoritmo4 Cálculo dos pesos TfIdf para cada palavra de cada documento (calculateTfIdf)

Por cada iterar $it \in TodasPalavras$, **repetir**
$palavra \leftarrow it$
$valorcorpopalavra \leftarrow buscaTodasPalavras(palavra)$
$valorpalavra \leftarrow buscaPalavra(palavra)$
*tfidf= valorpalavra[1] * valorcorpopalavra[1]*
valorpalavra[2]= tfidf
*vec = vec + tfidf * tfidf*
colocarpalavra(palavra,valorpalavra)
Fim
$vec = \sqrt{vec}$
construirDocumentos()
ImprimirDados()

Algoritmo5 Guardar o valor do peso TfIdf correspondente para cada palavra (printData)

Por cada iterar $it \in TodasPalavras$, **repetir**
$palavra \leftarrow it$
$valorcorpopalavra \leftarrow buscaTodasPalavras(palavra)$
$valorpalavra \leftarrow buscaPalavra(palavra)$
EscreverFicheiro(palavra, valorcorpopalavra[1]*valorpalavra[1])
Fim
FecharEscritaFicheiro()

Assim, a medida tf-idf permite avaliar a importância de uma palavra (term) para um documento, onde se define a importância como a proporcionalidade do número de vezes que uma palavra aparece no documento, permitindo estabelecer um padrão de frequência. Dessa forma pode-se constatar que a medida tf-idf utiliza palavras comuns existentes em cada documento (Lee et al. 2011).

O resultado da modelação de dados será um documento estruturado para cada utilizador, ou seja um documento que contém os *tokens* mais relevantes bem como os respetivos pesos tf-idf. Para cada um dos documentos gerados, as palavras mais utilizadas em conjunto com os seus pesos tf-idf, permitirão posteriormente o cálculo da similaridade entre utilizadores, denominado de *user similarity*, mencionado na próxima fase.

5.4 User Similarity

Para determinar os tópicos mais semelhantes entre dois utilizadores, é utilizado o algoritmo do modelo de *user similarity*. Este caracteriza-se como sendo uma função que permite medir a similaridade entre dois documentos *u* e *v*, possibilitando encontrar os utilizadores com maior valor de similaridade, a partir do cálculo da similaridade.

O modelo de *user similarity* é classificado de três formas:

1. Recomendações baseadas no conteúdo, em que o utilizador será recomendado de itens que gostava no passado;

2. Recomendações colaborativas, onde o utilizador será recomendado por outros utilizadores com preferências idênticas no presente;

3. Recomendações hibridas, caracterizada pela junção das recomendações baseadas no conteúdo e pelas recomendações colaborativas.

Por outro lado, o *Collaborative Filtering* (CF) tem como vantagem em relação às abordagens mencionadas anteriormente, de que não depende do conteúdo dos itens, mas apenas na preferência dos utilizadores. Estas podem ser explícitas, ou seja valores numéricos de *rating* ou implícitas, que descrevem o comportamento do utilizador, ou seja as suas ações, como por exemplo a compra de um item (Sun et al. 2012).

A fase de *user similarity* é utilizada de acordo com o CF descrito anteriormente. Este caracteriza-se como sendo uma aplicação a um sistema de recomendação, de modo a ajudar os utilizadores a descobrir quais os seus itens favoritos, podendo esta ser agrupada em duas classes: *memory-based* e *model-based*.

Para este sistema será utilizada a classe *memory-based*, geralmente utilizada em sistemas de recomendação comerciais.

O *memory-based* CF permite identificar a similaridade entre dois utilizadores, com base no *rating* dos itens que ambos os utilizadores tenham escolhido. Existem várias formas de cálculo para esta classe, entre as quais o cosseno da similaridade, o PCC, a abordagem modificada do cosseno, entre outros. A abordagem escolhida para medida de similaridade entre dois documentos foi o cosseno da similaridade.

A fórmula (7) demonstra a fórmula do cálculo do cosseno da similaridade entre dois utilizadores (Udagawa 2013) :

$$sim(u,v) = \frac{\sum_{i=1}^{N} w_{i,j} * w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} * \sqrt{\sum_{i=1}^{N} w_{i,q}^2}} \quad (7)$$

Esta classifica-se por calcular o cosseno do ângulo da similaridade entre dois utilizadores.

Algoritmo6 Identificação de documentos similares (similarDocuments)

```
TreeMap <String,Double> similarDocs
Por cada iterar it ∈ documentos, repetir
        doc ← it
        Se (document = doc)
            continua
        Fim
        Adiciona a similarDocs (doc, userSimilarity(document,doc))
Fim
TreeMap <String,Double> sortSimilar
 Adiciona similarDocs a  sortSimilar
 Retorno de sortSimilar
```

Algoritmo7 Obtenção do valor referente á similaridade entre dois documentos (similarD)

```
Por cada iterar it ∈ documentos, repetir
        doc ← it
        Se (document = doc)
            continua
        Fim
        userSimilarity(document,doc))
Fim
Retorno de userSimilarity
```

Algoritmo8 Cálculo do cosseno da similaridade entre eles (userSimilarity)

```
Por cada iterar it ∈ TodasPalavras em doc1, repetir
        palavra ← it
```

Se $(existe\ em\ doc1\ a\ palavra\ em\ doc2)$
$similaridade = similaridade + [\ doc1(palavra)[2] * doc2(palavra)[2]\ ]$
Fim
Fim
$similaridade = similaridade\ \div [\sqrt{(doc1)^2}\ *\ \ (doc2)^2\ ]$
Retorno de similaridade

O modelo de *user similarity* assume que, se houver uma sobreposição significativa entre os utilizadores sobre duas palavras, então implica que existe uma estreita relação entre os tópicos.

Esta fase permitirá como resultado obter os documentos mais similares para cada um dos documentos existentes, identificando os utilizadores com maior semelhança.

5.5 Clustering

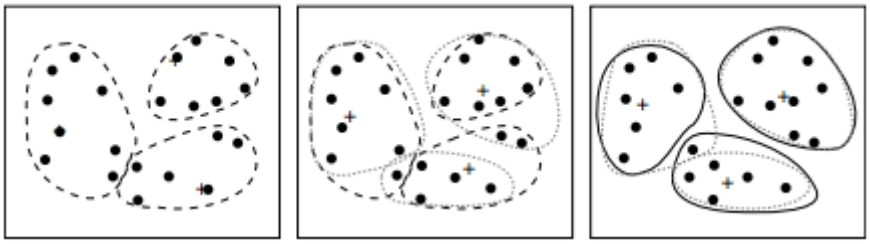
Nesta fase, os modelos de dados criados anteriormente devem estar finalizados, para que se possa proceder á construção de agrupamentos de dados, segundo o seu grau de semelhança.

Este será feita a partir da aplicação do método de *clustering* com uma aprendizagem do tipo *unsupervised learning*, dado que os algoritmos criam automaticamente grupos de classificação atribuindo uma descrição generalizada do conteúdo da base de dados. Caracteriza-se por efetuar a partição do conjunto de instâncias recebidas num pequeno número de subconjuntos de *clusters*, de modo a que instâncias similares pertençam ao mesmo grupo e instâncias distintas pertençam a grupos diferentes.

Para o treino do modelo de classificação será utilizada a ferramenta WEKA, com aplicação do método *k-Means*. Este método caracteriza-se por ser uma heurística de agrupamento não hierárquico que permite minimizar a distância entre os elementos, começando inicialmente por uma partição aleatória, continuando a atribuir aos *clusters* novos padrões com base na similaridade entre o padrão e o cluster, até atingir um critério de convergência.

Caracterizada como sendo numa ferramenta de aprendizagem de vários algoritmos e de processamento de dados, a ferramenta WEKA disponibiliza suporte para todo o processo experimental de *Data Mining*, incluindo na preparação dos dados de entrada, avaliando estatisticamente esquemas de aprendizagem e visualiza, quer os dados de entrada, quer o resultado da aprendizagem (Frank et al. 2005).

A figura 20 demonstra a aplicação da técnica de submissão de dados *clustering*, a partir do método *k-Means*, mostrando todas as fases realizadas até ao agrupamento final dos documentos (Jiawei & Kamber 2001) .



(a) *Clustering* Inicial (b) Iteração (c) *Clustering* Final

Figura 20 Aplicação de *clustering* com método *k-Means*

Este cálculo vai permitir identificar para cada objeto qual o *cluster* mais similar.

Nesta fase foi analisada a relação da similaridade entre dois documentos para o conjunto total. Para estruturar a informação e realizar a análise pretendida para este cenário será utilizada a análise de *clustering*. Uma vez que as técnicas de *clustering* existentes na ferramenta WEKA não suportam a medida de similaridade entre o cosseno do ângulo entre dois documentos, revelou-se necessária a adaptação do valor já obtido da similaridade para realizar o algoritmo *k-Means* escolhido.

Dessa forma, foi criada uma classe em Java que implementa a função da distância Euclidiana dada esta ser, por defeito, a distância que a técnica utiliza, para que seja possível a utilização do valor do cosseno da similaridade entre dois documentos já realizado. Esta recebe dois documentos e o respetivo valor de similaridade, retornando o valor adaptado para o algoritmo. Tendo em conta que o valor da distância Euclidiana se traduz na obtenção da distância entre dois pontos, tendo como resultado a menor distância entre eles, no caso do cosseno da similaridade esta representaria documentos poucos similares num algoritmo *k-Means*. Para que haja coerência no resultado obtido, é calculada a diferença entre o valor máximo do intervalo e o valor da similaridade. Assim, o valor correspondente será ajustado ao cálculo da distância do algoritmo. O resultado serão *clusters*/grupos de perfis semelhantes que permitirão identificar e perceber se os perfis em estudo, são ou não da mesma pessoa. Para ilustrar o conjunto de *clusters*/grupos resultantes da aplicação do algoritmo *k-Means*, é implementado um método que converte o resultado para o formato suportado pela ferramenta Graphviz. Esta permitirá identificar os *clusters* com perfis semelhantes, bem como o valor de convergência utilizado para cada um.

Atendendo ao caso de estudo em análise, procedeu-se aos seguintes passos até se obter o modelo final:

- Aplicação do método *k-Means* para diferentes valores de k (concretamente k=3, k=4, k=5 e k=6) com vista a avaliar qual o número de grupos que permite uma maior diferenciação com interpretação pertinente e coerente no âmbito do sector do turismo;

Considerou-se o método com valor k=6, por este ter um melhor resultado na diferenciação, na coerência e com interesse interpretativo no que respeita á identificação de perfis de utilizador com similaridades idênticas.

A aplicação do método *k-Means* com seis grupos estabelecidos *a priori*, com vista à convergência do algoritmo, resulta na seguinte análise:

1. Atribuição dos indivíduos aos grupos criados: Esta atribuição refere-se a que grupo pertence cada indivíduo, permitindo definir os indivíduos com maior convergência em cada grupo.
   Pela extensão dos documentos em estudo, será apenas demonstrado o gráfico que refere o resultado da aplicação do algoritmo *k-Means* para o valor de k=6, identificando para cada *cluster* o valor de convergência e o respetivo centroide selecionado. Será colocado apenas um excerto da solução, dado o grande número de perfis existentes na seção de Anexos, Anexo 5, para melhor compreensão da mesma.

2. Tabela referente ao utilizador: todas as variáveis referidas contribuem fortemente para a definição dos grupos.

No conjunto das variáveis existentes destacam-se as selecionadas para definir o perfil de utilizador, nomeadamente:

- POI: Pontos de interesse turísticos;
- Date_of_birth: Data de nascimento do utilizador;
- Gender: Género do utilizador, identificado como Masculino ou Feminino;
- Action_Timestamp: Ano e mês em que o utilizador realizou a pesquisa;
- Marital_Status: Estado civil do utilizador, identificado como “Married”, “Divorced”, “Single” ou “Widower”;
- Country: País de origem do utilizador.

A figura 21 demonstra um diagrama dos *clusters* gerados, ilustrando a distância média entre cada um deles. A distância média é dada pela média da distância entre todos os pares de pontos de dois *clusters*.

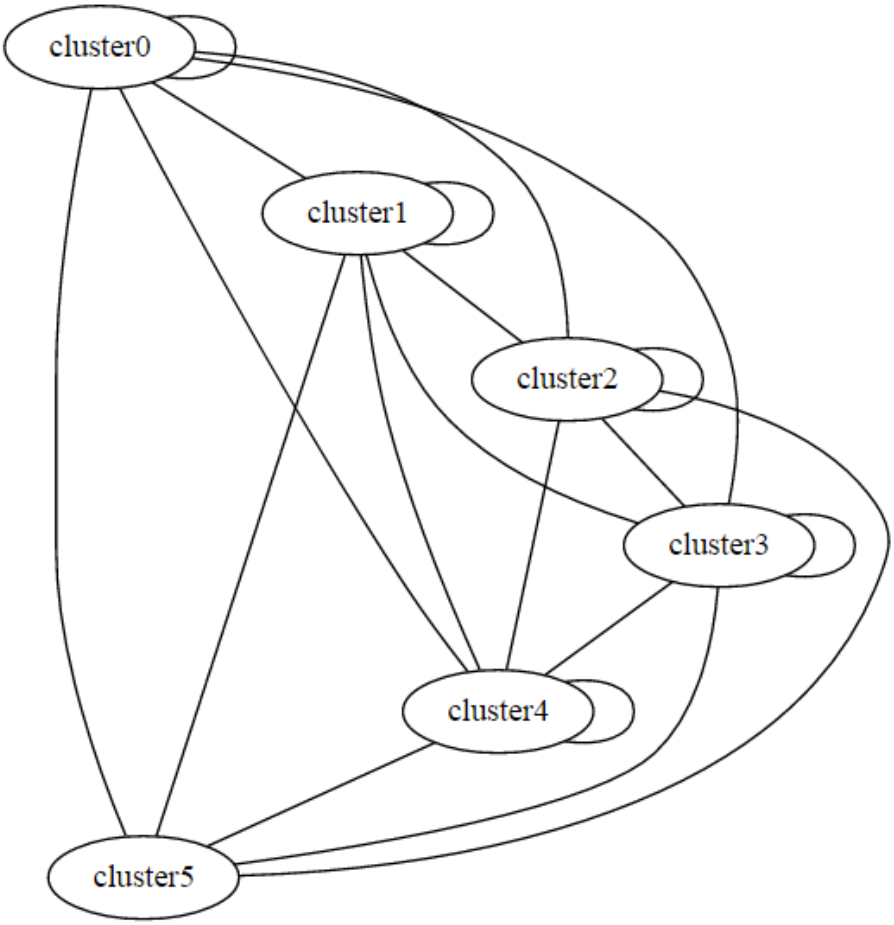


Figura 21 Diagrama com *clusters* usados na implementação do algoritmo *k-Means*

A apresentação deste método em particular deveu-se ao facto deste ser utilizado no sistema desenvolvido. A sua seleção, em detrimento de outros métodos existentes, teve por base o facto de este ser um método bastante utilizado no agrupamento de objetos, de utilizar vetores de pesos, sendo essenciais para receber o valor da similaridade entre os utilizadores, e pela sua facilidade de aplicação quando se está perante um elevado número de observações. Também pelo facto deste algoritmo se encontrar na ferramenta WEKA utilizada para a realização do *clustering*, foi outro dos motivos pelo qual este método foi o escolhido.

5.6 Validação do sistema

Nesta seção apresenta-se a validação do sistema com um conjunto de dados previamente selecionados, permitindo ilustrar e demonstrar procedimentalmente o processo do sistema.

Não foram realizados testes ao sistema pelo facto da informação disponibilizada ser informação que não garante a consistência nem a fiabilidade dos dados extraídos, sendo impossível o teste ao sistema por via de dados reais e fidedignos. Dessa forma a melhor solução para comprovar o funcionamento do sistema é a sua validação a partir de uma seleção de documentos prévia como *input* para o sistema.

Assim sendo, para validação do sistema, dado o grande volume de documentos existente, são analisados dez documentos para ilustrar e validar todo o processo do sistema, bem como a consistência dos dados em estudo.

De entre 121 documentos resultantes da fase de extração foram selecionados aleatoriamente os documentos:

- 99.csv, 97.csv, 40.csv, 106.csv, 27.csv, 121.csv, 18.csv, 1.csv, 96.csv e 15.csv.

As figuras 22, 23 e 24 ilustram o processo de Data Extraction para os primeiros três documentos selecionados para a validação (99.csv, 97.csv e 40.csv), onde se referenciam todas as palavras existentes em cada documento, identificando também o número de documentos em que cada palavra ocorre.

```
Data Extraction
 lda. 1.0
 s.a. 1.0
 s.a.    2.0
 sgps 2.0
1959-06-22 1.0
1971-04-22 1.0
1972-12-15 1.0
1974-01-22 2.0
1978-10-22 1.0
1979-04-06 1.0
1991-08-14 1.0
1995-10-07 1.0
2014-07-29 1.0
2015-01-22 1.0
2015-03-08 1.0
2015-03-09 1.0
2015-03-10 1.0
\n 8.0
adega vi ├‖ va 1.0
ancipa - associa ├ª ├úo nacional de comerciantes e industriais de produtos alimentares 1.0
apima ôçô associa ├ª ├úo portuguesa das industrias de mobili ├írio e afins 1.0
assador t ├¡pico 1.0
associa ├ª ├úo de estudantes da faculdade de belas artes da universidade do porto 1.0
auto via ├ª ├úo landim 1.0
auto via ├ª ├úo pacense 2 1.0
baco 1.0
bali-hai polinesian bar 2.0
bar das cardosas 2.0
boombap store 1.0
cacicambra  1.0
```

Figura 22 Excerto da fase de Data Extraction (1ªparte)

```
carristur - yellow bus official sightseeing tours 1.0
casa aberta 1.0
centro de sa ├ ‖ de de campanh ├ú 1.0
chafariz do jardim de s. l ├ízaro 1.0
consulado da noruega 2.0
consulado geral do brasil 1.0
correios campanh ├ú 2.0
correios conde de ferreira 1.0
correios pedro hispano 1.0
c ├ómara de com ├®rcio e ind ├ ‖ stria luso-mexicana 1.0
d. pedro v  1.0
douro acima 1.0
escola francesa do porto 2.0
escola superior de biotecnologia da universidade cat ├ | lica portuguesa 1.0
faculdade de desporto da universidade do porto 1.0
faculdade de economia da universidade do porto 1.0
faculdade de engenharia da universidade do porto 1.0
farm ├ícia alves da silva 1.0
farm ├ícia alves moreira 1.0
farm ├ícia campo alegre 1.0
farm ├ícia carneiro barbosa 1.0
farm ├ícia central do porto 1.0
farm ├ícia confian ├ ª a 1.0
farm ├ícia contumil 1.0
farm ├ícia lago 1.0
farm ├ícia lemos 1.0
farm ├ícia mafalda 1.0
farm ├ícia pasteleira 1.0
farm ├ícia ramos 1.0
farm ├ícia s ├í 1.0
feira da vandoma 1.0
female 4.0
fep junior consulting 1.0
foot locker 1.0
galeria fernando santos 1.0
garage&stage - music store - loja foz 1.0
```

Figura 23 Excerto da fase de Data Modeling (2ª parte)

```
garrafeira velha 1.0
grupo soares da costa 2.0
hospital da boavista 1.0
instituto superior de administra ├ª ├║o e gest ├║o 1.0
instituto superior de educa ├ª ├║o e trabalho 1.0
irm ├║os unidos 1.0
junta de freguesia de miragaia 1.0
laborat ├ | rio de processos de separa ├ª ├║o e rea ├ª ├║o-lsre-faculdade de engenharia da universidade do porto  1.0
lexus 1.0
lunar 1.0
l ├¡der 1.0
male 6.0
married 5.0
monte carlo 1.0
mundauto 1.0
oporto british school      1.0
oporto house 1.0
ourivesaria coutinho 1.0
park hotel porto aeroporto 1.0
perfumaria castilho 1.0
portugal 7.0
renex  1.0
rui vieira-joalheiros 1.0
serafim caetano pereira 1.0
single 4.0
spain 2.0
top ├ízio 1.0
well domus - fitness spa  2.0
xafariz artesanato  1.0
zenith lounge 1.0
```

Figura 24 Excerto da fase de Data Modeling (3ª parte)

Depois de selecionados os dez documentos resultantes da fase de Data Extraction, procede-se á fase de Data Modeling para a modelação de cada documento analisado.

Como resultado é gerado um *output* para cada documento estruturado referente a cada utilizador, contendo os pesos tf-idf para cada palavra.

Para a fase de Data Modeling serão também identificados os três primeiros ficheiros usados na validação do sistema, contendo para cada documento as palavras extraídas bem como o cálculo do peso tf-idf.

As figuras 25, 26 e 27 demonstram para cada documento, o processo de Data Modeling. Para cada palavra/termo são identificados os valores de cálculo utilizados.

Os diferentes valores para cada linha referem-se a:

- 1º Argumento- Termo/palavra do documento;
- 2º Argumento- Número de vezes que o termo/palavra aparece no documento (df);
- 3º Argumento- Frequência com que um termo/palavra aparece num documento (tf);
- 4º Argumento- Cálculo do Inverse Document Frequency (idf);
- 5º Argumento- Valor do cálculo referente ao TfIdf.

```
Filename :99
Term: 1974-01-22        df:1.0  tf:0.16666666666666666  idf:1.6094379124341003  TfIdf:0.26823965207235
Term: \n        df:2.0  tf:0.3333333333333333   idf:0.22314355131420976 TfIdf:0.07438118377140325
Term: male      df:1.0  tf:0.16666666666666666  idf:0.5108256237659907  TfIdf:0.08513760396099845
Term: married   df:1.0  tf:0.16666666666666666  idf:0.6931471805599453  TfIdf:0.11552453009332421
Term: portugal  df:1.0  tf:0.16666666666666666  idf:0.3566749439387324  TfIdf:0.05944582398978873
```

Figura 25 Cálculo de TfIdf para o documento 99.csv

```
Filename :97
Term:  s.a.     df:1.0  tf:0.016129032258064516 idf:1.6094379124341003  TfIdf:0.025958676007001618
Term:  sgps     df:1.0  tf:0.016129032258064516 idf:1.6094379124341003  TfIdf:0.025958676007001618
Term: 1995-10-07        df:10.0 tf:0.16129032258064516  idf:2.302585092994046   TfIdf:0.3713846924183945
Term: 2015-01-22        df:10.0 tf:0.16129032258064516  idf:2.302585092994046   TfIdf:0.3713846924183945
Term: bali-hai polinesian bar   df:1.0  tf:0.016129032258064516 idf:1.6094379124341003  TfIdf:0.025958676007001618
Term: bar das cardosas  df:1.0  tf:0.016129032258064516 idf:1.6094379124341003  TfIdf:0.025958676007001618
Term: consulado da noruega      df:1.0  tf:0.016129032258064516 idf:1.6094379124341003  TfIdf:0.025958676007001618
Term: correios campanh├ú        df:1.0  tf:0.016129032258064516 idf:1.6094379124341003  TfIdf:0.025958676007001618
Term: escola francesa do porto  df:2.0  tf:0.03225806451612903  idf:1.6094379124341003  TfIdf:0.051917352014003236
Term: faculdade de desporto da universidade do porto    df:1.0  tf:0.016129032258064516 idf:2.302585092994046   TfIdf:0.03713846924183945
Term: female    df:10.0 tf:0.16129032258064516  idf:0.9162907318741551  TfIdf:0.14778882772163793
Term: grupo soares da costa     df:1.0  tf:0.016129032258064516 idf:1.6094379124341003  TfIdf:0.025958676007001618
Term: junta de freguesia de miragaia    df:1.0  tf:0.016129032258064516 idf:2.302585092994046   TfIdf:0.03713846924183945
Term: portugal  df:10.0 tf:0.16129032258064516  idf:0.3566749439387324  TfIdf:0.057528216764311677
Term: single    df:10.0 tf:0.16129032258064516  idf:0.9162907318741551  TfIdf:0.14778882772163793
Term: well domus - fitness spa  df:1.0  tf:0.016129032258064516 idf:1.6094379124341003  TfIdf:0.025958676007001618
```

Figura 26 Cálculo de TfIdf para o documento 97.csv

```
Filename :40
Term: 1972-12-15        df:2.0  tf:0.16666666666666666  idf:2.302585092994046   TfIdf:0.3837641821656743
Term: 2014-07-29        df:2.0  tf:0.16666666666666666  idf:2.302585092994046   TfIdf:0.3837641821656743
Term: farm├ícia ramos   df:2.0  tf:0.16666666666666666  idf:2.302585092994046   TfIdf:0.3837641821656743
Term: female    df:2.0  tf:0.16666666666666666  idf:0.9162907318741551  TfIdf:0.15271512197902584
Term: married   df:2.0  tf:0.16666666666666666  idf:0.6931471805599453  TfIdf:0.1155245300932421
Term: spain     df:2.0  tf:0.16666666666666666  idf:1.6094379124341003  TfIdf:0.26823965207235
Filename :96
Term: 1991-08-14        df:1.0  tf:0.16666666666666666  idf:2.302585092994046   TfIdf:0.3837641821656743
Term: \n        df:2.0  tf:0.3333333333333333   idf:0.22314355131420976 TfIdf:0.07438118377140325
Term: male      df:1.0  tf:0.16666666666666666  idf:0.5108256237659907  TfIdf:0.08513760396099845
Term: portugal  df:1.0  tf:0.16666666666666666  idf:0.3566749439387324  TfIdf:0.05944582398978873
Term: single    df:1.0  tf:0.16666666666666666  idf:0.9162907318741551  TfIdf:0.15271512197902584
```

Figura 27 Cálculo de TfIdf para o documento 40.csv

Por fim, apresentam-se nas figuras 28, 29 e 30 os três documentos selecionados, identificando as cinco palavras/termos que ocorrem com maior frequência em cada documento bem como o conjunto de documentos mais similares a esse documento.

```
99.csv
-------------------------------------------
Best words in document
1974-01-22
married
male
\n
portugal


Similar documents with this words
106.csv
15.csv
18.csv
27.csv
40.csv
```

Figura 28 Lista de palavras mais usadas e documentos similares no documento 99.csv

```
97.csv
----------------------------------------
Best words in document
1995-10-07
female
portugal
escola francesa do porto
faculdade de desporto da universidade do porto


Similar documents with this words
121.csv
1.csv
27.csv
40.csv
106.csv
```

Figura 29 Lista de palavras mais usadas e documentos similares no documento 97.csv

```
40.csv
----------------------------------------
Best words in document
1972-12-15
spain
2014-07-29
female
married


Similar documents with this words
18.csv
1.csv
121.csv
97.csv
106.csv
```

Figura 30 Lista de palavras mais usadas e documentos similares no documento 40.csv

Os resultados obtidos permitem a identificação do conjunto de documentos utilizados para a validação, identificando em primeiro lugar as palavras que ocorrem com maior incidência num documento. Dessa forma, obtém-se as palavras-chave de cada documento, permitindo a obtenção de uma perspetiva de como será o perfil de utilizador em análise.

Depois obtém-se os documentos com maior valor de similaridade, garantindo o conjunto de documentos com maior semelhança e características identificando-se assim, perfis de utilizador relacionados.

Como resultado será construído um diagrama e um gráfico, sendo que o último permitirá visualizar o resultado da implementação do algoritmo *k-Means*, contendo para os três *clusters* escolhidos previamente, os documentos que obtiveram convergência com o algoritmo.

As figuras 31, 32 e 33 demonstram o conjunto de perfis semelhantes que fazem parte de cada *cluster*, mencionando o respetivo centroide.

O valor "Distance" representado identifica o valor de convergência do *cluster*.

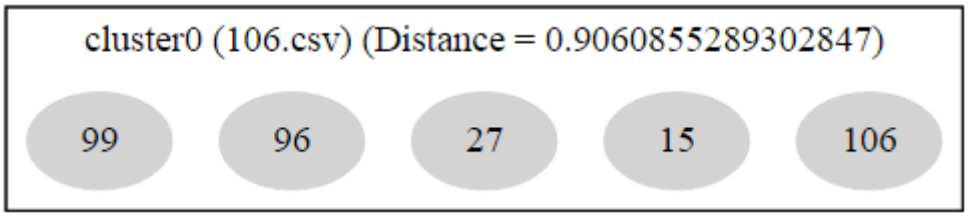


Figura 31 Validação do sistema- Primeiro *cluster*

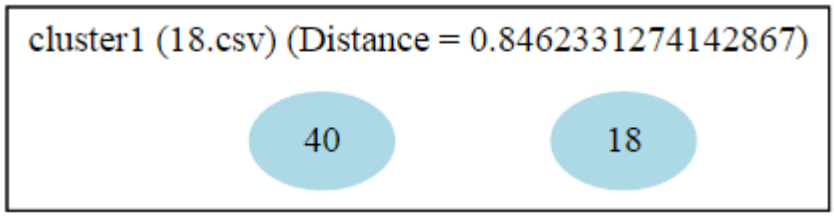


Figura 32 Validação do sistema- Segundo *cluster*

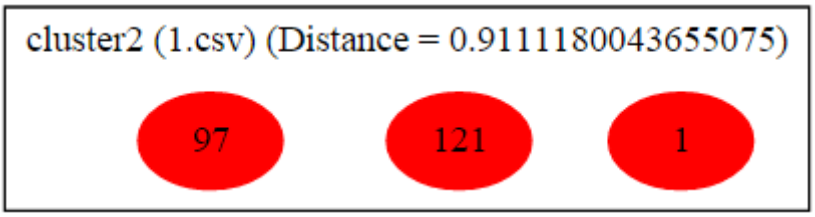


Figura 33 Validação do sistema- Terceiro *cluster*

Foram selecionados previamente três *clusters* pela coerência e estabilidade que este conjunto de *clusters* garantia, para este caso de estudo em concreto.

Depois de identificada a composição de cada *cluster*, a figura 34 representa um diagrama com os *clusters* em estudo*,* ilustrando também a distância média entre eles.

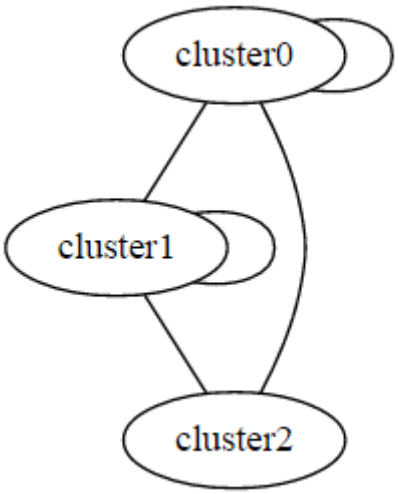


Figura 34 Diagrama com *clusters* usados na validação do sistema

Este diagrama permite entender de que forma os *clusters* estão relacionados, dando uma perspetiva geral da distância entre eles, ou seja existe menor distância entre o cluster0 e cluster1 do que a distância entre o centro do cluster0 e cluster2.

Dado que um mau *cluster* se caracteriza por distâncias curtas entre objetos de *clusters* diferentes pode-se concluir que, neste caso, a formação de *clusters* resultou em bons *clusters,* estando eles a uma distância aceitável, sendo a distância média entre o cluster0 e cluster1 cerca de 0.962, entre o cluster1 e cluster2 cerca de 0.972 e entre o cluster1 e cluster2 cerca de 0.967.

6 Conclusões

"Portugal é menos icónico do que os países famosos, mas oferece uma riqueza de oportunidades aos viajantes: aldeias de charme, comida excelente, música regional fascinante, oportunidades culturais, uma costa bonita e até surf de categoria mundial" (USA Today, 2015).

Esta afirmação, entre muitas outras, permite que se possa afirmar que Portugal é um destino em ascensão no que toca ao turismo e com uma ampla visibilidade internacional. A cidade do Porto foi considerada em 2014 como o melhor destino Europeu e, nesta perspetiva, revela-se crucial o estudo e classificação das pessoas que visitam o nosso país, de modo a captar e motivar potenciais e reincidentes turistas (Destinations, 2014).

A utilização de BI permitiu uma melhor abordagem na extração de dados, permitindo a obtenção de dados uteis e tornar o sistema desenvolvido, num sistema de apoio á decisão. Também permitiu modelar os dados numa perspetiva em que os dados referentes a um perfil de utilizador estivessem disponíveis de forma clara e rápida, possibilitando a aquisição de conhecimento e numa tomada de decisões mais qualitativa.

Sistemas de recomendação nas mais diversas áreas de atuação revelam-se fundamentais e atrativos, e na área do turismo não é exceção. A implementação de um sistema hibrido permite assim obter as vantagens existentes no sistema baseado no conteúdo e no sistema colaborativo.

Enquanto o sistema baseado em conteúdo permite identificar num utilizador as características pessoais e seus itens, num sistema colaborativo permite identificar a similaridade entre dois

utilizadores. Neste caso, com base na informação do utilizador e dos objetos com que ele interagiu, possibilitou obter o nível de similaridade entre eles, permitindo o agrupamento daqueles que têm uma similaridade idêntica.

O estudo e desenvolvimento de um sistema que permitisse extrair informação pessoal de um utilizador, bem como os objetos turísticos e a respetiva classificação de um utilizador, contribuiu para que fossem definidos agrupamentos com base no seu grau de semelhança. Como resultado permitiu definir perfis de utilizador relacionados, possibilitando uma caracterização mais detalhada e eficaz de cada utilizador, relativamente às suas preferências e aptidões.

A escolha do método tf-idf, como cálculo para determinar os pesos dos termos para cada documento, foi pela possibilidade em determinar, a partir do número de ocorrências que uma palavra aparece num grupo de documentos, como sendo uma métrica que permite constatar a importância e relevância que essa palavra tem nesse mesmo grupo de documentos. Isso terá posteriormente reflexo na similaridade entre dois documentos.

Como indicado na seção 2.1.2, a escolha da medida do cosseno de similaridade entre dois documentos foi a eleita, dado esta medida ser muito utilizada em documentos bem como em *information retrieval* e na técnica de *clustering*, tendo esta sido utilizada no sistema desenvolvido. Tendo também como característica principal a independência no tamanho dos documentos esta revelou-se importante, dada a recomendação colaborativa ter o problema de dispersão dos dados. Independentemente de alguns registos referentes ao perfil de utilizador não terem alguns atributos preenchidos, esta medida permite a comparação com outros perfis existentes, mesmo estes serem de tamanhos diferentes.

Assim o sistema desenvolvido contribuiu para que fosse possível identificar e definir, a partir de um conjunto de informações de diferentes fontes existente na base de dados Toursplan, perfis de utilizador, sendo estes construídos a partir de informação selecionada e considerada relevante.

Dessa forma torna-se possível caracterizar as ações e preferências de cada utilizador, sem o identificar individualmente, permitindo saber qual o valor de similaridade entre os perfis analisados, de modo a que sejam feitos agrupamentos para cada um deles, determinando quais os perfis relacionados.

Com isto permite-se ter, não só uma perspetiva de como os perfis de utilizador existentes no sistema estão relacionados, como também identificar métricas que possam caracterizar os utilizadores que nos visitam e quais as preferências/aptidões de potenciais turistas no nosso turismo, obtendo respostas e feedback das suas intenções quando desejam visitar o nosso país no caso de ser estrangeiro, ou região no caso de viver em Portugal.

Como resultado exibe uma base analítica para obtenção de conhecimento face aos turistas e aos mercados alvo.

6.1 Objetivos atingidos

Para a construção do sistema foi realizada a análise, definição e implementação dos métodos de extração de conhecimento, a partir de dados partilhados pelos utilizadores/turistas, recolhidas das mais diversas fontes tecnologias de apoio (por exemplo portais web, aplicações móveis e quiosques).

Este processo terá como auxilio uma base de dados existente (Toursplan), que permitiu ser uma base de apoio nos dados recolhidos. Esta base de dados contém atualmente incluídos módulos para a recolha de dados interativos do turista, recomendação de produtos de turismo, planeamento de trajetos, e apresentação de "*dashboards*" com base nos dados recolhidos.

Para suporte foi adicionada informação, de forma a sustentar e a reforçar o perfil de um utilizador, bem como as suas atividades turísticas, locais e pontos de interesse, sem nunca os identificar pessoalmente.

Após a extração e tratamento da informação será possível identificar um perfil de utilizador, a partir das diferentes iterações realizadas, mostrando a similaridade entre eles.

A análise destas relações, juntamente com as ferramentas já existentes, irá disponibilizar uma base analítica para a obtenção de conhecimento sobre o turista face aos respetivos mercados alvo.

O sistema apresenta várias funcionalidades colocados ao dispor, entre as quais serão referenciadas as mais importantes e relevantes:

- Extração da informação relevante da base de dados Toursplan para a construção do perfil de utilizador, criando múltiplos ficheiros csv para cada utilizador;
- Armazenamento do valor do peso tf-idf para cada um dos termos existentes de cada documento, para análise da importância que cada termo tem;
- Identificação para cada documento, aqueles que são similares a este, indicando o respetivo valor de similaridade;
- Integração do sistema com a ferramenta WEKA;
- Substituição da distância euclidiana para o cosseno de similaridade para a realização do *clustering*, dado que a ferramenta WEKA não suporta previamente o cosseno da similaridade entre dois documentos;
- Criação do *clustering*, utilizando a ferramenta WEKA, a partir do algoritmo *k-Means*;
- Visualização gráfica do algoritmo *k-Means*, demonstrando todos os *clusters* adicionados previamente, bem como os objetos que dele fazem parte.

6.2 Limitações

Algumas limitações surgiram ao longo do projeto em que na fase inicial do desenvolvimento do projeto, o estudo e decisão da melhor abordagem para responder aos objetivos pretendidos, foi um processo que demorou algum tempo até que este permitisse a iniciação do desenvolvimento da implementação de forma concreta e assertiva.

Como referido anteriormente, a realização de testes ao sistema revelou-se num ato desproporcionado tendo em conta que os dados recolhidos e que constituem o perfil de um utilizador, tais como o seu identificador numérico, POI, data de nascimento, género, *action timestamp*, estado civil e país não são suficientes para a realização de testes consistentes e coerentes.

Outra limitação que surgiu, deveu-se á fase de *clustering* na integração do sistema com a ferramenta WEKA. Esta por não conter a função de similaridade entre dois documentos e também por ter um formato de ficheiro próprio, foi necessário perceber como este é composto e de que forma seria necessário abordar esta ferramenta para responder ao pretendido.

Existiu a necessidade em utilizar uma ferramenta que permitisse a visualização gráfica dos *clusters* resultantes da aplicação do algoritmo *k-Means*, dada a necessidade em converter a distância euclidiana que, por defeito é a utilizada na ferramenta *k-Means*, para o cosseno da similaridade já anteriormente calculado. Esta conversão fez com que se inviabilizasse a geração gráfica dos *clusters* a partir da ferramenta WEKA, tendo em conta que o passo do cálculo do algoritmo seja realizado no sistema após o momento da conversão da distância pretendida, o que vai contra os procedimentos da ferramenta, já que ela recebe os valores para posteriormente implementar o algoritmo pretendido.

6.3 Trabalho futuro

O sistema está desenvolvido apenas para receber modelos de dados baseado em texto, o que permite apenas a receção de dados provenientes de repositórios de dados e de arquivos. Assim sendo, para abranger uma maior diversificação de informação disponibilizada, numa modelação de dados baseada na rede, (Network-based modeling) os dados oriundos da web permitiriam dar resultados mais precisos sobre o utilizador, a partir da fase de aprendizagem da máquina que esta modelação incluiria.

Permitiria receber da fase de Data Modeling os documentos estruturados de cada utilizador, realizando a classificação de cada um a partir de modelos preditivos, resultando na escolha dos documentos com maior precisão.

Outra característica a ser desenvolvida será a criação de *dashboards*, com base nos dados recolhidos, para que estes ilustrem o funcionamento do sistema bem como os resultados que se obtém ao ser efetuado a técnica de *clustering* nos documentos modelados.

7 Referências bibliográficas

(Adomavicius & Tuzhilin 2005) Adomavicius, G. & Tuzhilin, A., 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), pp.734–749.

(Aggarwal & Reddy 2013) Aggarwal, C.C. & Reddy, C.K., 2013. Data Clustering: Algorithms and Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series),

(Balaji 2012) Balaji, S., 2012. WATEERFALLVs V-MODEL Vs AGILE : A COMPARATIVE STUDY ON SDLC. WATEERFALL Vs V-MODEL Vs AGILE : A COMPARATIVE STUDY ON SDLC, 2(1), pp.26–30.

(Cazella et al. 2010) Cazella, S.C., Nunes, M. & Reategui, E., 2010. A Ciência da Opinião: Estado da arte em Sistemas de Recomendação. *CSBC XXX Congresso da SBC Jornada de Atualização de InformáticaJAI*, pp.161–216. Available at: http://www.dcomp.ufs.br/~gutanunes/hp/publications/JAI4.pdf.

(Chang 2014) Chang, K.C., 2014. Towards a Social Media Analytics Platform : Event Detection and User Profiling for Twitter. In M. Gupta, R. Li, & K. Chang, eds. Seoul, Korea, pp. 193–194.

(Chien & Chen, 2008) Chien, C.-F., & Chen, L.-F. (2008). Data mining to improve personnel selection and enhance human capital A case study in high-technology industry. *Expert Systems with Applications*, pp. 280-290.

(Destinations, 2014) European Best Destinations, http://www.europeanbestdestinations.com/top/europe-best-destinations-2014/ [último acesso: 2015]

(Doni 2004) Doni, M.V., 2004. Análise de Cluster: Métodos Hieráquicos e de Particionamento. , p.92.

(Fayyad et al. 1996) Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework. *Proc*

2nd Int Conf on Knowledge Discovery and Data Mining Portland OR, pp.82–88. Available at: http://www.aaai.org/Papers/KDD/1996/KDD96-014.

(Fesenmaier & Ricci 2003) Fesenmaier, D. & Ricci, F., 2003. DIETORECS: Travel advisory for multiple decision styles. *Information and communication technologies in tourism.*, pp.232–241.

(Figueiredo 2010) Figueiredo, A.M.N. de A.B., 2010. Sistemas de Apoio à Decisão aplicados à área do Turismo.

(Frank et al. 2005) Frank, E. et al., 2005. Weka. , pp.1305–1314.

(Gangadharan & Swami, 2004) Gangadharan, G. R., & Swami, S. N. (6 de 2004). Business Intelligence Systems: Design and Implementation Strategies. *2th Int. Conf. Information Technology Interfaces IT1*, pp. 139-144.

(Gazzana & Silveira 2009) Gazzana, P.P. & Silveira, S.R., 2009. Sistema de Recomendação Para a Área de Turismo.

(Goel & Aggarwall, 2013) Goel, N., & Aggarwal, A. (2013). A Survey of Cloud Business Intelligence. pp. 519-521.

(Graphviz, 2015) Graphviz- Graph Visualization Software, http://www.graphviz.org/ [último acesso: 2015]

(Gualtar 2011) Gualtar, C. De, 2011. Instituto Superior de Engenharia do Porto. , pp.33940–33943.

(Heracles, 2015) Heracles Information Integration Research Group, http://www.isi.edu/integration/Heracles/ [último acesso: 2015]

(Huang 2008) Huang, A., 2008. Similarity measures for text document clustering. Proceedings of the Sixth New Zealand, (April), pp.49–56.

(Jiawei & Kamber 2001) Jiawei, H. & Kamber, M., 2001. Data mining: concepts and techniques, pp. 453.

(Kimball & Wiley , 2004) Ralph Kimball, Joe Caserta Wiley, 2004, The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data, Chapters 3, 4, 5, and 6.

(Lee et al. 2011) Lee, K. et al., 2011. Twitter trending topic classification. Proceedings - IEEE International Conference on Data Mining, ICDM, pp.251–258.

(Luz et al. 2015) Luz, N., Anacleto, R. & Almeida, A., 2015. Tourism Mobile and Recommendation Systems - A State of the Art. CSREA EEE 2010:277-283.

(Maimon & Rokach 2010) Maimon, O. & Rokach, L., 2010. *DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK*

(Manning et al. 2009) Manning, C.D., Raghavan, P. & Schutze, H., 2009. An Introduction to Information Retrieval. Online, (c), p.569.

(Mathioudakis & Koudas 2010) Mathioudakis, M. & Koudas, N., 2010. Twittermonitor: trend detection over the twitter stream. SIGMOD '10 Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pp.1155–1158.

(MATLAB, 2012) MATLAB Central, http://www.mathworks.com/matlabcentral/fileexchange/24616-kmeans-clustering [ultimo acesso: 2015]

(Metz & Monard 2005) Metz, J. & Monard, M.C., 2005. Clustering hierarquico : uma metodologia para auxiliar na interpretação dos clusters. , (September), pp.1170–1173.

(MySQL, 2015) MySQL Workbench, https://www.mysql.com/products/workbench [último acesso:2015]

(NetBeans, 2015) Netbeans, https://netbeans.org/features/index.html [último acesso: 2015]

(Olszak & Ziemba, 2007) Olszak, C. M., & Ziemba, E. (2007). Approach to Building and Implementing Business Intelligence Systems. *Interdisciplinary Journal of Information, Knowledge, and Management*, pp. 135-148.

(Ranjan, 2009) Ranjan, J. (2009). Business Intelligence: Concepts, Components, Techniques and Benefits. *Journal of Theoretical and Applied Infor. Technology*, pp. 60-70.

(Renko, 2011) Sanda Renko,2011. Supply Chain Management - New Perspectives, Chapter 28.

(Rithme, 2015) Rithme Business Intelligence Solutions, http://www.rithme.eu/ [último acesso: 2015]

(Rygielski, Wang, & Yen, 2002) Rygielski, C., Wang, J.-C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in Society*, pp. 483-502.

(Schiff, 2010) Schiff, M.A., 2010. Business intelligence: Improving your company efficiency and effectiveness, no matter its size. SAP White paper, (July 2010), Available from http://www.silicon.com/white-papers/view/business-functions/business-intelligencea-guide-for-midsize-companies-improving-your-companys-efficiency-and-effectiveness-no-matter-its-size-61306040/

(Sharef 2013) Sharef, N.M., 2013. Design and Architecture of Agent-based and Adaptive Hybrid Recommendation System.

(Sun et al. 2012)	Sun, H.-F. et al., 2012. JacUOD: A New Similarity Measurement for Collaborative Filtering. *Journal of Computer Science and Technology*, 27(6), pp.1252–1260. Available at: http://link.springer.com/10.1007/s11390-012-1301-5.
(Thuraisingham 1998)	Thuraisingham, B., 1998. Data mining: technologies, techniques, tools, and trends.
(Tintarev & Masthoff 2011)	Tintarev, N. & Masthoff, J., 2011. *Recommender Systems Handbook*, Available at: http://www.springerlink.com/index/10.1007/978-0-387-85820-3.
(Travel-Buddy, 2015)	Travel-Buddy, http://www.travbuddy.com/ [último acesso: 2015]
(TripAdvisor, 2015)	TripAdvisor, http://www.tripadvisor.com [último acesso: 2015]
(Tuncay & Belgin, 2010)	Tuncay, E. G., & Belgin, Ö. (2010). Effects of Business Intelligence Techniques on Enterprise Productivity. pp. 3-6.
(Twitter, 2015)	Twitter, https://about.twitter.com/company/ [último acesso:2015]
(Tyagi & Sharma 2012)	Tyagi, N. & Sharma, S., 2012. Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page. , (3), pp.441–446.
(Udagawa 2013)	Udagawa, Y., 2013. Source Code Retrieval Using Sequence Based Similarity. , 3(4), pp.57–74.
(USA Today, 2015)	USA Today, http://www.10best.com/awards/travel/best-european-country/#thumbnail-1260-click/ [último acesso: 2015]
(Weka, 2015)	Weka, http://www.cs.waikato.ac.nz/ml/weka/ [último acesso:2015]

8 Anexos

8.1 Anexo 1 - Tecnologias utilizadas

Nesta seção apresentam-se as tecnologias adotadas para a criação e construção do sistema. Indica-se o tipo de ferramenta e a sua utilidade.

8.1.1 MySQL Workbench

MySQL Workbench é uma ferramenta *OpenSource* que fornece modelação de dados, desenvolvimento de SQL e ferramentas de administração abrangentes para configuração do servidor, administração de utilizadores, *backup*, entre outros. Oferece também ferramentas para criar, executar e otimizar consultas SQL (MySQL, 2015).

Esta ferramenta foi escolhida para o acesso á informação da base de dados já existente, referente á aplicação Toursplan, permitindo também implementar os scripts para seleção dos dados mais relevantes, de modo a criar em tempo de execução um documento para cada utilizador em formato csv. A grande vantagem em utilizar esta ferramenta reflete-se essencialmente por esta ser gratuita e pelas características positivas que tem relacionadas com o código aberto e as fortes características ligadas á Web, tais como a velocidade de acesso a grandes quantidades de dados, utilização em aplicações Web, capacidade de inserção rápida de grandes volumes de dados e a sua disponibilidade. Também pelo facto de ser uma ferramenta de visual que integra algumas características tais como administração da base de dados e migração de dados, são características positivas na sua utilização.

8.1.2 Netbeans IDE

NetBeans IDE é um ambiente de desenvolvimento que permite que seja rapidamente e facilmente desenvolvido uma área de trabalho Java, móvel e aplicações web, bem como aplicações HTML5 com HTML, JavaScript e CSS. O IDE também fornece um grande conjunto de ferramentas para desenvolvedores PHP e C / C ++. É também livre e *open source* e tem uma grande comunidade de utilizadores e desenvolvedores de todo o mundo (NetBeans, 2015) .

A escolha deste IDE deveu-se não só ao facto de este ser gratuito, como também pelas suas capacidades de desenvolvimento e também pela integração de um módulo já desenvolvido, tendo em conta que suporta um grande conjunto de linguagens de desenvolvimento.

8.1.3 WEKA

Weka é um conjunto de algoritmos de aprendizagem de máquina para tarefas de *data mining*. Os algoritmos podem ser aplicados diretamente para um conjunto de dados ou chamado a partir de seu próprio código Java. Weka contém ferramentas para pré-processamento de dados, classificação, regressão, *clustering*, regras de associação, e visualização (Weka, 2015).

Esta ferramenta foi escolhida pela necessidade da utilização do algoritmo de *clustering*, a partir dos dados dos utilizadores anteriormente modelados.

Também por ser uma ferramenta de fácil utilização e de *open source*, esta ferramenta foi considerada uma boa opção a adotar.

8.1.4 Graphviz

Graphviz é uma ferramenta *open source* de visualização gráfica, permitindo representar a informação estrutural com diagramas de gráficos abstratos (Graphviz, 2015).

Com esta ferramenta será possível a representação dos *clusters* previamente definidos no algoritmo *k-Means*, em conjunto com o peso de cada cluster e os respetivos documentos convergidos.

8.2 Anexo 2 – Algoritmo geral

Nesta seção apresenta-se o algoritmo geral considerando as fases de desenvolvimento do sistema, a leitura de cada um dos documentos existentes, extração das palavras-chave, o cálculo dos pesos Tf-Idf para todas as palavras existentes em cada documento.

Depois de calculados e identificados todos os documentos similares é realizado o *clustering*, por via do algoritmo *k-Means*, utilizando a distância do cosseno da similaridade.

Como resultado serão obtidos cinco *clusters*, valor definido previamente, contendo para cada um deles os documentos resultantes do critério de convergência realizado, permitindo definir aquele que melhor se aproxima do valor do centro de cada cluster.

Algoritmo1 Método principal (Main)

```
NovoFicheiro f(filename)
NovoFicheiro f1(filename)
NovaConexaoDB()

CriarDiretorio(f)
CriarDiretorio(f1)

Se (naoExiste(f) ∩ naoExiste(f1)) então
    ConetarDB()
    CriarFicheiroEmDiretoria(f)
    CriarFicheiroEmDiretoria(f1)
 Senao
       LimparDiretorio(f)
       LimparDiretorio (f1)

ConexaoDB()
TfIdf tf(filename)

Por cada iterar it ∈ TodasPalavrasemTfIdf, repetir
        palavra ← it
        corpopalavra ← buscaTodasPalavrasemTfIdf(palavra)
Fim
construirDocumentos()
ImprimirDados()

Por cada iterar it ∈ documentosemTfIdf, repetir
        palavra ← it
        guardarFicheiro1(palavra)
Fim

Para cada i = 0, i < documentosSimilaresemTfIdf, repetir
        guardarFicheiro2(palavra)
        guardarFicheiro2(documentosSimilares)
        guardarFicheiro2(valorSimilaridade)
Fim
    Fim se
Fim
```

8.3 Anexo 3- Classes do sistema

Seguidamente serão identificadas as classes que compõe a aplicação, sendo em cada uma indicados os seus métodos.

8.3.1 DBConnection.java

Esta classe permite estabelecer a ligação á base de dados Toursplan, de modo a aceder á informação armazenada e assim ser possível iniciar o sistema de criação de perfis de utilizador.

Para que seja possível é realizada a chamada ao procedimento desenvolvido, sendo este executado na base de dados de dados Toursplan, retornando os campos escolhidos para definir um perfil de utilizador.

8.3.2 Main.java

Classe principal do sistema tem como objetivo executar o sistema, criando inicialmente um diretório para armazenar os documentos não estruturados, gerados a partir da execução do SP criado para o efeito. De seguida passará por cada fase descrita na solução do sistema, ligando-se á base de dados Toursplan para a extração e coleção de dados de cada utilizador, passando pela modelação dos dados e cálculo do nível de semelhança de cada utilizador, até á fase de *clustering*.

8.3.3 TfIdf.java

Contém as operações necessárias para o cálculo dos pesos tf-idf. Uma instância tf-idf contém TreeMaps para documentos e para o corpus.

Os métodos existentes nesta classe caracterizam-se por:

- Aceitar apenas documentos em formato csv;
- Carregar o corpo de cada documento e atualiza-lo, se necessário;
- No caso de ser necessário a sua atualização, permite passar por cada palavra existente no documento e mudar a sua frequência;
- Calcular os valores tf-idf de um ou vários documentos;
- Incrementa a ocorrência de uma palavra num documento;
- Calcula a similaridade entre dois utilizadores, a partir do cosseno da similaridade entre dois documentos, e armazena o valor de similaridade de cada um dos documentos em comparação, num diretório previamente criado (“C:/TfIdf”);

- Verifica documentos similares, a partir do método que calcula o cosseno da similaridade, em que caso sejam similares estes serão listados;
- Comparação dos documentos e listagem dos mais similares.

8.3.4 Document.java

Classe que representa um documento de texto, mantendo o controlo do número de vezes que uma palavra aparece no texto, a frequência do termo (tf), bem como a frequência inversa do termo (idf) para encontrar palavras-chave importantes no documento.

Uma instância Document contém TreeMaps para o nome do documento e o respetivo valor.

Os métodos existentes nesta classe classificam-se por:

- Verifica, linha-a-linha, o tamanho de cada palavra em cada documento em que, caso esta tenha menos de dois caracteres, não é considerada uma palavra. Caso contrário contabiliza a palavra na variável do tipo TreeMap, que permite armazenar a palavra e o valor desta;
- Inserção de cada um dos documentos na instância documents do tipo Document, que contém um TreeMap, de modo a permitir iterar sobre cada um deles;
- Atualização do Idf, de acordo com alterações que possam surgir num dos documentos em causa;
- Carregamento de todos os documentos, a partir da instância documents do tipo Document que contém um TreeMap, de modo a permitir calcular os pesos tf-idf de cada documento;
- Cálculo dos pesos tf-idf de cada documento;
- Retorno do número de palavras que aparecem num documento;
- Retorno das palavras mais importantes no documento;
- Comparação entre nomes em cada documento, de modo a determinar as palavras mais importantes.

8.3.5 Clustering.java

Classe que contém o cálculo do *clustering*, a partir do algoritmo k-Means.

Funcionalidades:

- Recebe o ficheiro em formato csv e define previamente o número de *clusters* que o algoritmo deve conter;

- Implementa o algoritmo *k-Means*, com base na nova distância calculada;

- Salva o resultado do algoritmo num ficheiro em formato DOT, para posterior leitura da ferramenta Graphviz.

8.4 Anexo 4 - MySQL Stored Procedure

Para permitir a geração automática de múltiplos documentos por utilizador, contendo a informação que o caracterize bem como os objetos turísticos com que ele interagiu, foi necessária a criação de um *SP*. Este permite a execução do código que selecionará os campos considerados mais importantes, de um conjunto de tabelas existente na base de dados, em que para cada utilizador reconhecido a partir do seu identificador numérico, será criado um documento com essa mesma informação filtrada e identificada por um valor incrementado numérico.

Os valores de cada campo apresentado no documento gerado é delimitado por vírgulas e aspas.

A figura 35 demonstra o excerto de código utilizado no MySQL Workbench para a criação do *SP* createUserDataToMultipleCSVs.

```
CREATE DEFINER=`root`@`localhost` PROCEDURE `createUserDataToMultipleCSVs`()
BEGIN
DECLARE done int default 0;
    DECLARE theUser int default 0;

    -- declare cursor for user fetching
    DECLARE myC CURSOR FOR select profile_id from toursplan_db.tbl_profile;

    -- declare the NOT FOUND handler (triggered at end of all users fetched)
    DECLARE CONTINUE HANDLER FOR NOT FOUND SET done = 1;

    OPEN myC;

    theLoop: LOOP
        FETCH myC INTO theUser; -- fetch 1st or next user into variable

        IF done = 1 THEN
          LEAVE theLoop; -- bail out we are done
        END IF;

        -- send out to separate CSV file
        -- set @sql:=concat(@sql, theUser = theUser+ 1);
        set @sql:=concat("SELECT poi.name as POI, user.date_of_birth as DATE_OF_BIRTH,
            (case when user.is_male=1 then 'Male' else 'Female' end) as GENDER, poi.action_timestamp as ACTION_TIMESTAMP,
            user.marital_status_description as MARITAL_STATUS, user.country_description as COUNTRY INTO OUTFILE \'C:\\\\CSV\\\\",theUser,".csv\'");
        set @sql:=concat(@sql," FIELDS TERMINATED BY \',\' OPTIONALLY ENCLOSED BY \'\"\' LINES TERMINATED BY \'\n\'");
        set @sql:=concat(@sql," FROM (SELECT tbl_profile.*, tbl_marital_status.description as marital_status_description,
        tbl_country.description as country_description FROM tbl_profile
        LEFT JOIN tbl_marital_status ON tbl_profile.marital_status_id = tbl_marital_status.marital_status_id
        LEFT JOIN tbl_country ON tbl_profile.country_id = tbl_country.country_id WHERE tbl_profile.profile_id = ",theUser,") as user
        LEFT JOIN (SELECT tbl_profile_action.*, tbl_item.name FROM tbl_profile_action, tbl_item
        WHERE tbl_profile_action.concept_id = tbl_item.item_id AND tbl_item.item_type_id = tbl_profile_action.item_type_id
        AND tbl_profile_action.concept_type_id = 2 AND tbl_profile_action.profile_id = ",theUser,") as poi ON user.profile_id=poi.profile_id");
        prepare stmt from @sql;
        execute stmt;
    END LOOP;
    CLOSE myC; -- clean up
END
```

Figura 35 SP createUserDataToMultipleCSVs

8.5 Anexo 5- Gráfico de clusters

Neste anexo é identificado na figura 36 um excerto dos perfis de utilizador semelhantes em cada *cluster*.

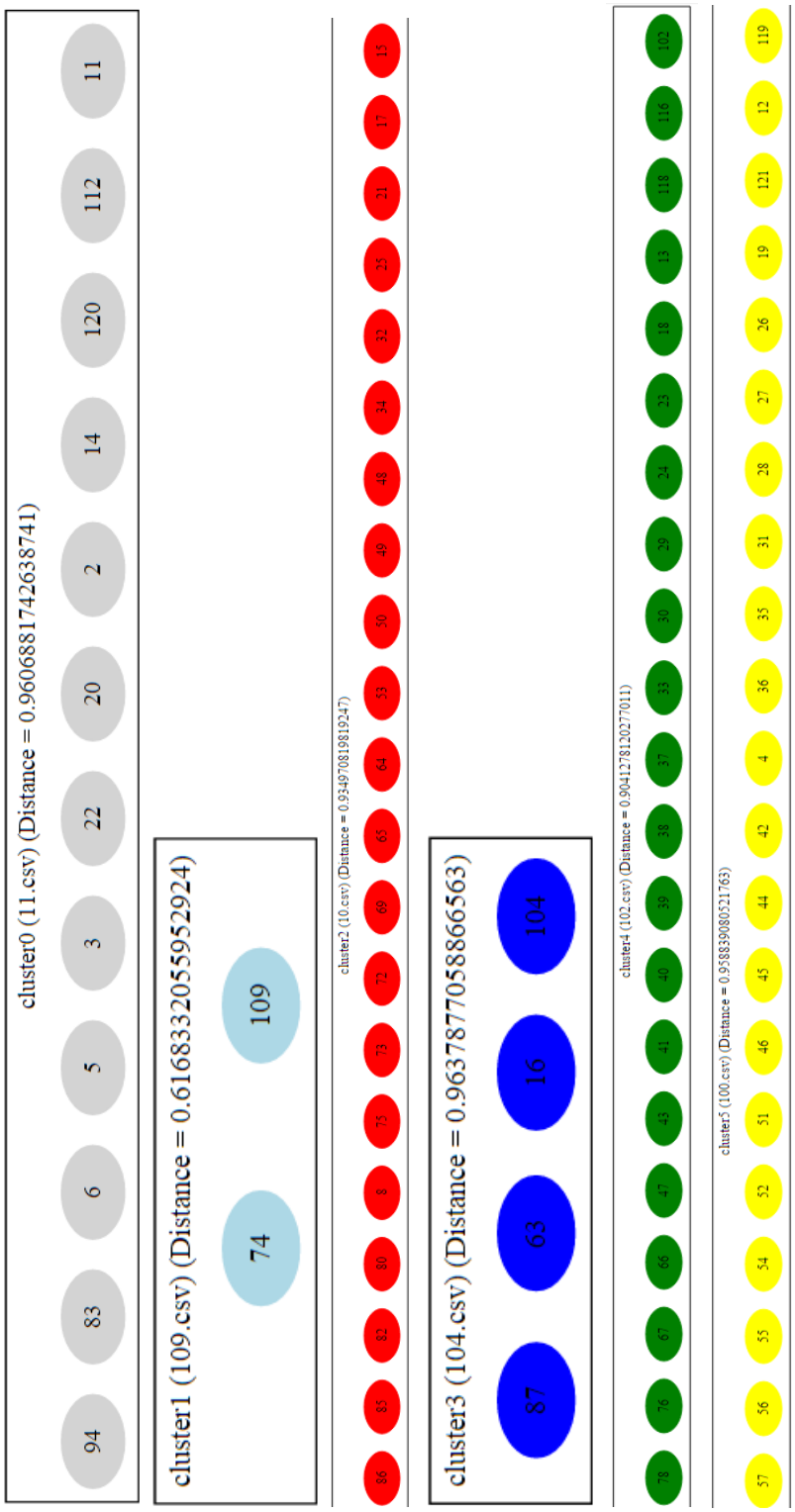


Figura 36 Conjunto de *clusters*/grupos de perfis semelhantes do sistema

8.6 Anexo 6 - Modelo de dados do Toursplan

Neste anexo é demonstrado na figura 37 o modelo de dados completo, referente á base de dados utilizada.

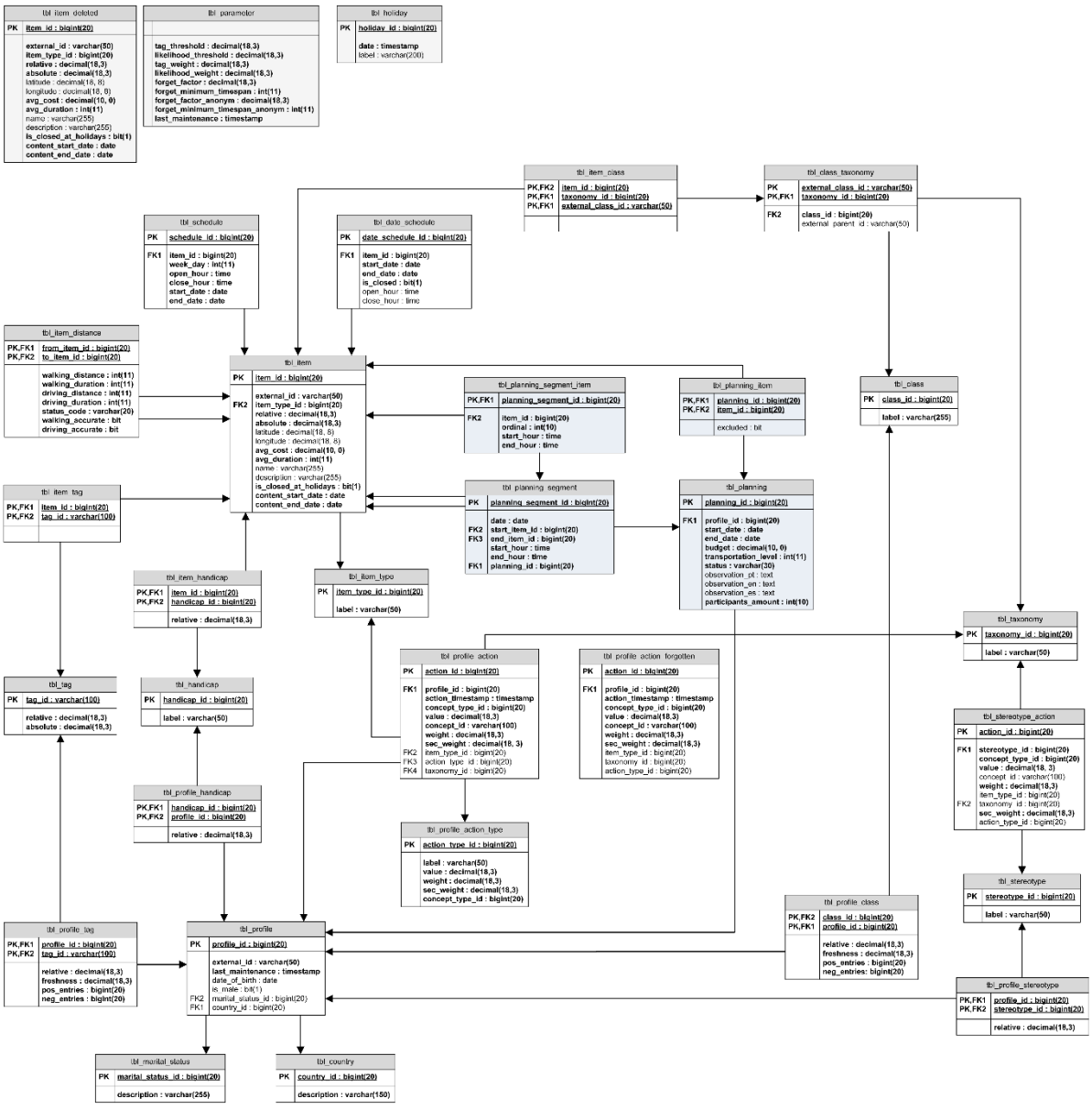


Figura 37 Modelo de dados do Toursplan