Uso da Wikipédia na Investigação em Informática

Sérgio Nunes

Faculdade de Engenharia
Universidade do Porto

Academia Wikipédia
16 Abril 2010 | Porto, Portugal

Sumário

1. Investigação sobre/com a Wikipédia.
2. Tópicos e áreas de investigação.
3. Projecto WikiChanges.
4. Recursos e ferramentas existentes.

Investigação / Wikipédia

Estudos sobre a Wikipédia

Estudos **focados na compreensão da Wikipédia**, agrupados em dois tipos: (1) produção e qualidade do conteúdo, (2) aspectos sociais (ex. uso, edição).

Estudos com base na Wikipédia

Estudos que **fazem uso dos recursos disponibilizados** para resolver problemas gerais, não limitados à Wikipédia. Alguns exemplos: classificação de textos, sumariação, tradução automática, etc.

Qualidades da Wikipédia (1)

Dimensão e Diversidade

Mais de 15 milhões de artigos em mais de 250 idiomas (3,2 em Inglês; 0,5 em Português). Mais de 270 mil sub-categorias organizadas sob 11 categorias principais.

Qualidade

Grande diversidade na qualidade dos artigos. Pode ser estimada com base no número e conteúdo das revisões, e nas categorias a que um artigo pertence.

Qualidades da Wikipédia (2)

Importância e Impacto

É um fenómeno de importância mundial, justificando, por si só, a investigação. Os dados disponíveis permitem não só estudos ao nível do conteúdo dos artigos mas também ao nível das interacções humanas associadas às revisões.

Acessibilidade

Fácil acesso aos dados através de arquivos disponibilizados pela Wikipédia ou terceiros (ex. DBpedia). Acesso programático a dados e funções com a API do MediaWiki.

Qualidades da Wikipédia (3)

Reprodutibilidade

A reprodutibilidade é um aspecto crucial do método científico. O uso da Wikipédia permite a repetição das experiências por outras equipas e a comparação dos resultados.

Áreas de Investigação

Processamento de Linguagem Natural

Uso da Wikipédia para resolver tarefas associadas ao processamento automático de linguagem natural: proximidade semântica de termos, desambiguação semântica, ...

Recuperação de Informação

Uso da Wikipédia para melhorar a recuperação de informação noutros contextos: expansão de interrogações, sumariação de documentos, resposta a perguntas, categorização, ...

Extracção de Informação

Uso da Wikipédia para extracção de informação estruturada a partir de dados não estruturados: identificação de entidades mencionadas, ...

Alguns Exemplos

Tradução Automática

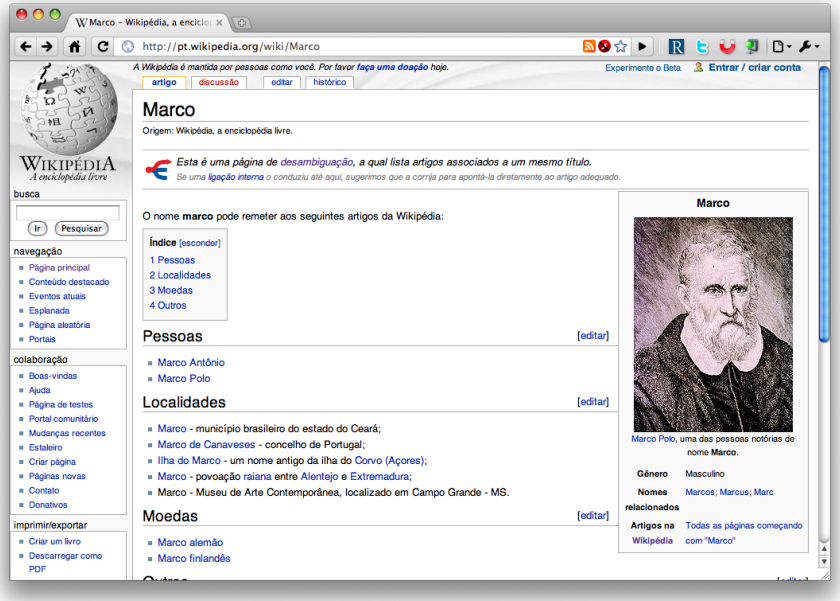
A existência de colecções paralelas em diferentes idiomas torna possível desenvolver sistemas de tradução automática.

Maracujá

パッションフルーツ

Desambiguação Semântica Automática

As páginas de redireccionamento e desambiguação da Wikipédia são recursos importantes na resolução automática de problemas de polissemia e sinonímia.



Extracção de Informação

A Wikipédia é rica em elementos estruturados que possibilitam a prospecção automática de conceitos, relações e factos.

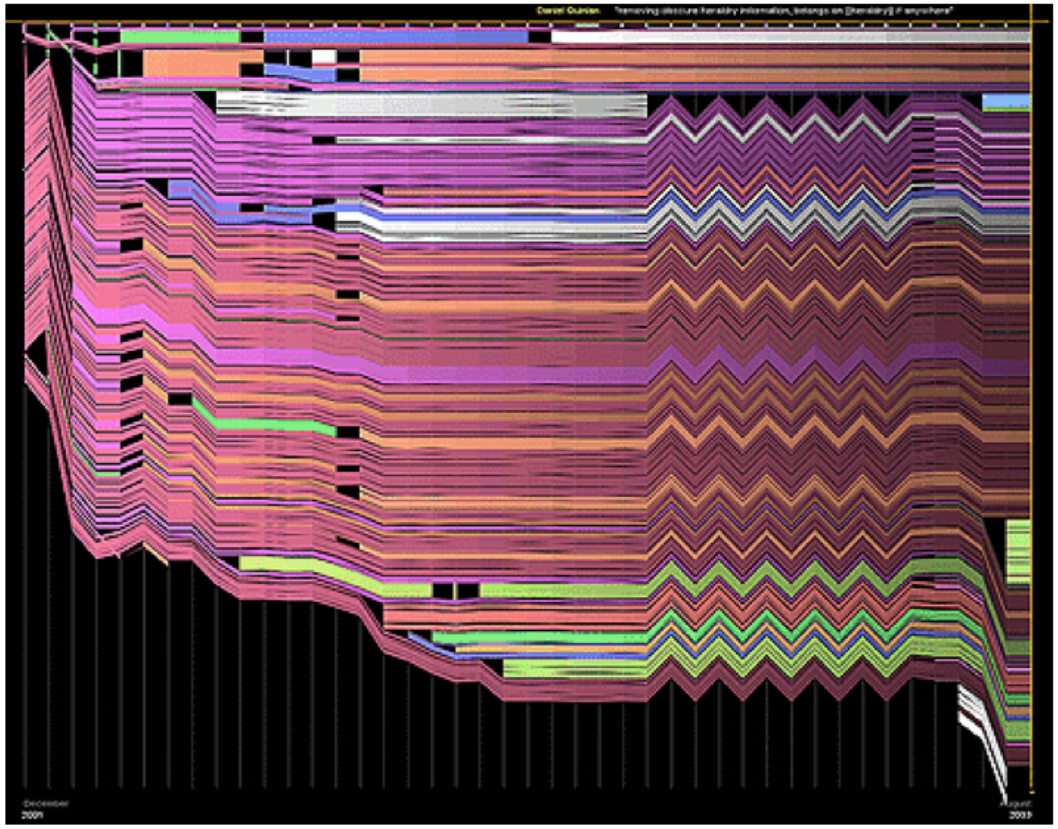
- Artigos
- Páginas de desambiguação
- Redireccionamentos
- Hiperligações *internas e externas*
- Estrutura de categorias
- *Predefinições* e *Infocaixas*
- Páginas de discussão
- Historial de revisões

Exemplo: *infocaixas*

As predefinições existentes nas páginas da Wikipédia, como as *infocaixas*, permitem a identificação e extracção automática de factos.

Visualização de Informação

Técnicas de visualização têm sido exploradas para compreender a dinâmica interna da Wikipédia.



Viégas, F., Wattenberg, M., Dave, K. (2004) – *Studying Cooperation and Conflict between Authors with history flow Visualizations.* ACM CHI'04.

Projecto WikiChanges

Projecto WikiChanges

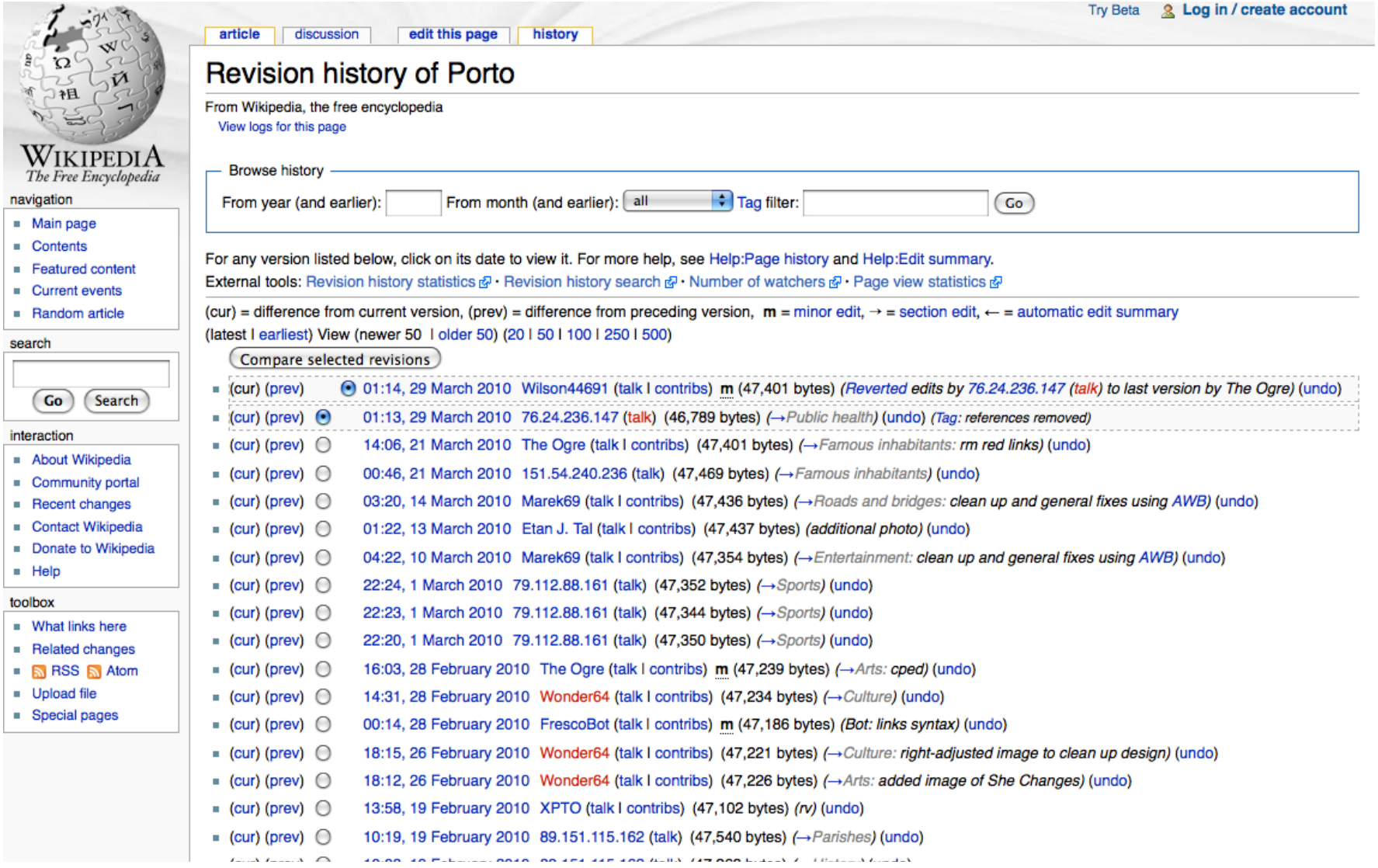
Ideia Base

Explorar as propriedades temporais para melhorar a recuperação de informação em colecções (ex. pesquisa).

Questões Iniciais

- Como evolui o conteúdo de um documento?
- As edições a um documento seguem um padrão *neutro*?
- É possível extrair informação do perfil de edições?

Historial de Edições a um Artigo



Try Beta Log in / create account

article | discussion | edit this page | history

Revision history of Porto

From Wikipedia, the free encyclopedia

View logs for this page

Browse history

From year (and earlier): From month (and earlier): all Tag filter: Go

For any version listed below, click on its date to view it. For more help, see Help:Page history and Help:Edit summary.

External tools: Revision history statistics · Revision history search · Number of watchers · Page view statistics

(cur) = difference from current version, (prev) = difference from preceding version, **m** = minor edit, → = section edit, ← = automatic edit summary

(latest | earliest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

Compare selected revisions

- (cur) (prev) 01:14, 29 March 2010 Wilson44691 (talk | contribs) **m** (47,401 bytes) *(Reverted edits by 76.24.236.147 (talk) to last version by The Ogre)* (undo)
- (cur) (prev) 01:13, 29 March 2010 76.24.236.147 (talk) (46,789 bytes) *(→Public health)* (undo) *(Tag: references removed)*
- (cur) (prev) 14:06, 21 March 2010 The Ogre (talk | contribs) (47,401 bytes) *(→Famous inhabitants: rm red links)* (undo)
- (cur) (prev) 00:46, 21 March 2010 151.54.240.236 (talk) (47,469 bytes) *(→Famous inhabitants)* (undo)
- (cur) (prev) 03:20, 14 March 2010 Marek69 (talk | contribs) (47,436 bytes) *(→Roads and bridges: clean up and general fixes using AWB)* (undo)
- (cur) (prev) 01:22, 13 March 2010 Etan J. Tal (talk | contribs) (47,437 bytes) *(additional photo)* (undo)
- (cur) (prev) 04:22, 10 March 2010 Marek69 (talk | contribs) (47,354 bytes) *(→Entertainment: clean up and general fixes using AWB)* (undo)
- (cur) (prev) 22:24, 1 March 2010 79.112.88.161 (talk) (47,352 bytes) *(→Sports)* (undo)
- (cur) (prev) 22:23, 1 March 2010 79.112.88.161 (talk) (47,344 bytes) *(→Sports)* (undo)
- (cur) (prev) 22:20, 1 March 2010 79.112.88.161 (talk) (47,350 bytes) *(→Sports)* (undo)
- (cur) (prev) 16:03, 28 February 2010 The Ogre (talk | contribs) **m** (47,239 bytes) *(→Arts: cped)* (undo)
- (cur) (prev) 14:31, 28 February 2010 Wonder64 (talk | contribs) (47,234 bytes) *(→Culture)* (undo)
- (cur) (prev) 00:14, 28 February 2010 FrescoBot (talk | contribs) **m** (47,186 bytes) *(Bot: links syntax)* (undo)
- (cur) (prev) 18:15, 26 February 2010 Wonder64 (talk | contribs) (47,221 bytes) *(→Culture: right-adjusted image to clean up design)* (undo)
- (cur) (prev) 18:12, 26 February 2010 Wonder64 (talk | contribs) (47,226 bytes) *(→Arts: added image of She Changes)* (undo)
- (cur) (prev) 13:58, 19 February 2010 XPTO (talk | contribs) (47,102 bytes) *(rv)* (undo)
- (cur) (prev) 10:19, 19 February 2010 89.151.115.162 (talk) (47,540 bytes) *(→Parishes)* (undo)

navigation

- Main page
- Contents
- Featured content
- Current events
- Random article

search

Go Search

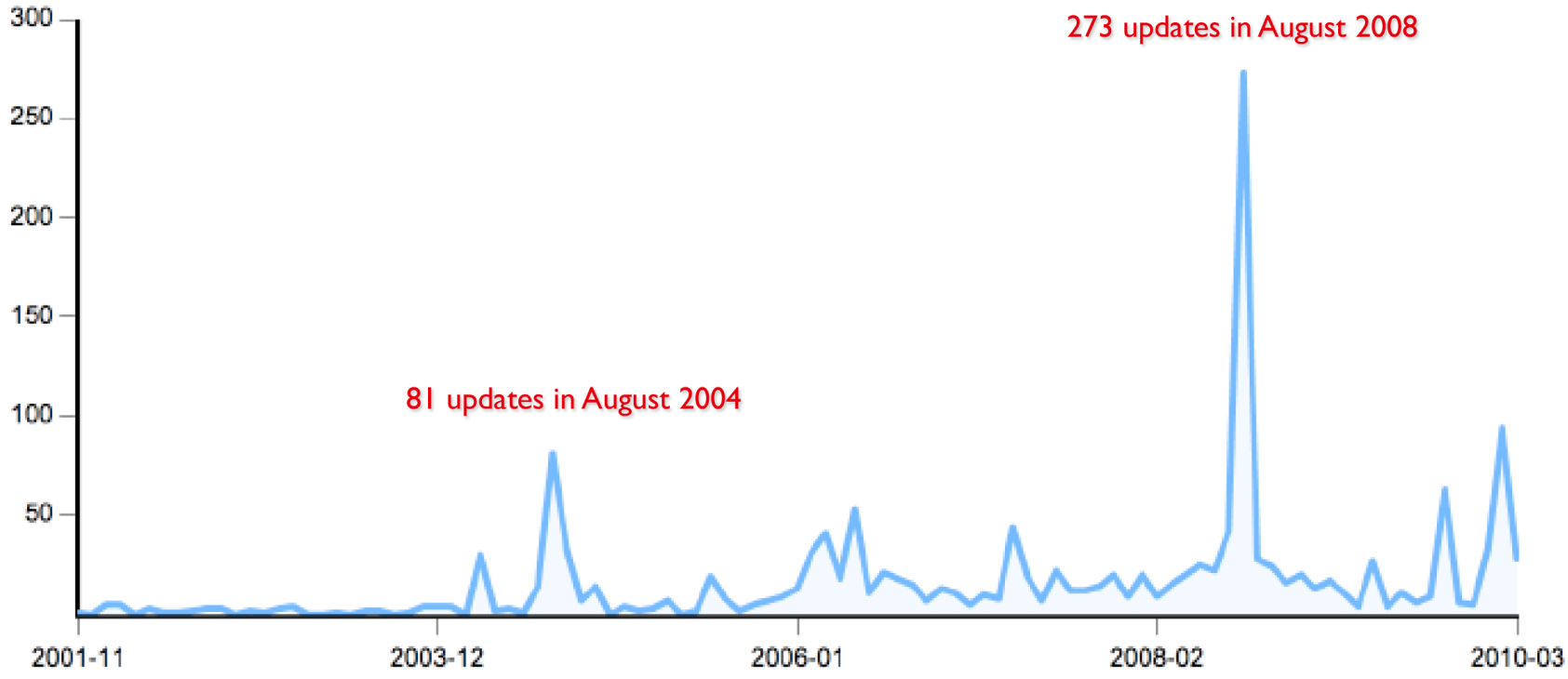
interaction

- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

toolbox

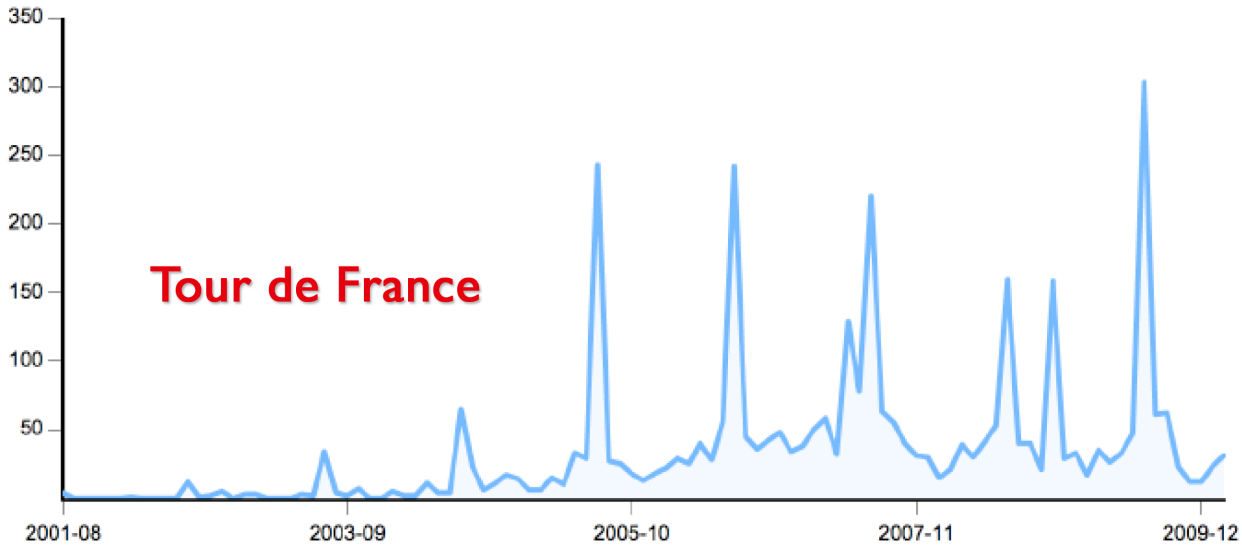
- What links here
- Related changes
- RSS Atom
- Upload file
- Special pages

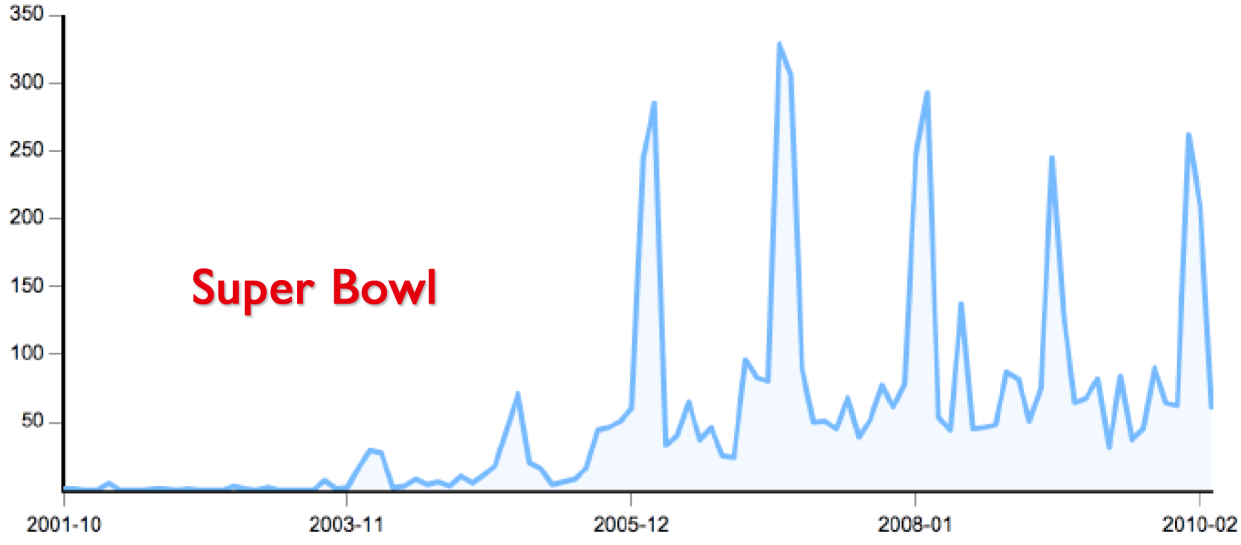
Perfil de Edições ao Longo do Tempo



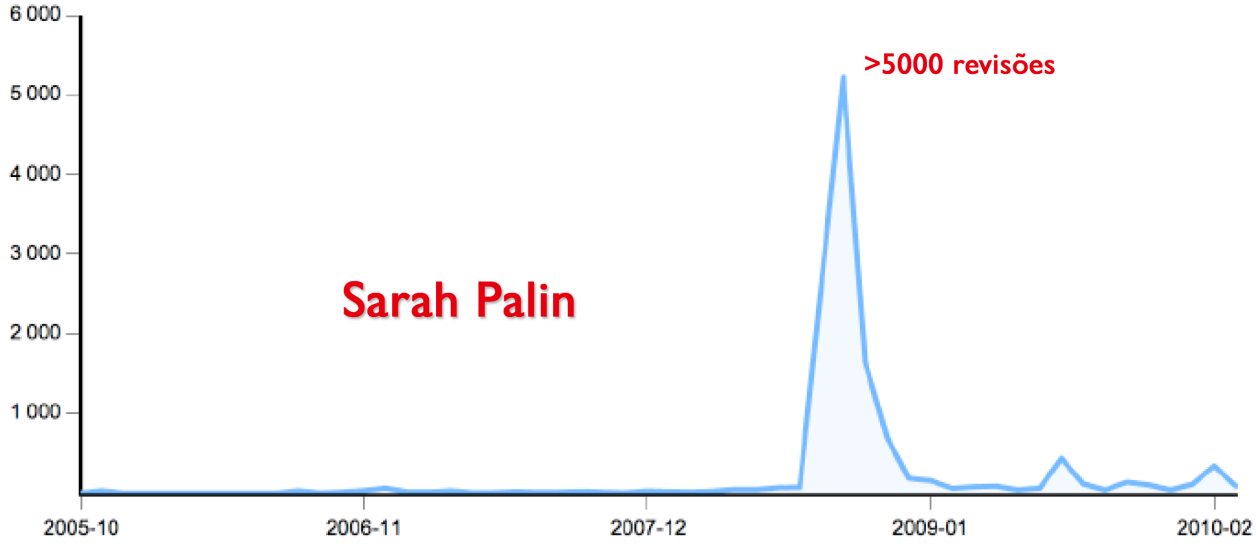
Número de edições mensais ao artigo Summer Olympic Games.

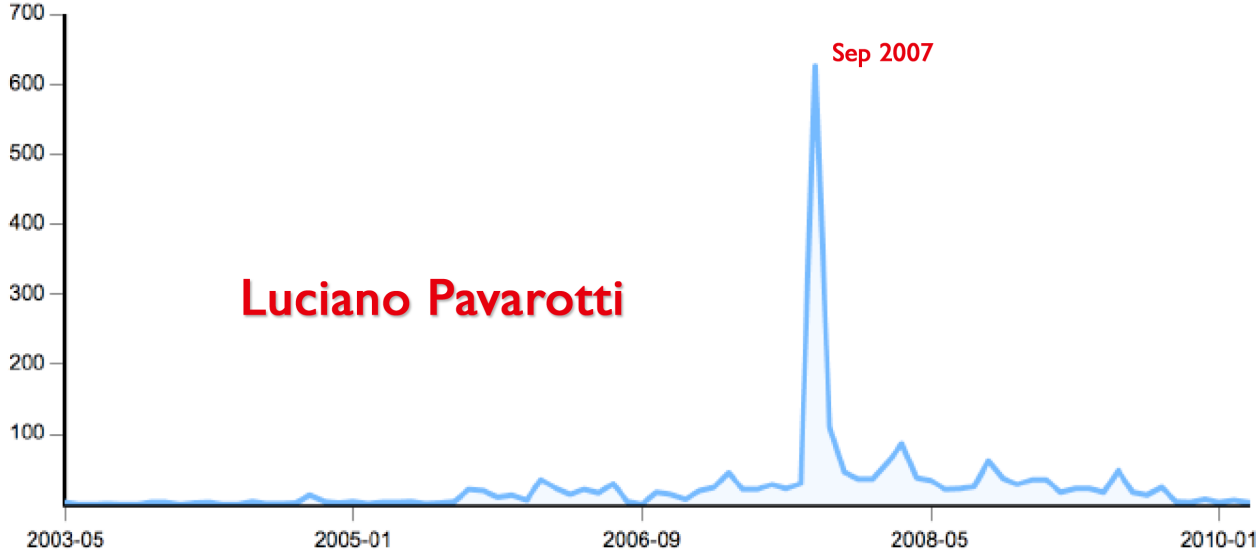
Acontecimentos Recorrentes



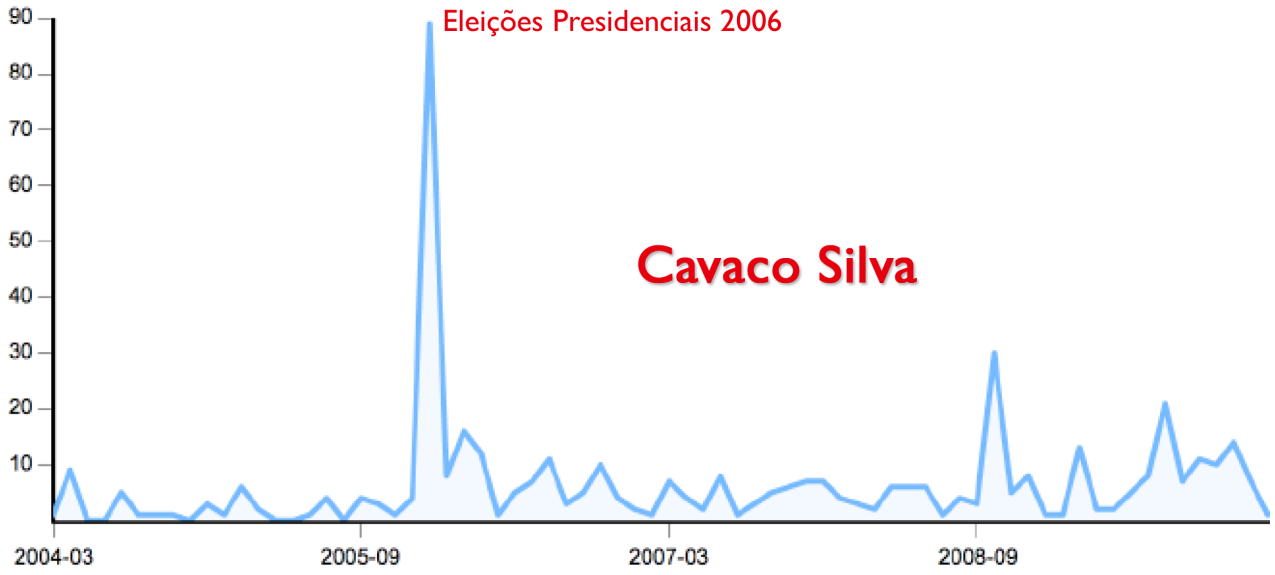


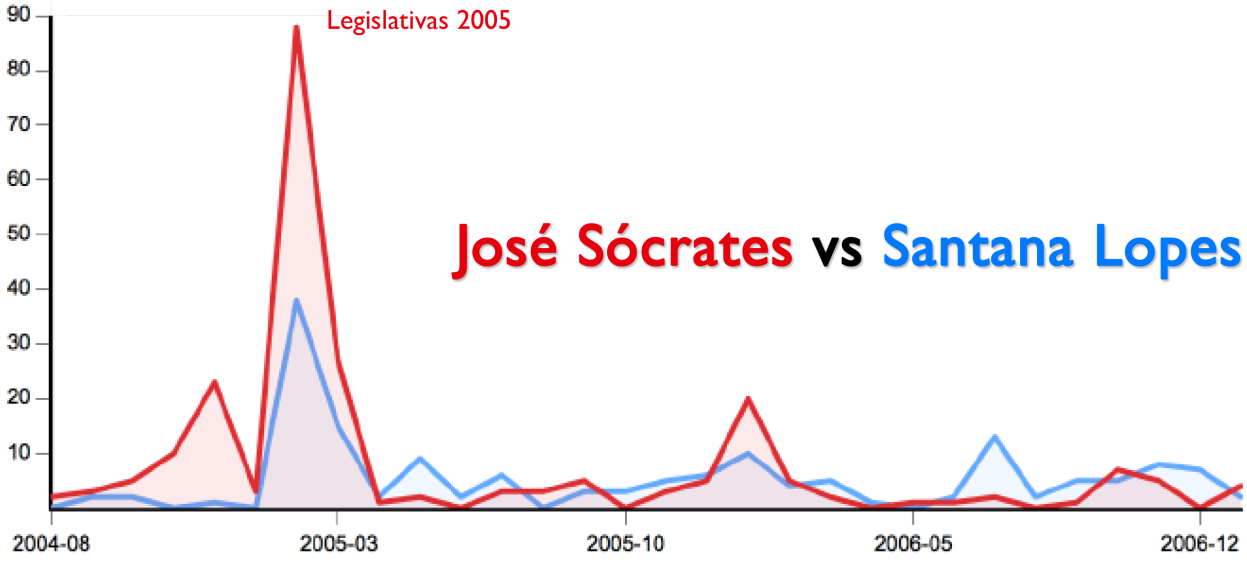
Acontecimentos Inesperados



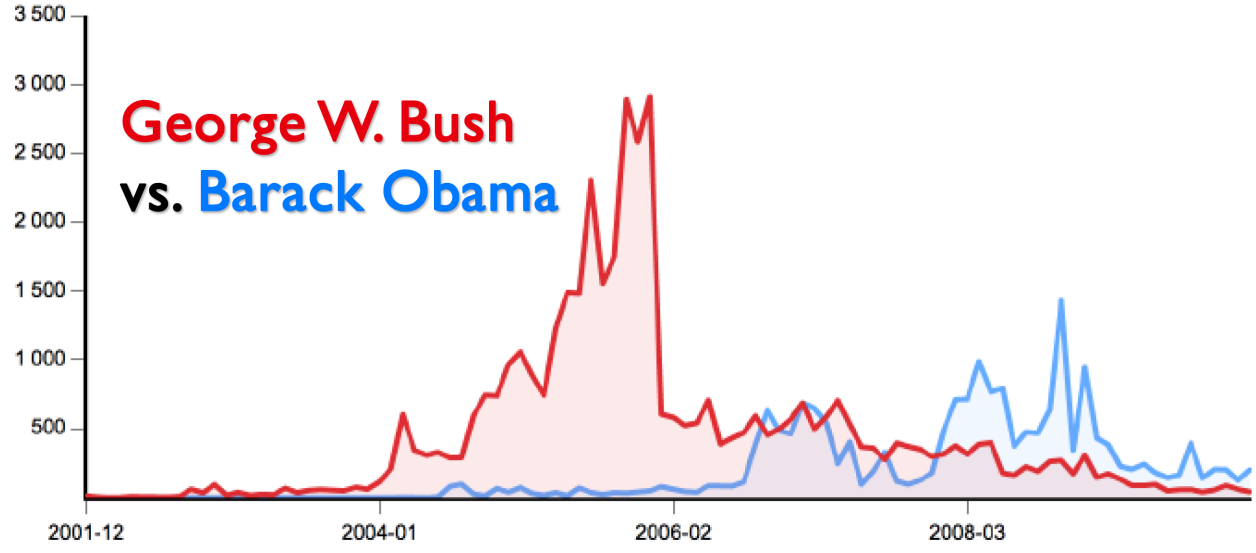


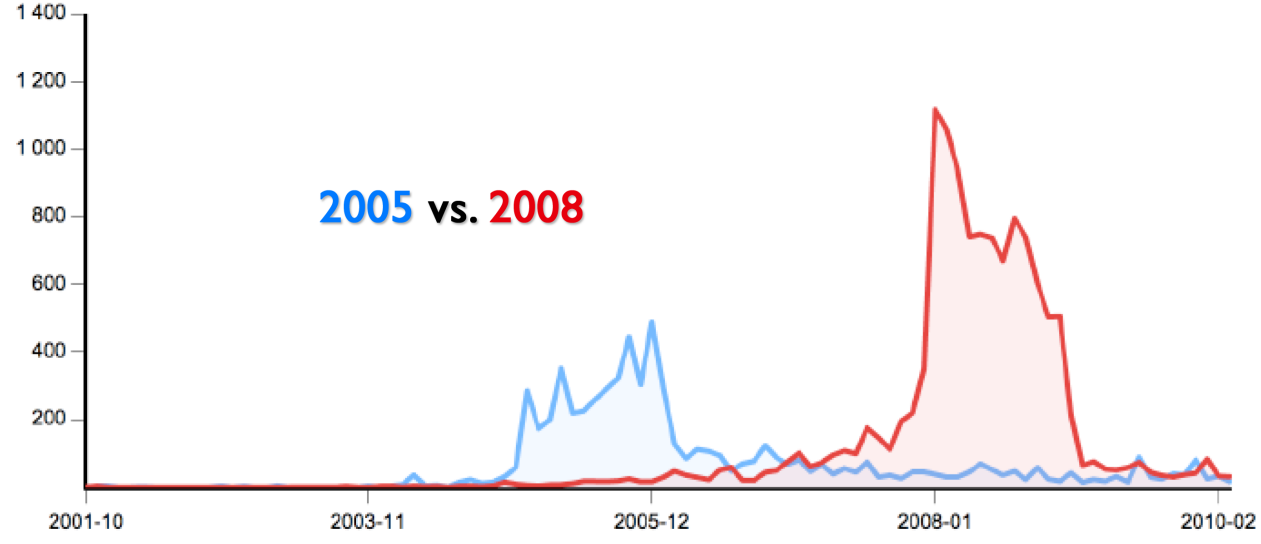
Acontecimentos Regionais



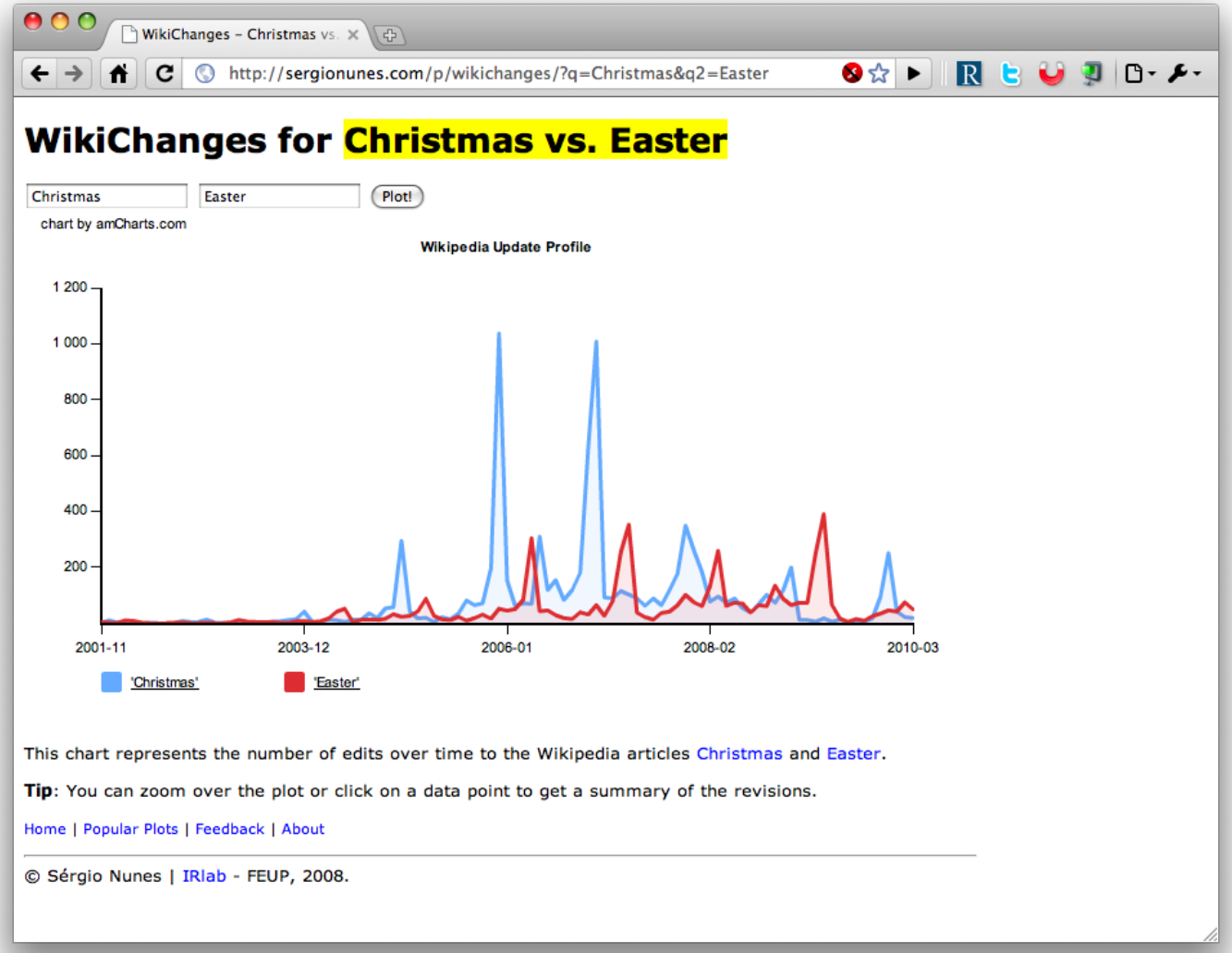


Comparações Lado a Lado



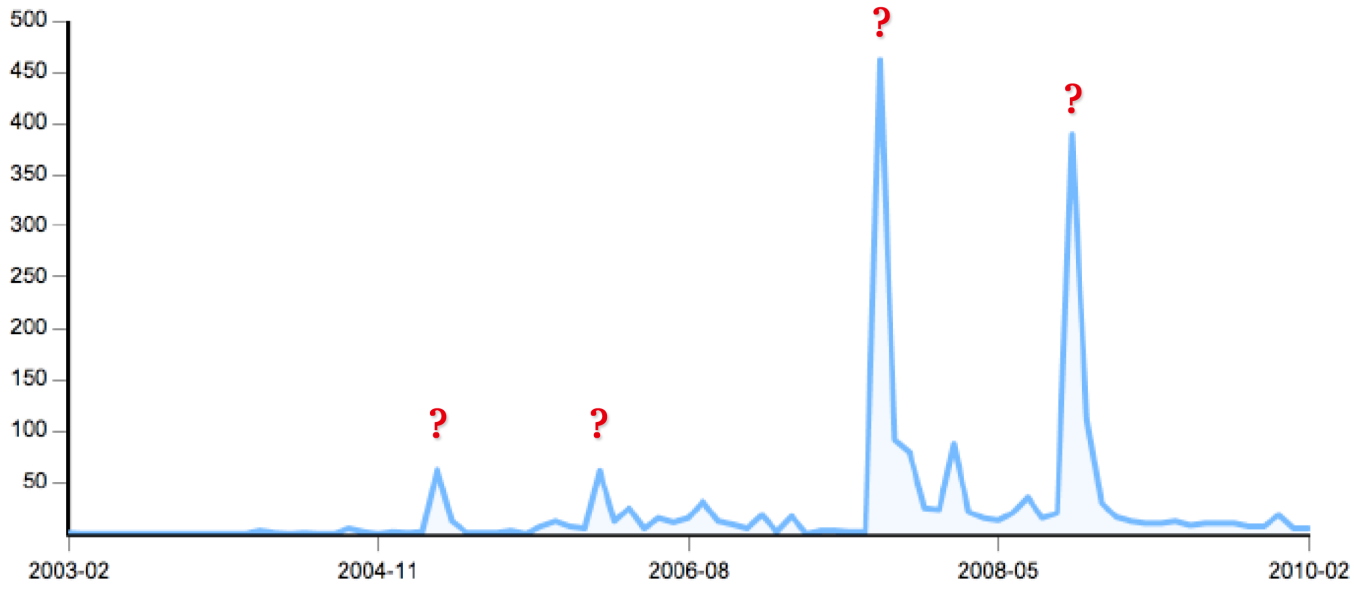


WikiChanges



http://irlab.fe.up.pt/p/wikichanges

Como Sumariar Edições?



O que significam os picos?
Como produzir sumários automáticos?

Ferramentas e Recursos

Acesso aos Dados

Arquivos de Texto

`http://download.wikimedia.org`
Cópia dos dados disponíveis na Wikipédia (*download* ou *cd*).
Artigos da Wikipédia Inglesa (>100 GB) + historial (5.6 TB!).

API

`http://en.wikipedia.org/w/api.php`
A plataforma MediaWiki inclui uma API que permite aceder programaticamente aos dados de forma parcelar.

Outros

Colecções preparadas e processadas por terceiros. Por exemplo: DBpedia (factos), YAGO (ontologia) ...

Projectos de Investigação

Wikify

`http://wikifyer.com`

Anota automaticamente o texto de qualquer página web com ligações para artigos da Wikipédia.

Koru

`http://www.nzdl.org/koru`

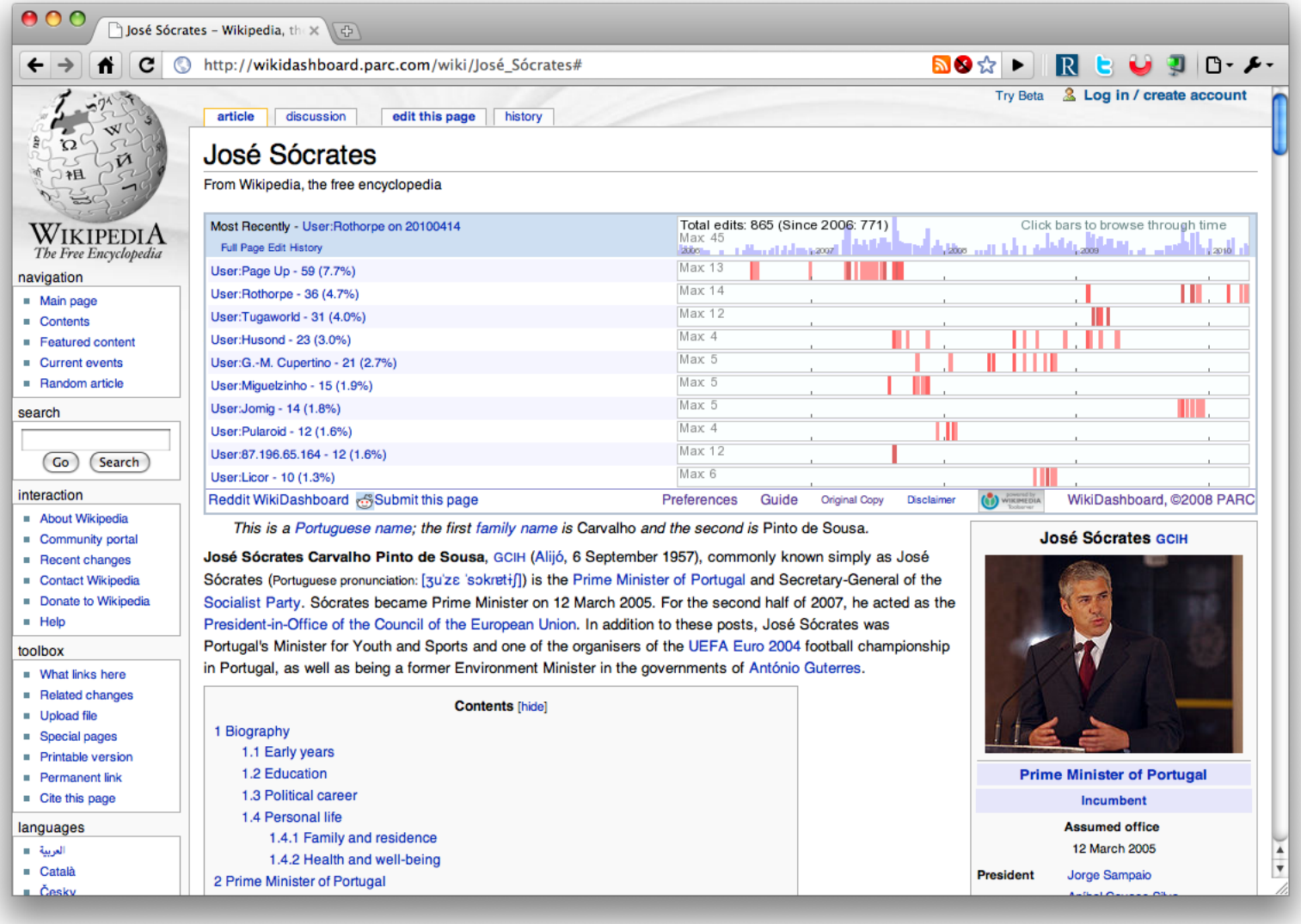
Motor de pesquisa que incorpora conhecimento extraido da Wikipédia (ex. tópicos, relações, etc).

WikiDashboard

`http://wikidashboard.parc.com`

Ferramenta de visualização e análise da dinâmica editorial dos artigos da Wikipédia.

WikiDashboard



Questões?

Referências

- **Academic studies about Wikipedia**

  `http://en.wikipedia.org/wiki/Academic_studies_about_Wikipedia`

- **Academic studies of Wikipedia**

  `http://en.wikipedia.org/wiki/Wikipedia:Academic_studies_of_Wikipedia`

- **Mining meaning from Wikipedia**
  Medelyan, O., Milne, D., Legg, C., Witten, I. H. (2008)

  `http://arxiv.org/abs/0809.4530`