Metabrain – Conhecimento na era do PetaByte

João Teixeira Gabriel Barata Daniel Gonçalves
Dep. Engª. Informática, IST
Av. Rovisco Pais, 1000 Lisboa
{joao.teixeira,gabriel.barata}@ist.utl.pt, daniel.goncalves@inesc-id.pt

Sumário
Hoje em dia, a Internet é uma fonte enorme de informação sobre os mais diversos ramos do conhecimento. No entanto, este conhecimento encontra-se disperso por muitos sítios, sem qualquer ligação ente eles, o que torna difícil inter-relaciona-lo e compreendê-lo. O objectivo deste trabalho é identificar e avaliar métodos de extracção de informação de forma simples e eficaz, sem recorrer a parsing de língua natural ou bases complexas de conhecimento anotado. Queremos mostrar que é possível extrair informação de forma implícita na Web através de métodos estatísticos. Para além disso, dados em bruto são normalmente de difícil compreensão. Como tal, procurámos também estudar como permitir aos utilizadores usar técnicas de visualização de informação de forma simples e eficaz para melhor os analisar e entender. Neste trabalho, propomos uma nova abordagem onde os utilizadores podem criar os seus próprios extractores de informação e respectivas visualizações, sem precisar de escrever uma única linha de código, de uma maneira fácil e altamente flexível, através de uma interface especialmente desenvolvida para o efeito. Um protótipo deste sistema, MetaBrain, foi desenvolvido e encontra-se em fase de testes e afinações.

Palavras-chave
Extracção de conhecimento, inferência probabilística, data mining, visualização, Internet.

1. INTRODUÇÃO

A versatilidade da Internet é também o seu Calcanhar de Aquiles. Qualquer pessoa pode criar páginas das mais díspares formas e conteúdos. Isto permitiu o seu rápido crescimento e que hoje em dia seja um repositório enorme de informação. No entanto, a Internet é um meio muito pouco estruturado, tornando difícil a extracção de informação. Apenas a leitura por um humano permite efectuar esta tarefa com algum grau de sucesso e, mesmo assim, a dispersão da informação obriga a consultar muitas fontes, nem sempre consistentes.

Para além disto, há todo um conjunto de informação não explícita ao nosso dispor. As páginas e os seus conteúdos reflectem um "inconsciente colectivo". Os temas versados e a forma em como o são revelam tendências e padrões que podem não ser imediatamente aparentes.

Tenta-se recorrer a ferramentas relativamente sofisticadas para fazer face a toda esta complexidade e extrair informação semanticamente relevante da Web. Desde processamento de língua natural a redes semânticas, de tudo se tem usado. Mas essas técnicas são frágeis e os resultados obtidos são válidos apenas em contextos muito restritos. A necessidade de uma base de conhecimento bem estruturada e anotada para o bom funcionamento deste tipo de análises é outro dos grandes problemas.

Existe no entanto, outra forma de abordarmos o problema. A utilização de métodos de análise estatística permitem-nos extrair informação relevante da Web. Por exemplo, o simples número de resultados de uma pesquisa no Google fornece informação, sem que seja necessário compreender o conteúdo das páginas retornadas. Imagine-se que queríamos saber se o Barak Obama é ou não a favor do encerramento da prisão de Guantanamo Bay. "Barack Obama Guantanamo keep open" devolve 96500 resultados e "Barack Obama Guantanamo close" 1230000. É bastante evidente qual a sua opinião sobre esse assunto, e não foi preciso fazer *parsing* de nada nem compreender língua natural.

Assim, o objectivo deste trabalho é estudar a melhor forma de explicitar o conhecimento que existe na Internet de forma implícita, fazendo-o de forma robusta e independente do domínio. Por outro lado, é nosso desejo permitir que tal seja fácil de fazer, sem necessidade de conhecimentos especiais de programação. Assim, procurámos também desenvolver uma interface gráfica eficaz e usável para a extracção da informação relevante.

Finalmente, a informação apresentada será tão mais útil quanto mais fácil seja a sua compreensão. Estudámos pois qual a melhor forma de permitir a utilizadores comuns usar técnicas avançadas de visualização de informação para explorar e apresentar os dados recolhidos. Um protótipo, o Metabrain, foi criado para demonstrar a validade da nossa abordagem.

O artigo encontra-se dividido da seguinte forma: primeiro é feito um resumo dos trabalhos analisados mais importantes, seguido de uma descrição da solução proposta e seu desenvolvimento. Por fim são descritos os métodos de avaliação propostos.

2. TRABALHO RELACIONADO

As soluções analisadas partilham entre si a utilização da Internet e os serviços nela disponibilizados como fonte dos seus métodos de análise. Não existe a necessidade de catalogar ou estruturar manualmente informação nem são utilizadas bases de conhecimento especialmente criadas para o funcionamento destes trabalhos.

O trabalho The Secret Lives of Numbers[1] é possivelmente um dos primeiros trabalhos nesta área. Este permite visualizar a popularidade de um enorme conjunto de números inteiros num dado motor de pesquisa. Esta popularidade é estimada mediante a contagem dos resultados de uma simples pesquisa num motor de busca por cada número. A mesma técnica é em Twitter Venn[2] com o objectivo de analisar o relacionamento entre vários conceitos, com base nas redes sociais. Aqui, são registados o número de resultados onde várias combinações de palavras-chave co-ocorrem.

O projecto Prism[3] permite criar paletas de cores com base num dado conceito. A técnica utilizada é semelhante à anterior mas utiliza listas pré-definidas de cores e palavras relacionadas com saturação e luz. A paleta de cores é criada com base no número de resultados das várias combinações entre o conceito dado pelo utilizador e as listas existentes (maçã+verde+claro, etc.).

Recentemente foram tornados públicos os resultados de alguns projectos particulares como o Simplistic Sentiment Mining from Tweets[4] e [Kramer10] que utilizam APIs de redes sociais como base para uma análise de co-ocorrência de conceitos. O primeiro procura pelas palavras que mais ocorrem juntamente com o nome de alguns dispositivos comerciais. Este método simples produz resultados que permite, por exemplo, aos comerciantes verificar a popularidade dos seus produtos. O segundo trabalho analisa tudo o que é dito diariamente e faz uma contagem de palavras positivas vs palavras negativas, criando assim um gráfico de felicidade. Nos resultados, é fácil verificar que épocas como o Natal e a Páscoa, entre outras épocas festivas, correspondem a picos de felicidade.

Trabalhos como [Bollegala07], [Cimiano04] e [Etzioni04] utilizam o número de resultados devolvido pelos motores de pesquisa como fonte de informação para realizar inferências, através de por exemplo conectores linguísticos como os descritos em [Banko08]. Utilizando o conector "é um" e um conceito dado pelo utilizador, é possível realizar uma pesquisa, como por exemplo: "Portugal é um", e verificar qual a palavra que mais ocorre de seguida. Desta forma a palavra que mais deve ocorrer será "país", sendo assim possível descobrir de forma automática a que categoria pertence um determinado conceito. O contrário também possível com a utilização do conector "tais como". Com este conector é possível fazer uma pesquisa por "países tais como" e verificar quais as palavras que ocorrem de seguida nos resultados obtidos.

Todos estes trabalhos incluem restrições do domínio de execução, desenvolvimento fechado, restrições no método de visualização, impossibilidade por parte do utilizador de usar vários métodos em conjunto e por fim a necessidade, por parte do programador de desenvolver de raiz os métodos em questão.

3. SOLUÇÃO

Dado o nosso objectivo de permitir mesmo a utilizadores não experientes em programação a extracção e visualização de informação implícita na Web, desenvolvemos o Metabrain de forma modular e altamente personalizável, sem descurar a sua usabilidade e facilidade de utilização. Isto foi feito tanto para a extracção de informação como para a sua visualização, ao fornecer vários modos de visualização de informação os quais podem ser utilizados e personalizados de acordo com as necessidades do utilizador.

A arquitectura da nossa solução é composta por três camadas: biblioteca de extracção de informação; gestão de conjuntos de dados; visualização de dados, descritas em seguida.

3.1 Biblioteca Metabrain

A biblioteca Metabrain foi desenvolvida com o intuito de proporcionar um conjunto de métodos de análise e extracção de informação da Web prontos a usar de uma forma simples sem que os seus utilizadores tenham que se preocupar com a complexidade por detrás destes.

A biblioteca abstrai o seu utilizador do modo como a informação é obtida e pesquisada. Sendo incluídos na biblioteca métodos de acesso a várias APIs públicas, listas de palavras anotadas (por exemplo, palavras relacionadas com sentimentos) e técnicas de construção de *queries* complexas de forma automática.

Por exemplo, existem métodos que devolvem o número de resultados encontrados de várias palavras de uma lista, outros que analisam os resultados obtidos e conseguem identificar quais as palavras mais usadas e relacionadas com a nossa pesquisa inicial. Existem ainda outros que incluem técnicas especiais de geração de *queries*, os quais nos permitem extrair significado da Web. Um exemplo de uma das *queries* seria "Cidades tais como *", através da análise dos resultados obtidos é possível extrair da Web uma lista de cidades. Para além disto, estes métodos podem ser afinados de acordo com as necessidades do utilizador através de opções avançadas. Estas podem ser bastante úteis como por exemplo durante uma análise de sentimento. Seleccionada a opção de filtro de afirmações negativas durante a geração de *queries* de pesquisa estende a *query* normal 'I like x' para 'I like x – "I don't like"', o que permite melhorar os resultados, excluindo alguns resultados negativos.

[1] http://turbulence.org/Works/nums/ (visitado a 03/06/2010)

[2] http://neoformix.com/Projects/TwitterVenn/view.php (visitado a 03/06/2010)

[3] http://nodebox.net/code/index.php/Prism (visitado a 03/06/2010)

[4] http://neoformix.com/2009/SimplisticSentimentMiningFromTweets.html (visitado a 03/06/2010)

Tudo isto permite que o utilizador só tenha que preocupar com qual a informação que pretende extrair da Web e não em como essa extracção será realizada.

3.2 Aplicação Metabrain

A aplicação Metabrain é construída com base na biblioteca Metabrain. A interface gráfica desta aplicação foi desenvolvida para permitir que um comum utilizador consiga realizar pesquisas de informação simples ou avançadas e consiga analisar os resultados aí obtidos de uma forma fácil através de vários modos de visualização.

A utilização desta aplicação pode ser levada a cabo em 3 passos: gestão de conjuntos de dados (*datasets*); escolha de informação a visualizar e seu modo de visualização; e personalização da visualização e exportação desta.

3.2.1 Conjuntos de dados

O primeiro passo na utilização da aplicação é a obtenção de informação implícita na Web. Para tal foi criada uma interface que dá ao utilizador uma grande liberdade na criação de novos conjuntos de dados.

O principal desafio no desenvolvimento desta interface foi permitir que o utilizador consiga executar diferentes tipos de queries de modo tão flexível quanto ao utilizar directamente a API da biblioteca Metabrain. Cada pesquisa tem como resultado uma tabela, com cada coluna correspondendo a uma faceta (número de resultados da pesquisa, por exemplo). Para alem disto é possível encadear pesquisas, tornando possível, por exemplo, extrair uma lista de cidades da Web e de seguida encontrar quais as palavras que mais ocorrem junto a estas.

A solução encontrada para permitir o encadeamento de pesquisas foi possibilitar a realização de pesquisas hierárquicas. Durante uma pesquisa o utilizador pode decidir usar como parte da sua *query* um valor proveniente do resultado de uma outra utilizando uma variável que tomará sucessivamente os vários valores de uma das colunas do resultado da pesquisa original. A Figura 1 Exemplo de pesquisa de informação.Figura 1 exemplifica uma possível pesquisa hierárquica para exemplo dado anteriormente. Neste o utilizador decide obter uma lista de palavras relacionadas com cada elemento dos resultados da variável %1, que corresponde à extracção, da Web, de uma lista de cidades.

Como durante a criação de pesquisas encadeadas avançadas o utilizador pode necessitar de executar algum nó da pesquisa de forma a validar os seus resultados em separado, cada pesquisa da hierarquia é acompanhada na interface por um botão *preview*.

Este mecanismo de pesquisa foi desenvolvido com o auxílio de tecnologias DHTML (CSS, JavaScript) de modo a criar um ambiente de maior dinamismo e rapidez de utilização. A junção destes vários elementos cumpre o nosso objectivo de disponibilizar ao utilizador uma grande liberdade na criação de novos conjuntos de dados.

Tendo em conta que alguns dos métodos de análise necessitam de fazer, por vezes, milhares de pesquisas online e que a maior parte dos serviços online usados impõe limites de utilização, foi desenvolvida uma cache que permite utilizar qualquer um dos resultados obtidos em pesquisas anteriores como fonte para novas pesquisas. Todos os resultados ficam também disponíveis para posterior visualização (**Error! Reference source not found.**).

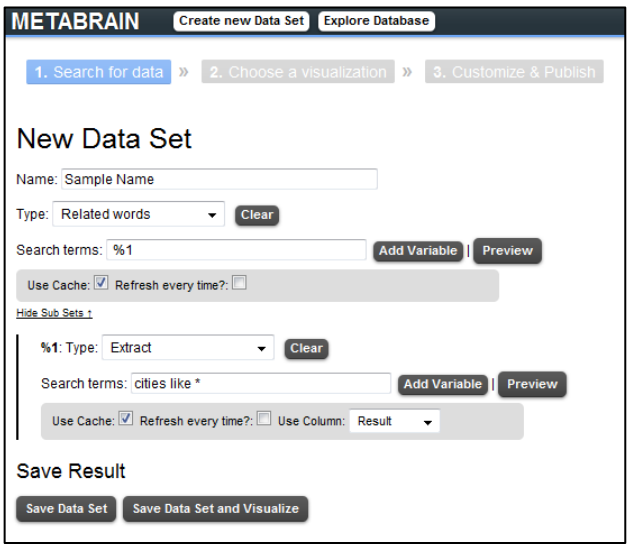

Figura 1 Exemplo de pesquisa de informação.

Figura 2 Listagem de conjuntos de dados guardados na base de dados.

3.2.2 Visualização

O objectivo da secção de visualização é permitir que utilizador consiga visualizar a sua informação para que esta tenha um maior significado, mais uma vez dando um elevado grau de liberdade na sua personalização.

De forma a cumprir este objectivo a interface é dividida em duas secções: escolha dos dados a visualizar e modo de visualização; e pré-visualização mais personalização. Cada conjunto de dados, como já referido, tem a forma de uma tabela com várias colunas ou facetas.

Existem vários modos de visualização disponíveis, desde gráficos de linha até *treemaps*. Cada uma destas visualizações é caracterizada por um certo número de graus de liberdade, que definem em cada visualização quais os elementos desta que podem variar. Depois de analisado um grande número de visualizações concluímos que estes graus de liberdade podem ser encapsulados num destes quatro campos: colunas (eixo X), linhas (eixo Y), cor/paleta de cores, tamanho (largura/altura). Os dois primeiros campos são usados na selecção dos dados a visualizar e os outros dois são usados para personalizar a visualização.

Tendo isto em conta e sabendo que o utilizador ao criar uma visualização já tem em mente quais os dados que pretende usar, mas nem sempre sabe qual a melhor forma

de os visualizar, decidimos criar uma interface que o auxilie nesta tarefa. A interface foi então desenvolvida de modo a que à medida que o utilizador selecciona as facetas que pretende visualizar uma lista de sugestões seja dinamicamente actualizada.

As facetas que o utilizador pode seleccionar são recuperadas do conjunto de dados seleccionado e listadas na interface. De forma a facilitar a selecção destas facetas, a sua listagem é automaticamente separada em categorias, de acordo com o conteúdo de cada uma. Os elementos desta lista podem ser arrastados pelo utilizador para os campos que representam os eixos X e Y, indicando que devem ser usados na visualização.

De acordo com as facetas seleccionadas e sua localização nos eixos, todas as visualizações incluídas na aplicação são verificadas. No fim desta verificação, todas as visualizações cujo número de graus de liberdade seja compatível com a selecção do utilizador será apresentada a este como uma sugestão de visualização. O utilizador poderá então seleccionar qual a visualização que pretende usar.

Depois de seleccionada uma visualização, esta é apresentada automaticamente utilizando os dados seleccionados. Nesta altura são também adicionados novos campos na interface que permitem uma maior personalização da visualização seleccionada. Aqui o utilizador poderá decidir, no caso de um gráfico de barras por exemplo, que o tamanho das barras e/ou sua cor é definido pelo valor de outra coluna, simplesmente arrastando essa coluna da lista anteriormente referida para o campo de personalização em questão.

Os métodos de visualização usados nesta aplicação são desenvolvidos utilizando o *toolkit* Protovis [Bostock09] que proporciona uma grande flexibilidade na criação de visualizações em Javascript, permitindo assim a exportação de visualizações criadas para outros websites.

Na Figura **3** é apresentado um exemplo de visualização através de um gráfico de barras.

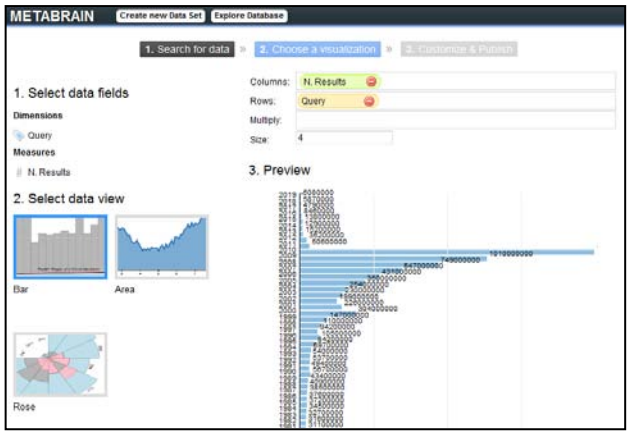

Figura 3 **Exemplo de visualização de um conjunto de dados.**

3.3 Futuro

Como trabalho futuro, está em primeiro lugar a finalização da aplicação, no que toca à personalização das visualizações, seguida do desenvolvimento de novos módulos de análise e novos modos de visualização.

4. AVALIAÇÃO

Apesar de estarmos confiantes na qualidade da solução desenvolvida, esta será avaliada oportunamente, para demonstrar inequivocamente a sua adequação ao problema em causa. Em primeiro lugar, faremos uma avaliação heurística para descobrir os principais problemas de usabilidade da interface. Uma vez corrigidos esses problemas, terão lugar testes com utilizadores em que estes deverão desempenhar um conjunto de tarefas de extracção e visualização de informação. Serão medidos tempo e erros, e avaliada a sua satisfação e a facilidade de aprendizagem e compreensão da interface.

5. CONCLUSÃO

Durante o desenvolvimento deste trabalho foram analisados vários trabalhos que utilizam métodos estatísticos, análise de padrões e tendências como meio para aceder ao "inconsciente colectivo" que é a Internet. Desta análise foi possível descobrir os vários métodos existentes actualmente e quais as melhores fontes na Web onde estes podem ser usados. As várias abordagens analisadas, apesar de se basearem na mesma ideologia, apresentam alguns problemas e nenhuma tira proveito de todos os métodos de extracção analisados. Para além disso, também restringem o domínio de execução e apresentam pouca flexibilidade na visualização dos resultados.

De modo a responder aos problemas encontrados surge a necessidade da criação de uma solução que disponibilize ao utilizador uma vasta quantidade de métodos de extracção de informação, sem que este se tenha de preocupar com análise de linguagem natural ou criação de bases de dados gigantes de conhecimento anotado. A solução proposta materializa-se sob a forma da biblioteca Metabrain e uma aplicação complementar que fornece ao utilizador um modo personalizável de obter e visualizar a informação a qual, de outro modo, seria difícil obter e analisar.

6. REFERÊNCIAS

[Bollegala07] Bollegala, Danushka, Yutaka Matsuo, e Mitsuru Ishizuka. "Measuring semantic similarity between words using web search engines." ACM, 2007. 757--766.

[Banko08] Banko, Michele, e Oren Etzioni. "The Tradeoffs Between Open and Traditional Relation Extraction." Association for Computational Linguistics, 2008. 28--36.

[Bostock09] Bostock, Michael, e Jeffrey Heer. "Protovis: A Graphical Toolkit for Visualization." *IEEE Transactions on Visualization and Computer Graphics* (IEEE Educational Activities Department) 15 (2009): 1121--1128.

[Cimiano04] Cimiano, Philipp, e Steffen Staab. "Learning by googling." *SIGKDD Explor. Newsl.* (ACM) 6 (2004): 24--33.

[Etzioni04] Etzioni, Oren, et al. "Web-scale information extraction in knowitall: (preliminary results)." ACM, 2004. 100--110.

[Kramer10] Kramer, Adam D. "An unobtrusive behavioral model of gross national happiness." ACM, 2010. 287--290.