

Instituto Politécnico de Setúbal



Escola Superior de Ciências Empresariais

**Recomendação de Navegação em
Portais da Internet
como um serviço suportado em
ferramentas *Web Analytics***

João Fernando dos Anjos Gomes

Dissertação apresentada para cumprimento dos requisitos
necessários à obtenção do grau de
Mestre de Sistemas de Informação Organizacionais

Orientadores: Cláudio Sapateiro
Hernani Mourão

Setúbal, 2016

Índice

1	Introdução.....	1
1.1	Motivação.....	1
1.2	Objetivos	3
1.3	Metodologia	4
2	Revisão da Literatura.....	5
2.1	Conceitos Gerais.....	5
2.2	Propósitos e Métodos	6
2.3	Abordagens à Recomendação e Personalização	7
2.4	A Perspetiva <i>Web Analytics</i>	9
2.5	Ferramentas de Recomendação e <i>Web Analytics</i>	11
2.5.1	Ferramentas de Recomendação.....	11
2.5.2	Ferramentas <i>Web Analytics</i>	13
2.5.3	Google Analytics.....	15
2.6	Discussão	20
3	Recomendação suportada em <i>Web Analytics</i>	21
3.1	Modelo Proposto.....	22
3.2	Lógica de Notificação	23
3.3	Síntese do Modelo.....	25
4	Protótipo Desenvolvido.....	26
4.1	Arquitetura Geral	27
4.2	<i>GoRecommend</i>	28
4.3	<i>GoRecommendJS</i>	30
4.4	Descrição Funcional de Aplicação	32
5	Análise e Discussão dos Dados Obtidos	35
5.1	Abordagem e Resultados	35
5.1.1	1º Caso – Alteração do objetivo	36
5.1.2	2º Caso – Adição de um novo objetivo.....	37
5.2	Discussão e Análise de Limitações.....	39
6	Conclusões e Trabalho Futuro.....	40
6.1	Conclusões.....	40
6.2	Trabalho Futuro.....	41

Índice de Figuras

Figura 1 - Diagrama Web Analytics	10
Figura 2 - Integração entre componentes - PredictionIO	12
Figura 3 - Integração entre componentes – Google Prediction API	13
Figura 4 – Adobe Analytics	14
Figura 5 – Piwik	14
Figura 6 - Google Analytics	14
Figura 7 - Ferramentas de Web Analytics	15
Figura 8 - Número de conversões vs número de visitas	16
Figura 9 - Google Analytics como uma ferramenta de gestão	18
Figura 10 - Número de visitas mensais agrupadas pela sua origem.....	19
Figura 11 - Síntese do modelo proposto	25
Figura 12 - Arquitetura do modelo proposto	27
Figura 13 - Exemplo de resposta do serviço GoRecommend	30
Figura 14 - Exemplo de implementação da biblioteca GoRecommendJS	32

Índice de Tabelas

Tabela 1 - Exemplo de um relatório que conjuga métricas e dimensões (Google Analytics).....	20
Tabela 2 - Recomendações para a página objetivo '/contacts'	36
Tabela 3 - Recomendações para a página objetivo '/portfolio'	37
Tabela 4 - Recomendações para a página objetivo '/pt/contactos'	38
Tabela 5 - Recomendações para as páginas objetivo '/pt/contactos' e '/pt/clientes'	39

Palavras-chave

Data Mining, Digital marketing, Google Analytics, Recommendation systems, User Experience, Web Analytics, Web Usage Mining

Siglas e Acrónimos

API - Application Programming Interface

GA - Google Analytics

HTTP - Hipertext Transfer Protocol

KPI - Key Performance Indicator

OLAP - Online Analytic Processing

SAAS - Software as a Service

UML – Unified Modeling Language

Resumo

Com o constante crescimento da utilização da *Internet* o número de *websites* e respectivas páginas continua a evoluir também, por este motivo, verifica-se uma necessidade de alinhar a experiência de utilização com os objetivos gerais de um *website*. Para satisfazer esta necessidade o sistema de recomendação proposto sugere páginas ao utilizador que possam ser do seu interesse com base em perfis de navegação de um *website* em geral. A maioria dos sistemas de recomendação são baseados em regras de associação ou palavras chave (quando o conteúdo é considerado). No entanto, quando os dados não são suficientes ou são muito dispersos e a ordem é considerada, uma abordagem tradicional pode ser inadequada. Por outro lado, assumindo outro paradigma, a área de *Web Analytics*, tem obtido um crescimento considerável, através de ferramentas robustas que permitem a recolha e análise de dados da *internet*, a fim de compreender e otimizar eficiência e eficácia do *website*. O presente artigo propõe o desenvolvimento de um sistema de recomendação baseado na ferramenta *Google Analytics*. O protótipo é composto por dois componentes principais que são: 1) um serviço responsável pela construção e lógica associada à criação das recomendações; 2) uma biblioteca incorporável em qualquer *website* que providenciará um *widget* de recomendação configurável. Avaliações preliminares constataram que a implementação segue a lógica do modelo proposto.

Abstract

As the Internet usage keeps increasing, the number of web sites and hence the number of web pages also keeps increasing, so there is a need to align the user experience with the overall websites purposes. Toward this requirement, the proposed recommendation systems suggest the user pages that might be of its interest based on past navigation profiles of overall site usage. Most of existing recommendation systems are based on association rules or based on keywords (when content is considered). However, on usage data shortage or sparse data and if sequential order is to be considered such traditional approaches may become unsuitable. Conversely, the Web Analytics arena, assuming other paradigm, has experienced a considerable growth through mature tools that allow the collection and analysis of internet data in order to understand and optimize website efficiency and efficacy. This work proposes the development of a recommendation system based on the Google Analytics tool. The prototype is constituted by two main components which are: 1) a service responsible for the construction and associated logic that underlies recommendations generation; 2) an embeddable library on any website that will furnish website with a configurable recommendation widget. Preliminary evaluations had showed that the implementation follows the logic of the proposed model.

1 Introdução

1.1 Motivação

Sendo que a *Internet* apresenta-se atualmente no dia a dia das pessoas e organizações, os fornecedores adotam novas formas de melhorar a informação e serviços disponibilizados ao consumidor considerando também novas formas de interação, hoje ao alcance dos utilizadores.

Torna-se fundamental, configurando uma vantagem competitiva, entender as necessidades e preferências dos utilizadores/clientes. Este conhecimento pode hoje ser acionado de diferentes maneiras, tais como, melhoramento do próprio serviço/produto e/ou canais mediadores, ou até mesmo dimensionamento de novos. No comércio não digital, as fontes de informação são, principalmente, estudos de mercado realizados por meio de pesquisas e entrevistas a consumidores. Tais métodos, embora úteis, são muitas vezes caros e limitados na amostra.

Os dados relacionados com a navegação *Web* tornam-se uma fonte principal para explorar e dar sentido a grandes quantidades de diferentes tipos de dados. Os servidores *Web* não só fornecem os respetivos serviços, como também constituem uma fonte de informação útil para os proprietários dos *websites*, dado que registam as visitas às páginas *Web*. Durante alguns anos, os *Web Logs* foram analisados de forma automatizada para acompanhar os utilizadores do *website* e as respetivas sessões.

De facto, atualmente o mundo digital abre novos horizontes ao nível da compreensão do comportamento do consumidor/utilizador. No entanto, os registos de atividade nem sempre são fáceis de processar, e converter esta análise em resultados viáveis é muitas vezes desafiante (a reconstrução de um *site* é muitas vezes com base em tentativa e erro). Infelizmente, nem todos os proprietários têm conhecimento ou recursos para realizar esta análise e convergir os resultados com os objetivos.

Para além disso, enquanto os *Logs* do servidor são alvo de uma análise intensiva, muita pesquisa foi anexada a estes, nunca tanto trabalho foi realizado ao redor da navegação em geral. Segundo (Chiarandini, 2014), a navegação pode ser definida de forma diferente do que o caso de pesquisa; enquanto o utilizador navega o objetivo é menos claro do que uma pesquisa dirigida porque a navegação não pode ser expressa através de texto. Também pode acontecer que o utilizador não tenha de todo um objetivo definido - navegação intensiva. Adicionalmente (Baeza-Yates, Ribeiro-Neto, & others, 1999) classificou ainda que a tarefa em que os utilizadores de um sistema de recuperação de informação estão envolvidos, pode classificar-se de duas formas: informação ou dados de consulta (pesquisa) e navegação. Ambas são formas em que os utilizadores acedem à informação, sendo a primeira de acesso aleatório e a segunda de acesso sequencial. Pode-se então assumir que esta distinção também exista na estrutura das sessões *Web*.

Atualmente, uma atenção considerável é colocada nas aplicações de *Data Mining*, análise preditiva e prospeção de negócios. Na maioria das aplicações, onde normalmente a expectativa vem destas previsões há necessidade de recuperar ficheiros de registos de acessos ao *website* em questão. Tal como já foi mencionado, extrair conhecimento a partir destes dados acarreta várias exigências e desta forma, novos métodos emergiram, tal como é o caso das ferramentas de *Web Analytics* discutidas na seção 2.3.2 que em certa medida fundamentam com rigor a abordagem deste trabalho.

A partir de uma perspectiva funcional, cada *website* precisa de integrar a lógica que melhora a concretização dos objetivos pretendidos que fundamentam a sua própria existência. Tais objetivos, podem variar consoante as visitas de um utilizador a uma página específica para converter potenciais interessados.

Deste modo, sistemas de recomendação foram criados de forma a fornecer orientação, melhorar a experiência de utilização e finalmente atingir as metas pretendidas.

As recomendações podem ajudar os utilizadores do *website* a focarem-se no objetivo pretendido para evitar dispersões. Podem ser fornecidas algumas indicações diretas, tais como, onde é que se pode obter determinada informação? Como é que se pode alcançar o objetivo? E também associar conteúdo relacionado ou produtos/serviços.

Com o crescimento contínuo e proliferação do *e-commerce*, o volume de dados do utilizador, recolhidos pelas organizações nas suas operações diárias atingiu proporções astronómicas.

Analisar estes dados, pode ajudar as organizações a determinar o valor do tempo de vida de estratégias de *marketing* e produtos/serviços, assim como avaliar a eficácia das campanhas promocionais, otimizando a funcionalidade das aplicações baseadas na *Web*, fornecendo um conteúdo personalizado/segmentado aos utilizadores e mais eficaz na estrutura lógica do *website*.

Este tipo de análise envolve a identificação automática de padrões significativos e relações de uma vasta coleção de dados semiestruturados muitas vezes armazenados na *Web*, registos de acesso e bases de dados operacionais.

Mining de utilização da *Web* refere-se à identificação automática e análise de padrões no fluxo de cliques subjacente à navegação, assim como dados recolhidos ou gerados como um resultado das interações do utilizador com os recursos da *Web*, num ou mais *websites* (Liu, Mobasher, & Nasraoui, 2007). O objetivo é captar o modelo e analisar os padrões de comportamento e os perfis de utilizadores nas suas interações. Os padrões identificados, são geralmente representados como coleções e/ou sequências de páginas, objetos ou recursos frequentemente consultados por grupos de utilizadores com interesses ou necessidades comuns.

Para efetuar tais recomendações, por norma é solicitada determinada informação ao utilizador. Posteriormente, será relacionada com algum segmento ou perfil de utilizador, construído com base no histórico de navegação/utilização do *website*, produzindo a recomendação processada.

Existem alguns pontos fracos nesta lógica global, pois nem sempre é possível identificar o utilizador (esta identificação é feita através de *cookies* que podem não estar disponíveis ou ser permitidos), a informação mais detalhada sobre os utilizadores muitas vezes requer preenchimento completo do perfil, o mesmo utilizador pode navegar em dispositivos diferentes e diferentes utilizadores podem navegar num mesmo dispositivo (considerado o mesmo utilizador devido à chave guardada no *cookie*), o mesmo utilizador embora possa ter alguns interesses a longo prazo pode também ter interesses a curto prazo, que se podem alterar de forma mais dinâmica (por exemplo, o utilizador muitas vezes pode obter recomendações sobre produtos similares aos que já adquiriu).

Por exemplo, em sites de *e-commerce*, uma recomendação desvirtuada pode ter um resultado significativo numa compra futura do utilizador (Cho, Kim, & Kim, 2002). Recomendações pouco válidas podem ter duas características de erros: falsos negativos, que são produtos que não são

recomendados, embora o cliente possa manifestar interesse neles; falsos positivos, que são produtos que são recomendados mesmo que o cliente não apresente qualquer tipo de interesse.

Além disso, deve-se reconhecer que o resultado final dos sistemas de personalização nem sempre coincide com a percepção do utilizador final de tal personalização (Li, 2016). No entanto, (Parra & Brusilovsky, 2015) mostrou que permitindo ao utilizador alguma forma de controlar o grau de personalização, mantendo este processo o menos intrusivo possível, produz um melhor envolvimento entre o utilizador e a experiência geral. Isto pode vir como resultado em que além dos grandes portais em que o perfil de utilizador é atualizado regularmente e os mecanismos de acompanhamento estão disponíveis devido à identificação do utilizador através de autenticação, em inúmeros *websites* é difícil recolher os dados pessoais e, portanto, conhecer o utilizador. Desta forma, a identificação de um utilizador recorrente no *website* e o seu contexto atual é quase impossível de obter.

Portanto, a orientação deste trabalho segue uma linha de identificação e utilização de padrões de navegação sem a preocupação de uma personalização ao utilizador. Isto é, não interessa tanto a sua identificação, histórico e captação de perfil; mas sim, o emparelhamento da sua navegação no momento (no seu atual contexto de necessidades) com os padrões de navegação semelhantes que tiveram sucesso considerando o objetivo definido para o *website*. O alinhamento das necessidades dos utilizadores e objetivos para o *website* tem de ser considerado pelo seu administrador, o âmbito deste trabalho abrange justamente garantir que uma vez definidos o encaminhamento dos utilizadores, este é eficaz.

1.2 Objetivos

No presente trabalho de dissertação, pretende-se construir uma ferramenta de recomendação de navegação incorporável em qualquer *website*. Tendo em conta que, por vezes um utilizador por inúmeras razões possa não compreender o fluxo de um *website* (por exemplo por questões de *design* ou complexidade de um *website*) esta ferramenta pode orientá-lo de acordo com os objetivos (para este problema concreto a solução passa por visitar determinada página) definidos pelo administrador do *website* que contemple a estratégia do mesmo.

A ferramenta proposta tem por base a utilização do histórico de navegação disponibilizado pela ferramenta: *Google Analytics*, sendo as recomendações a realizar orientadas pelos objetivos definidos pelo administrador do *website* no *Google Analytics*.

O objetivo deste trabalho passa por gerar tais recomendações considerando o utilizador no seu contexto de navegação e não o utilizador *per se*. Assim, a sessão do mesmo utilizador não é uma componente necessária de contemplar no desenvolvimento do trabalho, uma vez que esta inclusive ainda mantém alguns desafios relativamente à identificação da sua duração, que têm sido abordados ao longo dos anos numa base heurística. Estudos recentes sobre o assunto apontaram para um limiar de 60 minutos (Halfaker et al., 2015), ou a indexação à média de tempo nas páginas do *website* (Dinucă, 2012), contudo, na maioria das vezes o limiar original estabelecido de 30 minutos é aplicado como exemplo (Berendt, Mobasher, & Spiliopoulou, 2001).

Em termos funcionais, o objetivo do trabalho corrente passa pelo desenvolvimento de uma biblioteca com base na linguagem *Javascript* que disponibiliza um *widget* ao utilizador quando

esta seja incorporada num *website*. Este *widget* tem como função a exibição das recomendações e algumas funcionalidades extra descritas com maior detalhe no capítulo do protótipo.

A biblioteca *Javascript* consome um serviço, também proposto no desenvolvimento aplicacional numa vertente de *software as a service* (SaaS) responsável por consultar os dados reunidos pela ferramenta *Google Analytics* com os parâmetros de pesquisa identificados como necessários, produzindo recomendações com base nos utilizadores que atingiram os objetivos definidos pelo administrador do *website*. Este serviço é o responsável por alimentar a biblioteca *Javascript* mencionada anteriormente com as recomendações produzidas e indicar qual é o momento certo a exibir as mesmas, tendo em conta os tempos médios de abandono de uma determinada página. A lógica implementada no serviço é devidamente detalhada no decorrer do trabalho.

1.3 Metodologia

Para a realização da presente dissertação é utilizada uma metodologia de trabalho em que primeiramente foi analisado o estado da arte que se divide em revisão da literatura, ferramentas de recomendação e *Web Analytics*.

No capítulo de estado da arte são identificados os problemas e diferentes abordagens implementadas na área de recomendação/ personalização de utilizadores *Web*. Relativamente ao capítulo de Ferramentas de Recomendação e *Web Analytics* são apresentadas e discutidas algumas ferramentas de *machine learning* utilizadas para recomendação e ferramentas de análise *Web* com objetivo de recolha e criação de relatórios estatísticos do tráfego de um *website*. Após introduzido o termo *Web Analytics* e o seu conceito, é apresentada com maior detalhe a ferramenta base deste projeto o *Google Analytics* e os seus objetivos.

Após integração com o estado da arte e abordagens de recomendação/*Web Analytics* segue-se a proposta de recomendação suportada em *Web Analytics* onde está descrito o modelo proposto e toda a sua lógica subjacente. Uma vez estabelecido o modelo procede-se a uma implementação do respetivo protótipo funcional. Para tal, define-se a sua arquitetura geral e relação entre os diferentes componentes do mesmo. Após conclusão desta fase de análise concretiza-se a implementação/especificação do serviço *GoRecommend* que é o responsável pela criação das recomendações e base da biblioteca *GoRecommendJS*, que está implementada seguidamente, responsável pela camada representacional do protótipo.

Dando-se por concluída a implementação é possível realizar a descrição funcional da aplicação através de um caso prático e *screenshots* representativos, que se inserem neste documento.

Com o protótipo concluído é descrita a abordagem de avaliação que melhor demonstra o seu potencial através de recolha de dados reais (subjacentes por questões de confidencialidade) confrontados com o modelo. Discutem-se os resultados e limitações identificadas.

Por último realiza-se uma reflexão relativamente à contribuição que este trabalho possa trazer à comunidade e o estudo que pode ser desenvolvido com base neste conceito.

2 Revisão da Literatura

2.1 Conceitos Gerais

Extrair conhecimento sobre os utilizadores da *Internet* e o seu histórico de navegação tem sido uma área de pesquisa para muitos domínios de aplicação, desde a procura por melhorias na experiência de utilização (Ex: recomendação e personalização), à procura da eficácia no esforço de *marketing* digital segmentado e desta forma alcançar o utilizador com conteúdos relevantes. Independentemente do objetivo, as diferentes abordagens que têm sido adotadas enfrentam vários desafios comuns. Neste capítulo em que se aborda uma caracterização global do trabalho relacionado na área, serve a presente subsecção para estabelecer a base sobre conceitos e terminologia a ser considerada nas subsecções seguintes.

O termo *Web Mining* tornou-se muito popular principalmente devido à aplicação de técnicas de prospeção de dados para extrair conhecimento a partir de dados da *Web*, incluindo documentos, *hiperlinks*, registos de utilização de *websites*, etc. (Mehtaa, Parekh, Modi, & Solanki, 2012).

Atualmente, a área de *Web Mining* compreende 3 categorias principais:

- 1) *Web Mining* de conteúdo;
- 2) *Web Mining* de estrutura;
- 3) *Web Mining* de utilização (ou navegação).

Web Mining de conteúdo é o processo de extração de informação semântica a partir de conteúdo de documentos *Web*, tais como, textos ou imagens transmitidas ao utilizador final. Tal processo engloba a descoberta de recursos, categorização e agrupamento.

Web Mining de estrutura foca-se na estrutura de *hiperlinks*, normalmente envolvendo a análise dos *hiperlinks* de entrada e saída de páginas *Web*, este método é usado pelos motores de busca para classificação de resultados (Mehtaa et al., 2012). Este tipo de *Mining* permite extrair conhecimento sobre a organização intra e inter páginas.

Web Mining de utilização, consiste principalmente na aplicação de técnicas de *Data Mining* para compreender e servir as necessidades dos utilizadores das páginas *Web* (Srivastava, Cooley, Deshpande, & Tan, 2000). Tipicamente, nos dados reunidos para tal análise inclui-se endereços de IP, navegador utilizado, páginas visitadas, origem e tempo em página, procurando a eliciação de padrões de utilização para informar domínios de aplicação tais como: personalização, recomendação (orientada pela navegação, não a baseada em conteúdo) ou interfaces adaptativas.

O processo de *Web Mining* compreende o fluxo típico associado às iniciativas tradicionais de *Data Mining*: recolha de dados, pré-processamento (limpeza, formatação, relacioná-los e alinhá-los para que se revelem compatíveis/acionáveis considerando a utilização pretendida), aplicar o(s) método(s) de análise de eleito(s); e quando não se trata de uma iniciativa somente exploratória e/ou descritiva, produzir consequentemente um modelo e utilizá-lo para classificação e/ou previsão atendendo aos objetivos finais definidos, por exemplo, sistemas de recomendação e/ou personalização.

As principais fontes de dados usadas para *Web Mining* são os *Logs* de acesso a servidores *Web* e *Logs* aplicativos. Fontes de dados adicionais também são complementarmente utilizados tanto para a preparação dos dados como para seu processamento na identificação de padrões, incluindo ficheiros do *website*, metadados e bases de dados operacionais. Em alguns casos podem estar disponíveis dados adicionais devido à recolha dos mesmos no lado do utilizador, por exemplo, na informação persistida em *cookies*, ou disponibilizada em contas que contenham perfis de utilizadores (contemplando, e.g. dados demográficos e preferências). Também serviços globais de agregação de dados e monitorização de tendências como: *ComScore* (www.comscore.com), *NetRatings* (www.nielsen-netratings.com) e *Acxiom* (www.acxiom.com), constituem potenciais fontes adicionais a compor ou cruzar na exploração dos dados.

2.2 Propósitos e Métodos

O corpo do trabalho em *Web Mining* pode ser enquadrado em cinco grandes âmbitos: 1) Análise de sessões, 2) Análise de agregação e segmentação de utilizadores, 3) Análise de associações e correlações, 4) Análise sequencial (temporal) e de padrões de navegação e 5) Classificação e previsão baseada nas transações do utilizador (Liu et al., 2007).

Análise de sessões é uma abordagem comum para gerar medidas estatisticamente agregadas sobre sessões. Técnicas padrão de estatística são utilizadas para adquirir conhecimento sobre o comportamento geral dos utilizadores, considerando, por exemplo, hora do dia, páginas frequentemente acedidas ou de entrada e saída, tempo em página, etc. No entanto, outra forma de análise de sessões é a análise integrada de dados de utilização através da abordagem OLAP (*Online Analytical Processing*) (Alves, Belo, & Costa, 2007). A abordagem OLAP assenta numa estrutura mais integrada para análise com maior flexibilidade e em diferentes níveis de granularidade. A fonte de dados para a análise de OLAP recai geralmente na construção de um (híper) cubo. Um cubo de dados que pode ser visto como uma estrutura de matriz multidimensional, em que cada dimensão representa um conjunto de atributos e cada célula armazena, o valor agregado de uma medida, como contagem, soma, etc. Este tipo de representação de dados suporta uma navegação exploratória através das dimensões e respetivos atributos. Os dados são carregados através de uma base de dados multidimensional que mantém um modelo de dados dimensional desnormalizado integrando as dimensões selecionadas para análises (co)relacionadas. As ferramentas OLAP permitem hierarquias e alterações nos níveis de agregação ao longo de cada dimensão durante a análise. Análise de dimensões neste tipo de estruturas pode ser baseada nos vários atributos disponíveis; atributos do modelo dimensional e.g. duração, domínio, recurso utilizado, navegador, origem; obtidos a partir de fontes originais (como por exemplo *Logs*) que carregaram segundo o processo ETL (Extract-Transform-Load) o modelo dimensional (Kimball & Ross, 2015). Assim permite-se que a análise seja efetuada em partes do *Log* correspondentes a um período de tempo ou a um nível diferente de abstração com respeito a qualquer dimensão.

A abordagem OLAP pode ser estendida a outros âmbitos de *Web Mining*. Por exemplo, a integração de dados de *e-commerce*, quando presentes, explora ainda mais o potencial desta orientação analítica para providenciar métricas importantes para o negócio (exemplo paradigmático no domínio de *Business Intelligence*, que pode abranger os vários âmbitos).

Também é comum que resultados da exploração e análise OLAP sejam utilizados para informar a seleção de métodos complementares de *Data Mining* utilizados nos restantes âmbitos de *Web Mining*.

A análise por agrupamento de dados (*clustering*) constitui uma das técnicas de *Data Mining* que agrupa um conjunto de dados que tenham características semelhantes. No domínio da análise de utilização/navegação de um *website*, existem 2 tipos principais de agrupamentos interessantes que podem ser identificados: grupos de utilizadores e grupos de páginas (Hassan & Karim, 2011). Agrupar registos de utilizadores (sessões ou transações) é uma das tarefas mais utilizadas em *Web Mining* de utilização e *Web Analytics* (que será abordado na secção seguinte), que procura estabelecer grupos de utilizadores que contêm padrões de navegação semelhantes.

Por outro lado, o agrupamento por páginas pode ser efetuado com base nas características de conteúdo associado (palavras chave ou atributos). No caso do agrupamento com base no conteúdo, os resultados podem ser grupos de páginas ou produtos relacionados com o mesmo tópico ou categoria.

Uma outra abordagem suporta-se na descoberta de regras de associação e análise de correlação tendo como objetivo encontrar grupos de itens ou páginas frequentemente consultadas ou adquiridas em conjunto. Por sua vez, permite organizar o conteúdo de *websites* e/ou navegação de forma mais eficiente, ou promover recomendações transversais com eficácia.

Na verdade, uma das principais aplicações de *Mining* em regras de associação, quer em utilização *Web*, assim como dados *e-commerce* é a recomendação (que será discutida especificamente na subsecção seguinte).

Se for introduzida, a dimensão tempo, tem-se uma orientação diferente da análise. Até aqui, foi considerada a orientação por semelhanças ou associações. A análise de padrões temporais é uma técnica específica de *Mining* que tenta encontrar tendências maioritariamente inter-sessões. Tal análise permite, por exemplo, otimizar o *marketing Web* prevendo a natureza de visitas futuras. Outros tipos de análise temporal que podem ser efetuados sobre padrões sequenciais incluem deteção de pontos de mudança (de tendência) e análise de similaridades e singularidades.

Muitos dos métodos acima referidos, permitem em última análise informar modelos de previsão. Existem vários métodos cuja sua natureza é a classificação e previsão. A classificação consiste em mapear um conjunto de dados em várias classes predefinidas (e.g. previamente identificadas segundo alguma das abordagens anteriores). No domínio *Web* tal consiste em desenvolver o perfil dos utilizadores que pertencem a uma classe ou categoria em particular. O que requer a extração e seleção de características que melhor descrevem as propriedades de uma determinada classe ou categoria. A classificação pode ser efetuada por meio de algoritmos de aprendizagem e supervisionados, tais como, árvores de decisão, classificadores Bayesianos (*Naive Bayes networks*), máquinas de vetores de suporte (*Support Vector Machines*). Uma vez treinado, é avaliado o classificador, e este, poderá ser utilizado para previsão.

2.3 Abordagens à Recomendação e Personalização

Conforme introduzido na secção anterior a abordagem mais comum à recomendação suporta-se na descoberta de associações tipicamente baseadas numa das muitas variantes do algoritmo

A priori original. O objetivo destes algoritmos é encontrar grupos de itens que juntos ocorrem frequentemente em muitas transações, com a consideração de diversos critérios complementares (Slimani & Lazzez, 2014). Embora muito utilizados e com resultados satisfatórios para a identificação de associações, a recomendação com base em regras de associação inerentemente enfrenta degradação se o conjunto dos dados for disperso (Kasliwal & Katkar, 2015) (Dinucă, 2012). Este é frequentemente o caso em *Web Mining* de utilização pois as visitas dos utilizadores cobrem apenas uma pequena fração dos itens disponíveis e, portanto, muitas vezes é difícil encontrar um número suficiente de itens comuns nos diferentes perfis de utilizadores. Existem várias propostas para superar tais limitações (Liu et al., 2007)], por exemplo, redução do número de dimensões (mas podem estar a ser descartadas dimensões interessantes para os padrões finais); ou em alternativa, reduzindo o lado esquerdo das regras no universo de regras globais, considerando o caminho percorrido pela sessão ativa, reduzindo a restrição de uma completa correspondência com a regra no seu lado esquerdo.

Uma área com grande aplicação de *Web Mining* é de facto a recomendação e ainda a sua frequente utilização para potenciar personalização com objetivo de melhorar a experiência do utilizador. Embora sejam áreas (recomendação e personalização) distintas *per se*, não é incomum utilizar recomendações na personalização. De forma geral, os sistemas de personalização *Web*, dada a sua finalidade e complexidade podem ser enquadrados de acordo com (Suguna & Sharmila, 2013) nos principais tipos seguintes: 1) memorização, a forma mais simples de personalização, onde as informações do utilizador como o nome e histórico de navegação são armazenadas (por exemplo, usando *cookies*), para mais tarde reconhecer o utilizador quando este voltar; 2) Personalização, recolher informação sobre o utilizador através de formulários com o objetivo de personalizar conteúdo e estrutura do *website* (o processo global tende a ser no máximo semiautomático, pois exige sempre alguma resposta por parte do utilizador); 3) Orientação ou recomendação, recomendar/auxiliar navegação ou conteúdos considerados relevantes para os interesses do utilizador (implicitamente inferidos, como por exemplo, perfil de navegação, ou explicitamente, como por exemplo, formulários de preferências dos utilizadores) a fim de facilitar o acesso à informação/produto num grande *website*; 4) Apoio ao desempenho de tarefas, executando ações em nome do utilizador.

Além dos algoritmos de *Mining* já mencionados acima, com particular relevância para as regras de associação em sistemas de recomendação, existem duas abordagens complementares utilizadas na recomendação e personalização: a recomendação baseada em filtragem de conteúdo e a recomendação baseada na filtragem colaborativa.

Na técnica de filtragem baseada em conteúdo, a recomendação de páginas considera os itens em que o utilizador tenha classificado anteriormente com base no seu perfil. O perfil do utilizador pode ser construído examinando respostas a pequenos questionários (por exemplo formulários de registo que explicitamente recolhem preferências), classificação de itens ou histórico e padrão de navegação do utilizador. Este método de recomendação depende fortemente da experiência passada pelo utilizador e mesmo quando possivelmente saber mais acerca do mesmo. A sua caracterização contextual continua a ser um desafio, dados os interesses dinâmicos e oportunistas (por exemplo após uma compra específica, o seu interesse pode mudar a curto prazo ou até mesmo o foco no *website*) (Nadi, 2011).

Na abordagem de filtragem colaborativa e.g. (Riecken, 2000), páginas *Web* (ou produtos) são recomendadas a um utilizador particular quando outros utilizadores semelhantes também preferem as mesmas páginas. A definição de "semelhança" entre utilizadores depende das

aplicações que utilizam os sistemas de recomendação, por exemplo, podem ser definidos utilizadores semelhantes quando estes atribuem classificações próximas a páginas *Web* ou *websites*, assim como, utilizadores com padrões de navegação semelhantes. Esta abordagem também é alvo de alguma crítica apontando duas desvantagens específicas: 1) Dispersão e 2) Escalabilidade (Suguna & Sharmila, 2013). Sendo a dispersão, a quantidade de classificações obtidas anteriormente, podendo ter ordens de grandeza menor que o número de classificações que necessitam realmente de ser previstas. Em relação à escalabilidade, decorre que, em geral cálculos muito longos podem ser necessários quer pelo crescimento do número de utilizadores, quer pelo número de recomendações (por exemplo produtos/páginas).

2.4 A Perspetiva *Web Analytics*

A associação de *Web Analytics* define a mesma como a “medida, recolha, análise e comunicação de dados da *Internet* com o objetivo de compreender e otimizar a utilização da *Internet*”.

De acordo com (Nakatani & Chuang, 2011), *Web Analytics* é utilizado para compreender os clientes *online*, assim como os seus comportamentos, a influência do *design* para que seja possível tomar decisões benéficas para o negócio e alcançar os objetivos da organização. Os autores ressaltam, no entanto, que o objetivo da análise *Web* não é apenas otimizar páginas do *website*, mas de apoiar a concretização dos objetivos globais de uma organização.

O campo de análise *Web* começou no início dos anos 90 com a expansão da *Internet*. Ao longo de todos estes anos evoluiu e tornou-se uma parte vital no desenvolvimento de um *website*. Dos 500 *websites* mais consultados, 91% implementam *Web Analytics* para guardar informações sobre os seus utilizadores (Hamel, 2012). Apesar de ser utilizado há mais de 2 décadas a análise da *Web* ainda está em fase de crescimento inicial do seu ciclo de vida, pois existe uma evolução contínua com métodos mais avançados a serem continuamente introduzidos.

Tal como referido anteriormente *Web Analytics* tenta compreender os utilizadores de um *website* através da análise das métricas recolhidas para que a eficácia dos *websites* seja melhorada. Via *Web Analytics* pode-se inferir comportamento dos utilizadores face aos cliques realizados, às suas pesquisas dentro de determinado *website*, permitindo ao analista detetar os locais exatos de cliques, a sua frequência e fluxo.

Os resultados destas análises são fornecidos por parte de especialistas sob a forma de tabelas, gráficos ou diagramas, cujo resultado da análise levará potencialmente a redesenho de imagem, estrutura e conteúdo de secções do *website*.



Figura 1 - Diagrama *Web Analytics*

(fonte: <http://www.canadianschoolguide.com/what-is-web-analytics.php>)

Em suma, *Web Analytics* não é apenas um processo para medir o tráfego na *Internet*, mas sim um instrumento acionável para vários objetivos. Alguns exemplos são: a pesquisa de negócios e/ou mercados, avaliação e melhoria na eficiência e eficácia de um *website*, ou na quantificação da eficácia de campanhas de *marketing*.

Efetivamente é possível de forma estruturada e sistematizada:

- Avaliar quantitativamente a qualidade de um *website*;
- Realizar análises comparativas;
- Identificar padrões do *website*;
- Demonstrar alinhamento com objetivos do negócio;
- Avaliar experiência de utilização.

As aplicações *Web Analytics* podem auxiliar na avaliação de resultados relativamente ao impacto de campanhas mesmo em canais tradicionais (e.g. televisão), pois cada vez mais estas remetem para as plataformas digitais que se monitorizadas, alimentarão indicadores passíveis de ser contrastados com os objetivos da campanha.

Web Analytics é utilizado para compreender os padrões de uso do site e de navegação dos utilizadores e fornecer aos gestores de *marketing* um mecanismo perspicaz para melhorar o desempenho do *website*.

Complementarmente, ter acesso a informações sobre os utilizadores do *website* ajuda os próprios programadores a tomar decisões apropriadas sobre o desenvolvimento do mesmo. Quanto mais conhecimento disponível sobre os utilizadores, mais fácil será satisfazer as suas necessidades. (Phippen, Sheppard, & Furnell, 2004) afirmam que a análise *Web* é fundamental para o sucesso dos *websites*. Baseado no seu estudo, as organizações que adotaram e integraram a análise *Web* têm vindo a colher benefícios múltiplos e a descobrir informações valiosas sobre os seus clientes. O conhecimento do comportamento dos utilizadores permite que um site possa criar uma relação forte com o mesmo, pois trará satisfação à sua utilização. Como já referido as ferramentas de *Web Analytics* permitem a recolha de informações muito detalhadas sobre os visitantes, incluindo a quantidade de tempo gasto em uma determinada página, a localização física de um utilizador, o caminho de navegação do mesmo, etc. No entanto pode ser sensível tirar conclusões definitivas sobre os dados reportados pelas ferramentas. (Kaushik, 2008) dá o seguinte exemplo: qualquer ferramenta de *Web Analytics* indica que

páginas de um *website* são mais visitadas, mas com base em dados de fluxo de cliques é difícil dizer se essas páginas têm realmente o conteúdo que os utilizadores procuram. A razão para visitas frequentes pode ser apenas um sistema de navegação enganoso ou desequilibrado. O que complica ainda mais o problema é a diversidade de utilizadores, pois os *websites* servem vários grupos de interessados ao mesmo tempo e diferentes grupos têm razões diferentes. A interpretação das métricas *Web* pode inclusive ser muito simplificada quando o efeito de variáveis intervenientes não é considerado, estas não devem ser examinadas separadamente, mas sim em conjunto com outras métricas do negócio.

2.5 Ferramentas de Recomendação e *Web Analytics*

Nesta seção são abordadas inicialmente algumas ferramentas de recomendação, objetivos das mesmas e alguns exemplos. Posteriormente as ferramentas de *Web Analytics* com maior destaque para a ferramenta de *Web Analytics* base deste trabalho o *Google Analytics*.

2.5.1 Ferramentas de Recomendação

As ferramentas de recomendação alteram a forma como os utilizadores dos *websites* interagem com estes. Ao invés de se cingir a uma experiência estática em que os utilizadores pesquisam determinados conteúdos/produtos, dotam-nos de uma interatividade adicional para providenciar uma experiência melhorada de utilização. As ferramentas de recomendação preveem as necessidades e expectativas de um utilizador específico com base no seu histórico de visualizações, pesquisas, comportamento e preferências, constituindo e considerando o seu perfil, o perfil de outros utilizadores e caracterização dos conteúdos/produtos.

Por norma, as listas de recomendação produzidas nestas ferramentas suportam-se numa de duas abordagens principais: filtragem colaborativa e filtragem com base em conteúdo, cujo princípio e lógica subjacentes foram já expostos na secção de revisão da literatura.

Segue-se uma breve apresentação de algumas ferramentas representativas nesta área, focando os seguintes aspetos: como estas ferramentas são disponibilizadas, através de um serviço ou como uma aplicação local, os componentes principais que as constituem e a sua lógica principal de funcionamento.

PredictionIO

O *PredictionIO* é um serviço de *Machine Learning OpenSource* fornecido por um servidor que disponibiliza uma application programming interface (API) à qual é possível fazer pesquisas com base nos dados carregados que deram origem ao modelo de recomendação.

PredictionIO é composto pelos seguintes componentes:

- Plataforma *PredictionIO* – Conjunto de ferramentas *OpenSource* para compilar, avaliar e produzir resultados com base em algoritmos de *machine learning*;
- Servidor de Eventos – Camada de *Analytics* para compilar eventos de múltiplas plataformas;
- Galeria de *Templates* – Onde estão disponíveis os diferentes algoritmos de *machine learning*.

Como funciona:

O servidor de eventos coleciona continuamente os dados reportados pelas diferentes plataformas. Posteriormente os dados são tratados pelos algoritmos de *machine learning* escolhidos da galeria de *templates* e são construídos modelos preditivos que são compilados para um serviço *Web* que recebe filtros como parâmetros e responde com previsões em *real-time*. A arquitetura lógica está ilustrada na figura abaixo.

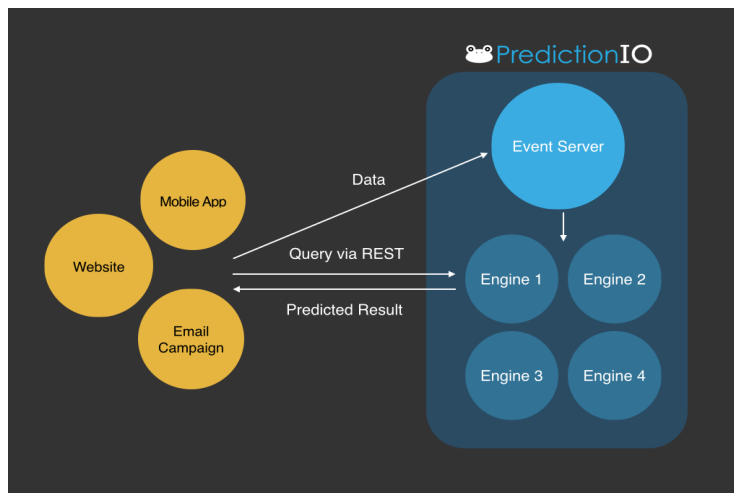


Figura 2 - Integração entre componentes - PredictionIO

(fonte: <http://predictionio.incubator.apache.org/start/>)

O *PredictionIO* providencia um motor de recomendação baseado na ferramenta *Apache Spark MLlib's* que por sua vez aplica o algoritmo de filtragem colaborativa por defeito.

Google Prediction API

O *Google Prediction API* é um serviço todo ele baseado na *cloud* que permite construir modelos de *machine learning*. Providencia um conjunto de ferramentas de previsão que podem ajudar a analisar os dados de um determinado modelo para alimentar outras aplicações, tais como análise de sentimentos do cliente, detecção de *spam*, sistemas de recomendação, etc.

Google Prediction API é composto pelos seguintes componentes:

- *Google cloud storage* – Nuvem de armazenamento para a qual o modelo de dados definido pelo utilizador é carregado;
- API – Serviço através do qual o utilizador pode carregar dados para a *cloud storage*, formatar esses mesmos dados e posteriormente efetuar pesquisas sobre os mesmos.

Como funciona:

Em primeiro lugar deve ser criado um projeto na plataforma *Google cloud*. Após o projeto criado é possível criar vários modelos de treino, para isso basta criar um novo “*bucket*” com um nome único e carregar os dados de treino (o *google* disponibiliza também uma *API*), através da qual é possível carregar um modelo dados previamente definido), assim que o modelo for carregado é

possível consultar o estado do treino, sabendo se este já foi concluído ou não e obter informação acerca do grau de precisão dos resultados e outros metadados.

Após o treino dos dados é disponibilizada uma API onde é possível pesquisar por previsões em tempo real com base nos modelos treinados, tendo como vantagem o grande nível de *performance* pois a resposta da API nunca é superior a 200ms.

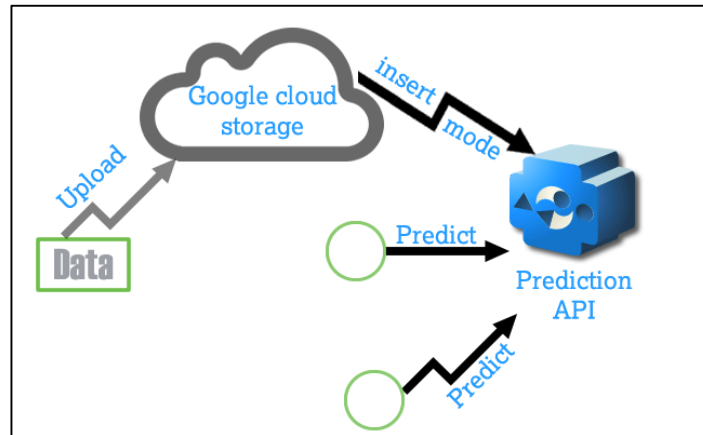


Figura 3 - Integração entre componentes – Google Prediction API

(fonte: <https://cloud.google.com/prediction>)

2.5.2 Ferramentas *Web Analytics*

Website, blog ou aplicação móvel, em qualquer dos casos é importante medir e acompanhar o desempenho e o impacto de cada uma destas plataformas relativamente à utilização/consumo da oferta de conteúdos/ produtos, considerando não só o conteúdo como também a navegação subjacente.

Através do acompanhamento das estatísticas e da interpretação destes dados é possível conhecer os perfis dos utilizadores (por exemplo, que tipo de conteúdos de leitura apreciam ou não) e, conseqüentemente, implementar melhorias que permitam tornar o *website, blog* ou aplicação móvel mais atrativos, indo ao encontro dos seus interesses com uma melhor usabilidade e experiência de navegação. Há diversas ferramentas no mercado intuitivas que transformam a informação recolhida num relatório fácil de consultar. Existem diferentes tipos de licenciamento de utilização, sendo algumas destas, pelo menos relativamente a um conjunto de funcionalidades, gratuitas. Em baixo apresenta-se alguma das ferramentas melhor cotadas de acordo com a Figura 7 - *Ferramentas de Web Analytics*, para analisar o tráfego do *website*.

Adobe Analytics (Pago) – O *Adobe Analytics* é mais completo que o *Google Analytics*, no entanto não existe uma versão gratuita desta ferramenta. O *Adobe* utiliza relatórios de *marketing*, de vendas, relatórios de todo o comportamento do utilizador, entre outros. Trata-se de uma ferramenta que permite a análise em tempo real, funcionando muito bem em dispositivos como os *tablets*, *smartphones* e *desktop*.



Figura 4 – Adobe Analytics

(fonte:
<https://www.adobe.com/marketing-cloud/web-analytics.html>)

Piwik (Gratuito) - é um *software OpenSource*, que permite acompanhar e analisar vários *websites* em tempo real. Para além de fornecer gráficos e tabelas detalhadas sobre as características dos utilizadores e da sua utilização, esta plataforma permite ainda criar objetivos



Figura 5 – Piwik
(fonte:
<https://piwik.org>)

e acompanhar campanhas. Uma das principais características do *Piwik* é que é uma ferramenta bastante personalizável. Por exemplo, o painel principal é constituído por uma série de *widgets* que podem ser trocados ou reorganizados de acordo com as preferências de cada um. Para além disso, ao contrário do que acontece com o *Google Analytics*, os dados são propriedade do administrador, fazendo com que estes não possam ser partilhados com terceiros.

Google Analytics (Gratuito) – O *Google Analytics* é a ferramenta mais conhecida. A assinatura da *Google* é, muitas vezes, motivo suficiente para esta ser a primeira escolha das empresas. A acrescentar a isto está o facto de ser gratuito, flexível e fácil de usar, permitindo monitorizar em tempo real um *blog*, *website* ou aplicação móvel e aceder a gráficos e tabelas com informações detalhadas sobre os



Figura 6 - Google Analytics

(fonte:
<https://www.google.com/analytics>)

utilizadores, origem do tráfego, etc. Esta plataforma disponibiliza uma série de relatórios padrão sobre as características e comportamentos do público-alvo, que podem ser personalizados de acordo com os interesses e objetivos de cada um.

Uma das principais vantagens do *Google Analytics*, é permitir estabelecer objetivos, ou seja, configurar uma série de *links* com objetivo de consulta no *website*. Desta forma é possível ajustar e avaliar estratégias relativamente a campanhas de *marketing* digital e/ou experiência de utilização, facilitando o acesso do utilizador até estas páginas.

O *Google Analytics* também dispõe de um “mapa de calor” definido como *Heatmap* que representa graficamente dados, onde os valores assumidos por uma variável num mapa bidimensional são representados por cores. O objetivo é analisar as áreas onde os utilizadores estão focados, através de cliques e movimentos do rato.

Tendo neste trabalho adotado a utilização do *Google Analytics*, como se justifica na secção discussão, no final deste capítulo, a sua estrutura, abordagem e conceitos são descritos com maior detalhe na secção seguinte.

Seguidamente, podemos verificar uma tabela com a relação frequência de pesquisa e cotação das ferramentas de *Web Analytics*, com origem na plataforma líder *Trustradius* (<https://www.trustradius.com>), onde o objetivo passa pela partilha de críticas / opiniões sobre experiências organizacionais na adoção de determinadas ferramentas:

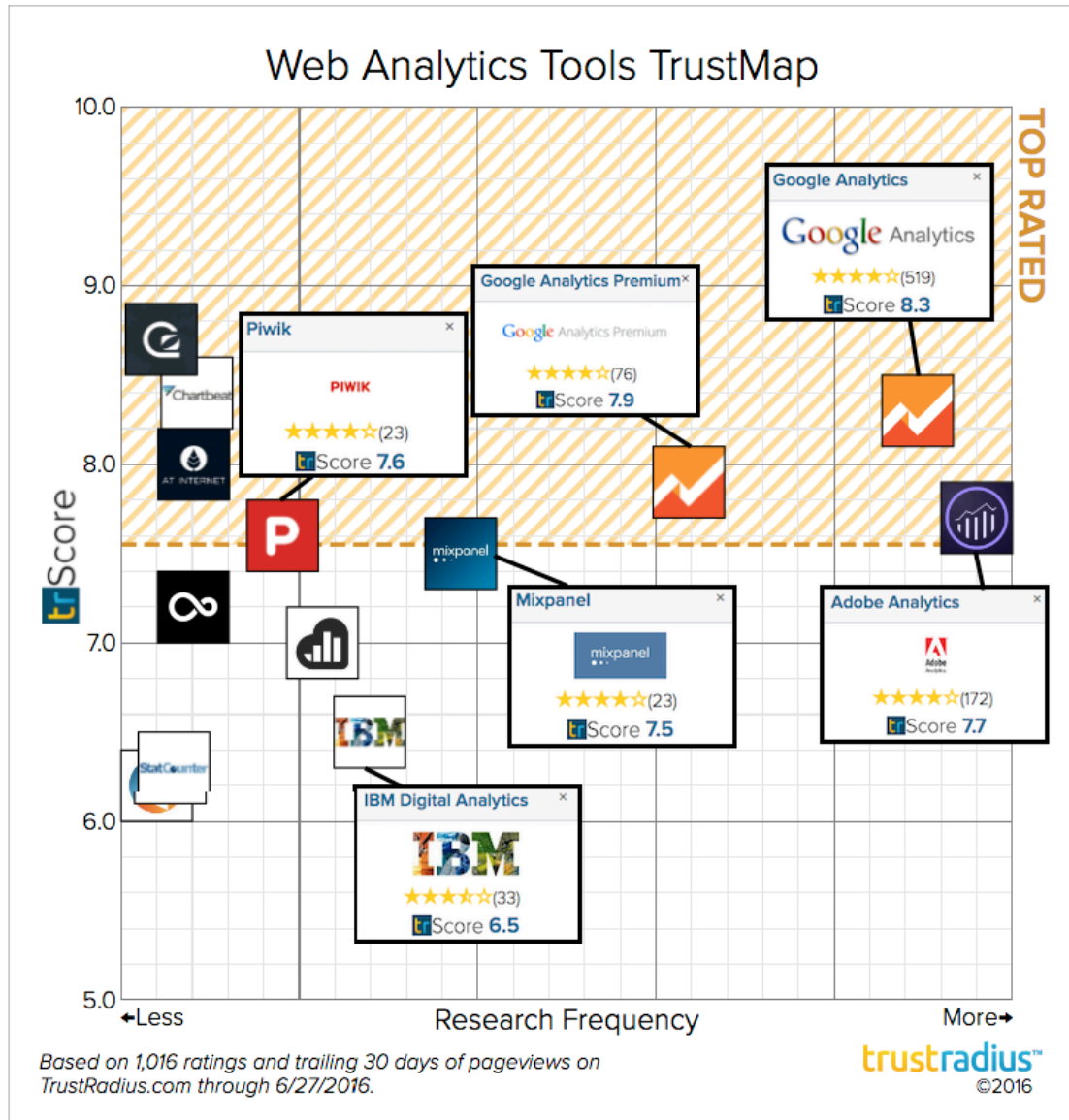


Figura 7 - Ferramentas de Web Analytics

(fonte: <https://www.trustradius.com/web-analytics>)

2.5.3 Google Analytics

As bibliotecas *JavaScript* do *Google Analytics* utilizam *cookies HTTP* (*hypertext transfer protocol*) para identificar o utilizador num *website*. Esta identificação ocorre armazenando um identificador único para que o site possa ler esse identificador se o utilizador voltar novamente e desta forma reconhecê-lo com base nos dados colecionados anteriormente. O *Google*

Analytics armazena também informações sobre as páginas, visitas do utilizador, a frequência com que este utiliza o site e quanto tempo permanece no mesmo.

Objetivos

A missão principal do *Google Analytics* é perceber de uma forma geral como é que os utilizadores se comportam ao navegar pelas diversas páginas e secções de cada site e não apenas saber quantos utilizadores o fazem.

Por exemplo, é com base nas informações colecionadas através do *Google Analytics* que podemos analisar se os visitantes de um determinado site estão a ter um comportamento desejado como resposta a uma dada campanha de *marketing online*.

O *Google Analytics* apresenta vários instrumentos para quantificar essas avaliações, que vão para além do simples carregamento e monitorização do código de acompanhamento nas páginas. Um desses instrumentos do *Google Analytics* são os objetivos. Estes permitem monitorizar um destes elementos:

- quantas pessoas chegam a uma certa página (predefinida);
- quantas pessoas ficam no *site*/pagina um determinado intervalo de tempo;
- quantas pessoas viram um número mínimo de páginas no *site*;
- quantas pessoas desencadearam um determinado evento (e.g. preencher um formulário).

Porque é que os objetivos devem ser criados?

Os objetivos dão-nos uma boa quantidade de informação útil. Com objetivos podemos monitorizar se as pessoas que estão num determinado *website*, estão a fazer aquilo que é suposto estarem a fazer (visitar uma página, uma venda, ver um vídeo, preencher um formulário, etc). A navegação que leva à concretização dos objetivos origina as conversões, i.e., dos visitantes quantos atingiram os objetivos estabelecidos na monitorização. Na figura abaixo podemos verificar a correspondência entre o numero de visitas e total de objetivos cumpridos.

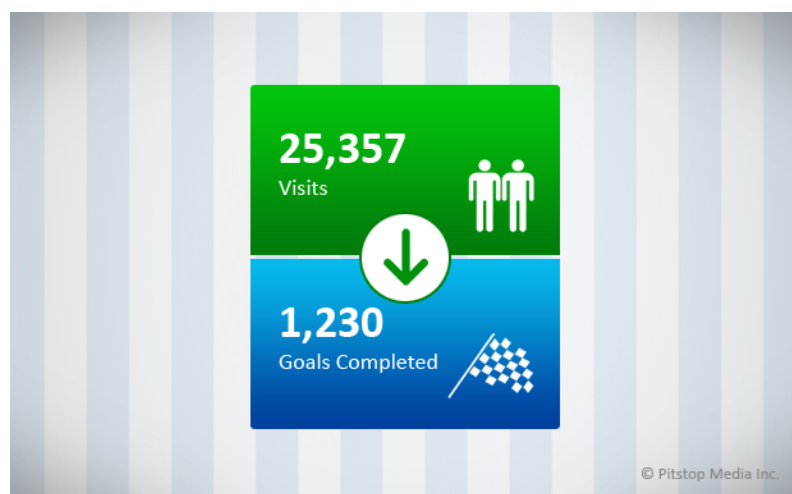


Figura 8 - Número de conversões vs número de visitas

(fonte: <http://www.pitstopmedia.com/sem/visits-goal-report-google-analytics>)

Os objetivos podem ser criados na secção de *administração do painel do Google Analytics*. O utilizador deve clicar em Objetivos, e para criar um objetivo, clica em Novo Objetivo. Irá dispor de um total de 20 objetivos. Se o utilizador pretender mais do que 20 objetivo terá de pagar pelo serviço de *Google Analytics*. Na descrição do objetivo, deve escolher o nome e tipo de objetivo.

Devem ser criados objetivos que sejam úteis e relevantes, o que requer um trabalho prévio de alguma análise e identificação de candidatos aos mesmos. Há muitas coisas que o utilizador pode monitorizar, mas podem revelar-se irrelevantes para obter informação que lhe vá permitir ter lucro ou notoriedade para o seu *website*/ negócio.

É aqui que se revela a pertinência de uma estratégia bem pensada quer na sua implementação quer na sua monitorização. Terá de ser pensado sobre aquilo que se deseja que os visitantes façam quando estão em determinado *website*.

Dimensões:

O conceito de dimensão é implementado no *Google Analytics* como um atributo descritivo ou uma característica de um objeto que pode receber valores diferentes. Por exemplo, um local geográfico poderia ter uma dimensão chamada Latitude, Longitude ou Nome da cidade. Os valores para a dimensão Nome da cidade poderiam ser São Francisco, Berlim ou Lisboa.

Browser, *Página de saída*, *Dispositivo* e *País* são exemplos de dimensões exibidas por padrão no *Google Analytics*. As dimensões podem ser exibidas em todos os relatórios, ainda que se possam ver algumas dimensões diferentes, dependendo do relatório específico e sua customização. Estas devem ser usadas para ajudar a organizar, segmentar e analisar os dados.

Com o *Google Analytics*, é possível também criar dimensões customizadas para guardar tipos adicionais de dados que são enviados por meio de um código de acompanhamento, usando a importação de dados ou a *Google Analytics API*.

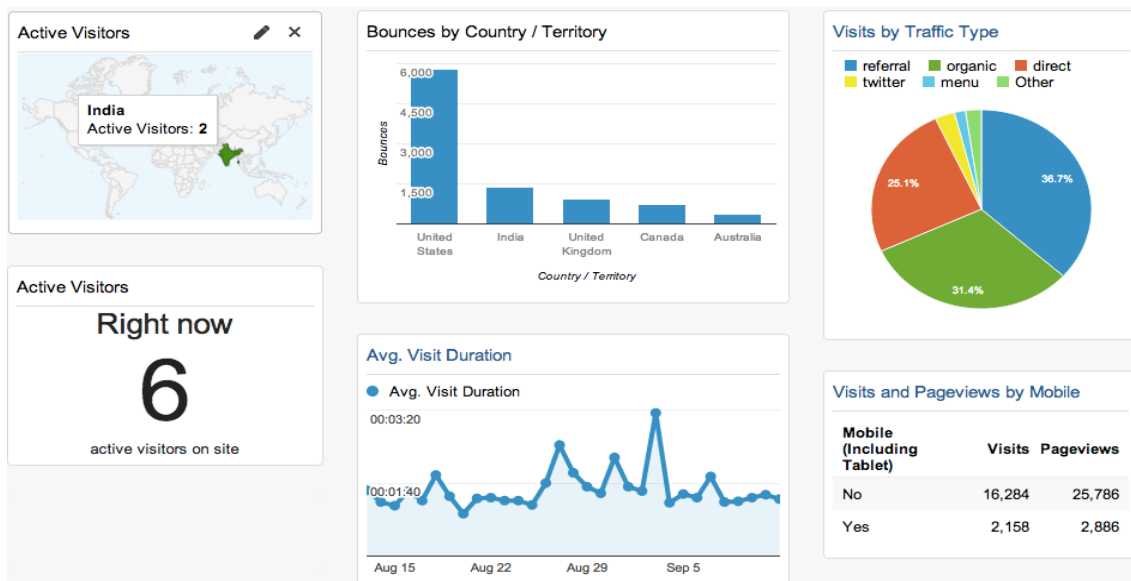


Figura 9 - Google Analytics como uma ferramenta de gestão

(fonte: <https://www.google.com/analytics/>)

Como exemplo de uma dimensão pode ser analisada a origem do tráfego e várias podem ser as análises feitas relativamente a cada fonte de tráfego. Por exemplo, a taxa de abandono pode ser analisada tanto para as pessoas que chegam através do *Google Search Engine*, como também para aquelas que acedem pela via dos *hyperlinks* (tráfego referenciado ou *referral*). As origens mais comuns são:

- **Tráfego Direto:** são as pessoas que acedem ao website colocando o endereço no motor de busca ou acedendo aos favoritos.
- **Hyperlinks (outros sites):** também chamados de tráfego de referência ou *referral*. Mostra todos os websites de terceiros que levaram visitantes para o seu website. Nesta lista, entram também as redes sociais, como o *Facebook* e o *Twitter*.
- **Procura Orgânica:** mostra todas as palavras chave que trouxeram os visitantes para o site através do *Google* e outros motores de pesquisa.

Na figura abaixo podemos verificar o número de visitas mensais agrupadas pela origem dos utilizadores:

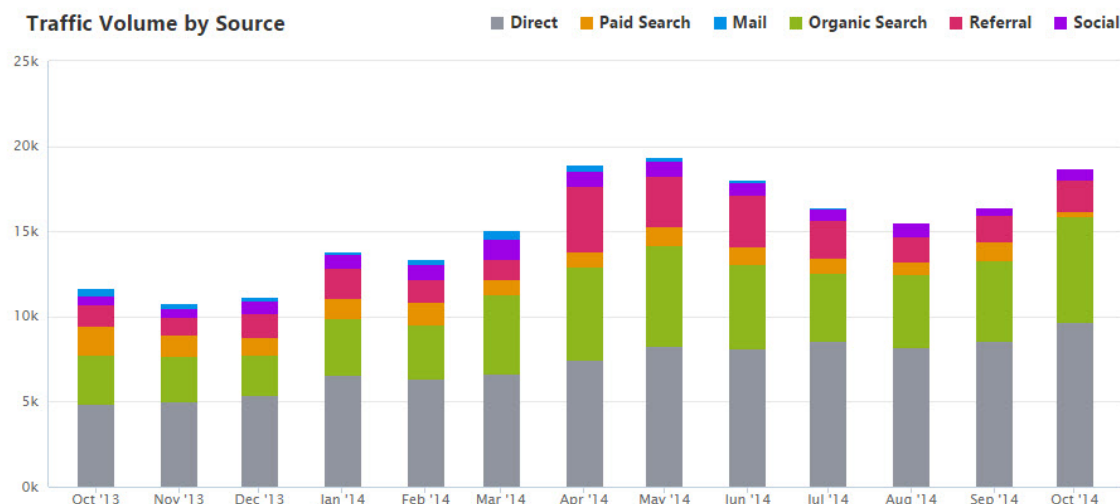


Figura 10 - Número de visitas mensais agrupadas pela sua origem

(fonte: <https://www.rankranger.com/documentation/analytics-traffic-sources>)

As métricas

É uma medida quantitativa dos dados das dimensões. As métricas do *Google Analytics* podem ser fornecidas em percentagens ou números exatos. Nº de Páginas por sessão, Duração média da sessão e tempo em página são exemplos de métricas no *Google Analytics*. Abaixo serão descritos alguns exemplos de métricas disponíveis:

Visitas e visitantes são dois termos que podem facilmente ser confundidos. O número de visitantes (ou visitantes únicos) é a quantidade de pessoas que visitaram determinado website, independentemente da quantidade de vezes que o fizeram. O número de visitas mostra quantas vezes houve acesso ao website, sem levar em consideração a quantidade de pessoas que o acederam. Por exemplo, se o mesmo visitante aceder a um determinado site 3 vezes durante o mesmo dia, o *Google Analytics* só irá contabilizar 1 visitante e 3 visitas. Esta métrica é importante de ser acompanhada pois reflete diariamente a audiência que um site ou um *blog* possui. Por exemplo, um site que mantém sempre estável o número de visitantes e visitas pode mostrar em análise que não está a atrair novos visitantes, ou que as pessoas que estão a aceder ao site são as mesmas.

O número de visualizações das páginas ou as chamadas *Pageviews*, indicam quantas páginas do website foram visitadas, no entanto esta métrica por si só não é específica. Por exemplo, ter 50.000 *Pageviews* e 100 visitantes num *website*, é completamente diferente de ter 50.000 *Pageviews* e 50.000 visitantes. Ambos os casos são extremos e diferentes. Uma métrica que é derivada das visualizações de páginas e que traz informações mais relevantes em relação ao comportamento dos visitantes do *site* são as visitas por página, pois indica-nos o número de visualizações que uma página obtém.

A taxa de abandono mostra a percentagem de pessoas que acedem apenas a uma página do *website* e não continuam a navegação para uma segunda página.

A taxa de saída, por sua vez mostra a percentagem de visitantes que estavam numa determinada página e saíram do *website*, no entanto, ao contrário da taxa de abandono, a taxa de saída não tem em consideração se é, ou não, a primeira página do *website* visualizada por um visitante. Pelo que, cada uma das páginas mais importantes do *website* deve ter a taxa de saída analisada de forma frequente, para poderem ser feitas otimizações.

A percentagem de novas visitas mostra quantos dos visitantes num determinado período de tempo estão a visitar um *site* pela primeira vez. Para analisar esta métrica, é preciso alguns cuidados: uma taxa muito baixa significa que a retenção do *site* pode estar normalizada, porém, pode não estar a atrair novos visitantes, por sua vez uma taxa muito alta significa que o *site* pode estar a atrair muitas pessoas, contudo estas pessoas acabam por não regressar ao *site*. Assim o cruzamento com outras métricas pode ajudar a esclarecer. É ainda importante perceber que a identificação de visitantes novos e dos visitantes frequentes é feita através de um *cookie*, este é guardado no *browser* do visitante na primeira vez que ele acede ao *site*. Quando este visitante se torna frequente, o *Google Analytics* reconhece o *cookie* e sabe que aquele utilizador já esteve ali pelo menos uma vez. O que acontece é que este *cookie* pode ser apagado manualmente, ou desativado, e quando o visitante voltar ao *site*, é contabilizado como um visitante novo. Apesar de não ser algo comum, pode originar uma margem para erro.

Relatórios

No que diz respeito aos relatórios, todos são compostos por dimensões e métricas.

As tabelas na maioria dos relatórios do *Google Analytics* organizam os valores das dimensões em linhas, e as métricas em colunas. Por exemplo, esta tabela mostra uma dimensão (Cidade) e duas métricas (Sessões e Páginas/sessão).

DIMENSÃO	MÉTRICA	MÉTRICA
Cidade	Sessões	Páginas/sessão
São Francisco	5.000	3,74
Berlim	4.000	4,55

Tabela 1 - Exemplo de um relatório que conjuga métricas e dimensões (*Google Analytics*)

2.6 Discussão

Neste capítulo são discutidas as abordagens, métodos e ferramentas utilizados nos diferentes tipos de *Web Mining* e em particular aqueles que suportam a recomendação. Considerando os objetivos do trabalho e o seu enquadramento na vertente de *Web Mining* de navegação, consideram-se mais pertinentes os trabalhos que abordam a recomendação suportada em filtragem colaborativa ou em algoritmos de identificação de regras de associação. Contudo, ambos têm limitações inerentes que foram discutidas e que se prendem com a necessidade de volume de dados não dispersos e no caso das regras da associação ainda também com a

impossibilidade prática de contemplar a ordem dos fluxos de navegação (i.e., da sequência ordenada dos itens constantes na (esquerda) da regra). Adicionalmente qualquer abordagem suportada em algoritmos de *Data Mining* tem inerentemente um enfoque na extração de *features* que caracterizam grupos e/ou segmentos requerendo para tal (volume de) dados sobre muitas, nem sempre acessíveis, variáveis. Em alternativa, poderiam ser considerados algoritmos capazes de modelar dados com menor dimensão/volume, todavia, estes muitas vezes acabam por gerar modelos com fraco poder explicativo real.

As novas alternativas emergentes da tradicional abordagem OLAP focadas nas estatísticas de navegação (e não no utilizador) parecem constituir a alternativa considerando o âmbito do trabalho (Alves & Belo, 2004). Desta forma é possível inquirir os dados persistidos respeitantes à navegação, que ao contrário de alguns modelos obtidos por recurso a técnicas de *Mining*, estarão sempre alinhados com a utilização real atual. Relativamente à questão da *performance*, à medida que estes dados tenham mais dimensões (variáveis) e volume de registos, que é justamente a motivação para a sintetização que os algoritmos de *Mining* podem realizar, esta pode ser ultrapassada considerando restrições de dimensões e período temporal que se pretende em cada solicitação, atendendo às necessidades e contexto, facto que leva a um modelo, e conseqüente implementação, parametrizável como proposto a partir do capítulo seguinte.

Esta é a orientação base das ferramentas *Web Analytics* eleitas, em particular, o *Google Analytics*, adotado como base no desenvolvimento deste trabalho. Estas permitem a recolha e organização de dados assim como as questões relacionadas com a gestão do desempenho acima referida.

Assumidamente, com esta abordagem não se encontram necessariamente padrões ocultos, nem se sintetizam características de grupo de utilizadores e/ou conteúdo, mas realizam-se recomendações com base em padrões realmente existentes tenham esta qualquer ordem de representatividade. Fica assim o foco, de acordo com o âmbito do trabalho, no emparelhamento da navegação do utilizador (no seu contexto, qualquer que este seja) no padrão existente mais “conveniente” em cada momento à medida que se coleciona o seu próprio padrão. Sendo que a “conveniência” neste trabalho se refere à experiência de utilização melhorada face aos objetivos de navegação estabelecidos para o *website*.

3 Recomendação suportada em *Web Analytics*

Um utilizador quando procura determinado *website* pressupõe-se que o conteúdo do mesmo seja do seu interesse, no entanto, a forma como os conteúdos estão dispostos ou a navegação implícita do mesmo pode não estar de acordo com as necessidades ou expectativas dos utilizadores, para isso existem e podem ser definidos *KPI's* (*key performance indicator*) que permitem avaliar e tirar conclusões, relativamente ao caminho que o *website* está a levar e se este é o pretendido. Com base no *Google Analytics*, a ferramenta base deste trabalho, podemos dividir estes *KPI's* em 2 componentes:

As **Dimensões** são essencialmente as categorias dos dados, ou seja, definem características dos dados por categorias identificadas e customizadas, por exemplo algumas das dimensões disponíveis são o tipo de utilizador (novo ou frequente), a origem do utilizador (o *website* de onde o mesmo é proveniente), o navegador utilizado, dados demográficos do utilizador (país,

idioma, localidade), etc. Além das dimensões já disponibilizadas pelo *Google Analytics* é possível também criar dimensões customizadas e alimentá-las com uma lógica própria desenvolvida para o *website*.

As **Métricas** são os dados quantitativos do *website*, ou seja, o número de utilizadores que visitam uma determinada página, o tempo em página, o número de saídas, o número de conversões para um objetivo, o número de sessões, etc.

Com referido no subcapítulo 2.3.3, as dimensões e as métricas são utilizadas em conjunto para fornecer *KPI's* que suportem comparação, pois as dimensões indicam-nos a informação que estamos a tratar e as métricas, os números reais para comparação (Ex: o número de visitas de uma página à data de hoje comparativamente ao número de visitas após um determinado período de tempo).

3.1 Modelo Proposto

O objetivo do modelo proposto passa por encaminhar o utilizador através de um fluxo de páginas, tendo em conta a página em que este se encontra até à página/ páginas que o administrador do *website* tenha definido como objetivo. Em cada página visitada pelo utilizador é-lhe sugerida uma página seguinte que faça parte do fluxo e que o irá orientar até um objetivo, percorrendo o menor caminho possível.

Esta página sugerida é calculada através do histórico de utilizadores que tenham passado pela mesma e atingiram um objetivo. Com este modelo providenciamos ao administrador do *website* um dinamismo que não existe apenas com análise do *Google Analytics*, pois caso os utilizadores não estejam a atingir os objetivos do site, porque estes não se revelam evidentes ou haja uma mudança dos mesmos, implica que sejam tomadas atitudes para inverter a tendência. Esta pode passar por redesenhar o *website* para contemplar a nova tendência.

Com este modelo as recomendações serão sempre feitas tendo em conta os objetivos atuais, ou seja, a navegação do utilizador será sempre orientada para os mesmos e o utilizador não será desviado do caminho pretendido.

Note-se que, com esta abordagem a redefinição do *website* pode ser obviado, mas mesmo a revelar-se imprescindível existe um período temporal até à sua implementação e implantação cujo desalinhamento da navegação do existente com os objetivos atuais é minimizado, pelas recomendações que se reajustam.

Subjacente ao mecanismo de recomendação está a definição inicial pelo administrador do objetivo do *website*, materializando-os no *Google Analytics*. Com esta base, sempre que uma página é visitada por um utilizador a *GoRecommendJS*, que implementa o modelo (a implementação é descrita no capítulo 4), faz um pedido de recomendações (recorrendo à *api* do *Google Analytics*) para propor as páginas para o utilizador visitar seguidamente.

A lógica subjacente ao modelo proposto contempla a verificação se a página para a qual foram pedidas recomendações é já uma página objetivo e é então uma página de abandono natural, caso em que, não serão produzidas recomendações.

Caso contrário, são solicitadas ao *Google Analytics* as páginas seguintes à página atual, ordenadas pelo número de utilizadores por forma descendente (elegendo as 5 primeiras –

configurável) em que nesse caminho tenha ocorrido uma conversão para um objetivo. Esta solicitação ao GA é formulada atendendo a um conjunto de parâmetros (configuráveis) que podem enriquecer a lógica do modelo na elaboração das recomendações. Os parâmetros e o seu significado são apresentados seguidamente:

- (Opcional) Data de Início: baliza o período sobre o qual serão calculadas as recomendações;
- (Opcional) Tipo de dispositivo: define se as recomendações devem ser delimitadas pelo mesmo tipo de dispositivo ou não, o *Google Analytics* classifica os dispositivos em 3 categorias: *desktop*, *tablet* e *mobile*;
- (Opcional) Idioma: define se as recomendações devem ser delimitadas pelo mesmo idioma ou não;
- (Opcional) País: define se as recomendações devem ser delimitadas pelo mesmo país ou não;
- (Opcional) Chave da Dimensão customizada para calcular o tempo em página: esta variável é utilizada para calcular o tempo real em que o utilizador dedica a sua atenção ao *website*, o tempo que a página não está em foco não é contabilizado. Esta chave é obtida ao criar uma dimensão personalizada no *Google Analytics*;
- (Obrigatório) Chave da propriedade: também identificada como código de acompanhamento, é o código único de acompanhamento do *website* por parte do *Google Analytics*, todas as pesquisas serão feitas com base neste código.

3.2 Lógica de Notificação

O modelo desenvolvido contém uma preocupação adicional/complementar à geração das recomendações. Em termos de experiência de utilizador não é negligenciável, a intrusão que poderá causar a apresentação das recomendações ao utilizador.

Assim, foi igualmente desenvolvida uma lógica associada à apresentação/disponibilização das recomendações.

Para tal é solicitado ao *Google Analytics* quais os tempos de abandono da página atual, calculada a média e o seu desvio padrão, e com base nestes feito um cálculo para o tempo de atraso que condiciona quando as recomendações devem ser exibidas.

A implementação desta lógica é feita pela *GoRecommendJS* que quando recebe a resposta por parte do serviço *GoRecommend*, cria um evento com o atraso remetido pelo mesmo. Quando este evento é executado o *widget* da *GoRecommend* entra na página com uma animação de destaque para que suscite algum interesse sem ser demasiado intrusivo, ficando apenas com o ícone no final da animação.

Ao clicar nesse mesmo ícone do *widget* é disponibilizado ao utilizador uma lista com as recomendações existentes, ordenadas pela recomendação que mais converteu até à recomendação que menos converteu. Caso o utilizador já tenha clicado em alguma das recomendações estas descem para o fim da lista mantendo sempre a ordenação pelo maior número de conversões caso tenha clicado em mais que uma.

Além disso o *widget* providencia ao utilizador algumas opções de customização para que seja mais flexível de acordo com as preferências do utilizador.

Botões de + e -, que permitem ao utilizador diminuir o número de elementos da lista até um mínimo de 1 ou aumentar o número de elementos da lista até ao número de recomendações existentes.

Botões de aumentar e diminuir atraso com que a notificação é exibida com um mínimo de atraso de 0 ou com um máximo de atraso correspondente ao valor originalmente calculado (parametrizável em função da média e desvio padrão). Cada clique nos botões aumenta ou reduz em 10% do valor calculado.

O conjunto de parâmetros que permite a flexibilização na configuração da lógica de notificação é apresentado e descrito em baixo:

- (Opcional) Posição: por defeito o elemento é posicionado no canto superior direito, é possível colocar o elemento em qualquer um dos 4 cantos da página;
- (Opcional) Cor de fundo: permite definir a cor de fundo do *widget*;
- (Opcional) Cor de texto: permite definir a cor do texto e ícones;
- (Opcional) Título: par “chave – valor” que permite definir o título do *widget* que aparece no topo da listagem de acordo com o idioma, em que a chave é o código do idioma e o valor o título em questão.

3.3 Síntese do Modelo

Para sintetizar o modelo recorre-se a um diagrama de seqüências desenhado abaixo segundo a linguagem UML (Unified Modeling Language), que permite representar um sistema de forma padronizada, com o intuito de facilitar a compreensão antes da implementação.

O diagrama de seqüências por sua vez representa a seqüência de mensagens entre os participantes, onde o tempo decorre para baixo e mostra o fluxo de contacto entre os participantes. Através deste fluxo sequencial é permitido ao leitor visualizar de forma simples o conjunto de ações subjacente à lógica do modelo quer na geração de recomendações quer na sua disponibilização.

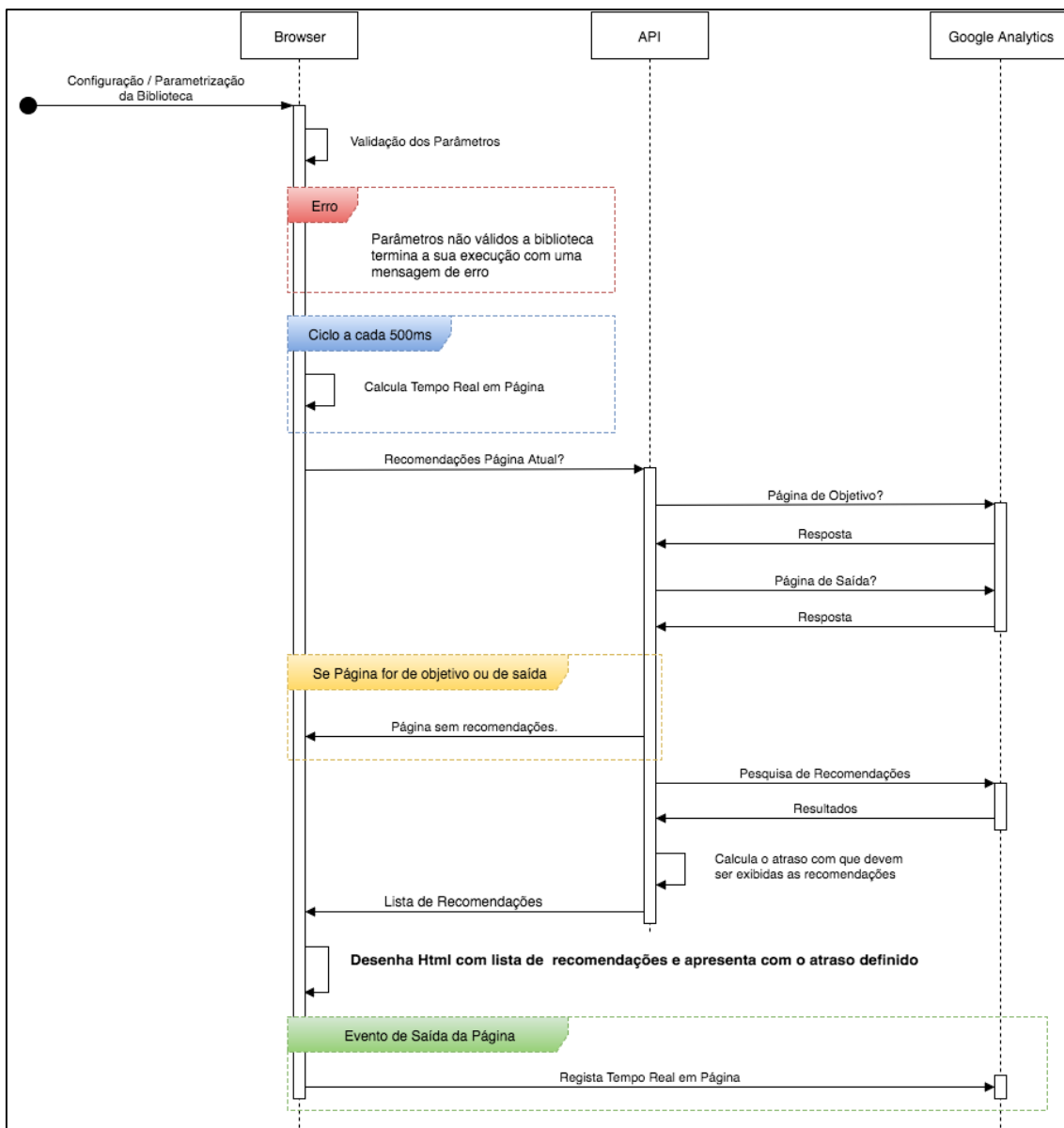


Figura 11 - Síntese do modelo proposto

4 Protótipo Desenvolvido

Tal como descrito anteriormente, o objetivo do modelo passa por encaminhar o utilizador através de um fluxo que o guie até uma página definida como objetivo pelo administrador do *website*, sugerindo a(s) página(s) seguinte(s). Este fluxo é calculado através do histórico de utilizadores que atingiram o objetivo, histórico este que é obtido recorrendo à ferramenta *Google Analytics*, após aferição do fluxo de navegação as recomendações são disponibilizadas ao utilizador através de um *widget* que deverá ser contemplado no *website*.

Antes de explicar a arquitetura da solução é importante ainda esclarecer algumas métricas do *Google Analytics* que no meu entender deviam ser complementadas e tiveram influência no desenho da solução.

O tempo em página é um desses casos, pois o GA calcula o tempo em página através do diferencial de tempo entre a entrada numa página e um clique para outra página do mesmo *website*, ou seja, mesmo que o utilizador esteja focado no conteúdo de uma página e demore alguns minutos na mesma caso não visite outra página o tempo contabilizado é igual a 0, este calculo não é feito no evento de saída da página ou do *browser* porque não é garantido que a notificação assíncrona seja entregue antes do fecho da página/*browser*, além disso existe ainda outra falha neste método de cálculo que passa por não ter uma noção exata do tempo real que o utilizador esteve em página pois este pode ter o *website* aberto numa aba e essa mesma aba estar em segundo plano ou minimizada, ou seja, o utilizador não está efetivamente a consultar o *website* apesar de ter o mesmo aberto. Existem algumas abordagens que passam não só por minimizar esta perda de dados, mas também otimizar os mesmos para que estejam o mais próximos possível da realidade.

Uma abordagem passa por criar eventos com intervalos de tempo e sempre que estes forem atingidos notificar o *Google Analytics* e desta forma é possível enquadrar os tempos em página em intervalos de tempo, ou seja, como exemplo são criados no *Google Analytics* 3 eventos, o intervalo de 0 – 10 segundos, 11 – 30 segundos e mais de 30 segundos, após a criação dos eventos é adicionada a devida lógica ao *website* que notifica o *Google Analytics* sempre que um intervalo é atingido.

Outra abordagem passa por utilizar a função *javascript "sendBeacon"*, que permite efetuar pedidos de forma assíncrona em segundo plano, quando o *browser* é fechado e desta forma conseguimos notificar o *Google Analytics*, a limitação desta função é a sua compatibilidade com os *browsers*, neste momento só é suportada em *Google Chrome e Firefox*, por isso deve ser complementada utilizando o evento *onBeforeUnload* chamado imediatamente antes do fecho do *browser* e efetuar a notificação do evento ao *Google Analytics* de forma síncrona.

A abordagem utilizada para o desenho da solução foi a última abordagem por se tratar de uma abordagem mais otimizada que permite chegar a valores reais e não a intervalos de valores, contemplando também uma maior cobertura de compatibilidades de forma a que a perda nos resultados seja mínima. Além disso o modelo terá também uma lógica de eventos em que só contabiliza o tempo em página, o tempo em que o utilizador está efetivamente a consultar a página e não o tempo em que esta se encontra em *background*.

4.1 Arquitetura Geral

Para o desenvolvimento da solução foi projetada uma arquitetura que se divide em 3 componentes:

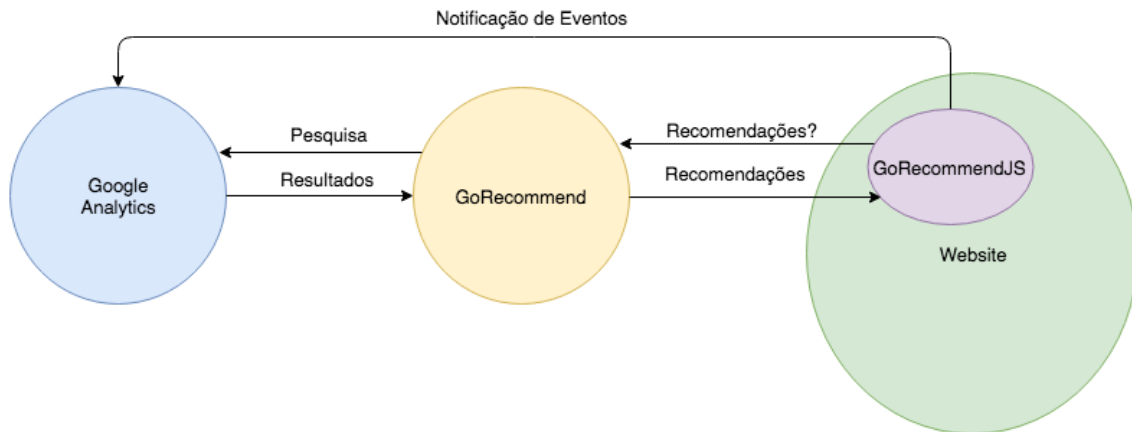


Figura 12 - Arquitetura do modelo proposto

GoRecommendJS:

- Objetivo: Biblioteca *Javascript* embutida no *website*, responsável por notificar o *Google Analytics* dos eventos que ocorrem no mesmo, em paralelo é também a responsável pela comunicação com o serviço *GoRecommend* desenvolvido, pedindo recomendações para o *website*.
- Especificações:
 - *Javascript*
 - Biblioteca *jQuery* (versão 1.9.1) (Dependência)

GoRecommend:

- Objetivo: Serviço responsável por toda a lógica de consulta no *Google Analytics* e formatação dos dados para construção das recomendações que são disponibilizadas através de uma *api*.
- Especificações:
 - *Apache Webserver*
 - *Php 5.5.23*
 - *Symfony 2.7.12*
 - *REST*

Google Analytics: repositório de dados do *website*, coleciona todas as notificações e eventos reportados pela *GoRecommendJS*, consultados mais tarde pelo serviço *GoRecommend*.

4.2 GoRecommend

Desenvolvida na linguagem de programação *PHP*, com recurso à *Framework symfony*.

Descrição do serviço *GoRecommend*:

Endpoint: `{{base_url}}/recommendation`

Método HTTP: GET

Parametrização Obrigatória:

- *path*: o caminho relativo da página sobre o qual será feita a recomendação ex.: `"/contact"`.
- *start_date*: a data de início por onde se devem começar a procurar resultados, por defeito inicia em 365 dias antes da data atual (*365daysAgo*).
- *language_code*: Código ISO 639 do idioma, parâmetro utilizado caso o *language_filter* seja ativado.
- *device_width*: A largura em pixels do dispositivo, este valor será enquadrado de acordo com a lógica do GA numa categoria (`< 600px = mobile`, `>= 600px e <= 1024 = tablet`, `>1024 = desktop`), este parâmetro é utilizado caso o *device_filter* seja ativado.
- *web_property_id*: também identificada como código de acompanhamento, é o código único de acompanhamento do *website* por parte do GA, todas as consultas serão feitas com base neste código.

Parametrização Opcional:

- *real_time_on_page_key*: nome da dimensão customizada do GA para consultar o valor real de tempo em página, caso não seja definido será utilizada a variável por defeito do *Google Analytics "timeOnPage"*.
- *country_filter*: define se os resultados devem ser delimitados pelo mesmo país, por defeito o valor é 0 (não filtrar), caso se pretenda filtrar atribuir o valor 1. Caso este parâmetro seja ativado será utilizada a biblioteca *Maxmind GeoIP* para calcular o país de acordo com o ip do utilizador.
- *device_filter*: define se os resultados devem ser delimitados pelo mesmo tipo de dispositivo, por defeito o valor é 0 (não filtrar), caso se pretenda filtrar atribuir o valor 1.
- *language_filter*: define se os resultados devem ser delimitados pelo mesmo idioma, por defeito o valor é 0 (não filtrar), caso se pretenda filtrar atribuir o valor 1.

Lógica de Recomendação:

Sempre que haja um pedido ao serviço *GoRecommend* este verifica se a página atual é uma página de abandono e não é uma página de objetivo, caso não satisfaça uma destas condições não haverá recomendação sendo enviado uma lista vazia. Caso satisfaça, o *GoRecommend* faz um pedido ao *Google Analytics* com os seguintes parâmetros:

- Dimensões:
 - *ga:previousPagePath* – Página anterior
 - *ga:pagePath* – Pagina atual
 - *ga:pageTitle* – Título da página atual
- Métricas:
 - *ga:users* – Número de utilizadores
- Filtros:
 - *ga:previousPagePath* igual a Página atual
 - *ga:pagePath* diferente da Página atual (Não há recomendação para a mesma página)

Ou seja, são solicitados ao *Google Analytics* todas as páginas anteriores que correspondam à página atual, páginas atuais que não correspondam à página atual (não recomendamos para a mesma página), títulos das páginas atuais, ordenados pelo número de utilizadores por forma descendente em que nesse fluxo tenha ocorrido uma conversão para o objetivo.

O serviço *GoRecommend* guarda os valores obtidos pelo *Google Analytics* diariamente em *cache*, ou seja, para uma determinada parametrização só é efetuado um pedido por dia mesmo que hajam múltiplos carregamentos diferentes de páginas do *website*.

Os dados são formatados para serem enviados e de seguida o *GoRecommend* pergunta ao *Google Analytics* quais os tempos de abandono da página atual, calcula a média e o seu desvio padrão e obtém o tempo de atraso sobre o qual as recomendações devem ser exibidas com base na seguinte fórmula:

$$\text{abs}(\text{MEDIA} - 0.6 * \text{DESVIO_PADRAO})$$

Valor absoluto da subtração da média por 60% do valor do desvio padrão

Finalmente é retornada a lista e o *delay* com que esta deve ser apresentada. Abaixo um exemplo de resposta do serviço.

Exemplo de Resposta do serviço *GoRecommend*:

```
{
  "suggestions": [
    {
      "url": "/example1",
      "title": "Exemplo1",
      "weight": "43"
    },
    {
      "url": "/example2",
      "title": "Exemplo2",
      "weight": "21"
    },
  ],
  "delay": 3.634
}
```

Figura 13 - Exemplo de resposta do serviço *GoRecommend*

4.3 *GoRecommendJS*

GoRecommendJS é uma biblioteca desenvolvida com base na linguagem de programação *Javascript*, esta biblioteca integra com o serviço *GoRecommend* que após configuração por parte do administrador do *website* com base num conjunto de parâmetros que serão descritos abaixo disponibiliza uma camada representacional com base num *widget* que indica as recomendações disponíveis para uma determinada página do *website* e alguns botões de configuração que permitem uma maior flexibilidade do mesmo perante o utilizador.

Localização: `{{base_url}}/js/go-recommend.js`

Parametrização Obrigatória:

- *web_property_id*: ver parâmetro *web_property_id* do ponto 4.2

Parametrização Opcional:

- *start_date*: ver parâmetro *start_date* do ponto 4.2
- *language_filter*: ver parâmetro *language_filter* do ponto 4.2
- *device_filter*: ver parâmetro *device_filter* do ponto 4.2
- *country_filter*: ver parâmetro *country_filter* do ponto 4.2
- *real_time_on_page_key*: ver parâmetro *real_time_on_page_key* do ponto 4.2
- *position_x*: posição do elemento no eixo do x, por defeito o elemento está encostado a direita, pode ser configurado para a esquerda com o valor *“left”*.

- *position_y*: posição do elemento no eixo do y, por defeito o elemento está encostado ao topo, pode ser configurado para o fim da página com o valor “*bottom*”.
- *background_color*: cor de fundo do *widget*, pode ser configurado com o código hexadecimal ou rgb de uma cor.
- *color*: cor do texto do *widget*, pode ser configurado com o código hexadecimal ou rgb de uma cor.
- *titles*: lista composta por pares chave-valor, em que a chave é o código ISO 639 do idioma e o valor o título do *widget* correspondente, caso a chave não seja fornecida, o título será assumido por defeito quando não haja um idioma correspondente, por exemplo: `[{value: “Recommendations”}, {key: 'pt-pt', value: “Recomendações”}]`, para os *navegadores* em que o idioma é pt-pt será exibido “Recomendações”, para todos os outros “Recommendations”, caso este atributo não seja fornecido será assumido sempre o valor “Recommendations”.

Lógica de Recomendação:

Sempre que a página é carregada, a *GoRecommendJS*, carrega todas as suas dependências e faz um pedido ao serviço *GoRecommend* por recomendações, caso a lista não venha vazia é criado um evento com o atraso definido pelo mesmo, quando este evento é executado o *widget* entra na página com uma animação.

Lógica de Configuração:

O *widget* providencia ao utilizador algumas opções de customização para que seja mais flexível de acordo com as opções do utilizador.

Botões de + e -, que permitem ao utilizador diminuir o número de elementos da lista até um mínimo de 1 ou aumentar o número de elementos da lista até ao número de recomendações disponíveis.

Botões de aumentar e diminuir atraso com que a notificação é exibida com um mínimo de atraso de 0, ou com um máximo de atraso correspondente ao valor calculado pelo *GoRecommend*. Cada clique nos botões aumenta/reduz em 10% do valor calculado pelo *GoRecommend*.

Lógica de Cálculo de Tempo em Página:

Quando a página é carregada, a *GoRecommendJS* começa a contar o tempo em página, a cada 500ms ocorre um evento de verificação do estado da página, caso esta esteja em segundo plano o contador para, no entanto os eventos continuam a ocorrer, caso o utilizador volte à página o mesmo evento ocorre e o contador continua a sua contagem. Quando a página for fechada quer esteja em segundo plano ou não, é enviado para o GA o valor real em página através dos métodos *sendBeacon* (caso seja suportado), ou através da combinação do método *beforeunload* com uma chamada síncrona ao GA, desta forma conseguimos garantir uma maior probabilidade de sucesso no envio dos dados.

Exemplo de utilização da biblioteca *GoRecommendJS*:

Importar o seguinte código no final da página:

```
<script src="//{{base_url}}/js/go-recommend.js"></script>

<script>
GoRecommend.init(
    {
        real_time_on_page_key:'ga:dimensionX',
        web_property_id:'UA-XXXXXXX-X',
        device_filter: 1,
        language_filter: 1,
        country_filter: 1,
        background_color: '#FFF',
        color: '#000'
    }
);
</script>
```

Figura 14 - Exemplo de implementação da biblioteca *GoRecommendJS*

4.4 Descrição Funcional de Aplicação

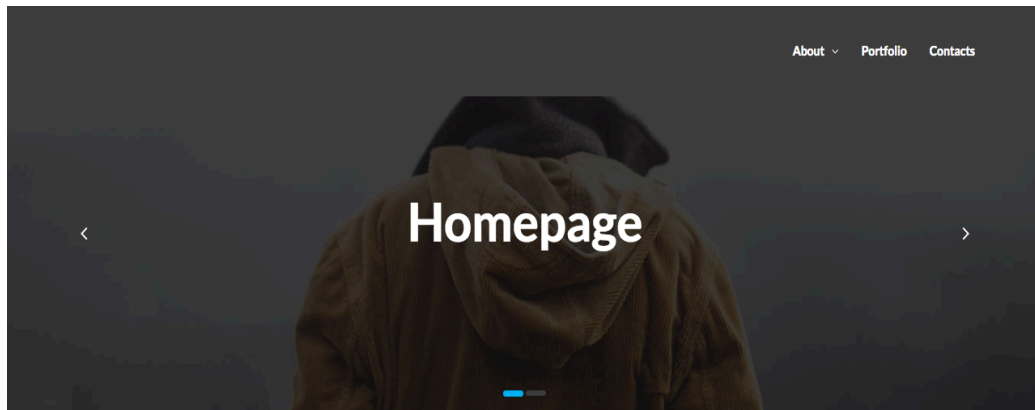
Para a demonstração da aplicação foram utilizados dados reais, contudo por questões de confidencialidade estes dados foram mascarados e incorporados num *layout* gratuito com a mesma estrutura de navegação que o original. Ao longo desta seção tem-se como objetivo apresentar a experiência do utilizador relativamente ao *widget* de recomendações apresentado.

A estrutura para a demonstração contempla as seguintes páginas: *Homepage*, 'Sobre nós', 'Contactos', 'Portfolio' e 'Forma'. Será também contemplada a página 'Portfolio' configurada pelo administrador do *website* no *Google Analytics* como página de objetivo.

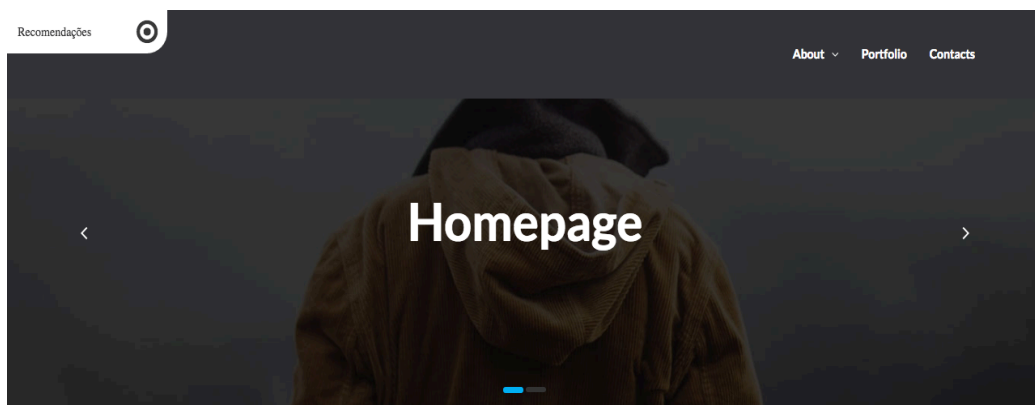
Inicialmente é apresentado o conjunto de animações que dão entrada ao *widget* na página, depois a dinâmica das recomendações nas diferentes páginas e finalmente o impacto da interação do utilizador com as ações de configuração disponíveis no *widget*.

Segue-se a apresentação dos *screenshots* e respetivas descrições:

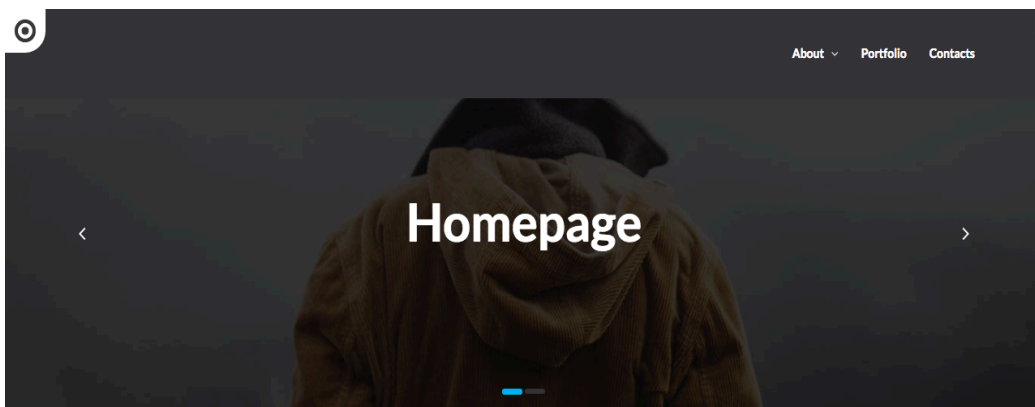
1. Página inicial antes do carregamento das recomendações.



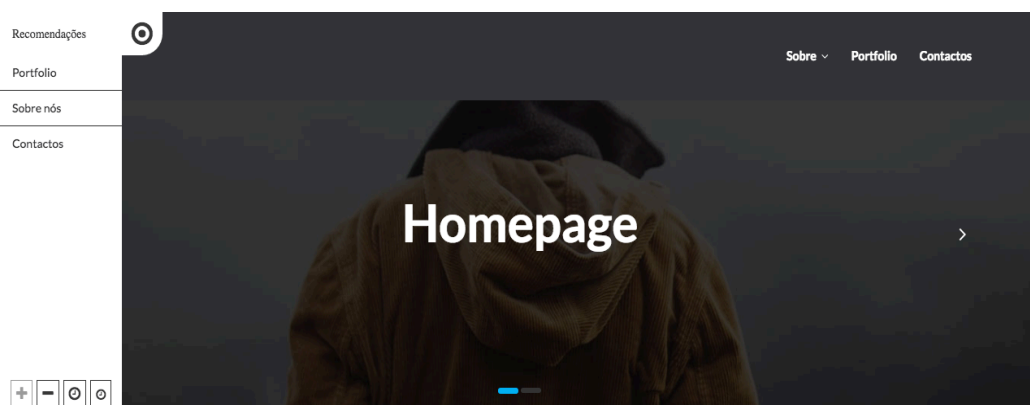
2. A apresentação do *widget* aumenta gradualmente a etiqueta até à total visibilidade do título e posterior diminuição até onde fica apenas o ícone. Para este caso o administrador do *website* configurou a *GoRecommendJS* com as cores de fundo branco, de letra a preto e posicionou-o no canto superior esquerdo segundo a parametrização descrita no ponto 4.3.



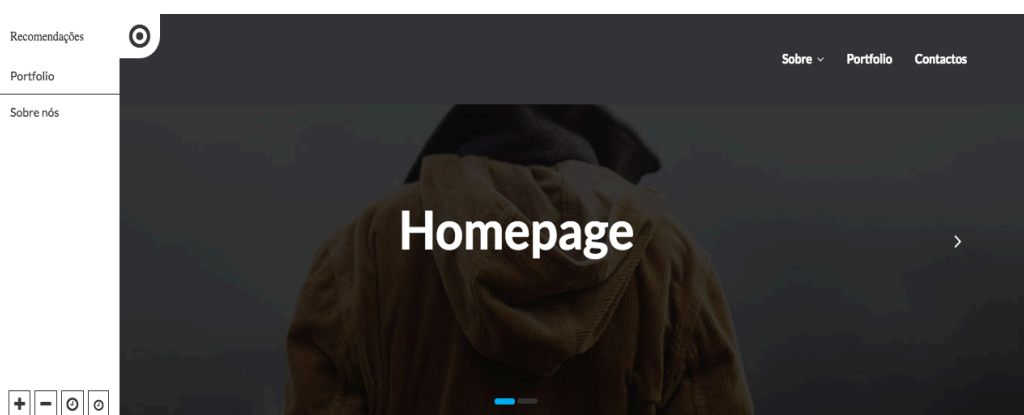
3. Estado final do *widget* após término da animação de entrada



4. As recomendações, onde o fluxo que corresponde em mais conversões a partir da homepage é para a página 'portfolio', sendo esta a página objetivo do *website* encontra-se na 1ª posição. De seguida as páginas com mais conversões são respetivamente a página 'Sobre nós' e 'Contactos'. Esta seria a recomendação da *GoRecommend* para a *Homepage*. Tal como podemos verificar ainda a página 'Sobre nós' não se encontra disponível à primeira vista no menu, um dos casos em que o *widget* de recomendação é uma mais valia.



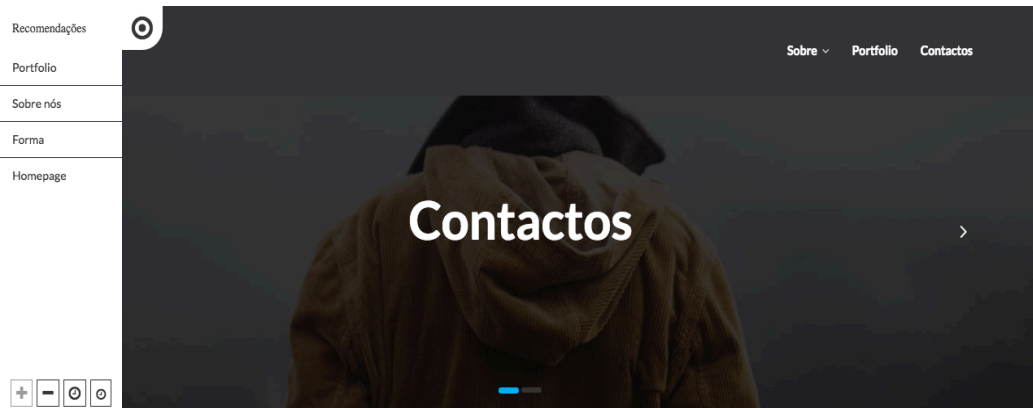
5. O mesmo caso da recomendação anterior, mas desta vez o utilizador utilizou o botão na configuração do *widget* indicando que pretende receber apenas as 2 primeiras recomendações disponíveis e não mais que isso. Passou também a ter o botão ativo, podendo assim repor as linhas disponíveis até ao número disponível de recomendações. Para os exemplos seguintes será restaurado o número de linhas para o número de recomendações disponíveis.




6. Neste exemplo é demonstrado a alteração nas recomendações num contexto diferente, ou seja, navega-se para a página 'Contactos' e é possível verificar que a página 'Portfolio' (Objetivo) consta na primeira posição das recomendações, sendo a que converte os utilizadores. A página 'Sobre nós' mantém-se na segunda posição, ou seja, as pessoas que mais convertem em segundo lugar visitam a página 'Sobre nós' posteriormente à página de 'Contactos'. É possível verificar também que a página 'Forma' é uma

recomendação que não constava na *Homepage*, o que pode levar a duas conclusões: 1) não existem visitas à página forma a partir da *Homepage*, 2) apesar de existirem visitas à página forma a partir da *Homepage* não existem conversões em que esta sequência ocorra no fluxo.

Por ultimo deixa-se de ter a página 'Contactos' como recomendação, pois é a própria e temos a *Homepage* como recomendação pois efetivamente existirão conversões onde o fluxo de navegação contenha a passagem pela *Homepage* com origem na página 'Contactos'.



Apenas não foi possível demonstrar as ações dos botões , que permitem aumentar ou diminuir atraso com que a notificação é apresentada pois trata-se de lógica de eventos não sendo possível demonstrar através de imagens, no entanto a lógica associada a estes botões está descrita no ponto 4.3.

5 Análise e Discussão dos Dados Obtidos

Nesta secção, num primeiro ponto é identificada a abordagem adotada na avaliação do trabalho, assim como, os resultados obtidos na mesma.

Num segundo ponto são discutidos os resultados e as limitações identificadas.

5.1 Abordagem e Resultados

A abordagem de avaliação adotada tem por base as seguintes premissas, abaixo descritas:

Tal como descrito em 3.1, sabem-se que o modelo proposto passa pela recomendação ao utilizador da página seguinte que conste de um fluxo que o guie para uma página objetivo. Assim sendo é avaliado o protótipo analisando a tendência das recomendações para o objetivo atual, ou seja, como a recomendação é calculada em tempo real a pesquisa é executada de forma dinâmica contemplando os objetivos que o administrador definiu atualmente para o *website*. Nesta abordagem são considerados 2 cenários, o primeiro em que o administrador do *website* substitui um objetivo por outro e um segundo cenário em que o administrador complementa o objetivo já adicionado anteriormente com um novo objetivo, ficando assim com 2 ou mais objetivos.

Pretende-se com esta abordagem constatar a alteração das recomendações tendo em conta as alterações que o administrador efetua nos objetivos ao longo do tempo.

Tal como referido na apresentação do protótipo foram utilizados dados reais da navegação de um *website* que, contudo, não esteve exposto ao protótipo.

Para cada um dos casos descrito será exibida uma tabela com 3 colunas, onde a 1ª coluna representa a página em que o utilizador se encontra, a 2ª coluna representa a recomendação e a última coluna representa o número de utilizadores que converteram para objetivo com esta recomendação. Os dados serão agrupados pela página em que o utilizador se encontra e ordenadas de forma descendente pelo número de utilizadores que converteram o objetivo, regras contempladas também para a recomendação do modelo.

5.1.1 1º Caso – Alteração do objetivo

No primeiro caso foram recolhidos os dados quando o objetivo era visitar a página '/contacts'. É possível verificar que a maioria das conversões ocorre diretamente para a página objetivo, ou seja, a 1ª recomendação feita seria para a própria página objetivo.

Página	Recomendação	Nº de Conversões
/	/contacts	1030
/	/forma	60
/	/portfolio	60
/	/sobrenos	60
/forma	/contacts	70
/forma	/portfolio	40
/forma	/	30
/forma	/sobrenos	20
/portfolio	/contacts	170
/portfolio	/sobrenos	110
/portfolio	/	20
/portfolio	/forma	10
/sobrenos	/contacts	120
/sobrenos	/portfolio	60
/sobrenos	/	30
/sobrenos	/forma	30
/sobrenos	/contacts	10

Tabela 2 - Recomendações para a página objetivo '/contacts'

60 dias após a alteração do objetivo '/contacts' para '/portfolio' foi feita uma nova recolha de dados verificando que a 1ª recomendação passaria a ser a página '/portfolio', ou seja, nota-se uma tendência clara das recomendações para o novo objetivo.

Página	Recomendação	Nº de Conversões
/	/portfolio	1920
/	/contacts	450
/	/sobrenos	170
/	/forma	20
/contacts	/portfolio	80
/contacts	/	70
/contacts	/sobrenos	60
/contacts	/forma	10
/forma	/portfolio	80
/forma	/	10
/forma	/sobrenos	10
/portfolio	/contacts	160
/portfolio	/sobrenos	120
/portfolio	/	90
/sobrenos	/portfolio	100
/sobrenos	/forma	70
/sobrenos	/	60
/sobrenos	/contacts	50

Tabela 3 - Recomendações para a página objetivo '/portfolio'

5.1.2 2º Caso – Adição de um novo objetivo

No segundo caso foram recolhidos os dados quando o objetivo era visitar a página '/pt/contactos'. É possível verificar aqui mais uma vez que o topo das recomendações ocorre diretamente para a página objetivo, no entanto e porque pode fazer sentido contextualizar o utilizador antes de o encaminhar para a página objetivo é possível verificar, ainda que com uma diferença substancialmente menor alguns utilizadores posteriormente à página inicial continuarem pela página '/pt/onde-estamos' e não na página '/pt/contactos'.

Página	Recomendação	Nº de Conversões
/pt/	/pt/contactos	560
/pt/	/pt/onde-estamos	100
/pt/	/pt/empresa/quem-somos	60
/pt/clientes	/pt/onde-estamos	60
/pt/empresa/a-nossa-frota	/pt/empresa/projectos-de-referencia	60
/pt/empresa/desempenho-da-empresa	/pt/empresa/a-nossa-frota	40
/pt/empresa/equipa	/pt/empresa/quem-somos	30

/pt/empresa/projectos-de-referencia	/pt/empresa/equipa	30
/pt/empresa/quem-somos	/pt/contactos	230
/pt/empresa/quem-somos	/pt/empresa/a-nossa-frota	40
/pt/logistica	/pt/frota	40
/pt/onde-estamos	/pt/contactos	110
/pt/onde-estamos	/pt/	40
/pt/onde-estamos	/pt/servicos	40
/pt/servicos	/pt/contactos	170
/pt/servicos	/pt/logistica	70
/pt/servicos	/pt/onde-estamos	40

Tabela 4 - Recomendações para a página objetivo '/pt/contactos'

60 dias após a criação de um novo objetivo '/pt/clientes' em complemento ao objetivo '/pt/contactos', ficando assim com 2 objetivos foi feita uma nova recolha de dados verificando mais uma vez a tendência das recomendações para os objetivos '/pt/clientes' ainda que em menor escala que '/pt/contactos', pois anteriormente o objetivo '/pt/clientes' não constava sequer nas recomendações e agora não só o próprio objetivo tem uma relevância muito maior para as recomendações assim como as páginas relacionadas com o mesmo.

Página	Recomendação	Nº de Conversões
'/pt/	'/pt/contactos	470
'/pt/	'/pt/clientes	240
'/pt/	'/pt/onde-estamos	120
'/pt/	'/pt/servicos	60
'/pt/	'/pt/logistica	40
'/pt/clientes	'/pt/contactos	60
'/pt/clientes	/pt/onde-estamos	40
'/pt/clientes	'/pt/empresa/quem-somos	40
'/pt/contactos	/pt/clientes	140
'/pt/contactos	'/pt/empresa/quem-somos	110
'/pt/contactos	/pt/	90
'/pt/contactos	'/pt/servicos	60
'/pt/empresa/a-nossa-frota	'/pt/empresa/projectos-de-referencia	70
'/pt/empresa/equipa	'/pt/empresa/quem-somos	70
/pt/empresa/projectos-de-referencia	'/pt/empresa/equipa	60
'/pt/empresa/quem-somos	'/pt/contactos	70
'/pt/empresa/quem-somos	/pt/clientes	60
'/pt/empresa/quem-somos	'/pt/empresa/equipa	60
'/pt/empresa/quem-somos	'/pt/empresa/quem-somos	60

'/pt/empresa/quem-somos	'/pt/empresa/desempenho-da-empresa	50
'/pt/empresa/quem-somos	/pt/	40
'/pt/frota	'/pt/clientes	90
'/pt/frota	'/pt/onde-estamos	40
'/pt/logistica	'/pt/frota	70
'/pt/logistica	'/pt/empresa/quem-somos	40
'/pt/logistica	'/pt/servicos	40
'/pt/onde-estamos	'/pt/contactos	140
'/pt/onde-estamos	'/pt/clientes	90
'/pt/onde-estamos	'/pt/empresa/quem-somos	50
'/pt/onde-estamos	'/pt/servicos	50
'/pt/servicos	'/pt/logistica	100
'/pt/servicos	'/pt/contactos	80
'/pt/servicos	/pt/clientes	60
'/pt/servicos	'/pt/onde-estamos	50
'/pt/servicos	'/pt/empresa/quem-somos	40

Tabela 5 - Recomendações para as páginas objetivo '/pt/contactos' e '/pt/clientes'

5.2 Discussão e Análise de Limitações

Tratando-se de um modelo cujo protótipo não foi aplicado numa utilização prática que permitiria avaliar o seu impacto efetivo nas recomendações obrigou à abordagem de avaliação descrita no ponto anterior, não sendo possível identificar a interferência causada ao utilizador e eventual resistência ou adoção à utilização do *widget* de recomendações. Também não foi possível testar as preocupações no âmbito da experiência de utilização contempladas na parametrização do modelo tais como: atraso na notificação, número de recomendações, posição do *widget*, etc.

Por último, como não foi possível aplicar o protótipo e desta forma avaliar a sua eficácia não existe termo de comparação com outros sistemas de recomendação, contudo, contacta-se da avaliação expressa na secção anterior que a implementação segue eficazmente os princípios lógicos subjacentes ao modelo proposto. O desempenho do protótipo não se mostra uma limitação, tendo a avaliação reportada tido por base ordem de milhares de registos, cujo tempo de resposta nas consultas ao *Google Analytics* é insignificante.

6 Conclusões e Trabalho Futuro

6.1 Conclusões

A abordagem do protótipo desenvolvido distinguiu-se das abordagens de *Data Mining* em particular as regras de associação por: rever o modelo em tempo real, ou seja, como este modelo tem por base a estatística através de parâmetros previamente definidos e contextualizados para pesquisa por recomendações com base na navegação dos utilizadores que converteram em objetivos é possível fazer recomendação constantemente atualizada, ao contrário das regras de associação de *Data Mining* que para recomendar em *real time* têm que refazer constantemente o modelo de dados, o que pode implicar grandes perdas na performance, caso contrário utiliza modelo desatualizado. Outra das vantagens do modelo proposto relativamente às regras de associação é o facto de este contemplar a ordem da sequência dos dados, pois para recomendação, a ordem em que as páginas são visitadas é de extrema importância pois seguem um contexto, nas regras de associação de *Data Mining* não é possível contemplar a sequência dado o inflacionar da ordem de complexidade da execução do algoritmo.

Assim sendo as novas alternativas focadas nas estatísticas de navegação (e não no utilizador) parecem constituir uma alternativa considerando ser possível inquirir os dados persistidos respeitantes à navegação, que ao contrário dos modelos de *Mining*, estarão sempre alinhados com a utilização real atual. Relativamente à questão da performance, à medida que estes dados tenham mais dimensões (variáveis) e volume de registos, que é justamente a motivação para a sintetização que os algoritmos de *Mining* podem realizar, pode ser ultrapassada considerando as dimensões, restrições e período que se pretende em cada solicitação.

Assumidamente, não se encontram padrões ocultos, nem se sintetizam características de grupo de utilizadores e/ou conteúdo, mas realizam-se recomendações com base em padrões realmente existentes tenham estes, qualquer ordem de representatividade. Considerando o foco no emparelhamento da navegação do utilizador, qualquer que seja o seu contexto no padrão existente, é providenciada uma experiência de utilização melhorada face aos objetivos de navegação estabelecidos para o *website* à medida que os padrões são identificados.

Esta dissertação assumiu como objetivo o desenvolvimento de uma ferramenta de recomendação de navegação incorporável em qualquer *website*, que se materializou através da construção de uma biblioteca *Javascript* parametrizável que através de consulta a um serviço *server side* produzido em paralelo e com base na interação com a ferramenta *Google Analytics* possa oferecer aos seus utilizadores recomendações de páginas para um determinado *website*. Estas recomendações são obtidas com base em histórico de utilizadores que frequentaram as mesmas páginas e visitaram as páginas objetivo definidas pelos administradores do *website*. O desenvolvimento da biblioteca e do serviço além de gratificantes pelo facto de considerar uma ferramenta orientadora e de bastante utilidade, permitiram consolidar o exercício de articulação do desenvolvimento com as linguagens *Css*, *Javascript* e *Php*, assim como a integração com serviços *Web* como era o caso do *Google Analytics*.

Em suma, considero este desenvolvimento um contributo para a identificação da importância de ferramentas de *Web Analytics* e utilizar o seu enorme potencial como guia para otimizar a experiência dos utilizadores. Dada a importância do tema considera-se que muito há ainda que percorrer nesta área sendo, portanto, um campo fértil de trabalho para outros investigadores.

6.2 Trabalho Futuro

A API de pesquisa do *Google Analytics* restringe a quantidade de dados que são retornados (*sampling*) quando uma consulta inclui um grande número de variáveis de pesquisa ou uma grande quantidade de dados (mais de 250 000), ou seja, se os dados dentro de um determinado intervalo excederem um determinado número de resultados os mesmos serão agregados e selecionados um subconjunto, chamados de amostra.

Para trabalho futuro considero relevante após utilização prática do protótipo, avaliar o impacto do *sampling* na precisão das recomendações em *websites* de maior consulta e caso necessário tomar medidas para prevenir/diminuir o impacto desta restrição, tais como: otimização/ alteração das pesquisas, aumentar parametrização da biblioteca apesar de uma das medidas já pensadas no desenvolvimento deste protótipo ter sido exatamente a possibilidade de configurar a cobertura temporal das pesquisas, permitindo manter atual a recomendação.

Outro ponto que considero também de extrema importância é o de desenvolver no futuro, ao nível da experiência de utilização, avaliar o grau de resistência ou adaptação dos utilizadores ao *widget* de recomendação, identificar se o seu objetivo está a ser compreendido e o seu *layout* aceite, assim como as suas funcionalidades compreendidas.

Referências

- Alves, R., & Belo, O. (2004). Mining Clickstream-Based Data Cubes. *International Conference on Enterprise Information Systems*.
- Alves, R., Belo, O., & Costa, F. (2007). Mining Clickstream-Based Data Cubes. In *11th Database Engineering and Applications Symposium* (pp. 120–128). IEEE.
- Baeza-Yates, R., Ribeiro-Neto, B., & others. (1999). *Modern information retrieval*. ACM press New York.
- Berendt, B., Mobasher, B., & Spiliopoulou, M. (2001). Measuring the accuracy of sessionizers for web usage analysis. *Proceedings of the Workshop on Web Mining at the 1st SIAM International Conference on Data Mining*, 7–14.
- Chiarandini, L. (2014). *Characterizing and Modeling Web Sessions with Applications*. PhD Thesis. Universitat Pompeu Fabra.
- Cho, Y. H., Kim, J. K., & Kim, S. H. (2002). A personalized recommender system based on web usage mining and decision tree induction. *Expert Syst. Appl.*, 23(3), 329–342.
- Dinucă, C. E. (2012). An application for clickstream analysis, *International Journal of computer and Communications* 6(1), 68-75.
- Halfaker, A., Keyes, O., Kluver, D., Nguyen, T., Shores, K., & Uduwage, A. (2015). User Session Identification Based on Strong Regularities in Inter-activity Time. *Proceedings of International World Wide Web Conference*, 410–418.
- Hassan, M. T., & Karim, A. (2011). Impact of behavior clustering on web surfer behavior prediction. *Journal of Information Science and Engineering*, 27(6), 1855–1870.
- Himanshu Sharma, H.S. (2016). Complete Guide to Dimensions and Metrics in Google Analytics. Disponível em: 9, Setembro, 2016, em: <https://www.optimizesmart.com/complete-guide-to-dimensions-and-metrics-in-google-analytics>
- Kasliwal, A. D., & Katkar, G. S. (2015). Web Usage mining for Predicting User Access Behaviour, *International Journal of Computer Science and Information Technologies* 6(1), 201–204.
- Kaushik, A. (2008). Web Analytics 2.0. The art of online accountability & science of customer centricity. *Redline Heidelberg*, 3(1), 1–14.
- Kimball, R., & Ross, M. (2015). Dimensional Modeling Tasks and Responsibilities. In *The Kimball Group Reader* (pp. 233–266). John Wiley & Sons, Inc.
- Li, C. (2016). When does web-based personalization really work? The distinction between actual personalization and perceived personalization. *Computers in Human Behavior*, 54, 25–33.
- Liu, B., Mobasher, B., & Nasraoui, O. (2007). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer. ISBN: 978-3-540-37881-5.
- Mehtaa, P., Parekh, B., Modi, K., & Solanki, P. (2012). Web Personalization Using Web Mining: Concept and Research Issue. *International Journal of Information and Education Technology*, 2(5), 510.

- Nadi, S. (2011). A Hybrid Recommender System for Dynamic Web Users. *International Journal Multimedia and Image Processing* 1(1), 3–8.
- Nakatani, K., & Chuang, T. (2011). A web analytics tool selection method: An analytical hierarchy process approach. *Internet Research*, 21(2), 171–186.
- Parra, D., & Brusilovsky, P. (2015). User-controllable personalization: A case study with SetFusion. *International Journal of Human Computer Studies*, 78, 43–67.
- Phippen, A., Sheppard, L., & Furnell, S. (2004). A practical evaluation of Web analytics. *Internet Research*, 14(4), 284–293.
- Rick Allen, R.A. (2012). Web Analytics: What Is It Good For?. Disponível em: 9, Setembro, 2016, em: <http://meetcontent.com/blog/web-analytics-what-is-it-good-for>
- Riecken, D. (2000). Introduction: Personalized Views of Personalization. *Communications of ACM*, 43(8), 26–28.
- Rita Frade, R.F. (2016). Web Analytics: As melhores ferramentas de web analytics. Disponível em: 9, Setembro, 2016, em: <http://comteudo.webtexto.pt/melhores-ferramentas-analytics>
- Slimani, T., & Lazzez, A. (2014). Efficient Analysis of Pattern and Association Rule Mining Approaches. *International Journal of Information Technology and Computer Science*, 6(3), 70–81.
- Srivastava, J., Cooley, R., Deshpande, M., & Tan, P.-N. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *SIGKDD* 1(2), 12–23.
- Stephane Hamel, S.H. (2012). Big Data – What It Means For The Digital Analyst. Disponível em: 25, Julho, 2016 em: <http://online-behavior.com/analytics/big-data>.
- Suguna, R., & Sharmila, D. (2013). An Efficient Web Recommendation System using Collaborative Filtering and Pattern Discovery Algorithms. *International Journal of Computer Applications*, 70(3), 37–44.
- Web Analytics Association, W.A.A. (2000). Web Analytics Definitions. Disponível em: 9, Setembro, 2016, em: http://www.digitalanalyticsassociation.org/Files/PDF_standards/WebAnalyticsDefinitions.pdf