

# Agregação e categorização de informação desportiva baseada em conteúdos gerados por utilizadores

Sérgio Carvalho<sup>1</sup>, Carlos Serrão<sup>2</sup>

1) ISCTE Instituto Universitário de Lisboa/ISTA/ADETTI-IUL, Lisboa, Portugal  
[mrsergiocar@hotmail.com](mailto:mrsergiocar@hotmail.com)

2) ISCTE Instituto Universitário de Lisboa/ISTA/ADETTI-IUL, Lisboa, Portugal  
[carlos.serrao@iscte.pt](mailto:carlos.serrao@iscte.pt)

## Resumo

O acesso a informação, em tempo útil, promove a procura de canais alternativos por parte de utilizadores que desejam obter conteúdos relevantes e personalizados. São inúmeros os que procuram informação desportiva *online*. Infelizmente, essa informação, está de uma maneira geral limitada aos conteúdos que as entidades responsáveis pelos mesmos desejam partilhar. Neste artigo, é apresentado um paradigma alternativo baseado em conteúdos desportivos gerados por utilizadores, utilizando as redes sociais, neste caso o Twitter, como base para agregação de informação desportiva para uma plataforma de acesso *online*. Neste artigo é igualmente identificada a arquitectura de suporte à plataforma que estará dividida em três etapas: a agregação, a categorização e a indexação de informação e de conteúdos. Por fim, discute-se o uso da plataforma a partir de dispositivos móveis, tanto na utilização da mesma para receber informação categorizada ou personalizada, assim como para alimentar a mesma com conteúdos gerados a partir dos próprios dispositivos móveis.

**Palavras chave:** Conteúdos gerados por utilizadores, Twitter, Modelos por Tópicos, desporto, eventos, dispositivos móveis.

## 1. Introdução

A obtenção e partilha de informação são uma das maiores necessidades da sociedade actual. O acesso a conteúdos digitais faz-se, nos nossos dias, em larga escala sendo que os mesmos são referentes a diversos temas, onde se inclui, por exemplo, o desporto.

Os conteúdos desportivos são, habitualmente, disponibilizados pelas entidades que gerem os desportos ou as aplicações desportivas. Como consequência, os utilizadores ficam limitados à informação que as mesmas decidem fornecer.

Como alternativa, pretende-se introduzir um novo paradigma que faça face a esta limitação, utilizando as redes sociais como base para um sistema de disponibilização de conteúdos desportivos, procurando deste modo usufruir de uma fonte mais alargada e rica em termos de conteúdos multimédia. Por outro lado, a plataforma permitirá contextualizar a informação recolhida, de forma a garantir aos utilizadores que usufruam dessa informação a possibilidade de parametrizar os conteúdos acedidos. A rede social utilizada será o Twitter e o processo passará por três etapas: agregação, categorização e indexação da informação recolhida. Apesar da rede social escolhida para este trabalho ter recaído no Twitter, o objectivo é que a arquitectura e sistema a desenvolver possa suportar múltiplos tipos de fontes de informação – como o Facebook ou blogs especializados em informação desportiva.

Em termos de aplicações futuras, a utilização deste sistema de agregação de conteúdos desportivos fará sentido em ambientes/dispositivos móveis, onde irá trazer uma nova perspectiva aos amantes do desporto que, por exemplo, por razões de localização ou fuso

horário não podem acompanhar algum evento do seu interesse, ou estando no próprio evento gostariam de ter uma perspectiva mais envolvente do mesmo. Assim, basta que os utilizadores possuam no seu *smartphone* uma aplicação móvel integrada com a plataforma de divulgação de conteúdos desportivos.

## 1.1 Redes Sociais

A utilização das redes sociais tem sido um fenómeno com enorme crescimento nos últimos anos a nível mundial. A infinita partilha de informação inerente a estas plataformas, incentiva os seus utilizadores a investirem tempo e dedicação à exploração das mesmas; no caso português, os internautas utilizam este meio de comunicação em larga escala para partilha de comentários, vídeos e fotografias e 20% desses utilizadores procuram conteúdos desportivos [Taborda et al. 2010].

Os serviços associados à partilha de conteúdos podem ser divididos em diversas categorias, entre as quais se destacam: *micro-blogging*, notícias, vídeo, fotos e música. Comparativamente, os serviços de *micro-blogging* têm uma importância superior, gerando mais conteúdos que os restantes, sendo o Twitter o maior contribuidor [Gupta et al. 2009]. Este facto, acabou por pesar na escolha do Twitter como base dos conteúdos a agregar para a plataforma, juntamente com a ideia do Twitter ser muito associado à divulgação de eventos, ou factos sobre eventos, em tempo real. Convocação de manifestações, informações relativas a catástrofes naturais ou comemorações de títulos desportivos são apenas alguns exemplos de utilização desta rede social.

## 1.2 Disponibilização de conteúdos (Entidades desportivas vs Utilizadores)

Uma das questões relevantes levantadas pela utilização das redes sociais como base para agregar conteúdos desportivos, prende-se com o facto da informação disponibilizada nas mesmas não seguir um conjunto de regras ou estruturas fixas, como normalmente acontece na informação obtida a partir de fontes desportivas especializadas. Cada utilizador do Twitter terá o seu próprio estilo para representar a informação a partilhar e as suas mensagens têm uma estrutura mais dinâmica se comparadas com as disponibilizadas por sítios dedicados. [Xu et.al. (2006)] num estudo sobre detecção de eventos desportivos ao vivo, definiram a existência de dois tipos de fontes de informação textual: texto bem estruturado sintacticamente e texto “livre”. O estudo baseou-se na utilização de palavras-chave para detecção de eventos e os autores chegaram à conclusão que o texto bem estruturado sintacticamente facilitava essa detecção. Fazendo o paralelo com os conteúdos existentes nas redes sociais, estes estarão incluídos no tipo de texto livre, tornando essencial que a plataforma disponha de um mecanismo de agregação e categorização que refute a ideia levantada pelos autores referidos em cima. Assim, os marcadores ou *hashtags*, existentes no Twitter, têm um papel fulcral na categorização das mensagens; ao termos uma palavra precedida pelo símbolo “#” significa que essa palavra é uma palavra-chave ou tópico da mensagem. Esta funcionalidade é muito utilizada no Twitter, e permite ao clicarmos numa palavra marcada ver todos os *tweets* que fazem parte da mesma categoria, ou seja, que foram marcados pelo mesmo *hashtag* (<https://support.twitter.com/articles/49309-what-are-hashtags-symbols>).

O artigo está organizado da seguinte forma: na secção 2 são apresentados trabalhos relacionados com as matérias em discussão; na secção 3 é apresentada a arquitectura que suporta a plataforma, dividida nas suas partes essenciais: agregação, categorização e indexação; por fim, na secção 4, apresentam-se conclusões e trabalho futuro.

## 2. Trabalhos relacionados

Diversos autores já procuraram debater-se sobre os problemas de agregação e categorização de conteúdos a partir das redes sociais, e mais especificamente do Twitter. Todos eles defenderam que a melhor forma de compreender e categorizar a informação partilhada nas plataformas sociais é a utilização de modelos por tópicos<sup>1</sup>.

[Ramage et al. 2010] apresentaram a implementação de um modelo de aprendizagem parcialmente supervisionado (Labeled LDA (Latent Dirichlet Allocation)) que tenta mapear os *tweets* em quatro dimensões: substância (relativa a eventos, ideias, coisas ou pessoas), estilo (ligada a tendências mais amplas de uso da linguagem), estado (associada a actualizações pessoais) e características sociais (ligada a fins de comunicação social). Este modelo, sendo uma extensão do LDA puro, permite a introdução de novas etiquetas que se aplicam apenas a um subconjunto das mensagens, o que permite aprender conjuntos de palavras associadas a etiquetas particulares (como *hashtags*). Os autores consideram ainda que as *hashtags* estão associadas à categoria substância (eventos e pessoas) e que embora a perspectiva comum seja diferente, a utilização da categoria substância é o dobro comparativamente à categoria estado (actualizações pessoais).

Noutro estudo [Hong et al. 2010], procurou-se perceber como treinar um modelo por tópicos eficazmente em ambientes onde as mensagens têm um tamanho mais reduzido, como é o caso do Twitter. Procuraram, entre outros objectivos, classificar mensagens e utilizadores segundo categorias, entre as quais desporto. Utilizaram dois tipos de modelos por tópicos, LDA e Autor-Tópico, e aplicaram diversos métodos para treinar os modelos, chegando a várias conclusões: os tópicos apreendidos usando diferentes estratégias de agregação de dados são substancialmente diferentes uns dos outros; treinar um modelo padrão de tópicos para mensagens agregadas de utilizadores produz um processo de treino mais rápido e com melhor qualidade; para a tarefa de classificação de mensagens e utilizadores, o modelo padrão LDA apresentou melhores resultados que o modelo Autor-Tópico.

Mais recentemente, [Zhao et al. 2011] procuraram comparar os conteúdos existentes no Twitter e os disponibilizados por um meio de comunicação tradicional, o jornal New York Times, utilizando um modelo não supervisionado por tópicos. Assim, para descobrir e classificar os conteúdos heterogéneos, definiram três conceitos: tópico, que representa um assunto discutido em um ou mais documentos; categoria de tópico, que agrupa tópicos pertencentes ao mesmo assunto (ex. Artes, Negócio, etc); tipos de tópico, que caracteriza a natureza do tópico – identificaram três tipos de tópicos: orientados a eventos, entidades e assuntos globais. Descobriram que o Twitter e os meios de comunicação tradicionais cobriam um número de categorias similares relativas a tópicos de informação, no entanto, as categorias mais importantes em cada um são totalmente diferentes. Enquanto na descoberta de tópicos, no New York Times foi aplicado directamente o modelo LDA tradicional, para o Twitter foi desenvolvida uma extensão a essa modelo, que os autores denominaram Twitter LDA. Enquanto o tradicional LDA pressupõe que cada documento tem associado um conjunto de tópicos, esta extensão define que um *tweet* diz respeito a um tópico em exclusivo.

Ainda, relativamente a conteúdos gerados por utilizadores, [Sarmiento et al. 2009] introduziram um método automático para criação, e actualização, de uma colecção de dados relativos a opiniões políticas; a abordagem pressupunha três fases: inicialmente recolheram opiniões colocadas por leitores de um jornal online; de seguida, aplicaram um conjunto de regras léxico-sintáticas (definidas manualmente) a esses comentários/opiniões, de forma a identificar um conjunto de frases que revelassem opiniões, positivas ou negativas, relativamente a entidades políticas relevantes; por fim, as opiniões identificadas são propagadas para as restantes frases

---

<sup>1</sup> Em aprendizagem automática e processamento de linguagem natural, modelos por tópicos são um tipo de modelo estatístico que procura descobrir tópicos abstractos com ocorrência numa colecção de documentos.

que fazem parte dos comentários que mencionam as entidades políticas, tendo por objectivo possuir um conjunto mais alargado de frases associadas a opiniões. Esta investigação estará, entre outras, relacionada com o projecto REACTION<sup>2</sup>, que permitiu o desenvolvimento do Twitómetro<sup>3</sup>, um instrumento que permitiu perceber os sentimentos revelados pelos utilizadores do Twitter, relativamente aos cinco líderes partidários portugueses mais relevantes nas eleições legislativas de 2011.

### 3. Arquitectura da plataforma

Na plataforma que está a ser desenvolvida como parte deste trabalho, o funcionamento da mesma ocorrerá segundo uma arquitectura baseada em três etapas: agregação, categorização e indexação de conteúdos. Na etapa da agregação, definem-se os mecanismos de interacção com a API<sup>4</sup> do Twitter para recolha de conteúdos desportivos (textuais e não-textuais). Ao nível da categorização dos conteúdos, é necessário identificar o algoritmo de categorização que define os conteúdos a pesquisar. Por fim, procede-se à indexação desses conteúdos, que consiste na forma como a informação ficará guardada para ser mais tarde disponibilizada aos utilizadores que interajam com a plataforma.

Como protótipo, da plataforma de agregação de conteúdos desportivos, será desenvolvido um mecanismo específico para identificação de conteúdos gerados por utilizadores que comunicam no Twitter usando a língua portuguesa.

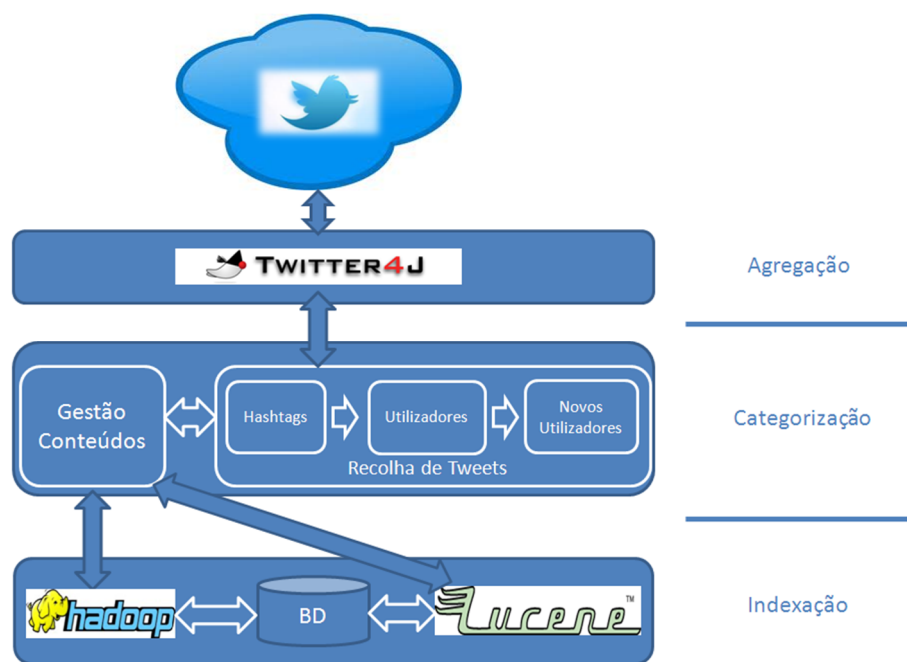


Figura 1 - Arquitectura da plataforma

#### 3.2 Agregação de conteúdos

A existência de inúmeras ferramentas para integração com a API do Twitter ligadas a linguagens de programação distintas facilita a tarefa dos programadores que desejam realizar tarefas de interacção com esta rede social. C++, Java, Python ou Ruby são apenas algumas das

<sup>2</sup> <http://xldb.di.fc.ul.pt/wiki/Reaction>

<sup>3</sup> <http://legislativas.sapo.pt/2011/twitometro/>

<sup>4</sup> API – Application Programming Interface

linguagens que possuem diversas ferramentas para o efeito. Na plataforma de integração de conteúdos desportivos, será utilizada a biblioteca Twitter4J<sup>5</sup>, ligada à linguagem Java. Esta ferramenta, desenvolvida por Yusuke Yamamoto e associada ao desenvolvimento em código aberto, permite interagir com o Twitter e realizar inúmeras tarefas como: ligação do utilizador à sua conta pessoal, realizar actualizações dessa conta através da própria API, obter todas as actualizações de outros utilizadores que siga no Twitter (accedendo às últimas mensagens na sua *timeline*) e realizar pesquisas de *tweets* segundo *queries*, permitindo identificar, por exemplo, autor e corpo dessas mensagens (<http://twitter4j.org/en/api-support.html>).

### 3.3 Categorização de conteúdos

A etapa de categorização pretende fazer a ponte entre os conteúdos que se agregam e a forma como estes serão mantidos na base de dados. A categorização será feita por fases e alimentará os conteúdos a agregar.

#### 3.3.1 Metodologia usada para categorização da informação

A metodologia de categorização utilizada segue muitas das ideias utilizadas no desenvolvimento do Twitómetro<sup>6</sup>, naturalmente adaptada à realidade desportiva.

A categorização da informação a pesquisar é feita segundo um conjunto de acções distintas: inicialmente são feitas pesquisas de acordo com um conjunto de *hashtags* definidas *a priori*. Estas *hashtags* dizem respeito a modalidades desportivas (ex. #Futebol, #Ténis, #Ciclismo), equipas (ex. #Benfica, #Porto, #Sporting) ou jogadores (ex. #Aimar, #Falcao, #Izmailov) – importante referir que serão definidas hierarquias entre *hashtags* (Modalidade > Equipa > Jogador) por forma a relacionar conteúdos e sugeri-los de acordo com preferências dos futuros utilizadores da plataforma.

Das pesquisas realizadas tendo por base as *hashtags* será feita uma recolha dos utilizadores portugueses que as efectuaram, através de duas técnicas: primeiro cruzando o *username* de cada utilizador com um léxico de nomes e apelidos portugueses<sup>7</sup>, e em segundo plano realizando uma análise do perfil do utilizador, de onde se podem extrair informações como localização, fuso horário ou língua.

Em segundo plano são feitas pesquisas baseadas nos utilizadores portugueses encontrados anteriormente, tendo dois objectivos: descobrir novos utilizadores portugueses (que tenham sido mencionados ou tenham tido mensagens “*retuitadas*” pelos utilizadores já descobertos) e descobrir novos tópicos de interesse desportivo existentes nas suas mensagens. Para o segundo objectivo, é usado o modelo por tópicos LDA, que permite descobrir tópicos abstractos que ocorrem num conjunto de documentos, neste caso *tweets*.

Para facilitar a tarefa do modelo por tópicos, será feito algum pré-processamento, no sentido de remover as chamadas *stop-words*, utilizando o modelo tf-idf (*term frequency-inverse document frequency*); este modelo permite avaliar a importância das palavras existentes num documento relativamente ao conjunto total dos documentos. A importância da palavra aumenta proporcionalmente ao número de vezes que aparece dentro de um documento, mas é compensada negativamente pela frequência de utilização da palavra no conjunto total dos documentos (<http://en.wikipedia.org/wiki/Tf%E2%80%93idf>).

Ainda, é necessário garantir que não se agregam mensagens duplicadas na plataforma; [Gomes et al. 2006] estudaram a presença de artigos duplicados em arquivos da Web e propõem um

---

<sup>5</sup> <http://twitter4j.org/en/index.html>

<sup>6</sup> [http://xldb.fc.ul.pt/xldb/publications/Silva.etal:NotasSobreA:2011\\_document.pdf](http://xldb.fc.ul.pt/xldb/publications/Silva.etal:NotasSobreA:2011_document.pdf)

<sup>7</sup> NomesLex-PT01 - <http://xldb.fc.ul.pt/wiki/NomesLex-PT01>

mecanismo para detecção de documentos duplicados antes dos mesmos serem guardados em disco. Baseando-se em resultados de experiências e trabalhos relacionados, concluíram que a duplicação de artigos em arquivos Web é considerável, porém não recomendam a eliminação de artigos parcialmente duplicados porque os sistemas de armazenamento existentes assumem a persistência dos identificadores dos documentos, o que pode levar a problemas de preservação quando aplicados a armazenamento distribuído em larga escala. Nesse sentido, procuraram eliminar apenas duplicados exactos através de um mecanismo simples baseado nas impressões digitais dos documentos. Assim, cada documento tem associada uma assinatura obtida através da aplicação de um algoritmo de *fingerprinting*.

No caso do Twitter, a informação a agregar tem as suas especificidades (ferramenta de *micro-blogging*) tendo um carácter menos complexo que aquele apresentado no problema de duplicação dos arquivos Web. Assim, propõe-se a utilização do coeficiente de similaridade de Jaccard, uma medida estatística que mede a similaridade entre dois conjuntos. Ou seja, dados dois conjuntos, o coeficiente é definido pela cardinalidade da intersecção desses conjuntos dividida pela cardinalidade da união dos mesmos (<http://www.infosci.cornell.edu/weblab/papers/Bank2008.pdf>). Um dos casos a ter em consideração, específico desta rede social, prende-se com a utilização da funcionalidade *retweet*, que consiste em encaminhar, ou retransmitir, uma mensagem. Deste modo, será necessário proceder não apenas à detecção de *tweets* duplicados exactos, mas também casos parcialmente idênticos.

### 3.4 Indexação de conteúdos

Na terceira e última etapa da arquitectura as principais preocupações são a eficiência, escalabilidade e performance da plataforma de agregação e categorização dos conteúdos desportivos. Estes objectivos serão obtidos utilizando duas ferramentas baseadas em código aberto, Apache Hadoop<sup>8</sup> e Apache Lucene<sup>9</sup>. A escolha destas duas ferramentas deveu-se não só ao seu carácter *open-source* mas também pela facilidade de integração com aplicações desenvolvidas em Java.

#### 3.4.1 Apache Lucene

Biblioteca *open-source* totalmente escrita em Java, pertencente à fundação *The Apache Software Foundation* (ASF), sendo uma ferramenta que realiza pesquisa e indexação de alta performance em recursos textuais. A nível de indexação possui um conjunto de funcionalidades interessantes, tais como capacidade de indexar 20 MB por minuto (num Pentium M a 1.5 GHz) e o tamanho dos índices gerados ocuparem apenas 20 a 30% comparativamente ao tamanho do texto indexado.

Consegue realizar pesquisas por classificação (devolve primeiro os melhores resultados), vários tipos de consultas (por frases, proximidade, intervalo, etc), pesquisa textual (ex. título, autor), permitindo ainda realizar pesquisas ao mesmo tempo que se actualizam os índices.

Outra funcionalidade a destacar, é a capacidade de fazer o “*merge*” de vários índices, criando um índice único, característica importante para sistemas de indexação distribuídos como o que se pretende desenvolver (<http://lucene.apache.org/java/docs/features.html>).

#### 3.4.2 Apache Hadoop

Igualmente desenvolvido pela fundação *The Apache Software Foundation*, é uma plataforma escalável cujo objectivo é realizar computação e armazenamento distribuído, processando

---

<sup>8</sup> <http://hadoop.apache.org/>

<sup>9</sup> <http://lucene.apache.org/>

grandes volumes de informação por vários sistemas computacionais, os chamados *clusters*. Existem três subprojectos directamente ligados ao Apache Hadoop: Hadoop Common, Hadoop Distributed File System (HDFS) e Hadoop MapReduce (<http://hadoop.apache.org/>).

- o Hadoop Common contém um conjunto de funcionalidades comuns aos outros projectos, podendo ser visto como a *framework* do Hadoop.
- o HDFS é um sistema de ficheiros distribuído, desenhado para suportar ficheiros com grande volume de informação (até perabytes) cujo principal objectivo é assegurar a replicação de dados para garantir tolerância a falhas. Um *cluster* HDFS segue uma arquitectura *master/slave* e consiste, sucintamente, num *NameNode* que gere os metadados do sistema de ficheiros e em *DataNodes* que gerem os dados armazenados no nó do *cluster* em que se encontram ([http://hadoop.apache.org/common/docs/current/hdfs\\_design.html](http://hadoop.apache.org/common/docs/current/hdfs_design.html)).
- o Hadoop MapReduce é um modelo de programação cujo objectivo é processar grandes quantidades de dados de uma forma paralela entre os vários nós do *cluster*. Possui uma arquitectura similar ao HDFS, tendo um nó principal, *JobTracker*, e múltiplos nós *slave*, chamados *TaskTracker*, existindo um por cada nó do *cluster*. O *JobTracker* é responsável por programar e monitorizar as tarefas dos nós *TaskTracker*, obrigando a uma re-execução em caso de falha. Os nós *TaskTracker* executam todas as tarefas ordenadas pelo nó *master* ([http://hadoop.apache.org/common/docs/current/mapred\\_tutorial.html](http://hadoop.apache.org/common/docs/current/mapred_tutorial.html)).

A utilização do Hadoop pode ser feita de três formas diferentes: localmente, funcionando tudo dentro do mesmo processo Java; de forma pseudo-distribuída, funcionando sob diversos processos Java, num conjunto pequeno de máquinas; ou totalmente distribuída, funcionando sob um conjunto alargado de *clusters*. Na plataforma de agregação e categorização de conteúdos desportivos, será utilizado o método pseudo-distribuído, com o desenvolvimento de um pequeno *cluster* para gerir toda a informação. Ainda, devido a preocupações inerentes ao enorme crescimento da plataforma de agregação de informações desportivas, pretende-se fazer uma gestão dos conteúdos, por forma a que apenas conteúdos com data inferior a um ano sejam mantidos no sistema.

#### 4. Conclusão

Este artigo aborda de uma forma clara uma das maiores preocupações da sociedade contemporânea - o acesso à informação. Nesse sentido, e tendo como base o acesso a informação de carácter desportivo, é apresentada uma proposta alternativa, sustentada numa arquitectura inovadora, que procura desenvolver uma plataforma de conteúdos desportivos relevantes para todos os utilizadores que venham a usufruir da mesma. A arquitectura faz uso de um conjunto de ferramentas ligadas ao desenvolvimento em código aberto, sendo o primeiro objectivo deste trabalho agregar conteúdos de utilizadores que utilizam a língua portuguesa.

Em termos de trabalho futuro, o primeiro passo será integrar o conjunto de ferramentas existentes na plataforma e discriminadas na arquitectura. Depois será desenvolvida uma aplicação para dispositivos móveis, que possibilite aos seus utilizadores interagir com a plataforma, não só para recolher conteúdos desportivos de acordo com as suas preferências pessoais mas possibilitando também que esses utilizadores alimentem a plataforma com fotografias, vídeos e comentários, podendo os mesmos ser realizados ao vivo nos eventos desportivos.

#### 5. Referências

Gomes, D., Santos, A., Silva, M.: Managing Duplicates in a web archive. In: Proceedings of the 2006 ACM symposium on Applied computing, Dijon, France (2006).

- Gupta, T., Garg, S., Mahanti, A., Carlsson, N., Arlitt, M.: Characterization of friendfeed-a web-based social aggregation service. In: Proc. AAAI ICWSM, San Jose, CA, USA (2009).
- Hong, L., Davison, B. D.: Empirical study of topic modeling in Twitter. In: Proceedings of the First Workshop on Social Media Analytics, Washington DC, USA (2010).
- Ramage, D., Dumais, S., Liebling, D.: Characterizing microblogs with topic models. In: International AAAI ICWM, Washington DC, USA (2010).
- Sarmento, L., Carvalho, P., Silva, M., Oliveira, E.: Automatic Creation of a Reference Corpus for Political Opinion Mining in User-Generated Content. In: 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement, Hong Kong, China (2009).
- Taborda, M., Cardoso, G., Espanha, R.: A Utilização de Internet em Portugal 2010. In: World Internet Project 2010 e LINI, Portugal (2010).
- Xu, C., Wang, J., Wan, K., Li, Y., Duan, L.: Live sports event detection based on broadcast video and web-casting text. In: Proceedings of the 14th annual ACM international conference on Multimedia, New York, USA (2006).
- Zhao, W., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: The 33rd European Conference on Information Retrieval, Dublin, Ireland (2011).