Faculdade de Engenharia da Universidade do Porto

Visualização de documentos partilhados em colecções dinâmicas

João Miguel Neves Pereira de Almeida

Relatório submetido no âmbito do

Mestrado Integrado em Engenharia Electrotécnica e de Computadores

Major de Telecomunicações

Orientador: Prof. Maria Cristina de Carvalho Alves Ribeiro

Co-orientador: Prof. Sérgio Nunes

Julho de 2009

U.PORTO

FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

MIEEC - MESTRADO INTEGRADO EM ENGENHARIA ELECTROTÉCNICA E DE COMPUTADORES 2008/20

A Dissertação intitulada

"VISUALIZAÇÃO DE DOCUMENTOS PARTILHADOS EM COLECÇÕES DINÂMICAS"

foi aprovada em provas realizadas em 22/Julho/2009

o júri

Presidente Professor Doutor Francisco José de Oliveira Restivo
Professor Associado do Departamento de Informática da Faculdade de Engenharia da Universidade do Porto

Professor Doutor José Paulo de Vilhena Geraldes Leal
Professor Auxiliar do Departamento de Ciência de Computadores da Faculdade de Ciências da Universidade do Porto

Professora Doutora Maria Cristina de Carvalho Alves Ribeiro
Professora Auxiliar do Departamento de Engenharia Informática da Faculdade de Engenharia da Universidade do Porto

Mestre Sérgio Sobral Nunes
Assistente do Departamento de Engenharia Informática da Faculdade de Engenharia da Universidade do Porto

O autor declara que a presente dissertação (ou relatório de projecto) é da sua exclusiva autoria e foi escrita sem qualquer apoio externo não explicitamente autorizado. Os resultados, ideias, parágrafos, ou outros extractos tomados de ou inspirados em trabalhos de outros autores, e demais referências bibliográficas usadas, são correctamente citados.

Autor - **JOÃO MIGUEL NEVES PEREIRA DE ALMEIDA**

Faculdade de Engenharia da Universidade do Porto

© João Miguel Neves Pereira de Almeida, 2009

Resumo

O conceito de colecção dinâmica generalizou-se nos últimos anos graças à evolução da tecnologia, nomeadamente em redes de computadores. As colecções dinâmicas têm como principal característica o facto de os seus conteúdos não serem estáticos, isto é, poderem evoluir ao longo do tempo. Associada a esta caraceterística, existe a possibilidade de múltiplos utilizadores poderem contribuir na construção dos documentos ou outros conteúdos, que se encontram alojados remotamente.

Existe uma ferramenta amplamente divulgada para o trabalho colaborativo, denominada *wiki*. O seu grande sucesso deve-se, sobretudo, às suas regras de funcionamento, nomeadamente no que diz respeito à livre edição de documentos, bem como à possibilidade de se poder manter um histórico das alterações dos mesmos. Para além disso, é possível consultar uma série de dados acerca dessas revisões, tais como o autor e a data. O exemplo mais mediático das *wikis* é a Wikipedia, uma enciclopédia online com milhões de artigos. A Wikipedia revela-se como um excelente objecto de estudo, pois a consulta dos dados referidos (e muitos outros) é rápida, de fácil acesso e sobretudo porque esses dados podem ser obtidos programaticamente. A Wikipedia é uma colecção de largo espectro e de grande dinamismo, uma vez que engloba as contribuições de muitos autores e é utilizada por numerosos leitores.

É objectivo deste trabalho investigar padrões de actividade em colecções dinâmicas, mais concretamente na Wikipedia, pelos motivos referidos anteriormente. Para tal foi desenvolvida uma ferramenta web, denominada WikiViz. O WikiViz serve de suporte à investigação da actividade na Wikipedia permitindo visualizar os resultados graficamente. É possível analisar o número de revisões feitas a um artigo e à respectiva página de discussão ao longo do tempo, o número de autores que efectuam essas mesmas revisões, a evolução do tamanho de um artigo, bem como informação mais específica sobre a contribuição de autores anónimos e registados. Todas estas características podem ser analisadas em diversos idiomas da Wikipedia para comparação de resultados.

Nos resultados obtidos foram detectados alguns casos interessantes, nomeadamente haver diversos artigos (o de Barack Obama, por exemplo) cuja actividade na página de discussão é, em certas alturas, superior à da página principal. Os autores anónimos que contribuem na revisão de artigos são, na sua maioria, oriundos dos Estados Unidos, independentemente do idioma da Wikipedia. Verificou-se ainda que o número de autores anónimos é significativo tendo em conta a totalidade dos autores da Wikipedia.

Os dados recolhidos nestas consultas são guardados numa base de dados relacional. Foi estudada a aplicação do conceito de um armazém de dados a este caso, de forma a que a pesquisa e análise de dados históricos seja mais fácil, rápida e diversificada.

Abstract

The concept of dynamic collection became very popular in the last years due to the technological evolution, particularly in the area of computer networking. Dynamic collections have as a distinguishing feature the fact that their contents are not static, but instead they can evolve through time. Another advantage is the possibility that multiple users can contribute to the creation and development of documents (or other contents), without the need of being in the same physical space (editing a document that is stored in a remote server).

The evolution of document's collaborative editing has lead to a very known tool - the wiki. Its business model turned it into a very successful platform, mainly because of its document edition policies and the possibility to rollback any changes made throughout over time. Each set of changes made by any given author in a moment in time is a revision. One of the most famous example of a successful wiki is Wikipedia, an online encyclopedia containing millions of articles on subjects covering almost all areas known to man. It turns out to be an excellent case study, as metadata is easily queried and accessed and can be programmatically obtained. Furthermore, any of Wikipedia's articles can be viewed by any person and edited by many users.

The purpose of this paper is to investigate activity patterns in dynamic collections, specially in the Wikipedia, for the reasons mentioned above. To accomplish this objective, a web based tool - WikiViz - was developed. It can be used to retrieve Wikipedia's articles metadata and display it graphically. It is possible to check how many revisions a certain article has, as well as its discussion page and how many authors contributed to that article. The data displayed is related to a configurable period of time. There is also the possibility of monitoring an article's size as well as the percentage of contributions from anonymous and registered users. All of these features can be analyzed in the different languages Wikipedia is written to allow a possible comparison of results.

The obtained results show some interesting cases, namely articles (related to Barack Obama and Sarah Palin, for example) where the activity on the discussion page is, some times, superior to the activity on the article itself. Also it can be seen that the majority of anonymous authors, who mostly contribute to the revision of articles, are United States natives, regardless of the language used in the Wikipedia article. Another fact is that the number of anonymous authors is very significant when considering the total number of Wikipedia authors.

All of the collected data in these searches is stored in a relational database. A data warehouse was implemented, in order to ensure that querying the data was easier and faster and to allow more diversified analyses.

Conteúdo

1 Introdução **1**
- 1.1 Objectivos . . . 2
- 1.2 Estrutura do trabalho . . . 2

2 Visualização de colecções dinâmicas **3**
- 2.1 Exemplos de ferramentas sobre a Wikipedia . . . 3
- 2.2 Outro exemplo . . . 11
- 2.3 Análise e Comentários . . . 12

3 Análise das propriedades da Wikipedia **15**
- 3.1 Análise de colecções dinâmicas . . . 15
- 3.2 Características e propriedades das *wikis* . . . 16
- 3.3 Escolha da Wikipedia como objecto de estudo . . . 17
- 3.4 Exploração da API . . . 17
- 3.5 Enquadramento das características nos trabalhos desenvolvidos . . . 19

4 Visualização de documentos da Wikipedia **21**
- 4.1 Funcionalidades e características . . . 21
- 4.2 Ferramentas para visualização de informação . . . 22
- 4.3 Análise de visualizações sobre a Wikipedia . . . 25

5 Integração com um armazém de dados **37**
- 5.1 Definição e objectivos de um armazém de dados . . . 37
- 5.2 Estruturas de um armazém de dados . . . 38
- 5.3 O armazém de dados do WikiViz . . . 38

6 A ferramenta de visualização **41**
- 6.1 Tecnologias escolhidas . . . 41
- 6.2 Potencialidades do WikiViz . . . 42
- 6.3 Lógica de funcionamento . . . 42

7 Conclusões **45**

A Anexo **47**

Referências **52**

Lista de Figuras

2.1 Interface gráfico do WikiChanges para o artigo Barack Obama da Wikipedia. 4
2.2 Interface gráfico do History Flow (retirado de [1]). 5
2.3 Técnica usada para a visualização do History Flow (retirado de [1]). 6
2.4 Visualização da informação disponibilizada pelo WikiDashBoard (retirado de [2]). 7
2.5 Visualização 2D e 3D da evolução do artigo Jolanda Di Savoia (retirado de [3]). 7
2.6 Revert Graph para o artigo Dokdo da Wikipedia [4]. 8
2.7 Exemplo de visualização das revisões para o artigo Gun Politics (retirado de [5]). 9
2.8 Número de visualizações de 4 diferentes artigos, usando o WikiRank. 10
2.9 Visualização da evolução do desenvolvimento da ferramenta Eclipse (retirado de [6]). 11

3.1 Representação esquemática da API da MediaWiki 18

4.1 Resultado da junção das duas ferramentas:AmChart e AmMap. 23
4.2 Intensity Map é uma das formas possíveis de visualização de informação no Google Visualization. 24
4.3 Motion Chart é outro formato disponível no Google Visualization. 24
4.4 Historial do número de revisões do artigo *Papa João Paulo II*. 26
4.5 Historial do número de revisões do artigo *Michael Phelps*. 26
4.6 Historial do número de revisões do artigo Volta a França. 27
4.7 Historial do número de revisões do artigo dos Óscares. 27
4.8 Página principal vs Página de Discussão do artigo *Barack Obama*. 28
4.9 Número de revisões da página principal vs número de revisões da página de discussão do artigo *Ataques de 11 de Setembro*. 29
4.10 Historial do número de revisões e do número de autores ao artigo *Ataques de 11 de Setembro*. 30
4.11 Evolução do tamanho do artigo Sarah Palin. 31
4.12 Distribuição geográfica dos autores no artigo Brasil. 32
4.13 Distribuição geográfica dos autores no artigo Aborto. 32
4.14 Nuvem de palavras com os nomes do utilizadores registados do artigo Aborto. 33
4.15 Historial do número de autores anónimos e de autores registados no artigo Vladimir Putin. 34
4.16 Historial do número de revisões do artigo Maria de Lurdes Rodrigues na Wikipedia inglesa. 34

4.17 Historial do número de revisões do artigo Maria de Lurdes Rodrigues na Wikipedia portuguesa. 35
4.18 Distribuição geográfica dos autores na Wikipedia polaca para o artigo Iraque. 35

5.1 Modelo em estrela de um armazém de dados. 39
5.2 Estrutura do armazém de dados no WikiViz. 39

6.1 Fluxograma da lógica de funcionamento do WikiViz. 42

A.1 Historial de revisões do artigo Gripe Suína na Wikipedia inglesa. 47
A.2 Historial de revisões do artigo Luciano Pavarotti na Wikipedia inglesa. . . . 48
A.3 Historial do número de revisões e do número de autores do artigo Comunismo na Wikipedia inglesa. 48
A.4 Historial do número de revisões e do número de autores do artigo Adolf Hitler na Wikipedia inglesa. 49
A.5 Distribuição geográfica dos autores anónimos no artigo Aníbal Cavaco Silva. 50
A.6 Distribuição geográfica dos autores anónimos no artigo Rainha Beatriz da Holanda. 50

Capítulo 1

Introdução

As colecções dinâmicas de documentos são conjuntos de documentos em constante evolução, graças à colaboração de múltiplos autores. Inicialmente os documentos eram escritos em papel e para efectuar uma alteração era necessária uma nova edição, tratando-se, portanto, de novos documentos estáticos. Os documentos dinâmicos, de um único autor, surgiram com o aparecimento dos computadores pessoais, podendo qualquer documento ser alterado à medida que fosse necessário. Mais recentemente, com o desenvolvimento das redes de computadores e todas as tecnologias inerentes a estas, surgiu o conceito de colecção dinâmica de documentos, em que múltiplos autores, em qualquer lado e com uma simples ligação à Internet, podem contribuir para a construção dos mesmos. Uma das tecnologias que suporta colecções de documentos são as *wikis* que, por guardarem todo o seu historial de revisões e outras importantes características, se tornam num alvo interessante de estudo para análise de comportamentos e padrões de actividade dos utilizadores.

As *wikis* são das ferramentas mais populares para trabalho em colaboração e têm como principal característica a possibilidade de qualquer pessoa poder editar o seu conteúdo. Para além disso, armazenam todas as revisões feitas, de modo a que, no caso de ser necessário, se possa recuar para uma versão anterior. Como tal, as *wikis* são adequadas ao armazenamento de grandes colecções de dados que estão em constante mudança. Estes dados são passíveis de ser abordados através das tecnologias de informação para a sua interpretação e posterior exposição visual, uma forma interessante de compreender as actividades em colecções dinâmicas ao longo do tempo.

A Wikipedia será o exemplo mais emblemático das *wikis*, com cerca de 10 milhões de artigos escritos por inúmeros utilizadores, em diversos idiomas [7]. Além da sua dimensão, o acesso a dados (como sejam, revisões anteriores e a meta-informação inerente a estas), através da API [8] disponibilizada, é bastante simples. A Wikipedia, para além das características próprias das *wikis*, disponibiliza ainda, para cada artigo, uma página de discussão que proporciona o debate sobre o mesmo, ficando toda a actividade nesta página registada e podendo também ser acedida através da API. Por estes motivos, a Wikipedia tem sido alvo de vários estudos, na detecção de diversos padrões de utilização, na análise

do comportamento dos utilizadores. Destacam-se entre outros trabalhos os de S. Nunes *et al* [9], F. B. Viégas *et al* [1], B. Suh *et al* [2], Aniket Kittur *et al* [4], Ulrik Brandes *et al* [5] e J. Gawryjo Lek *et al* [3].

Este trabalho surge na sequência do desenvolvimento de uma ferramenta designada por WikiChanges [9]. Esta ferramenta apresenta, em forma de gráfico, a evolução das revisões que um artigo da Wikipedia sofreu ao longo do tempo e permite adicionalmente disponibilizar na forma de nuvem de palavras, as palavras-chave mais inseridas entre a revisão seleccionada e a mais antiga desse dia ou mês.

1.1 Objectivos

O objectivo deste trabalho será explorar os temas suscitados pela análise ao WikiChanges. Alguns aspectos a explorar:

- Análise do comportamento dos autores dos artigos.
- Influência das páginas de discussão no número de revisões dos artigos.
- Análise dos meta-dados inerente às revisões, como sejam a data de revisão e o seu autor.
- Exploração de técnicas de visualização para a representação da informação que é extraída dos artigos.
- Implementação de uma ferramenta para a produção automática de visualizações.

1.2 Estrutura do trabalho

No Capítulo 2 é feito um levantamento do estado da arte sobre a visualização de documentos partilhados em colecções dinâmicas. No Capítulo 3 são descritas as principais características das colecções dinâmicas e, em particular, das *wikis*, os motivos da escolha da Wikipedia para objecto de estudo e a exploração da API usada neste trabalho. No Capítulo 4 são postas em evidência as funcionalidades da ferramenta desenvolvida, bem como os resultados obtidos, através do seu uso. No capítulo 5 é descrita a integração de um armazém de dados, no âmbito deste trabalho. No Capítulo 6 é referida a descrição técnica da ferramenta desenvolvida. O último capítulo é dedicado às conclusões e propostas para trabalho futuro.

Capítulo 2

Visualização de colecções dinâmicas

Muitos trabalhos foram desenvolvidos com o objectivo de estudar os padrões de actividade nas colecções dinâmicas. Uma parte significativa destes trabalhos foi desenvolvida sobre a Wikipedia pois, como já foi referido, a sua API disponibiliza muitos dados sobre a actividade lá desenvolvida. Para além disso, a sua enorme divulgação faz com que haja uma constante actividade, tornando mais significativas as análises efectuadas.

De seguida são apresentados os trabalhos mais significativos na área das colecções dinâmicas, e que incluem o desenvolvimento de ferramentas visuais para a apresentação dos resultados. De referir ainda que a última ferramenta apresentada não usa dados da Wikipedia, sendo no entanto um bom exemplo de visualização de outro tipo de colecção dinâmica.

2.1 Exemplos de ferramentas sobre a Wikipedia

WikiChanges

A WikiChanges [9] é uma ferramenta que tem como objectivo o estudo do historial de revisões de um artigo. Foi criada uma aplicação que se pode dividir em duas partes: Perfil de actualizações e Sumariação de revisões.

Uma das características das wikis é o armazenamento do historial de revisões dos artigos e aproveitando esse facto surge o *Perfil de Actualizações*, que consiste basicamente em mostrar o número de revisões do artigo ao longo do tempo. Esta informação é disponibilizada sob a forma de um gráfico.

Neste trabalho verificou-se que existe uma maior actividade nas revisões de um artigo nas alturas em que o conteúdo deste se tornou mais popular por qualquer motivo. Para ajudar a perceber a razão de tal acontecimento, foi desenvolvida a *Sumariação de Revisões*, que consiste na apresentação das palavras mais introduzidas pelos utilizadores entre a revisão seleccionada, de um determinado dia/mês, e a primeira revisão desse dia/mês,

de forma a simplificar o algoritmo. Esta informação é apresentada na forma de *nuvem de palavras*, isto é, a palavra mais vezes inserida aparece com o tamanho maior e assim sucessivamente para as restantes. É possível percorrer qualquer parte do gráfico e verificar o sumário para um dado período de tempo.

Esta ferramenta foi desenvolvida em Perl e, para a apresentação dos gráficos, foi utilizada a biblioteca Flash amCharts [10].

A Figura 2.1 representa o *interface* da WikiChanges, podendo ver-se na parte superior o *Perfil de Actualizações* e na parte inferior a *Sumariação de Revisões*.

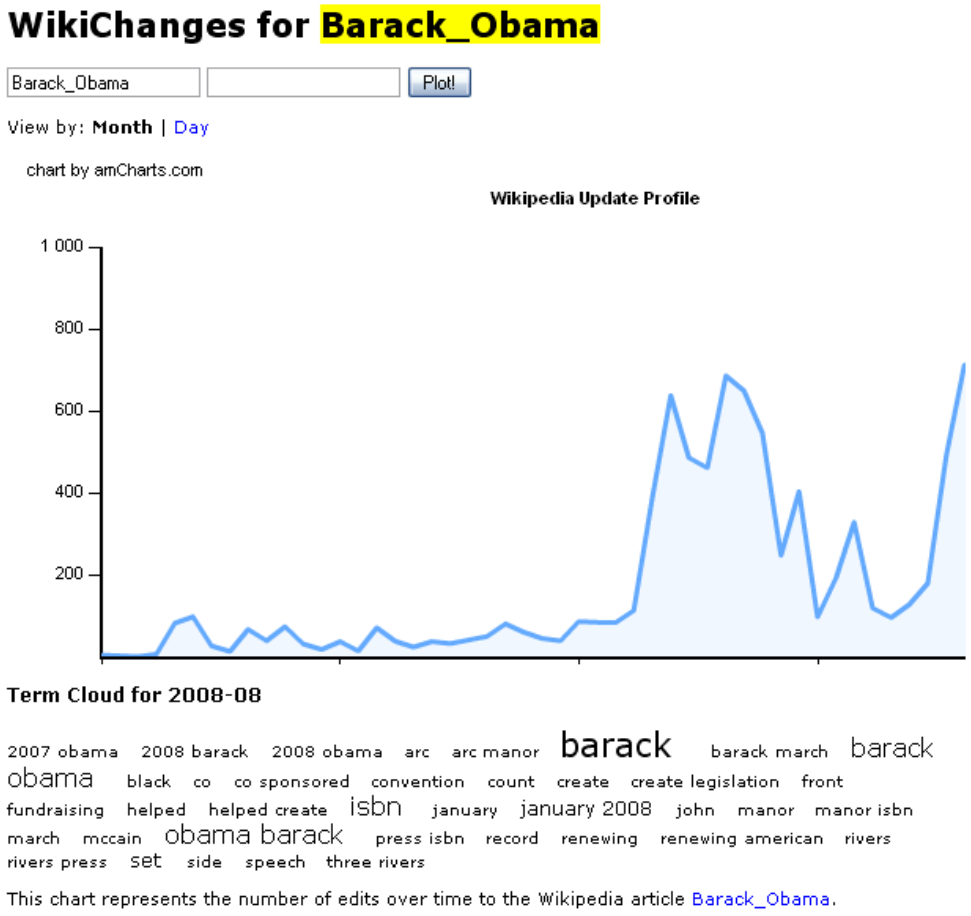

Figura 2.1: Interface gráfico do WikiChanges para o artigo Barack Obama da Wikipedia.

History Flow

A *History Flow* [1] é uma ferramenta de visualização de informação com o principal objectivo de mostrar as relações entre as múltiplas versões de um artigo na Wikipedia. De uma forma visual são representadas as partes de texto que se mantiveram entre versões consecutivas ao longo do tempo. O seu interface está representado na Figura 2.2.

Para explicar a técnica utilizada no processamento de dados desta ferramenta, consideremos o exemplo da Figura 2.3. Neste exemplo, existem 4 versões de um documento,

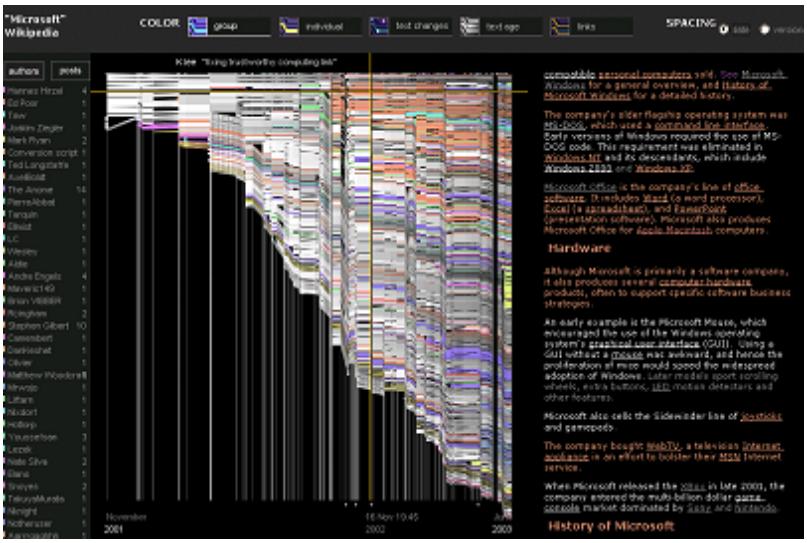

Figura 2.2: Interface gráfico do History Flow (retirado de [1]).

sendo cada uma delas representada por uma barra com diferentes cores, que correspondem a edições de diferentes utilizadores (Figura 2.3-A) . Esta informação isolada não tem muito valor, sendo portanto essencial ligar os pedaços de texto que se foram mantendo ao longo das diversas versões (Figura 2.3-B). Numa outra variante da técnica, demonstrada na Figura 2.3-C, o espaço entre as barras de revisão é proporcional ao tempo entre as datas de revisão, e portanto a revisão x é apresentada dessa forma.

Esta ferramenta permitiu detectar alguns padrões de colaboração, nomeadamente actos de vandalismo e a sua reparação, a negociação e a influência de os autores estarem autenticados ou não, ou seja, é possível perceber se autenticação influencia a qualidade da escrita de um documento. A Wikipedia é extremamente vulnerável a actos de vandalismo, tendo sido identificadas as seguintes variantes:

- Eliminação em massa - é eliminado todo o conteúdo de uma página.
- Cópia ofensiva - inserção de calúnias sobre o tema do artigo.
- Cópia falsa - inserção de texto que não está ligado ao tema do artigo.
- Redireccionamento falso - a ligação de redireccionamento pode ser maliciosa, fazendo a ligação para um artigo que não esteja minimamente relacionado.
- Cópia idiossincrática - inserção de texto de opinião parcial ou sem qualquer interesse.

Constatou-se que todos estes actos costumam ser rapidamente corrigidos pela comunidade da Wikipedia.

Outro dos padrões detectados pelo History Flow foi o de negociação, visualmente representada por uma forma em zig-zag, onde duas pessoas ou grupos alternam entre versões de um artigo.

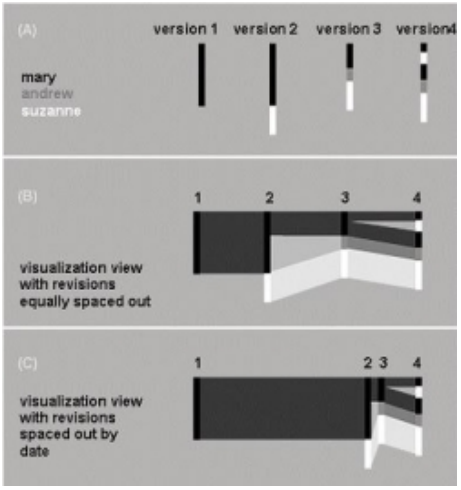

Figura 2.3: Técnica usada para a visualização do History Flow (retirado de [1]).

Verificou-se também que, consoante o tipo de artigo, pode haver mais edições por parte de utilizadores registados ou anónimos, mas não foi possível verificar se existe uma preferência por algum tipo de utilizadores. Finalmente, é de salientar o facto de não ter ficado comprovada uma ligação explícita entre os actos de vandalismo a artigos e o anonimato dos autores, como se poderia pensar antes do início do trabalho.

WikiDashBoard

A *WikiDashBoard* [2] é uma extensão para a Wikipedia, com o objectivo de investigar se os utilizadores têm maior confiança num artigo, se tiverem acesso a várias informações sobre o mesmo, nomeadamente quem o editou ou quantas vezes o fez, entre outras. Para além disso ajudará o utilizador a identificar diversos padrões de edição. Na Figura 2.4 está representada a interface da ferramenta.

Podemos dividir esta ferramenta em duas partes claras:

- *Article DashBoard*, onde é disponibilizada informação sobre as revisões do artigo na última semana, nomeadamente os autores que o editaram, quantas vezes o fizeram, o número total de edições do artigo e da respectiva página de discussão.
- *User DashBoard*, que surge depois de se clicar no nome de um dos autores, disponibiliza informação sobre a actividade semanal de edição do respectivo autor. Essa informação vai desde os artigos que editou até ao número de vezes que o fez em cada um deles.

JWikiVis

O JWikiVis é [3] muito semelhante ao History Flow, especialmente ao nível das funcionalidades. Apesar do seu interface 2D ser diferente, esta ferramenta conta com uma

Figura 2.4: Visualização da informação disponibilizada pelo WikiDashBoard (retirado de [2]).

componente 3D para visualização das alterações efectuadas num artigo ao longo do tempo, como se mostra na Figura 2.5.

Figura 2.5: Visualização 2D e 3D da evolução do artigo Jolanda Di Savoia (retirado de [3]).

Observando o quadro da Figura 2.5 verificamos que cada rectângulo corresponde a um parágrafo do documento. Para além disso, estão indicados o nome do autor e respectiva data de edição. No quadro as linhas representam o estado de um documento numa dada altura do tempo e as colunas representam as partes de texto que fazem parte desse artigo.

A mesma representação é feita num gráfico 3D, em que o terceiro eixo é o eixo dos autores. Cada autor é colocado num plano diferente, de modo a facilitar a sua distinção.

Revert Graph

O Revert Graph é uma ferramenta de visualização desenvolvida por Kittur *et al.* [4]. A informação apresentada baseia-se na divergência de opiniões ou conflitos dos utilizadores em determinados artigos, algo de inevitável num ambiente de colaboração como a

Wikipedia. A ferramenta desenvolvida, como mostra a Figura 2.6, permite de uma forma visual verificar o conflito entre autores de um artigo. Para calcular o grau de conflito foi criado um modelo que se rege pelos seguintes princípios: os *auto-reverts* são descartados e a quantificação da disputa entre dois utilizadores está relacionada com o número de *reverts* entre eles. Quando dois utilizadores A e B fazem *revert* a uma edição do utilizador C, assume-se que A e B fazem parte do mesmo grupo e quando uma página é revertida para uma versão mais antiga, apenas é contada a relação entre o autor do revert e o autor da edição imediatamente anterior. No caso da figura, está representado o resultado para o artigo Dokdo, onde facilmente identificamos todos os autores de revisões, agrupados em diferentes grupos, de acordo com as suas opiniões, respectiva motivação de edição e grau de conflito.

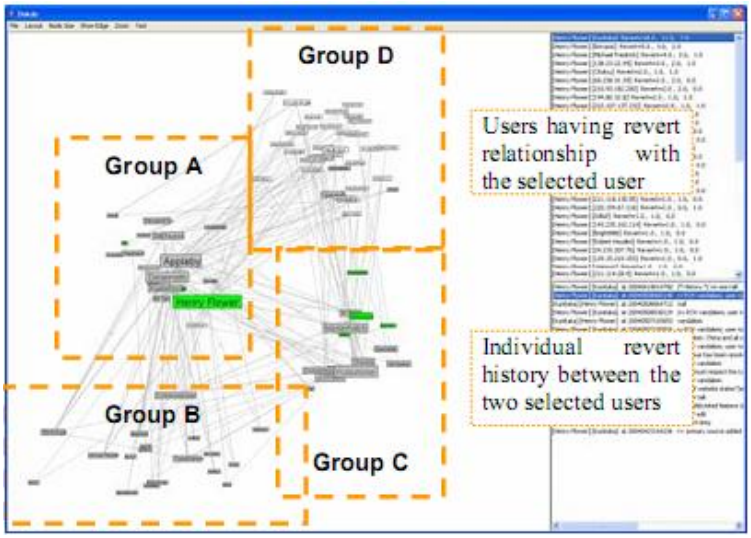

Figura 2.6: Revert Graph para o artigo Dokdo da Wikipedia [4].

Para além da informação mostrada na ferramenta, outros factos interessantes foram verificados, destacando-se o facto da actividade directa, isto é, o número de edições nos artigos, estar a diminuir e, pelo contrário, a actividade indirecta, como edições nas páginas de discussão dos artigos e actividades de manutenção (reversão de artigos e anti-vandalismo), estar a aumentar com o passar do tempo.

Who revises whom

No trabalho de Ulrik Brandes *et al* [5] foram desenvolvidas técnicas de visualização extremamente simples para o utilizador, que permitem obter diversas informações sobre as revisões de um artigo mais controverso da Wikipedia. É possível perceber quais os autores que entram em conflito na revisão de um artigo, se um autor tem um comportamento mais de revisor ou de revisto, o seu envolvimento na controvérsia do artigo e finalmente se um autor tem uma participação constante na revisão do artigo ou se isso apenas se verifica em

certos períodos de tempo. Para além destas informações, ainda se pode consultar o número de edições que o artigo sofreu ao longo do tempo. Na Figura 2.7 encontra-se um exemplo de visualização, mais concretamente para o tema *Gun politics* da Wikipedia. Cada nó na figura corresponde a um autor diferente, independentemente de ser um autor registado ou anónimo e quanto mais afastados estiverem dois autores, mais estes se reviram ao longo do tempo. De referir, ainda, que o diferente tamanho dos nós, representados na mesma figura, é directamente proporcional ao envolvimento do autor na controvérsia desse artigo e que a cor dos nós está relacionada com a frequência de edição desse mesmo autor.

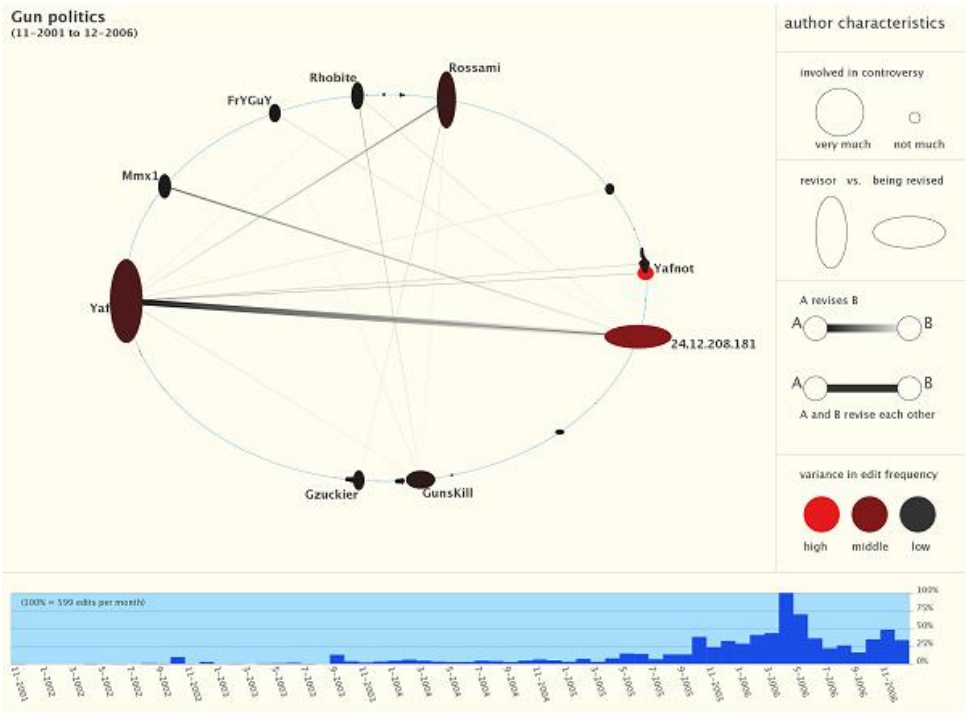

Figura 2.7: Exemplo de visualização das revisões para o artigo Gun Politics (retirado de [5]).

WikiChecker

A WikiChecker [11] encontra-se ainda em fase experimental(mas revela-se extremamente completa). É possível obter várias informações sobre um artigo, sobre um autor ou ainda sobre uma *guerra* de edições entre 2 ou mais autores num artigo.

Relativamente à informação que se pode de extrair de um artigo são usadas a sua data de criação, o número de revisões que foram feitas em média desde que este existe, o número total de utilizadores que contribuiu com revisões, o peso dos utilizadores anónimos na revisão do artigo, entre outros.

Quanto aos autores, é possível pesquisar pelo seu nome, sendo retornada uma série de informações sobre a sua actividade na Wikipedia. Destacam-se informações relevantes,

como: a data de registo, o número total de edições que fez, o tipo de actividades que se destacam nesse autor, as horas do dia e os dias da semana esse autor está mais activo, entre outros.

Finalmente, em relação à guerra de edições a página é actualizada regularmente identificando os artigos com maior guerra de edição, classificando-os com uma pontuação tanto maior quanto maior for a guerra de edição entre autores.

WikiTrends

O WikiTrends é uma ferramenta [12] bastante simples, informa o utilizador dos artigos mais populares na Wikipedia. A popularidade dos artigos é medida de acordo com o número de visualizações das páginas. A ferramenta disponibiliza a consulta dos artigos mais populares em diversos idiomas da Wikipedia. Desta forma é possível fazer uma comparação das tendências entre os vários países, e registar eventuais diferenças culturais.

WikiRank

O WikiRank [13] é um serviço disponível na *web* que se foca nas tendências da Wikipedia inglesa. Na sua página principal são facultados os nomes dos artigos mais vistos na Wikipedia, nos últimos 30 dias, devidamente ordenados e com o respectivo número de visualizações. Para além disso, são apresentados os artigos com mais visualizações nas últimas 24 horas. O WikiRank disponibiliza também um motor de busca de artigos da

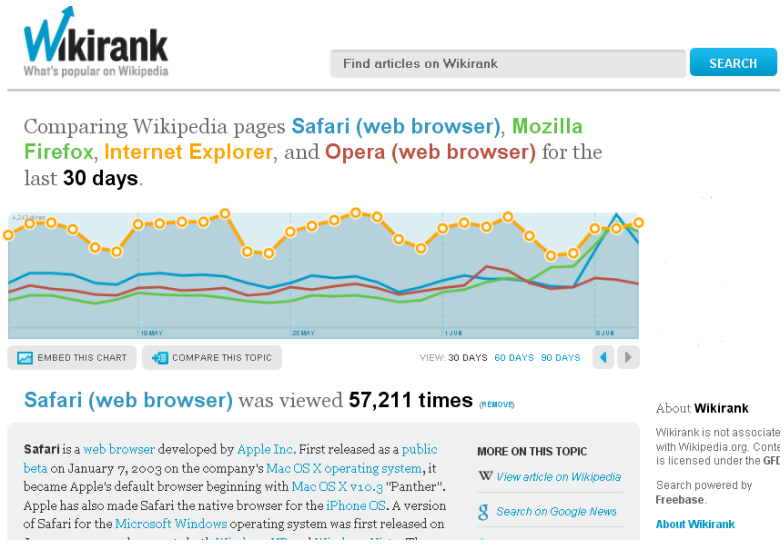

Figura 2.8: Número de visualizações de 4 diferentes artigos, usando o WikiRank.

Wikipedia, apresentando um gráfico com o número de visualizações ao artigo nos últimos 30, 60 ou 90 dias e parte do conteúdo desse mesmo artigo. De referir ainda que é possível

fazer a comparação do número de visualizações para 2 ou mais artigos à escolha do utilizador, como se pode ver na Figura 2.8 para os artigos dos *browsers* Safari, Mozilla Firefox, Internet Explorer e Opera.

2.2 Outro exemplo

Code Swarm

Na Figura 2.9 é apresentada uma visualização denominada Code Swarm [6] que demonstra, sob a forma de vídeo, as alterações feitas ao código de um projecto de software e posterior submissão para um repositório central de código. Nesta visualização são representados os responsáveis pelas alterações e os ficheiros submetidos. Estes ficheiros são representados em cores diferentes, consoante o seu significado, isto é, pode tratar-se de ficheiros de código ou simplesmente ficheiros de texto. Existem outros pormenores da visualização, quando um ficheiro é submetido, por exemplo, este desloca-se na direcção do respectivo autor. Por outro lado e de acordo com a participação de um utilizador, o seu nome fica mais ou menos brilhante.

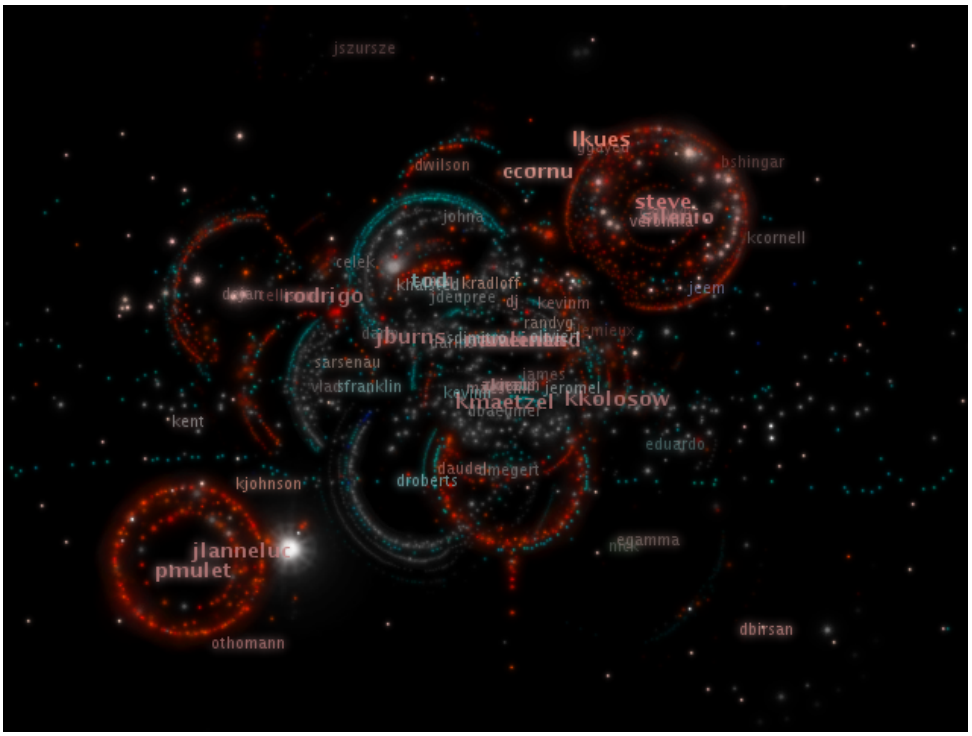

Figura 2.9: Visualização da evolução do desenvolvimento da ferramenta Eclipse (retirado de [6]).

2.3 Análise e Comentários

Depois da descrição de algumas das ferramentas de visualização em colecções dinâmicas e mais concretamente na Wikipedia, verificamos que nem todas têm os mesmos objectivos. Regem-se por diferentes parâmetros de actividade e, como tal, no final têm diferentes funcionalidades. Na Tabela 2.1 apresenta-se um resumo das funcionalidades e quais as mais aprofundadas em cada uma das ferramentas descritas, isto é, quantas mais vezes estiver representado o sinal (+) para uma ferramenta, tanto mais essa ferramenta se foca na propriedade.

Feita a análise, verifica-se que as ferramentas *Who Revises Whom* e *History Flow* e *JWikiVis* são as que mais parâmetros abrangem. A primeira é a que fornece mais informação ao utilizador e com a vantagem de apresentar toda a informação através de uma interface simples e intuitiva. O *History Flow* foca-se, sobretudo, no conteúdo das revisões e respectivos autores, no entanto a sua interface é um pouco confusa, especialmente quando o número de revisões é elevado. O *JWikiVis* tem as mesmas funcionalidades que o *History Flow*, mas a sua interface 2D já é um pouco mais apelativa e mais acessível, no entanto a abordagem 3D não é muito fácil de analisar, sobretudo quando se trata de um artigo com muitas revisões.

Como já foi referido e se pode verificar na tabela, a *WikiChanges* debruça-se, principalmente, sobre o número de revisões de um artigo e na sua evolução ao longo do tempo. Neste trabalho, seria interessante dotar a *WikiChanges* de outras funcionalidades, que já existam ou não noutras ferramentas, que a tornassem mais completa, através de uma nova interface que continue a ser intuitiva e de fácil análise. Nos próximos capítulos serão descritos, em maior detalhe, as funcionalidades que se pretende criar e o respectivo interface gráfico.

Propriedades	**FERRAMENTAS DE VISUALIZAÇÃO**				
	WikiChanges	*History Flow e JWikiVis*	*WikiDashBoard*	*Revert Graph*	*Who Revises Whom*
Nº de revisões a um artigo	+++		+		++
Conteúdo da revisão	+	+++		+++	++
Evolução dos artigos no tempo	+++	++	+		+
Autoria		+++	++	++	+++
Revisões na página de discussão			++		

Tabela 2.1: Funcionalidades das ferramentas descritas

Capítulo 3

Análise das propriedades da Wikipedia

3.1 Análise de colecções dinâmicas

Uma colecção dinâmica é um conjunto de conteúdos digitais, normalmente documentos de texto, que estão em evolução ao longo do tempo, devido à contribuição de um ou mais autores. Um caso típico que encaixa nesta descrição são, por exemplo, as *wikis*.

Neste secção será feito um levantamento das principais características das colecções dinâmicas de forma a perceber o seu funcionamento e o seu potencial.

Características e propriedades genéricas

Tendo em conta a definição de uma colecção dinâmica, faz sentido identificar as suas principais características. Podemos destacar as seguintes:

- Historial de revisões;
- Conteúdos guardados remotamente.

O historial de revisões é a característica mais significativa das colecções dinâmicas. Esta característica está relacionada com o facto de todas as revisões a um documento serem, frequentemente, guardadas numa base de dados. Desta forma é possível, à comunidade, substituir uma revisão por uma versão mais antiga, no caso de actos de vandalismo ou simplesmente por discordância do que foi alterado anteriormente. Ao facto de todas as revisões estarem guardadas, estão associadas diversas propriedades básicas relativas a essas revisões e que são passíveis de consulta, tais como o autor, a data, o seu conteúdo e ainda o comentário do autor à alteração efectuada. A possibilidade de extracção destes dados faz com que seja possível organizá-los de diversas formas, de acordo com os objectivos pretendidos. É, por exemplo, possível verificar, sob a forma de um gráfico, o número de revisões a um documento ao longo do tempo ou, então, o número de diferentes autores que

contribuem para a elaboração do mesmo, também ao longo do tempo. Todo o conteúdo descrito é guardado remotamente, de forma a facilitar o acesso a partir de qualquer ponto com acesso à Internet.

Exemplos de colecções e enquadramento nas características identificadas

Um exemplo de uma colecção dinâmica amplamente utilizada e conhecida é a Wikipedia. Trata-se de uma colecção de artigos que sofre várias alterações ao longo do tempo. Todas elas são guardadas, permitindo a sua reutilização caso seja necessário. A estas revisões estão associadas as propriedades já referidas, como a sua autoria, a data da alteração, etc. Para além destas propriedades, comuns à grande maioria das colecções dinâmicas, é possível extrair muitos outros dados relativos a toda a actividade da Wikipedia. O acesso a estes dados é público, bastando para tal o uso da API existente, que será descrita no Capítulo 4. Tal acesso potenciou o desenvolvimento de diversas ferramentas que exploram e organizam esses dados em informação mais clara, facilitando a sua análise por parte do utilizador.

Outro exemplo típico de uma colecção dinâmica é o dos repositórios de código de uma empresa. O código de um determinado projecto é desenvolvido, corrigido e reposto por vários trabalhadores dessa empresa, ao longo do tempo, até ser alcançado o produto final. Este é um caso típico de uma colecção dinâmica no mundo profissional. Como tal as propriedades identificadas na secção 3.1 são aplicáveis a este exemplo. Tendo acesso a esses dados podem ser desenvolvidas ferramentas cujas visualizações possam servir para o estudo da participação de cada um dos colaboradores e respectiva avaliação de desempenho.

3.2 Características e propriedades das *wikis*

As *wikis* são o exemplo mais conhecido de colecções dinâmicas. Para além das características comuns às colecções dinâmicas referidas no capítulo anterior, as *wikis* possuem outras características particulares, tais como:

- Todos os utilizadores têm os mesmos direitos de edição (salvo algumas excepções) na generalidade das *wikis*, ou seja, tanto utilizadores registados como não registados (anónimos) podem editar um artigo. No entanto, na Wikipedia existe uma organização hierárquica de categorias de utilizadores, cada qual com diferentes direitos [14].

- A sua expansão é feita de uma forma orgânica. Isto quer dizer que dois ou mais documentos com a mesma importância podem não ter o mesmo desenvolvimento, estando, portanto, tais documentos sujeitos às tendências e interesses da respectiva comunidade.

- Possuírem páginas de discussão associadas a artigos ou a autores, que têm como finalidade promover um debate, antes e durante, a alguma alteração ao artigo.

As propriedades passíveis de serem extraídas das *wikis*, foram já referidas no capítulo anterior, nomeadamente o autor, data, conteúdo e comentário à revisão em questão.

3.3 Escolha da Wikipedia como objecto de estudo

A escolha da Wikipedia como o elemento de estudo deste trabalho teve por base dois factores fundamentais.

O primeiro está relacionado com o facto de o acesso a todos os registos de actividade na Wikipedia serem públicos, acesso esse que é feito através do uso de uma API desenvolvida para o efeito. Pode ser extraída uma longa lista de dados, nomeadamente informação sobre um ou mais autores, sobre categorias de artigos ou ainda informação relativa às imagens dos artigos, para além dos já referidos dados inerentes às revisões. Na secção seguinte será feito um levantamento de todos os dados passíveis de serem extraídos.

O outro factor relevante é a dimensão da Wikipedia formada, actualmente, por milhões de artigos, nos mais diversos idiomas e, sobretudo, por ser uma colecção de grande dinamismo, existindo constantes alterações e discussões aos seus conteúdos.

3.4 Exploração da API

Uma API especifica como um utilizador que desenvolve determinada aplicação deve aceder ao comportamento e estado de um conjunto de classes e objectos. A que é usada neste trabalho foi criada pela MediaWiki [8] e foi também a base da maioria dos trabalhos apresentados, sobre a Wikipedia, no Capítulo 2.

A sua utilização é relativamente simples, funcionando à base de chamadas via *url*. Este *url* é construído de acordo com os dados que o utilizador pretenda extrair da Wikipedia, incluindo o formato dos resultados retornados, como se pode analisar na Figura 3.1. O formato escolhido, para o processamento dos dados, no desenvolvimento da ferramenta foi o XML, havendo a possibilidade de escolha de outros formatos, como JSON, ou mesmo em formato de texto.

A API da Wikipedia é muito extensa e como tal é possível extrair uma grande diversidade de dados. Por este motivo foi dividida em módulos, entre os quais se podem destacar: Query, Sitematrix, Login, Delete, Undelete e outros de menor relevo. No entanto, o módulo Query é o mais importante, uma vez que é neste que se concentra a maior parte dos dados de maior interesse para este trabalho, como se pode ver na Figura 3.1. Este módulo é divido em 3 partes distintas: Prop, List e Meta. De forma a se entender as potencialidades desta API é feito um levantamento exaustivo da informação possível de extrair destes 3 sub-módulos, sendo aprofundados os pontos de maior relevo para este trabalho.

O primeiro sub-módulo, Prop, é dedicado à extracção de diversas propriedades dos artigos da Wikipedia. As de maior relevância para este trabalho são as relativas às revisões,

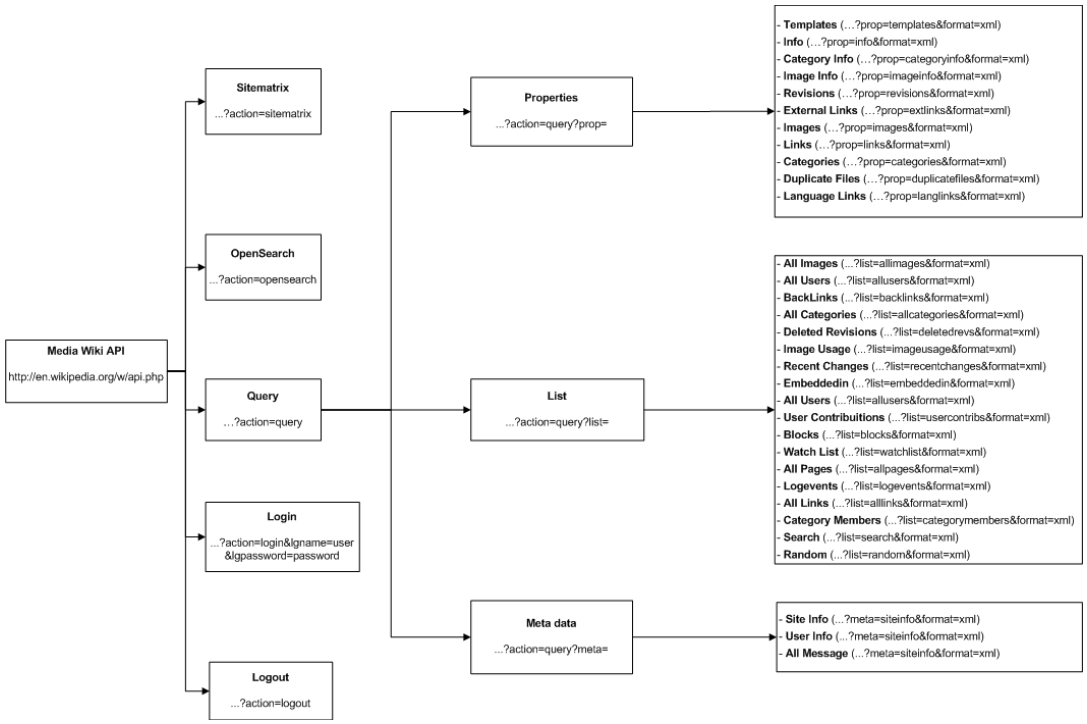

Figura 3.1: Representação esquemática da API da MediaWiki

nomeadamente a data, o autor, o conteúdo, o tamanho do artigo, o comentário e o número identificativo da revisão. A pesquisa de revisões pode ser filtrada de várias formas, tanto pela data de início como pela data final da enumeração dessas mesmas revisões ou ainda pelos seus números identificativos. De referir ainda que na filtragem das revisões é possível obter apenas aquelas que foram feitas por um determinado autor ou, pelo contrário, apresentar todas as revisões à excepção das desse mesmo autor. Para além disto, o utilizador pode ainda controlar o número de revisões retornadas, sendo o máximo possível 500. Por exemplo, a seguinte chamada à API:

`http://en.wikipedia.org/w/api.php?action=query&prop=revisions&format=xml&titles=Porto&rvlimit=500&rvprop=timestamp|user|size|comment`

Significa que se pretende obter as primeiras 500 revisões (parâmetro *rvlimit*) relativas ao artigo Porto (parâmetro *titles*) com as propriedades data, utilizador, tamanho e comentário (parâmetro *rvprop*).

Para além destas propriedades é possível extrair as ligações, os *url* externos e a informação sobre as imagens que se encontram num artigo. É possível ainda verificar as categorias a que um artigo pertence, bem como mais informações sobre essas mesmas categorias (número de subcategorias, número de artigos, etc).

Relativamente ao sub-módulo List é possível obter uma série de diferentes listas de dados. A que mais se destaca, no âmbito deste trabalho, é a listagem de todas as páginas sequencialmente, que pertençam a uma determinada gama de nomes. É possível fazer

vários tipos de filtragem das páginas que se desejam obter, nomeadamente a filtragem pelas letras inseridas. O endereço seguinte representa um pedido via API da MediaWiki:
`http://en.wikipedia.org/w/api.php?action=query&list=allpages&apfrom=B`
Neste caso é pedido que sejam apresentadas todas as páginas, utilizando o parâmetro *list*, cujo nome do título comece pela letra B, através do parâmetro *apfrom*.

Outras listagens estão disponíveis, tais como as ligações que apontam para uma determinada gama de nomes, as categorias de artigos existentes na Wikipedia, os autores registados e alguns dados inerentes a estes (contagem do número de revisões, os grupos a que pertencem, entre outros), ou ainda os utilizadores bloqueados por IP e respectivas propriedades (o utilizador que o bloqueou, a data e o motivo que levou ao bloqueio).

Para além desta informação, existe a possibilidade de obter outras listagens, nomeadamente as páginas que pertencem a uma dada categoria, as revisões apagadas, as mudanças recentes nos artigos, as edições feitas por um utilizador e respectivas propriedades (titulo do artigo, data, comentário), os endereços IP bloqueados, e ainda os utilizadores e as suas propriedades (informação de bloqueio, os grupos a que pertencem, somatório do número de edições realizadas, etc).

O sub-módulo List tem disponível a possibilidade de procurar por todas as páginas que incluam um título de um artigo ou que usem um dado título de imagem, de listar um determinado número de páginas que contenham, no seu título ou no seu conteúdo, o texto inserido pelo utilizador, de obter a listagem de todas as páginas, ou as suas mudanças recentes, dos favoritos de um utilizador autenticado, de enumerar as páginas que contenham um determinado *url*, assim como de apresentar um grupo aleatório de páginas.

Finalmente, o último sub-módulo, Meta, tem como finalidade facultar uma série de meta-dados relativos aos seguintes pontos: o site da Wikipedia, o utilizador actual e as mensagens do site do MediaWiki.

Na secção seguinte serão identificados os dados utilizados nos trabalhos já desenvolvidos e na Secção 5.1 serão referidos os dados que foram extraídos para alcançar os objectivos deste trabalho.

3.5 Enquadramento das características nos trabalhos desenvolvidos

Verificou-se que está disponível uma panóplia de dados sobre toda a actividade na Wikipedia e tal facto potencia o desenvolvimento de investigações, como as que já foram demonstradas. Em seguida são identificadas as características que cada um dos trabalhos desenvolvidos utilizou.

Para o desenvolvimento da ferramenta *WikiChanges*, foi necessário ter dados sobre a data de revisão de cada artigo para a construção do módulo *Perfil de Actualizações*. Para o módulo *Sumariação de Revisões*, para além da data de revisão, foi importante

ter em conta o conteúdo da revisão, de forma a poder verificar-se quais as palavras mais introduzidas entre a versão escolhida e a mais antiga.

No *History Flow* e no *JWikiVis* (ambos têm as mesmas funcionalidades) os dados extraídos da API da Wikipedia foram a data de revisão, o conteúdo da revisão, o tamanho da revisão e finalmente o autor da revisão.

No caso da *WikiDashBoard* foi preciso, para a construção do módulo Article DashBoard, determinar as datas das revisões e seus autores, bem como as datas de alterações nas páginas de discussão. No módulo User DashBoard são fundamentais dados relacionados com o utilizador, tais como os artigos que editou e quantas revisões fez nos mesmos.

Para o desenvolvimento do *Revert Graph* foi criado um algoritmo para a detecção de conflitos de revisões entre autores. Para tal, e como paramêtros de entrada, foram necessários dados sobre quem são os autores e o conteúdo de cada uma das revisões.

A ferramenta *Who Revises Whom* fornece vários tipos de informação, necessitanto portanto vários tipos de dados, nomeadamente a autoria das revisões, o número de vezes que o autor reviu determinado artigo, as datas em que foram feitas essas revisões e as datas de todas as revisões que um artigo sofreu.

O *WikiChecker*, que é uma ferramenta muito completa, à semelhança da *Who Revises Whom*, explora vários tipos de dados da Wikipedia. Para a visualização da informação sobre um artigo, o *WikiChecker* necessita de dados relativos às revisões, tais como a autoria e data. Quanto a visualização da informação sobre um autor da Wikipedia, são necessários dados sobre a totalidade de revisões feitas por essa pessoa, os artigos que editou e ainda a informação acerca desse autor ter sido ou não bloqueado. Todos estes dados são processados de diversas maneiras e transformados numa série de diferentes visualizações.

O *WikiTrends* e o *WikiRank* baseiam-se em informação [15] gerada por Domas Mituzas (faz parte da Wikimedia Foundation Board of Trustees [16] e trabalha com a tecnologia Wikipedia desde 2004) que desenvolveu um sistema capaz de fazer a contagem do número de visualizações de cada página da Wikipedia. Por este motivo estas duas ferramentas não usam qualquer propriedade da API da MediaWiki.

Capítulo 4

Visualização de documentos da Wikipedia

4.1 Funcionalidades e características

Esta investigação vem no seguimento de um outro trabalho que resultou no desenvolvimento do WikiChanges. No presente trabalho pretende-se desenvolver uma nova ferramenta com a implementação de novas funcionalidades.

Numa primeira fase, para além do gráfico das revisões existentes na WikiChanges, pretende-se saber qual o número de diferentes autores que fazem essas revisões e respectiva evolução ao longo do tempo, para o que são necessários dados sobre a autoria das revisões e as respectivas datas de ocorrência. Paralelamente, criar-se-á outro gráfico que mostre o número de comentários na página de discussão desse artigo ao longo do tempo. Desta forma, será possível perceber se esses comentários antecipam um pico de edições no artigo ou, se pelo contrário, apenas se verificam depois dos picos de edição.

Numa segunda fase, serão analisados outros aspectos das revisões nomeadamente o tamanho (em Kbytes) do artigo e o número de revisões designadas *reverts* (regresso à revisão anterior). Em ambos os casos serão apreciadas as respectivas evoluções com o passar do tempo, recorrendo-se a duas propriedades relativas às revisões: tamanho do artigo (*size*) e o comentário à revisão (indispensável para verificar se a revisão é *revert*).

É também objectivo do trabalho determinar para cada artigo a nacionalidade dos seus revisores anónimos, uma vez que estes são identificados pelo seu IP e agrupá-los por países, de maneira a poder-se determinar, de uma forma proporcional, as nacionalidades com mais influência na construção dos artigos. Através do IP e do uso do módulo Geo::IPfree [17] é possível fazer a sua geo-localização.

Em relação aos autores registados será efectuada uma contagem dos que mais participaram na edição de um artigo. O resultado desta contagem será visualizado na forma de uma nuvem de palavras, na qual o nome do autor com maior tamanho de letra corresponde

ao que mais contribuiu. Para ambos os casos, a propriedade da API a que é necessário recorrer é, simplesmente, o nome do autor da revisão.

Finalmente, pretende-se fazer a comparação do número de autores registados com o dos anónimos, que contribuíram na edição de um artigo, ao longo do tempo, para o que também é suficiente o acesso à autoria das revisões.

De salientar que todas estas funcionalidades estarão disponíveis não só para a Wikipedia inglesa como para todos outros idiomas existentes. Isto tem como objectivo principal a possibilidade de comparação de resultados, da pesquisa sobre um artigo, nos diversos idiomas. Para tal, basta alterar no url de chamada à API a sigla do idioma da Wikipedia. Por exemplo, a pesquisa do artigo Porto na Wikipedia portuguesa seria feita da seguinte forma:

`http://pt.wikipedia.org/w/api.php?action=query&prop=revisions&format=xml&titles=Porto&rvprop=timestamp|user|comment`

Como se pode observar, no início do endereço electrónico está indicada a sigla de Portugal (pt).

A ferramenta incluirá uma funcionalidade de um *auto-complete* para eventual ajuda ao utilizador na pesquisa de artigos da Wikipedia. Para cumprir esta tarefa é necessário aceder ao sub-módulo *List* da API, mais concretamente à secção *allpages*.

Pelo facto de todas as funcionalidades se traduzirem em diferentes visualizações sobre a Wikipedia, a ferramenta desenvolvida foi designada por WikiViz.

4.2 Ferramentas para visualização de informação

De forma a poder desenvolver o interface gráfico para demonstração de resultados, foi feita uma pesquisa de ferramentas que cumprissem os seguintes requisitos:

- Oferta de variadas formas de visualização;
- Formas de visualização que possam ser embutidas numa página Web e de forma dinâmica.

Várias ferramentas foram identificadas, das quais sobressaíram as seguintes: *XML/SWF Charts* [18], *Open Flash Chart* [19], *Black Box Chart* [20], *AmCharts* [10] e *Google Visualization API* [21].

O *XML/SWF Charts* é uma solução interessante, desenvolvida em Flash, uma vez que se trata de software livre e a ponte de comunicação entre cliente/servidor é XML. No entanto, suporta apenas gráficos muito simples e pouco apelativos como forma de visualização, tornando-se um pouco limitada nesse ponto. Por ter sido desenvolvida em Flash, esta ferramenta pode tornar-se um pouco mais lenta no seu processamento.

O *Open Flash Chart* é uma solução muito semelhante à anterior, inclusive na limitação de formas de visualização, mas a comunicação entre cliente/servidor é feita através de um ficheiro JSON (JavaScript Object Notation).

A *Black Box Chart* é uma solução também semelhante à primeira, mas um pouco mais limitada nos gráficos que disponibiliza. Partilha, portanto, também das mesmas desvantagens que a *XLM/SWF Charts*.

O *AmCharts* foi o software usado na implementação do WikiChanges. Trata-se uma ferramenta poderosa, desenvolvida também em Flash, capaz de criar diversos gráficos e várias formas de interacção. A forma de comunicação entre cliente e servidor é realizada através de um ficheiro XML, com a informação necessária à construção do gráfico. Existe, ainda, uma outra ferramenta designada por *AmMaps* [22], do mesmo criador do AmCharts, que permite a criação de mapas, e para além disso, é possível a junção das duas ferramentas, ou seja, a criação de um mapa e de um gráfico em simultâneo, como está representado na Figura 4.1. Estas ferramentas têm a desvantagem de serem um pouco *pesadas*, como é normal em todas as aplicações desenvolvidas em Flash.

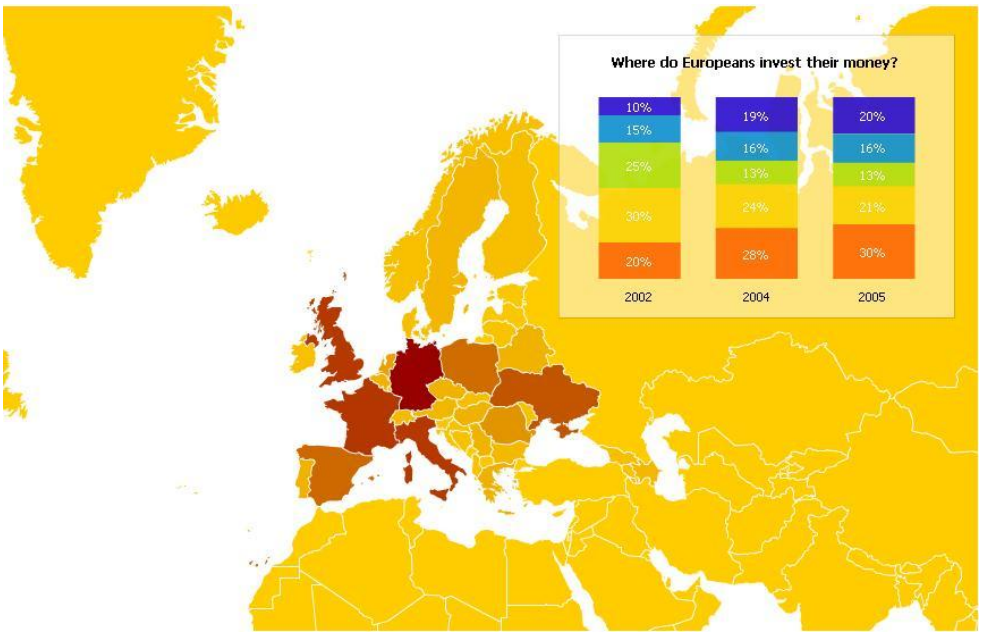

Figura 4.1: Resultado da junção das duas ferramentas:AmChart e AmMap.

O *Google Visualization* API [21] foi uma solução interessante encontrada, que preencheu os requisitos necessários. Trata-se de uma API, desenvolvida em JavaScript, para correr do lado do cliente e que dispõe de diversos formatos de visualização, para além dos já conhecidos gráficos de barras ou de linhas, como se pode ver nas figuras 4.2 e 4.3. É um software de uso gratuito como muitas outras aplicações da Google. Tem ainda a vantagem de a sua implementação ser bastante acessível e muito semelhante para todos os tipos de visualização disponíveis. Isto é, os valores são guardados numa tabela e posteriormente são transformados na respectiva visualização.

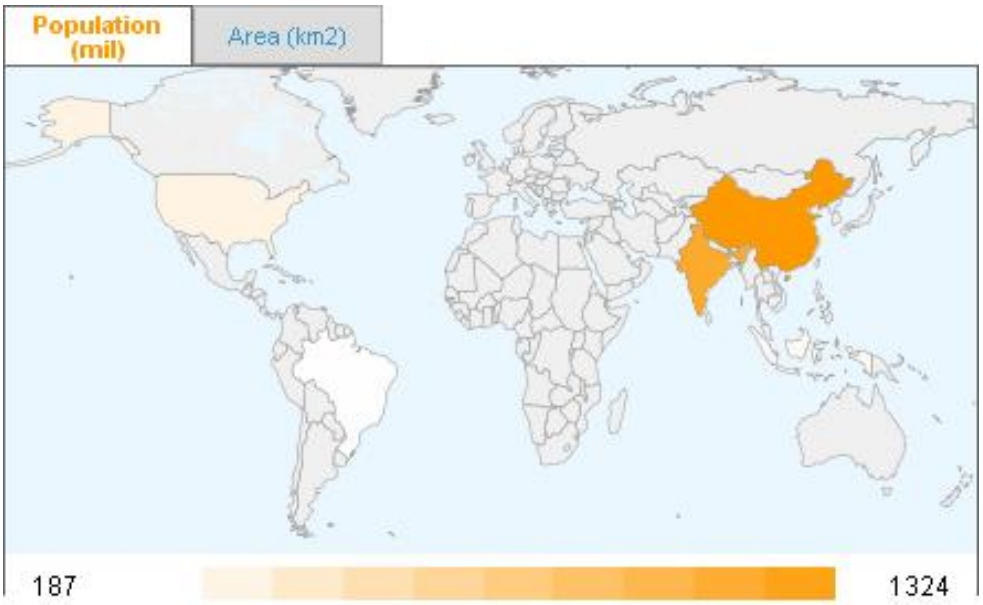

Figura 4.2: Intensity Map é uma das formas possíveis de visualização de informação no Google Visualization.

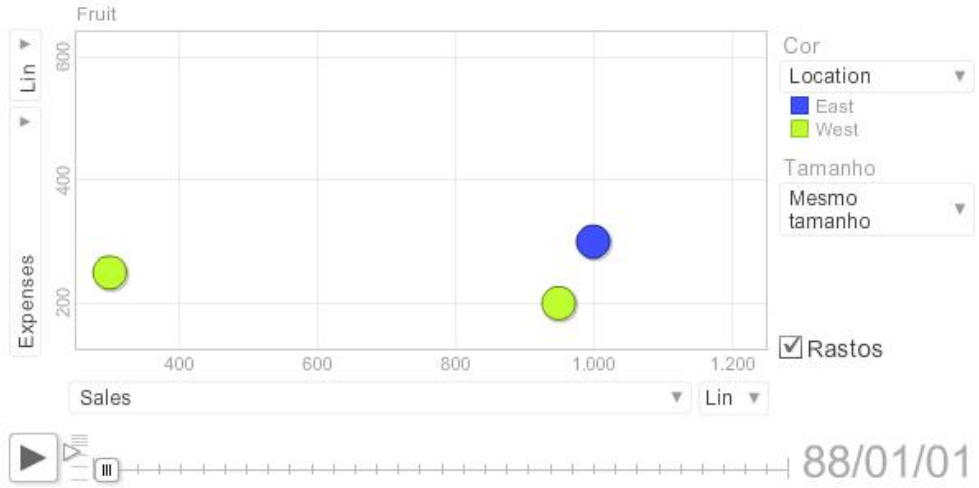

Figura 4.3: Motion Chart é outro formato disponível no Google Visualization.

4.3 Análise de visualizações sobre a Wikipedia

Nesta secção serão apresentadas e discutidas algumas das visualizações obtidas, usando o WikiViz. Em primeiro lugar, serão abordados alguns resultados sobre as revisões de artigos e respectivas propriedades. Posteriormente, serão postos em evidência os resultados sobre a autoria das revisões. Em ambos os casos a análise será feita na Wikipedia inglesa. Por fim, será feita uma comparação da actividade visualizada na Wikipedia inglesa com a da Wikipedia noutros idiomas.

Revisões e suas propriedades

Neste ponto serão analisadas, em primeiro lugar, algumas das tendências verificadas no número de revisões das páginas principal e de discussão. Depois serão apresentados os resultados observados das propriedades das revisões estudadas.

Revisões das páginas principal e de discussão

Desde cedo que foram efectuadas várias pesquisas no WikiViz, de forma a se poder detectar eventuais padrões de actividade. Uma característica detectada, logo nas primeiras análises, é o facto de os picos de revisões estarem associados a acontecimentos relevantes sobre esse artigo. São disto exemplos os artigos do *Papa João Paulo II* mostrado na Figura 4.4 e de *Michael Phelps* apresentado na Figura 4.5 - artigos com um pico de revisões anormal, tendo em conta o seu historial de actividade (com poucas revisões fora dessas datas).

No artigo do *Papa João Paulo II* o pico coincide com o seu falecimento (Abril de 2005), enquanto que no artigo de *Michael Phelps* coincide com a sua prestação notável nos Jogo Olímpicos em Agosto de 2008. Como estes casos existem muitos outros, como são os casos da Gripe Suína ou de *Luciano Pavarotti*, cujas visualizações são remetidas para os anexo A.1 e A.2. A ideia principal a reter, destes casos, é que a comunidade da Wikipedia está mais activa nas alturas de acontecimentos marcantes do momento actual.

Em artigos de eventos sazonais, foi particularmente observado que os picos de revisões são coincidentes com ocorrência do evento. O melhor caso que exemplifica esta situação é a *Volta a França* em bicicleta, que se realiza em Julho de cada ano, verificando-se picos de revisões nesse mês em todos os anos, como se pode observar na Figura 4.6. Outro exemplo não tão claro, mas que, ainda assim, reflecte este comportamento é o caso dos *Óscares* (Figura 4.7). Esta cerimónia não tem uma data fixa de realização, variando entre Fevereiro, Março ou mesmo Abril, no entanto é visível que é nestas alturas que ocorre um maior número de revisões.

Existem ainda os artigos que geram frequente discussão, casos do aborto ou a política de armas, ou mesmo de algumas figuras públicas (exemplo: *Barack Obama* e *Britney Spears*), que estão em constantes alterações, não sendo fácil explicar alguns dos picos de revisões existentes. De uma forma geral pode-se dizer que artigos que mexam com a

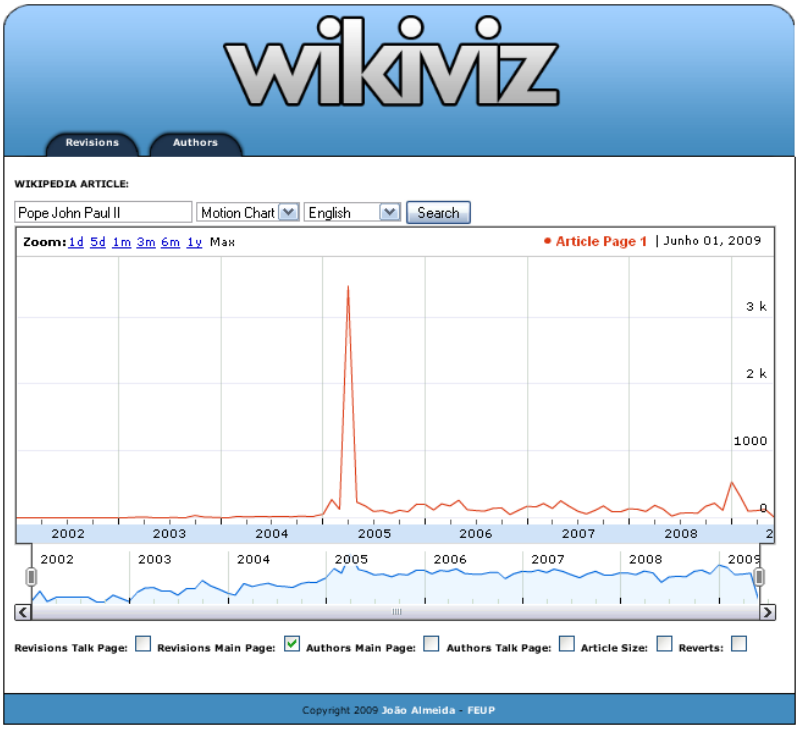

Figura 4.4: Historial do número de revisões do artigo *Papa João Paulo II*.

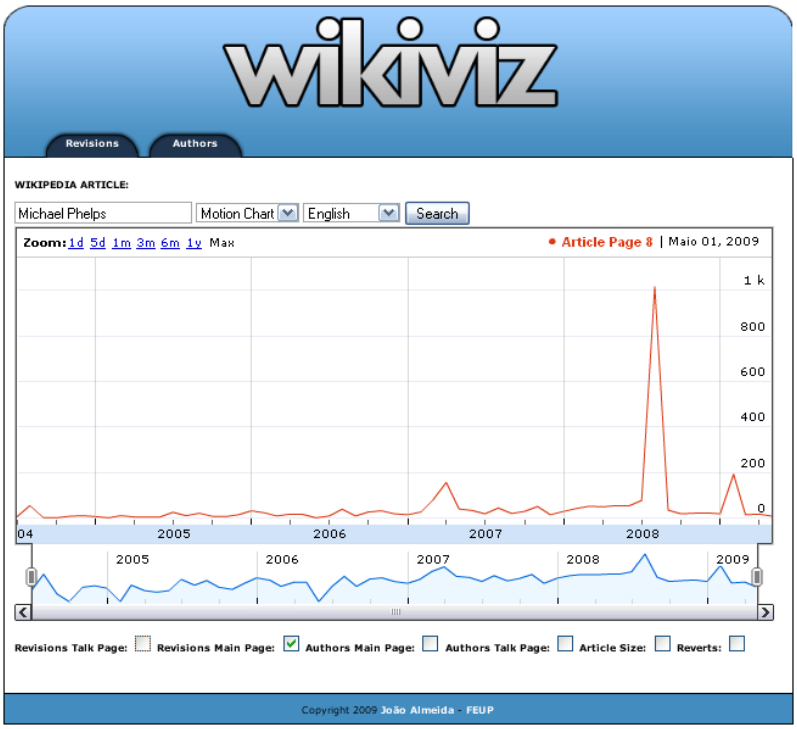

Figura 4.5: Historial do número de revisões do artigo *Michael Phelps*.

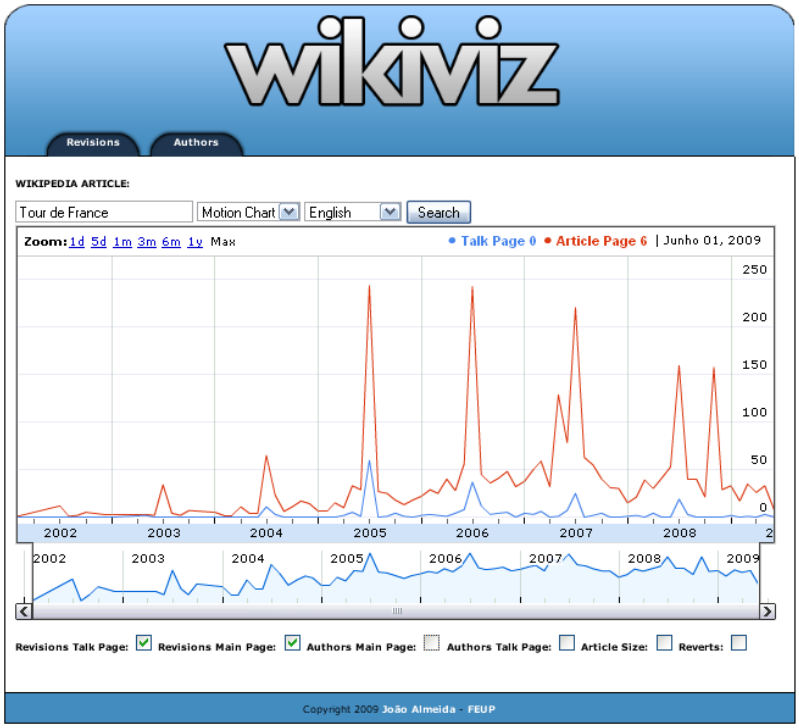

Figura 4.6: Historial do número de revisões do artigo Volta a França.

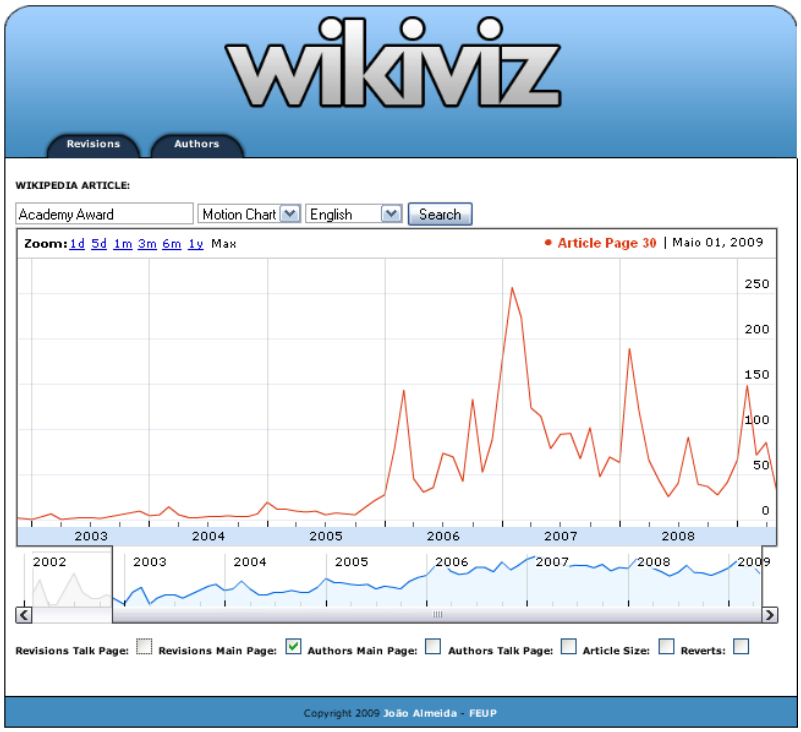

Figura 4.7: Historial do número de revisões do artigo dos Óscares.

opinião pública são alvos de muitas revisões por parte da comunidade da Wikipedia, como evidencia o exemplo da Figura 4.8 com o historial de *Barack Obama*.

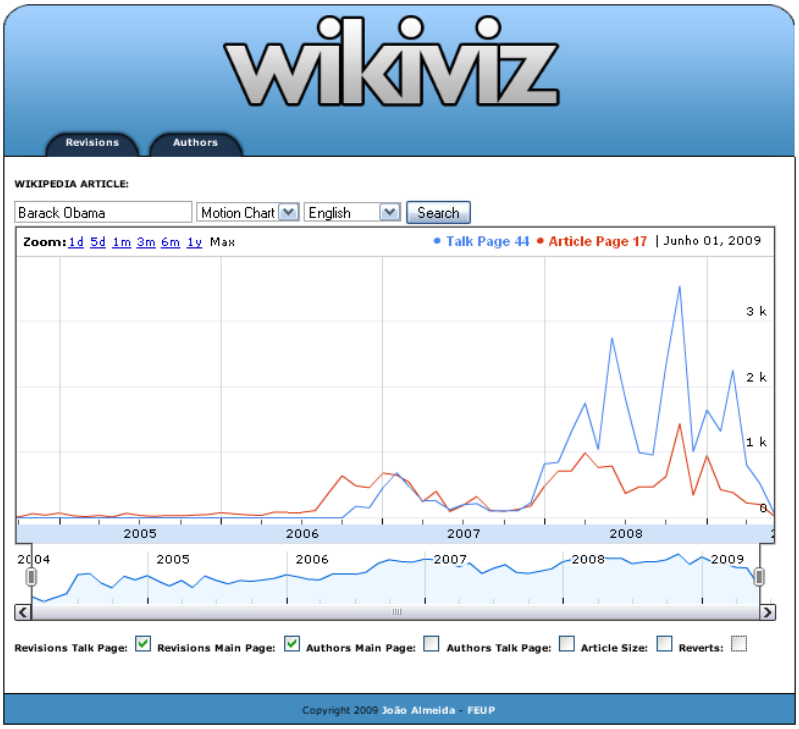

Figura 4.8: Página principal vs Página de Discussão do artigo *Barack Obama*.

Relativamente ao número de entradas nas páginas de discussão, é possível observar que artigos sobre temas mais polémicos ou sobre algumas figuras públicas têm mais discussão do que propriamente a edição da página principal. São exemplos deste tipo de actividade os artigos sobre *Barack Obama*, sobretudo a partir do momento em que se candidatou a presidência dos EUA, *Sarah Palin* ou ainda sobre o aborto e os ataques terrorista do 11 de Setembro, como se pode verificar nas Figuras 4.8 e 4.9.

Quanto a outros tipos de artigos não é fácil tirar conclusões imediatas sobre mais algum padrão de actividade. Existem artigos com alguns picos de revisões esporádicos, que estarão associados a acontecimentos da altura, e noutros simplesmente quase não existe quase qualquer discussão. No entanto, o que parece certo é a página de discussão, de vários artigos, ser cada vez mais utilizada. Na grande maioria dos exemplos exibidos, facilmente se repara que a página de discussão tem sido muito mais usada nos últimos anos.

Propriedades das revisões

O WikiViz permite ao utilizador visualizar a evolução do número de autores (anónimos e registados) das páginas principal e de discussão, do tamanho do artigo e do número de *reverts* executados ao longo do tempo.

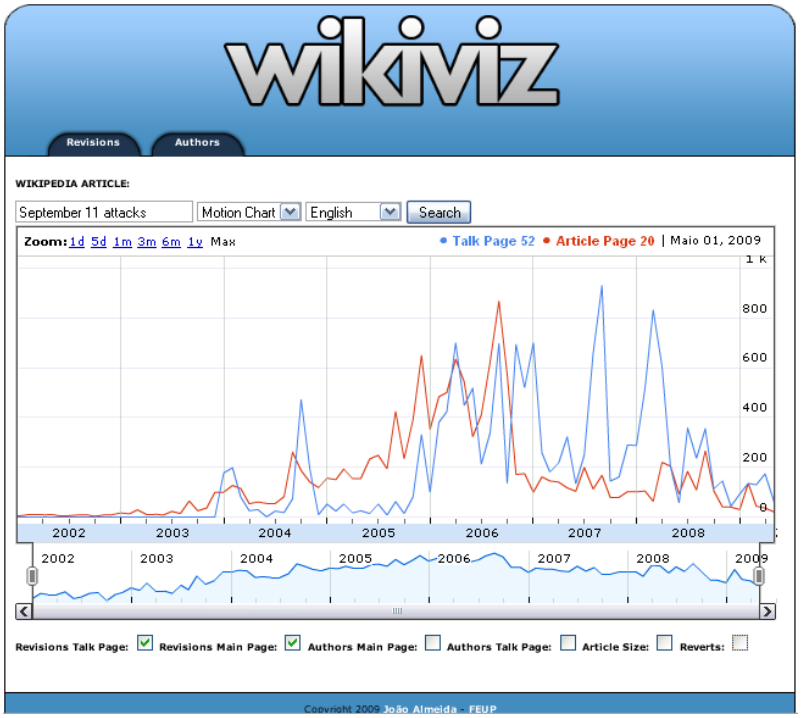

Figura 4.9: Número de revisões da página principal vs número de revisões da página de discussão do artigo *Ataques de 11 de Setembro*.

A distribuição do número de diferentes autores ao longo do tempo não tem um padrão sistemático, isto porque o número de revisões mensal tanto pode ser devido a um grande como a pequeno grupo de autores. No entanto, pode-se referir que nas alturas de acontecimentos marcantes sobre o tema de um determinado artigo, para além de aumentar o número de revisões, o número de autores também aumenta, o que significa o aumento do interesse da comunidade pelo tema em questão. Na Figura 4.10 encontra-se um exemplo que reflecte esta situação, ficando outros em anexo nas Figuras A.3 e A.4 no anexo A. Como podemos ver na figura, o número de autores acompanha o aumento ou a diminuição do número de revisões, para o artigo *Ataques de 11 de Setembro*.

Normalmente, com o aumento do número de revisões aumenta também o número de *reverts*, o que é compreensível, uma vez que o aumento do número de edições potencia um maior conflito entre autores e consequentemente a anulação de edições. Porém, a percentagem de *reverts* é menor em artigos que suscitam menos interesse à comunidade, independentemente do número de revisões aumentar ou não.

Igualmente através do WikiViz, podemos acompanhar a evolução do tamanho do artigo ao longo do tempo. A maior dificuldade para a análise desta propriedade é o facto alterações feitas antes de meados de 2007 não estarem registadas. Este *bug* da API não permite fazer uma análise muito rigorosa a muitos artigos, uma vez que muitos deles foram criados antes dessa data, pelo que não é possível ver toda a sua evolução. Pelo contrário, nos artigos cuja data de criação é posterior a 2007, pode fazer-se uma análise completa

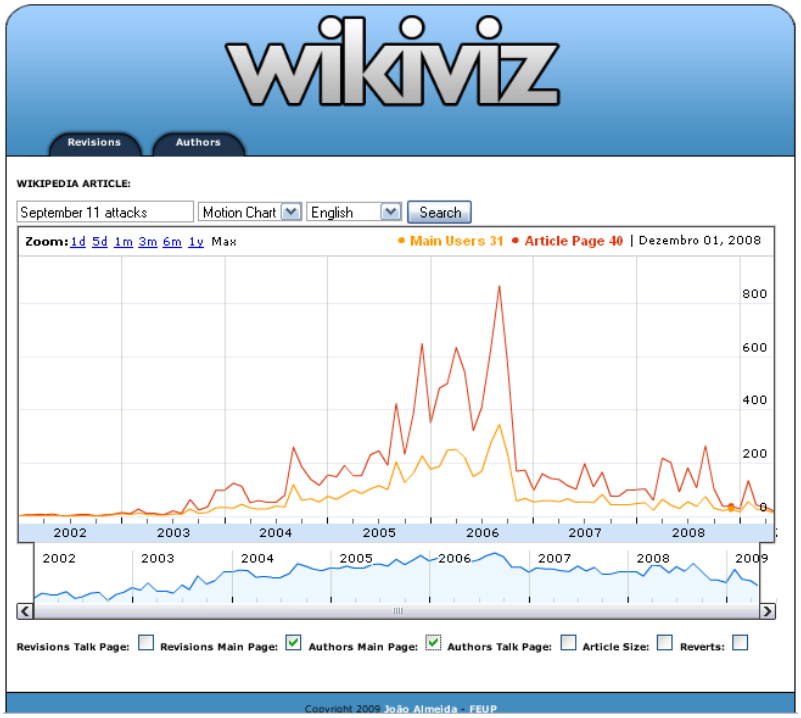

Figura 4.10: Historial do número de revisões e do número de autores ao artigo *Ataques de 11 de Setembro*.

da evolução do seu tamanho. O que se pode verificar na evolução desta propriedade em diversos artigos, é o aumento do tamanho a partir da altura de acontecimentos significativos. A título de exemplo podemos referir os artigos *Papa João Paulo II*, *Sarah Palin* (Figura 4.11) e *Barack Obama*. Em artigos sem grande impacto na sociedade o tamanho do artigo evoluiu de uma forma bastante uniforme, sem grandes mudanças.

Autoria das revisões

Na secção anterior foi abordada superficialmente a autoria das revisões. Nesta secção será feita uma análise mais detalhada, nomeadamente sobre os autores anónimos e registados, mas apenas da página principal.

Relativamente aos autores anónimos foi desenvolvida uma forma de visualização que faz a sua distribuição num mapa mundo, de acordo com a origem do autor. Após análise desta visualização para vários artigos chegou-se à conclusão que a grande maioria desses autores são oriundos dos Estados Unidos da América, quer para a Wikipedia inglesa quer para a Wikipedia noutros idiomas, como se verá mais a frente. Inglaterra, Canadá e Austrália são outros países cujos autores têm forte presença na contribuição a artigos. No entanto, com excepção dos Estados Unidos, verifica-se que o país com mais autores, corresponde ao que faz mais sentido atendendo ao tema do artigo. Por exemplo, sem analisar este tipo de visualização faria sentido admitir que a maior parte dos autores do

Figura 4.11: Evolução do tamanho do artigo Sarah Palin.

artigo Brasil fossem brasileiros e é o que realmente se verifica na Figura 4.12, pela cor mais escura com que o Brasil está representado.

Outros exemplos, remetidos para anexo, são os artigos sobre o Presidente da República Aníbal Cavaco Silva e sobre a Rainha Beatriz da Holanda, onde a maior parte dos autores é de Portugal e da Holanda, respectivamente. Porém existem artigos que não estão directamente ligados a um determinado país, participando autores oriundos de diversos países, nomeadamente os artigos que suscitam grande interesse como o *Aborto* (Figura 4.13), o *Google*, entre outros. Regra geral a Wikipedia inglesa recebe contribuições de autores distribuídos por todo o mundo.

Quanto aos utilizadores registados, a visualização criada foi uma nuvem de palavras, em que o tamanho da letra do nome do autor é directamente proporcional à sua contribuição nas revisões (Figura 4.14), podendo ter grande interesse para o utilizador. Para além disso, quando se clica num dos nomes pode-se verificar quais os artigos mais editados por esse autor, permitindo ao utilizador ter uma ideia dos interesses desse autor.

Finalmente foi criada uma visualização com o objectivo de se poder comparar o número de autores anónimos com o número de autores registados, ao longo do tempo. Esta visualização para alguns artigos é curiosa. Por exemplo, os artigos sobre Barack Obama ou Sarah Palin, quase não têm contribuições de autores anónimos. Pelo contrário, muitos outros artigos têm a contribuição de um grande número de autores anónimos, casos de José Sócrates, Portugal ou Vladimir Putin (Figura 4.15). Apesar da constatação destes dois factos, pode-se dizer que o peso dos utilizadores anónimos nas revisões aos artigos da

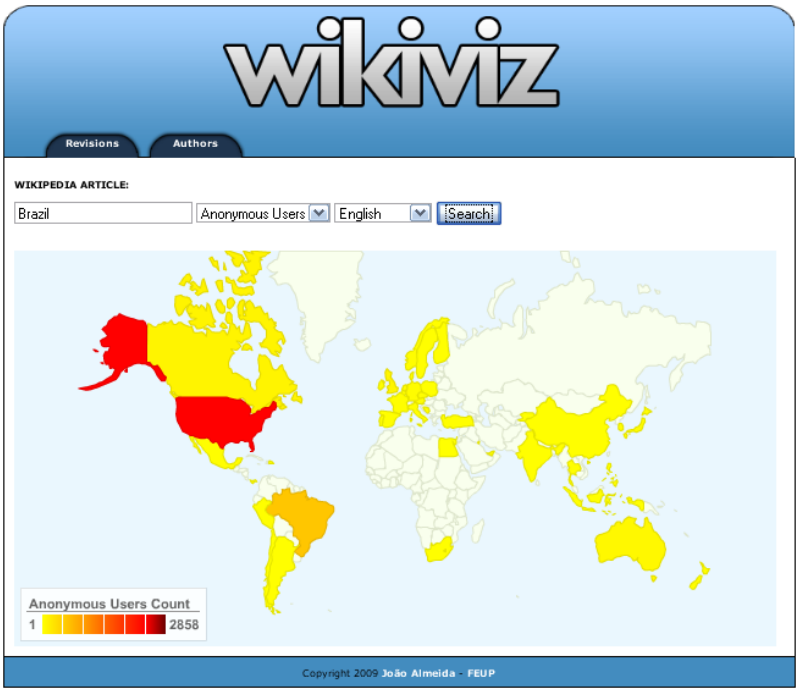

Figura 4.12: Distribuição geográfica dos autores no artigo Brasil.

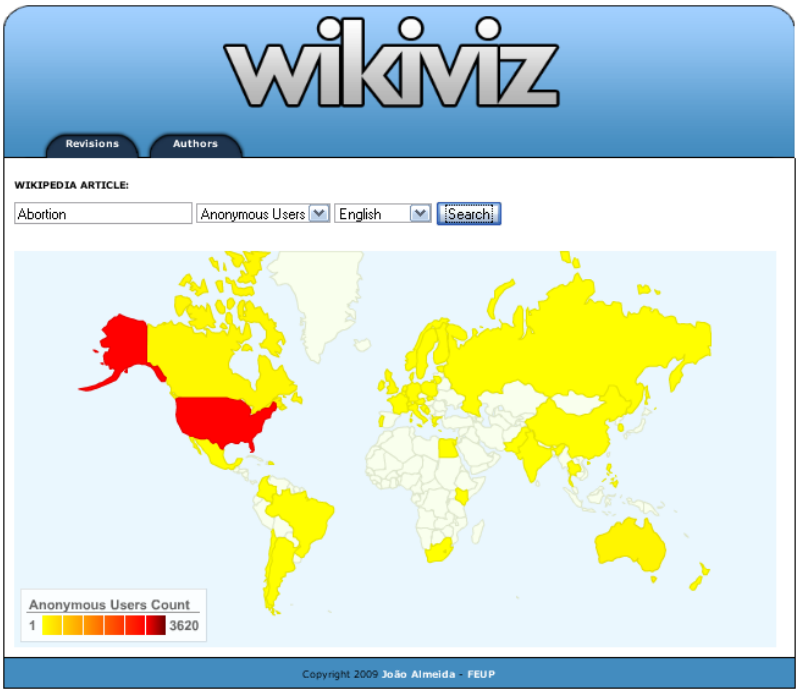

Figura 4.13: Distribuição geográfica dos autores no artigo Aborto.

Figura 4.14: Nuvem de palavras com os nomes do utilizadores registados do artigo Aborto.

Wikipedia é ainda significativo, por aquilo que foi pesquisado neste trabalho.

Análise nos diferentes idiomas

Esta funcionalidade tem como principal objectivo poder detectar diferenças de interesse entre as diferentes comunidades da Wikipedia, pelo que foram comparados diversos artigos em idiomas diferentes.

Na grande maioria das visualizações efectuadas, foi verificado que a Wikipedia inglesa é a mais activa no que diz respeito ao número de revisões a artigos e no uso das respectivas página de discussão. Nesta Wikipedia constatou-se, também, que o número de autores é superior em relação à Wikipedia noutros idiomas. Existem excepções a esta tendência, nomeadamente quando se trata de artigos cujo interesse é muito maior a nível de um país do que a nível internacional. A título de exemplo, pode-se referir o artigo de Maria de Lurdes Rodrigues onde o número de revisões é maior na Wikipedia portuguesa do que na inglesa, como se pode ver comparando as Figuras 4.16 e 4.17.

Quanto à geo-localização dos autores anónimos passa-se um fenómeno semelhante, ou seja, são em maior número na Wikipedia inglesa do que na Wikipedia nos outros idiomas. Independentemente do idioma pesquisado a maioria dos autores continua a ser oriunda dos Estados Unidos da América. De salientar, no entanto, que em países cujo idioma é menos falado pelo mundo, casos do holandês ou do polaco, a autoria das revisões resume-se praticamente ao próprio país (apesar de os autores dos Estados Unidos em alguns casos

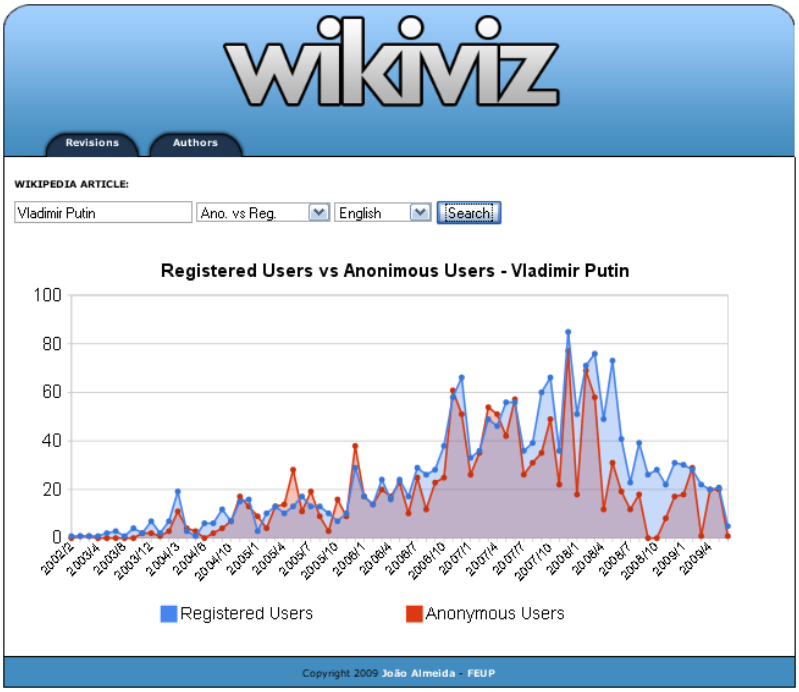

Figura 4.15: Historial do número de autores anónimos e de autores registados no artigo Vladimir Putin.

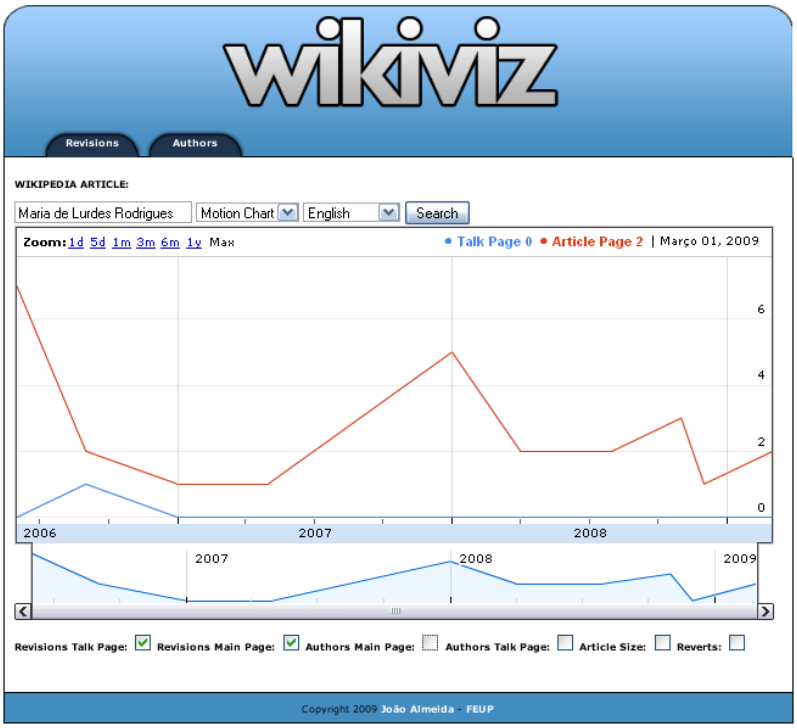

Figura 4.16: Historial do número de revisões do artigo Maria de Lurdes Rodrigues na Wikipedia inglesa.

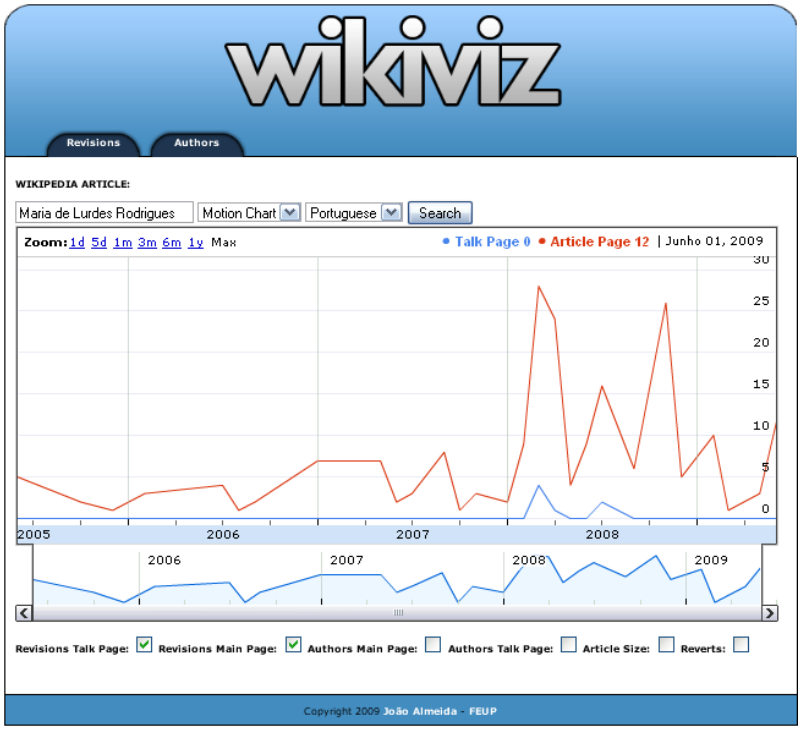

Figura 4.17: Historial do número de revisões do artigo Maria de Lurdes Rodrigues na Wikipedia portuguesa.

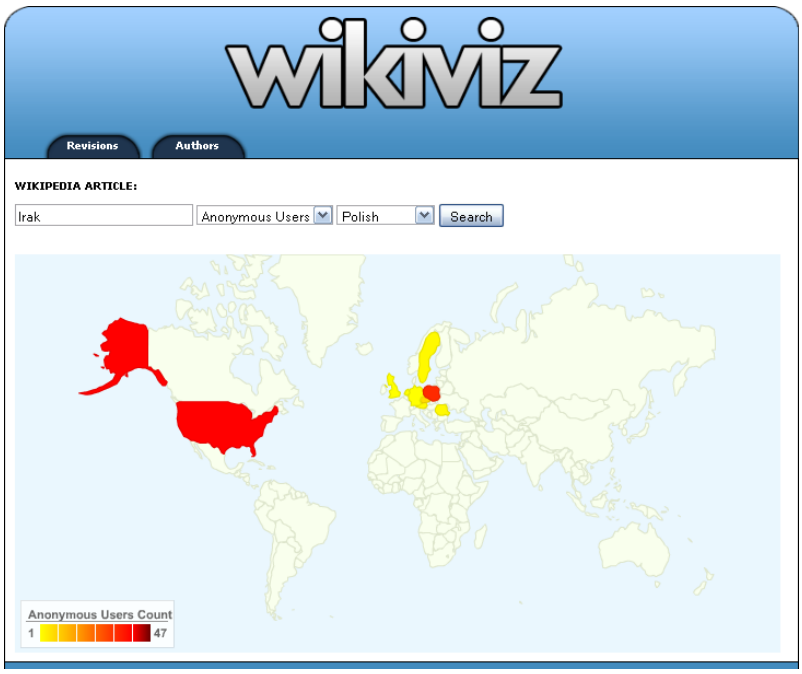

Figura 4.18: Distribuição geográfica dos autores na Wikipedia polaca para o artigo Iraque.

terem uma contribuição ainda considerável), como está representado na Figura 4.18, com o caso da geo-localização do artigo Iraque na Wikipedia polaca.

Capítulo 5

Integração com um armazém de dados

O desenvolvimento da ferramenta proposta neste trabalho levou, por questões de eficiência, a armazenar uma grande quantidade de informação histórica extraída da Wikipedia. Para este efeito foi usada uma base de dados relacional MySQL. Para sistematizar esta colecção auxiliar, foi estudada a aplicação do conceito de armazém de dados a esta colecção. Neste capítulo pretende-se expor esse estudo, explicando o que é um armazém de dados, os seus principais objectivos e o modelo a aplicar neste caso.

5.1 Definição e objectivos de um armazém de dados

Um armazém de dados é um repositório de dados, alojados remotamente numa máquina, dedicados a um tema, não voláteis e habitualmente associados a um período de tempo. Tipicamente, são usados em grandes empresas, onde se lida com grandes volumes de informação de estruturas complexas e que evolui muito ao longo do tempo. A reorganização desta informação numa estrutura típica de um armazém de dados facilita a análise dos mesmos, sendo uma importante ferramenta de apoio a decisões.

Os armazéns de dados têm como principais objectivos tornar a informação mais acessível, elevar a sua qualidade (tornando-a imune a modificações e mais completa) e constituir uma fonte robusta e adaptativa para novas pesquisas.

No caso do WikiViz, os dados guardados são todos relativos às revisões feitas aos artigos da Wikipedia. Estes dados inicialmente estavam organizados de uma forma menos eficiente, correspondendo cada tabela a uma diferente visualização. A reorganização dos dados num armazém de dados permitiu entre outras vantagens, reorganizar a base de dados, elevando a qualidade da informação e permitir diferentes visualizações históricas da informação, sem requerer novas *queries* via API. Adoptando esta estrutura é possível, por exemplo, visualizar apenas as revisões feitas a um artigo aos fins de semana ao longo de um ou mais anos, ou ainda mostrar se há mais revisões feitas de manhã ou à noite. A

possibilidade de diferentes visualizações seria de grande utilidade para pessoas ligadas à sociologia e outras áreas, que poderiam analisar detalhadamente a actividade num artigo da Wikipedia do seu interesse.

5.2 Estruturas de um armazém de dados

Os armazéns de dados (AD) são acedidos por *data marts* que são um subconjunto de um AD dedicado a uma área de interesse em particular (por exemplo: vendas, stock).

Os *data marts* são formados por tabelas de factos, que têm como principal característica a presença de dados redundantes, de forma a se obter um melhor desempenho no acesso aos dados. As tabelas de factos podem ser construídas de acordo com os passos seguintes [23]:

1. Definir um *data mart* e consequentemente escolhe de uma fonte de dados.
2. Definir a granularidade da tabela de factos, devendo ser tão fina quanto possível de forma a não se perder informação e a obter-se um desenho robusto.
3. Definição das dimensões a utilizar, que são determinadas pela granularidade escolhida. As dimensões representam todas as possíveis descrições que tomam valores singulares no contexto de cada medida. As dimensões são tabelas que estão associadas à tabela de factos, uma vez que as chaves estrangeiras destas são referentes às chaves primárias das tabelas dimensionais.
4. Definição dos factos, são as medidas para cada uma das linhas da tabela e devem estar de acordo com a granularidade escolhida.

Este tipo de tabela é usado nos modelos dimensionais existentes de um armazém de dados, nomeadamente o modelo em estrela, onde a tabela de factos se encontra no meio da representação e as outras tabelas, que a rodeiam, representam as suas dimensões, como se pode ver na Figura 5.1.

5.3 O armazém de dados do WikiViz

No WikiViz foi usado o modelo em estrela uma vez que, entre outras vantagens, é flexível para suportar a introdução de novos dados, bem como mudanças que ocorram no projecto. Por exemplo as tabelas de factos e as dimensões podem ser alteradas acrescentando novas colunas às tabelas. O modelo construído apresenta-se na Figura 5.2.

A tabela de factos construída seguiu os passos enumerados na secção anterior. Desta forma, inicialmente foi definido o *data mart*, isto é, o subconjunto do armazém de dados, que neste caso corresponde à totalidade dos dados, uma vez que apenas nos interessam as métricas relativas às revisões, feitas nos artigos da Wikipedia. Posteriormente foi definida a granularidade da tabela de factos, correspondendo assim cada linha desta tabela a apenas uma revisão a um artigo. Em seguida foram definidas as dimensões, que acabam por

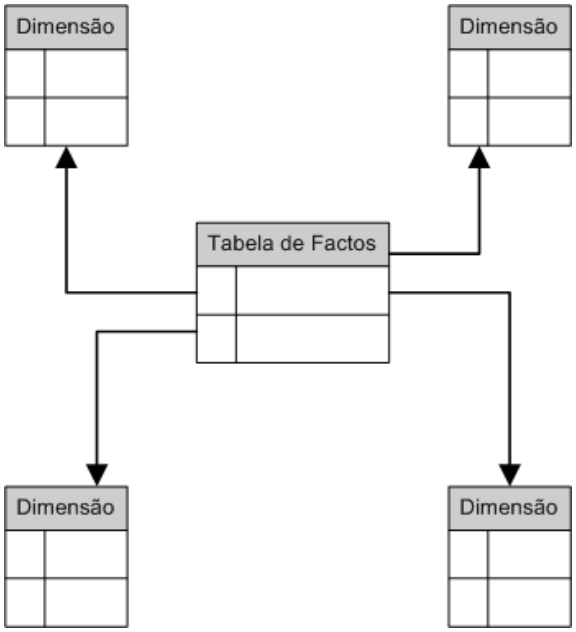

Figura 5.1: Modelo em estrela de um armazém de dados.

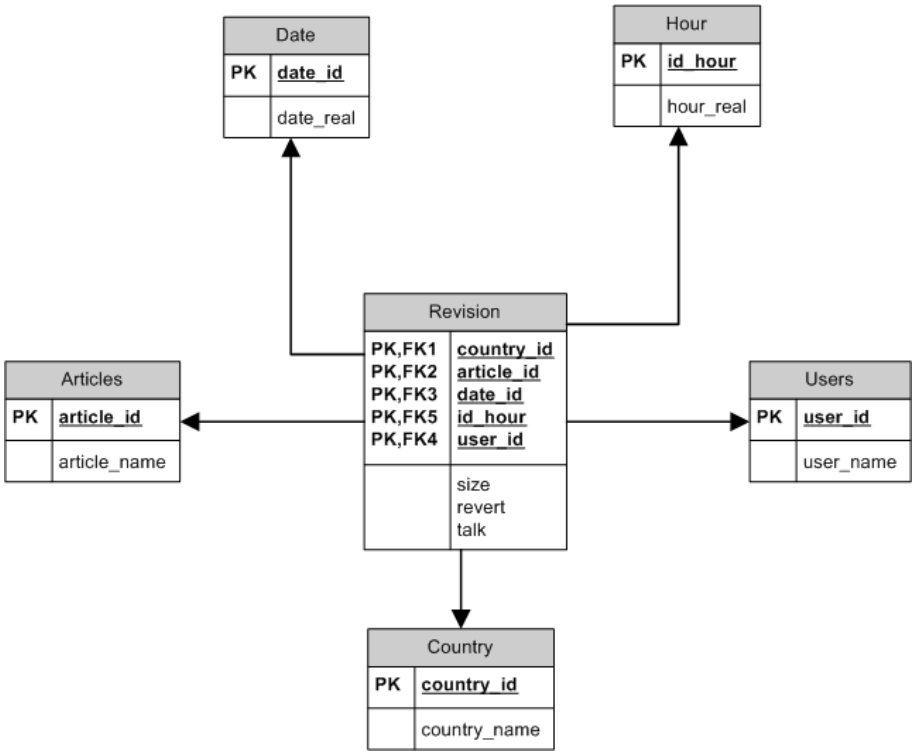

Figura 5.2: Estrutura do armazém de dados no WikiViz.

estar dependentes do nível de granularidade escolhida. Ou seja, para que cada linha seja *compreendida* são necessárias as seguintes dimensões: data, hora, nome do artigo, autor e país do autor da revisão. Finalmente foram definidos os factos para cada linha da tabela, sendo eles: o tamanho do artigo após a revisão (em Kbytes), se a revisão foi do tipo *revert*, ou seja, quando uma revisão é alterada para a anterior, e ainda se a revisão em questão é referente à página do artigo ou à sua página de discussão.

A grande vantagem deste tipo de estrutura é ser escalável. Isto significa que outras colunas podem ser adicionadas às tabelas dimensionais, sem qualquer tipo de prejuízo para os dados já guardados. Como foi referido no início deste capítulo, para fazer uma análise sobre se existiriam mais revisões à semana ou ao fim de semana, seria apenas necessário acrescentar uma coluna à tabela date que guardasse o dia da semana. Acrescentar este campo não degradaria a informação já guardada. Seria apenas necessário executar uma nova *query* que fizesse a leitura do dia da semana de cada revisão.

Capítulo 6

A ferramenta de visualização

Para dar inicio à fase de implementação do Wikiviz foi necessário escolher as tecnologias a usar e determinar a forma de como os dados recolhidos nas consultas iriam ser guardados. Posteriormente a essa escolha foi definida uma estrutura visual da ferramenta e consequentemente deu-se inicio à implementação das funcionalidades propostas anteriormente. A ferramenta desenvolvida está disponível *online* em: http://irlab.fe.up.pt/p/wikiviz/

6.1 Tecnologias escolhidas

O WikiViz tem por base uma página Web e por esse motivo a escolha das linguagens de programação recaiu sobre aquelas que estão mais vocacionadas para essa área. Do lado do cliente foi usado HTML, CSS e JavaScript, mais concretamente a biblioteca JQuery que simplifica a interacção com HTML bem como a escrita do código, permite a criação de animações mais facilmente, entre outras vantagens. Um exemplo concreto na implementação do WikiViz, que utilizou JQuery, foi a função de *auto-complete* na pesquisa de artigos da Wikipedia. Relativamente ao lado do servidor foram desenvolvidos scripts em Perl e em PHP. O primeiro foi usado para extracção dos dados, processamento de XML e na escrita dessa informação na base de dados. Já o PHP foi usado, essencialmente, para a construção das páginas Web.

A ferramenta de visualização seleccionada para o desenvolvimento deste trabalho foi a Google Visualization, uma vez que os resultados apresentados foram satisfatórios. São visualmente agradáveis, com possibilidade de interacção e o carregamento dos dados para a construção dos gráficos é bastante rápido. Para além disso a sua integração com o resto do trabalho foi relativamente fácil.

Relativamente aos dados recolhidos nas pesquisas efectuadas, estes são guardados em tabelas numa base de dados MySQL, como foi referido no capítulo anterior.

6.2 Potencialidades do WikiViz

Depois da visualização de resultados de várias consultas sobre artigos da Wikipedia, pode-se desde logo evidenciar algumas das potencialidades desta ferramenta. Deve-se salientar o facto de os resultados serem constantemente actualizados e se permitirem combinações de vários traçados, ao longo do tempo, relativos às propriedades das revisões, tais como o número de revisões da página principal ou de discussão, o respectivo número de autores, o tamanho do artigo e o número de vezes que os artigos foram revertidos. Para além disto, existe a possibilidade de analisar de que parte do mundo são os autores das revisões a um artigo. Este indicador não traduz, na realidade, a localização da totalidade dos autores, uma vez que apenas é possível para autores anónimos. No entanto, este tipo de autoria, em certos artigos, tem uma importante contribuição, pelo que acaba por funcionar como uma boa previsão da totalidade dos autores.

6.3 Lógica de funcionamento

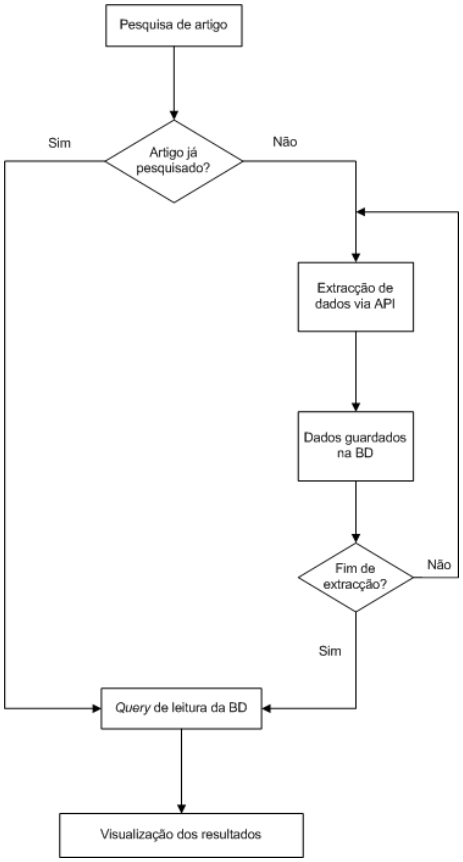

Figura 6.1: Fluxograma da lógica de funcionamento do WikiViz.

O Wikiviz tem uma lógica de funcionamento semelhante à do esquema apresentado na Figura 6.1. Isto é, após ser inserido o artigo a pesquisar, do lado do servidor é verificado se

este já foi pesquisado anteriormente ou não. No caso de ter sido, o retorno da visualização é imediato, uma vez que é apenas necessário fazer a leitura dos dados alojados na base de dados. Pelo contrário, se o artigo é pesquisado pela primeira vez terá de ser feita a extracção de todo o seu historial de revisões, recorrendo para tal à API, e posteriormente a sua escrita na base de dados. A visualização dos resultados pelo utilizador é, neste caso, mais lenta, devido a todo o processamento inerente, nomeadamente o de leitura de ficheiros XML e o da escrita dessa mesma informação na base de dados. Naturalmente que isto apenas acontece na primeira pesquisa de um artigo, sendo nas seguintes pesquisas quase instantânea a visualização dos resultados.

De realçar que após a primeira pesquisa de um artigo e o respectivo registo do mesmo na base de dados é feita uma constante actualização do seu historial das revisões, mais concretamente a cada hora, através do uso de uma tarefa automática (Cron) que executa um *script* desenvolvido para o efeito. Esta foi uma forma de assegurar que são facultados resultados constantemente actualizados ao utilizador e sem que este se aperceba de todo o processamento inerente à actualização dos dados.

Capítulo 7

Conclusões

Neste trabalho foi apresentado o WikiViz, uma ferramenta integrada numa página *web* para exploração de diferentes propriedades das revisões a artigos da Wikipedia. Com o auxilio do WikiViz apresentaram-se várias informações valiosas, que sem a sua ajuda, ficariam escondidas do utilizador da Wikipedia. Num mesmo gráfico é possível visualizar diferentes traçados, de forma a melhor atingir a actividade da comunidade na Wikipedia. Entre outros traçados, deve-se salientar a possível comparação entre o número de revisões a um artigo e as revisões à sua página de discussão. Ao mesmo tempo, existe a possibilidade de traçar o gráfico com o número de revisões que foram revertidas, quer por vandalismo, quer por simples discordância. Outra funcionalidade a salientar é a geo-localização dos autores anónimos. Esta visualização inovadora dá uma noção ao utilizador dos países que fazem parte da comunidade e quais deles têm maior peso na contribuição ao artigo. Apesar de representar apenas uma parte da totalidade dos autores existentes, esta visualização é significativa uma vez que em muitos artigos o peso dos autores anónimos é tão grande como a dos autores registados. Outra funcionalidade de grande valor do WikiViz é a possibilidade de todas as visualizações existentes poderem ser feitas sobre a Wikipedia nos diferentes idiomas disponíveis. A principal conclusão retirada é que a Wikipedia inglesa é a que mais contribuições sofre e a que reúne o maior número de autores. Este facto deve-se, essencialmente, à língua inglesa ser usada em grande parte do mundo. A estas visualizações pode-se acrescentar o facto de todas elas serem constantemente actualizadas, independentemente de os artigos serem pesquisados ou não, o que melhora a eficiência do WikiViz e para o utilizador, que não tem de esperar pela actualização.

A adopção de um armazém de dados para este trabalho revelou-se compensador, trazendo várias vantagens na fase de implementação, nomeadamente na simplificação de muitas das *querys*, no facto de os dados estarem guardados de uma forma mais intuitiva e de a estrutura adoptada ser escalável. Este último aspectos significa que podem ser adicionadas novas colunas às tabelas dimensionais sem que os dados guardados até então fiquem alterados, permitindo novas pesquisas.

Como trabalho futuro seria interessante investigar a actividade da comunidade ao nível

das categorias de artigos. Uma possibilidade seria perceber quais são as categorias com maior número de artigos e verificar alterações consoante a alteração do idioma.

Seria também interessante aproveitar o armazém de dados criado e serem implementadas novas visualizações. Um exemplo seria investigar se as revisões são efectuadas mais de dia ou à noite, se são feitas mais à semana ou ao fim de semana ou ainda em que semanas do ano existe mais actividade na Wikipedia.

A nível de implementação poder-se-ia experimentar outras formas de visualização usando para o efeito outras ferramentas disponíveis. Para além disso seria interessante fazer a extracção dos dados num outro formato, seja JSON ou ficheiro de texto, de forma a se poder verificar se o processo de leitura dos dados é mais rápida ou não.

Anexo A

Anexo

Na Figura A.1 está representado o historial de revisões do artigo Gripe Suína, cujo pico de revisões surge na altura da possibilidade de uma pandemia desta doença. Na Figura A.2 apresenta-se o historial de revisões do artigo de Luciano Pavarotti, em que o ponto alto de revisões coincide com a altura do seu falecimento. Em ambos os casos a actividade de edição era muito baixa, com a excepção das alturas relativas aos acontecimentos referidos.

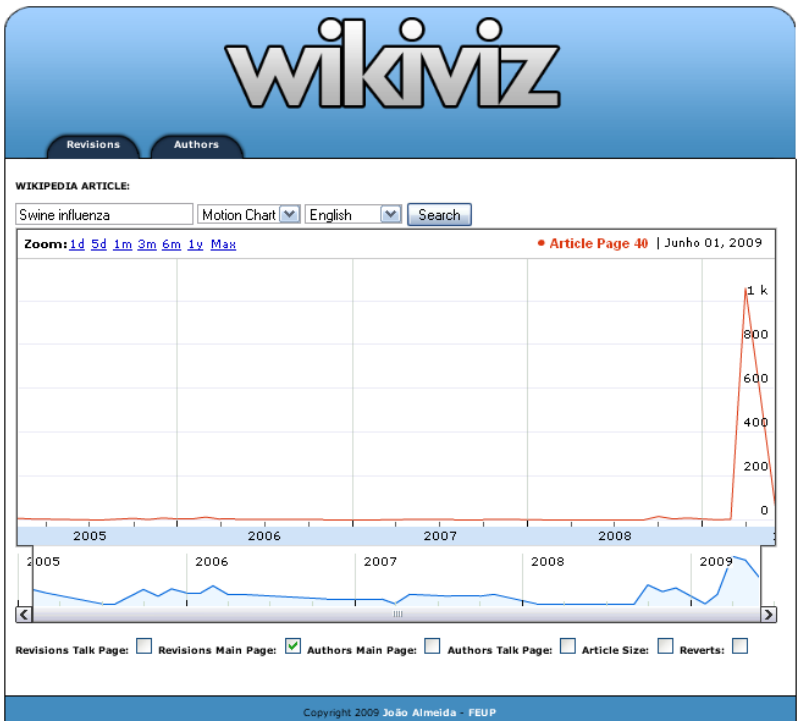

Figura A.1: Historial de revisões do artigo Gripe Suína na Wikipedia inglesa.

Nas Figuras A.3 e A.4 são apresentados dois exemplos onde existem picos significativos de revisões e o respectivo o número de autores nessas alturas também sobe significativamente. Esta tendência leva a crer que o interesse da comunidade da Wikipedia aumenta nas alturas de factos relevantes sobre o artigo em questão.

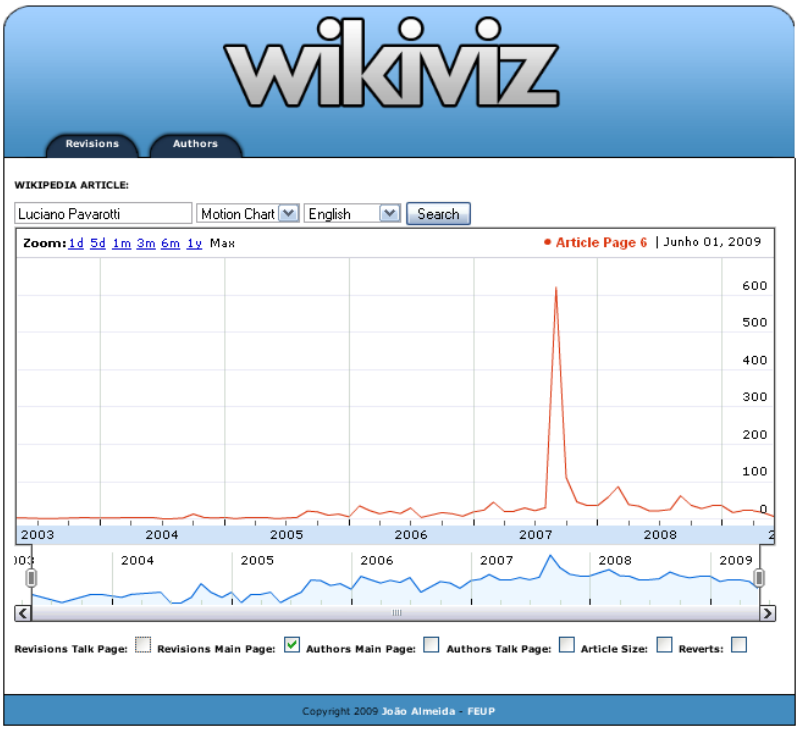

Figura A.2: Historial de revisões do artigo Luciano Pavarotti na Wikipedia inglesa.

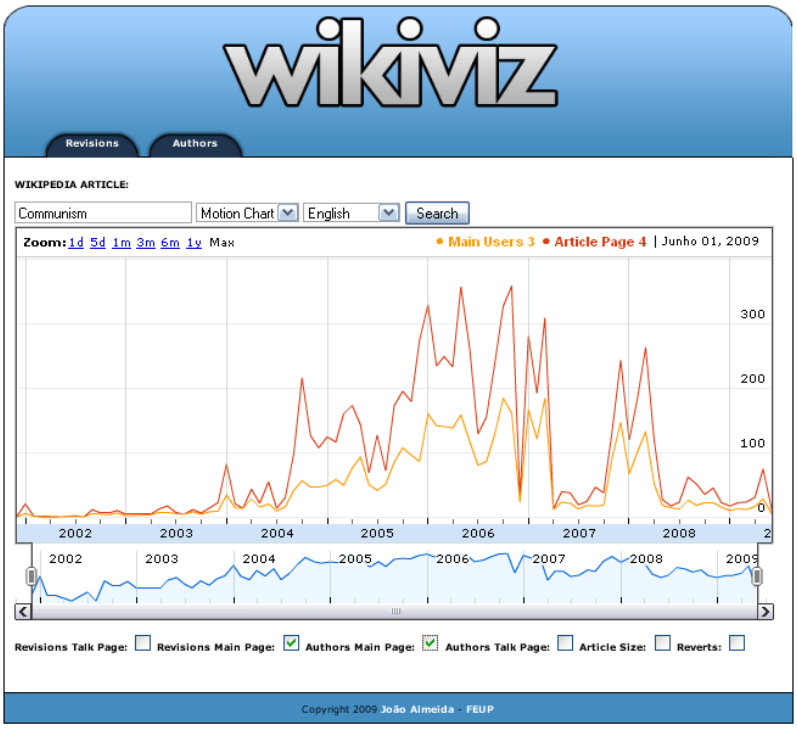

Figura A.3: Historial do número de revisões e do número de autores do artigo Comunismo na Wikipedia inglesa.

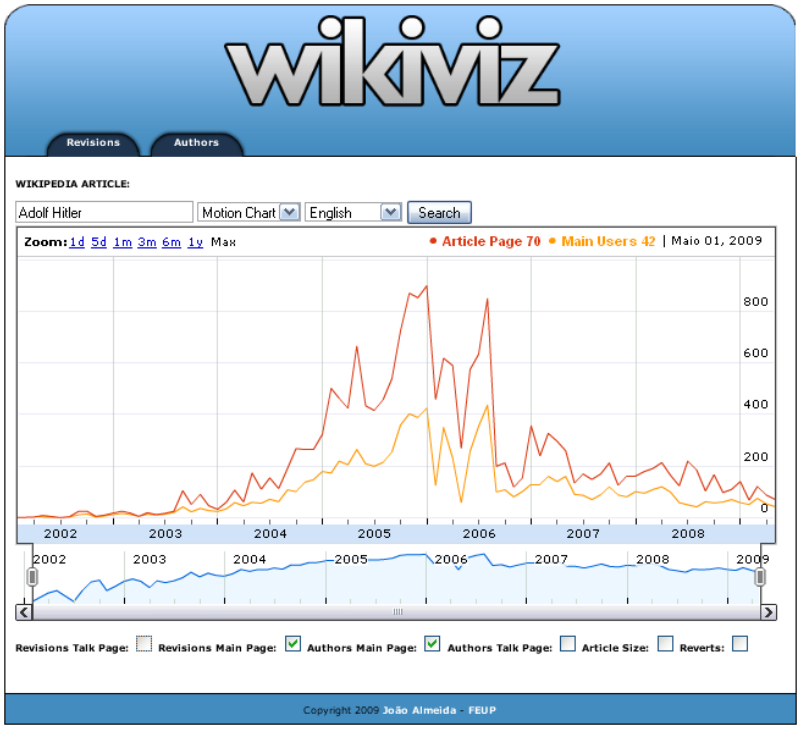

Figura A.4: Historial do número de revisões e do número de autores do artigo Adolf Hitler na Wikipedia inglesa.

Nas Figuras A.5 e A.6 podemos verificar a nacionalidade dos autores anónimos para os artigos Aníbal Cavaco Silva e Rainha Beatriz da Holanda, para a Wikipedia inglesa. Excluindo à partida os Estados Unidos da América, os países com mais autores são os esperados de início. Ou seja, no artigo do Presidente da República de Portugal, Aníbal Cavaco Silva, os autores são na sua maioria de Portugal. O mesmo acontece para o artigo da Rainha Beatriz da Holanda, sendo a maioria dos seus autores oriundos desse mesmo país. Para o primeiro caso existem 70 ocorrências oriundas dos Estados Unidos contra 22 de Portugal e para o segundo caso existem 260 ocorrências dos Estados Unidos enquanto que na Holanda existem 68.

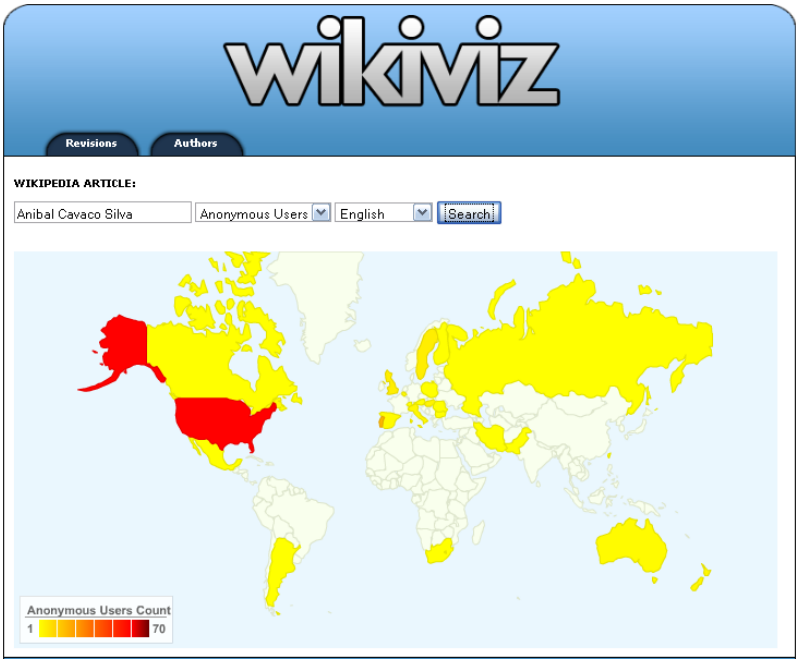

Figura A.5: Distribuição geográfica dos autores anónimos no artigo Aníbal Cavaco Silva.

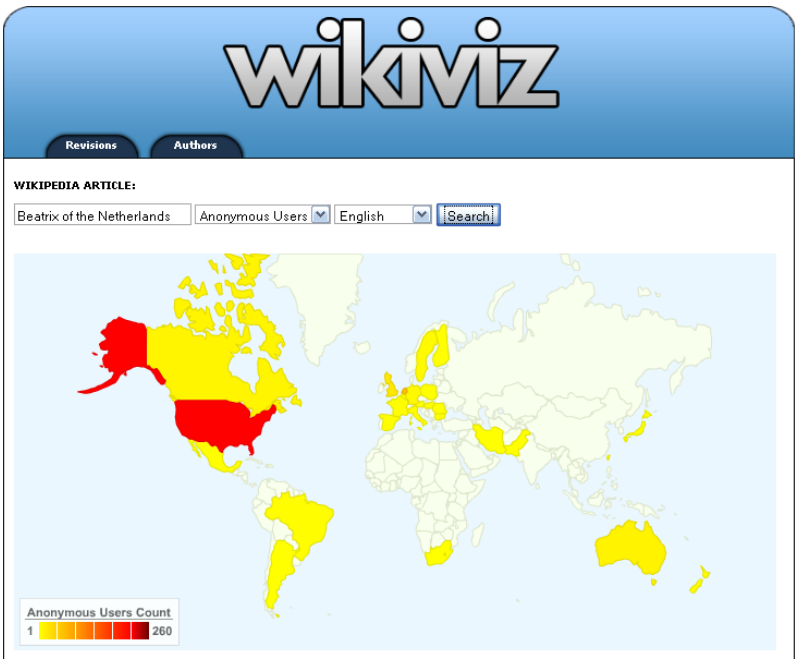

Figura A.6: Distribuição geográfica dos autores anónimos no artigo Rainha Beatriz da Holanda.

Referências

[1] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. *CHI '04: Proceedings of the 2004 conference on Human factors in computing systems*, ACM Press:575–582.

[2] B. Suh, E. H. Chi, A. Kittur, and B. A. Pendleton. Lifting the veil: improving accountability and social transparency in wikipedia with wikidashboard. *CHI'08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1037–1040, 2008.

[3] J. Gawryjo Lek and P. Gawrysiak. The analysis and visualization of entries in wiki services. *Advances in Intelligent Web Mastering*, pages 118–123, 2007.

[4] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: conflict and coordination in wikipedia. *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 453–462, 2007.

[5] Ulrik Brandes and Jürgen Lerner. Visual analysis of controversy in user-generated encyclopedias. *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST ' 07), to appear.*

[6] Michael Ogawa. Code swarm. [Consultado em 3 de Janeiro de 2009]. Disponível em http://vis.cs.ucdavis.edu/~ogawa/codeswarm/.

[7] Anonymous. Wikipedia - wikipedia, the free encyclopedia. [Consultado em 9 de Junho de 2009]. Disponível em http://en.wikipedia.org/wiki/Wikipedia.

[8] MediaWiki. Wikipedia api. [Consultado em 29 de Maio de 2009]. Disponível em http://en.wikipedia.org/w/api.php.

[9] S. Nunes, C. Ribeiro, and G. David. WikiChanges - exposing wikipedia revision history. *WikiSym'08*, ACM Press, 2008.

[10] Antanas Marcelionis. Amcharts. [Consultado em 3 de Janeiro de 2008]. Disponível em http://www.amcharts.com/.

[11] WikiChecker. [Consultado em 27 de Maio de 2009]. Disponível em http://en.wikichecker.com/.

[12] WikiTrends. [Consultado em 27 de Maio de 2009]. Disponível em http://users.student.lth.se/dt05jg2/wikitrends/.

[13] Small Batch Inc. WikiRank. [Consultado em 27 de Maio de 2009]. Disponível em http://wikirank.com/en.

[14] Wikipedia User Access Levels. [Consultado em 9 de Junho de 2009]. Disponível em http://en.wikipedia.org/wiki/Wikipedia:User_access_levels.

[15] Wikistats. [Consultado em 9 de Junho de 2009]. Disponível em http://dammit.lt/wikistats/.

[16] Wikimedia Foundation - Board of Trustees. [Consultado em 2 de Junho de 2009]. Disponível em http://wikimediafoundation.org/wiki/Board_of_Trustees.

[17] Geo::ip. [Consultado em 23 de Março de 2009]. Disponível em http://search.cpan.org/~borisz/Geo-IP-1.38/lib/Geo/IP.pm.

[18] Xml/Swf charts. [Consultado em 3 de Janeiro de 2009]. Disponível em http://www.maani.us/.

[19] Open Flash Chart. [Consultado em 3 de Janeiro de 2009]. Disponível em http://teethgrinder.co.uk/open-flash-chart-2.

[20] Black Box Chart. [Consultado em 3 de Janeiro de 2008]. Disponível em http://www.blackboxchart.com/documentation.php.

[21] Google. Google visualization. [Consultado em 23 de Maio de 2009]. Disponível em http://code.google.com/intl/pt-PT/apis/visualization/documentation/gallery.html.

[22] Antanas Marcelionis. Ammap. [Consultado em 3 de Janeiro de 2009]. Disponível em http://www.ammap.com/.

[23] Gabriel David. *Apontamentos da disciplina Armazéns de Dados*. Faculdade de Engenharia da Universidade do Porto, 2009.