Observatório da Web: Uma plataforma de monitoração, síntese e visualização de eventos massivos em tempo real

Walter dos Santos Filho, Gisele L. Pappa, Wagner Meira Jr, Dorgival Guedes, Adriano Veloso, Virgílio A. F. Almeida, Adriano M. Pereira, Pedro H. C. Guerra, Arlei L. da Silva, Fernando H. J. Mourão, Tiago R. de Magalhães, Felipe M. Machado, Letícia L. Cherchiglia, Lívia S. Simões, Rafael A. Batista, Filipe L. Arcanjo, Gustavo M. Brunoro, Nathan R. B. Mariano, Gabriel Magno, Marco Túlio C. Ribeiro, Leonardo V. Teixeira[1]
Altigran S. da Silva[2], **Bruno Wanderley Reis**[3], **Regina Helena Silva**[3]

[1]Departamento de Ciência da Comptuação
Universidade Federal de Minas Gerais (UFMG)
Av. Antônio Carlos, 6627 - Pampulha - Belo Horizonte, MG

[2]Departamento de Ciência da Computação
Universidade Federal do Amazonas (UFAM)
Av. Gal. Rodrigo Octávio Jordão Ramos, 3000 - Japiim
Manaus - AM

[3]Faculdade de Filosofia e Ciências Humanas- FAFICH
Universidade Federal de Minas Gerais (UFMG)
Av. Antônio Carlos, 6627 - Pampulha - Belo Horizonte, MG

{walter,glpappa,meira,dorgival,adrianov,virgilio,adrianoc}@dcc.ufmg.br

Resumo. *Este trabalho introduz o* Observatório da Web*, um portal criado para apresentar ao usuário uma síntese do que está sendo falado nas mais diversas fontes de conteúdo da* Web*, incluindo jornais, revistas, portais, redes sociais e o Twitter. Esse artigo descreve o modelo conceitual e arcabouço sobre o qual o Observatório foi criado, e mostra como ele pode ser utilizado para monitorar as Eleições Presidenciais de 2010. Os resultados ilustram como a informação é disponibilizada no portal, mostrando os índices de visibilidade e a polaridade dos* tweets *referentes a um político específico.*

Abstract. *This work presents the Web Observatory, a portal designed for providing to end users a synthesis of the current topics and trends in the various types of Web-based media, including newspapers, magazines, portals, social networks, and other applications such as Twitter. This paper describes the conceptual model and the framework on which the Observatory was built and shows how it may be used for monitoring the 2010 Brazilian Presidential Elections on the Web. Preliminary results show how the information is delivered in the portal, showing indices such as visibility and polarity of political personalities.*

1. Introdução

O surgimento da Web 2.0 fez com que os usuários passassem de meros consumidores a produtores ativos de conteúdo. Ferramentas como o Twitter[1], e redes sociais como o Face-

[1]www.twitter.com

book[2], Orkut[3] e Youtube[4], possibilitam que usuários comuns expressem suas opiniões sobre os mais diversos assuntos, e propaguem informação que consideram relevante em tempo real. É interessante notar como a interatividade e a avaliação do fluxo de informações em tempo real passou a ser um fator importante nas várias aplicações Web, o que se manifesta no interesse cada vez maior por fenômenos dinâmicos, substituindo o paradigma anterior da Web como repositório de dados e informações estático sobre os mais variados temas e personalidades. Essa mudança de paradigma fez da Web uma ferramenta ainda mais poderosa em áreas como a política e economia. Um exemplo recente é o impacto que a Web teve nas eleições americanas de 2008, tanto como fonte rica de informação atualizada quanto de arrecadações record de doações para campanha de Obama através do seu website.

Conforme cresce a quantidade de conteúdo disponível, cresce também nossa dificuldade de buscar as fontes certas de informação, e organizá-las de forma que tenhamos uma visão geral sobre o que se está falando na Web. Além disso, a dinamicidade da rede faz com que opiniões mudem e novos fatos apareçam a cada minuto. Buscando uma maneira de facilitar o acesso do usuário comum a informação disponível na Web, esse trabalho apresenta a arquitetura, discute a implementação e mostra resultados preliminares do *Observatório da Web*. O principal objetivo do Observatório é buscar informações nas mais diversas fontes de dados disponíveis na Web, sumarizar e exibir essas informações na forma de metáforas visuais ou indicadores.

O *Observatório da Web* foi criado a partir de um modelo conceitual, que envolve um contexto e uma série de entidades sendo monitoradas, entre outros. Nesse artigo, focamos no contexto das Eleições Presidenciais de 2010, e em como o Observatório monitora tanto os prováveis presidenciáveis, quanto outros políticos cuja influência na campanha presidencial é conhecida. O Observatório também provê indicadores que medem a *visibili dade* de um político na Web, bem como métodos que analisam a *polaridade* das notícias sobre um determinado político, visando determinar o quanto se está falando bem ou mal dele.

Do ponto de vista técnico, o arcabouço sobre o qual foi desenvolvido o *Observatório da Web* deve atender a alguns requisitos para que seja efetivo e eficiente. O primeiro requisito é processar a informação em tempo real ou próximo disso, de forma que mudanças no comportamento e na opinião nas várias fontes de acesso sejam percebidas rapidamente. O segundo requisito é a diversidade de fontes, uma vez que não basta focar em apenas uma única fonte, tendo em vista que elas podem trazer informações contraditórias, tendenciosas ou mesmo complementares. O terceiro requisito é a efetividade das metáforas visuais e indicadores, os quais devem sintetizar o sentimento coletivo das várias fontes a respeito da entidade. O quarto e último requisito é a escalabilidade, uma vez que o volume de dados tende a crescer e pode ser necessário armazenar e processar vários meses de dados, o que pode ser normalmente proibitivo se feito de forma isolada.

É interessante notar como o Observatório da Web condensa em uma única aplicação vários dos desafios de pesquisa elencados pela SBC em 2006. Ao tratar de grandes volumes de dados oriundos de várias fontes, ataca o primeiro desafio (Gestão da

[2]www.facebook.com
[3]www.orkut.com
[4]www.youtube.com

informação em grandes volumes de dados multimídia distribuídos). O provimento de um observatório para o público em geral, focado em assuntos de relevância para a população é uma forma de buscar mecanismos para o quarto desafio (Acesso participativo e universal do cidadão brasileiro ao conhecimento). A implementação do observatório traz vários desafios em termos de sistemas de computação, todos relacionados ao quinto desafio (Desenvolvimento tecnológico de qualidade: sistemas disponíveis, corretos, seguros, escaláveis, persistentes e ubíquos). Finalmente o estudo da sociedade e de como a Internet a influencia em temas do seu cotidiano demanda técnicas complexas que vão de encontro ao segundo desafio (Modelagem computacional de sistemas complexos artificiais, naturais e sócio-culturais e da interação homem-natureza).

Este trabalho está organizado da seguinte forma. A Seção 2 descreve algumas iniciativas anteriores para acompanhamento de eleições na Web, além das técnicas computacionais relacionadas. A Seção 3 descreve o modelo seguido para construção do Observatório da Web, enquanto a Seção 4 discute a instanciação do Observatório no contexto das eleições presidenciais de 2010. Os resultados obtidos até o momento são reportados na Seção 5, enquanto conclusões e trabalhos futuros são apresentados na Seção 6.

2. Trabalhos Relacionados

O acompanhamento de eleições na Web já foi feito anteriormente. Um exemplo de portal criado com esse objetivo é o Tendencias Politicas[5], do Chile. O portal acima citado tem objetivos parecidos com os aqui propostos, mas trabalha apenas com um índice de visibilidade dos candidatos, baseado em entropia. Nessa mesma direção está o sítio Polictfact[6]. Porém, ao invés de monitorar os candidatos durante a eleição, esse sítio analisa as opiniões omitidas por políticos em geral, e as classifica de acordo com seu "nível de veracidade".

Esta seção dá ênfase aos trabalhos relacionados a análise de polaridade ou análise de sentimentos de documentos ou *tweets*, já que esse é um dos grandes desafios do Observatório no momento. Existem vários trabalhos na direção de determinar a polaridade de notícias políticas e econômicas, análises de consumidores em relação a produtos comprados em mercados eletrônicos, etc.

[Dodds and Danforth 2009], por exemplo, criaram o índice de felicidade (*happiness index*, que classifica músicas e discursos políticos de acordo com a mensagem que esse passa a seus ouvintes ou leitores. O método é baseado em uma lista de palavras, denominada ANEW, que classifica vários substantivos da língua inglesa como tendo conotação boa ou ruim. Nessa mesma direção, [Jin et al. 2009] criou um arcabouço complexo que utiliza, entre outros, um algoritmo de aprendizagem para aprender uma lista de adjetivos, também na língua inglesa, capaz de distinguir boas revisões de produtos de revisões ruins. Porém, e a maioria dos trabalhos na área de análise de sentimentos ou opiniões foca em ferramentas de processamento de linguagem natural, a maioria delas está disponível apenas em inglês. Uma exceção é o trabalho de [Sarmento and Oliveira 2009], onde toda a análise considera palavras do português.

A análise polaridade feita no momento para o Observatório é baseada em um classificador associativo. No entanto, um sistema como o proposto por

[5]www.tendenciaspoliticas.cl

[6]http://www.politifact.com/truth-o-meter/statements/

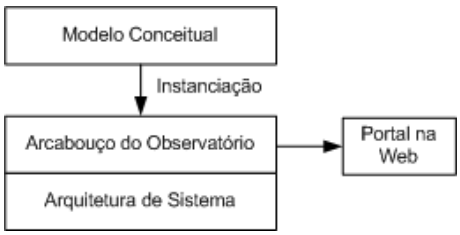

Figura 1. Organização da Arquitetura de Funcionamento do Observatório da Web

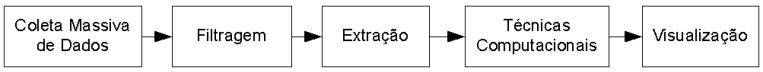

Figura 2. O arcabouço proposto

[Dodds and Danforth 2009] também foi testado, usando traduções de palavras do inglês para o português. Os resultados não foram tão promissores quanto os encontrados utilizando essa outra técnica.

3. O Observatório da Web

A concepção e implementação do Observatório da Web pode ser dividida em três partes principais: o modelo conceitual, a arquitetura do sistema, e o arcabouço de software necessário. A Fig. 1 ilustra como essas três partes interagem. O ponto de partida do Observatório de um cenário de observação é o modelo conceitual, que engloba sete definições básicas: contexto, entidade, fonte, autor, evento, tema e grupo.

O contexto descreve o que está sendo observado. Nesse artigo, focamos no contexto Eleições Presidenciais 2010. As entidades correlatas ao contexto são o alvo de monitoramento, e no caso das eleições correspondem aos pré-candidatos a presidência, e políticos com grande participação na eleição, como o presidente Lula. As fonte são o alvo da coleta, como o Twitter, Facebook, portais de notícias etc., e o autor é o responsável pelo conteúdo disponível na fonte. Eventos dizem respeito a acontecimentos importantes no contexto observado, tais como um debate, por exemplo, e que podem ter um grande efeito no conteúdo das fontes observadas. Temas são questões importantes sendo discutidas no contexto, tais como saúde, educação e economia, no caso das eleições. Por último, grupos representam organizações, conjuntos, agremiações, ou qualquer outra forma de agrupamento de entidades. No caso das eleições, partidos políticos são considerados grupos.

Esse modelo conceitual deve ser instanciado, e com base nas definições feitas, um conjunto de fontes, entidades, e temas, entre outros, são viabilizados como parâmetros de entrada para o arcabouço. O arcabouço executa sobre uma arquitetura de software complexa, criada para lidar com coleta e processamento de dados dinâmica e em tempo real, e têm como saída um conjunto de indicadores e metáforas visuais exibidos em um portal Web. As próximas seções descrevem o arcabouço e a arquitetura de sistema propostos.

3.1. O Arcabouço

A Figura 2 ilustra o arcabouço proposto, composto de cinco fases. Inicialmente, os dados são coletados e filtrados. A informação relevante passa então por um processo de limpeza e extração de entidades, que são nossos objetos de interesse. De posse das entidades,

técnicas computacionais, especialmente de mineração de dados, são aplicadas para produzir indicadores de interesse sobre os dados, que posteriormente são exibidos ao usuário final através de metáforas interessantes. As próximas seções descrevem como cada uma dessas cinco fases foi implementadas.

3.1.1. Coleta Massiva de Dados

A coleta de dados envolve quatro tipos de mídias na *Web*: (1) redes sociais, (2) Twitter, (3) jornais, revistas e portais de notícias, e (4) *blogs*. O tipo de informação proveniente dessas mídias representam tanto opiniões de cidadãos comuns quanto de formadores de opinião da sociedade. Enquanto os jornais e portais tem uma maior quantidade de conteúdo baseado em fatos , os outros ambientes contém informação rica sobre a opinião e a repercusão dos fatos reportados por esse primeiro tipo de mídia.

Toda a coleta é baseada em um vocabulário controlado, definido pelo usuário. As redes sociais estão sendo coletadas através de *crawlers*, enquanto no Twitter a coleta é feita utilizando a API disponível[7], que retorna um *stream* de dados. Considerando as páginas de jornais de notícias, portais, revistas e *blogs*, toda a coleta é baseada em RSS (*Really Simple Syndication*) [8]. Os dados coletados são armazenados em um banco de dados MySQL, e depois processados na fase de filtragem.

3.1.2. Filtragem

A fase de filtragem separa da base de dados informações irrelevantes para o contexto sendo estudado, e é também baseada em um vocabulário. Enquanto a coleta usa um vocabulário mais relaxado, devido as restrições da API do Twitter, que permite apenas buscas utilizando uma única palavra, i.e., nomes compostos não podem ser utilizados como filtro, a filtragem é mais específica, e define o que foi coletado e é realmente relevante para o contexto sendo observado. Apesar do vocabulário ser inicialmente definido por um humano, ele pode ser posteriormente estendido através de um método automático de aprendizado.

3.1.3. Extração

A fase de extração começa com a padronização da codificação dos caracteres, já que diferentes fontes podem user diferentes padrões (por exemplo, UTF8, ISO8859-1). Após a padronização, são eliminados das páginas coletada através de *feeds* o código HTML, cabeçalhos, e anúncios.

Numa segunda fase, métodos tradicionais de pré-processamento de textos [Manning et al. 2008], tais como remoção de *stop words* - palavras comuns, tais como “o”, “de”, “para”, etc, e *stemming* - que extrai os radicais das palavras do texto, são aplicados aos documentos.

[7]http://apiwiki.twitter.com/Streaming-API-Documentation
[8]http://www.rssboard.org/rss-specification

O último passo da extração, também considerado um dos mais importantes nesse processo, é a identificação e deduplicação de entidades [Ratinov and Roth 2009]. A identificação de entidades nos textos é feita através da ferramenta LBJ-Based Named Entity Tagger[9]. Essa ferramenta identifica as entidades de interesse nos documentos coletados.

Após a fase de identificação, segue uma fase de deduplicação de entidades. Isso porque os processos de filtragem e identificação de entidades não conseguem diferenciar "José Serra"e "fomos a serra", ou "Lula presidente"de "Lula Molusco". Nessa fase, um método de aprendizado é utilizado para aprender a associar entidades a determinados contextos.

3.2. Técnicas Computacionais

Os dados provenientes da fase de extração servirão de entrada para um conjunto de técnicas computacionais, criadas especialmente para medir aspectos relacionados aos componentes do modelo conceitual (tais como entidades) dos quais temos interesse. Entre as técnicas implementadas no Observatório da Web, destacamos: (i) visibilidade de uma entidade na mídia; (ii) análise de polaridade do que se anda falando sobre as entidades, (iii) recomendação de fontes e *tweets* para usuários interessados no mesmo tipo de assunto, (iv) análise da propagação de notícias no Twitter e nas mídias jornalísticas, entre outros. Aqui, descreveremos os métodos implementados para medir a visibilidade de uma entidade através de uma medida de entropia, e a polaridade de um *tweet* utilizando métodos de aprendizado de máquina.

A visibilidade de um político na mídia, tal como "José Serra"ou "Lula", está associada a quantidade de vezes em que o político aparece na mídia, e também na quantidade de diferentes mídias em que ele é citado. Por exemplo, a visibilidade de um candidato citado 10 vezes em um determinado portal não deve ser maior que a visibilidade de um candidato citado 6 vezes em três diferentes fontes de notícia, como um jornal e dois *blogs*. Uma boa métrica para capturar esse tipo de tendência é a entropia.

A entropia [Shannon 1948], na área de teoria da informação, é descrita como o grau de incerteza associado a uma variável aleatória. Ela é definida como na Eq. 1, onde n é o número de fontes e $p(x_i)$ é a probabilidade de ocorrência da entidade x na fonte i.

$$H(x) = -\sum_{i}^{n} p(x_i) \log_2 p(x_i) \tag{1}$$

A fórmula da entropia foi modificada para criar um indicador de visibilidade, definido como na Eq. 2, onde $f(x_i)$ representa a frequência absoluta com que a entidade x apareceu na fonte i.

$$Visibilidade(x) = \ln \sum_{i}^{n} f(x_i) \times H(x) \tag{2}$$

Já a análise de polaridade de tweets é baseada em um método de aprendizado de máquina baseado em classificadores associativos. Classificadores associativos são basea-

[9]http://l2r.cs.uiuc.edu/ cogcomp/LbjNer.php

dos em modelos de regra de associação de classe, onde uma regra é definida por uma relação $A \rightarrow c$ (A implica em c), onde A determina um conjunto de atributos e c uma classe. Classificadores associativos geram inicialmente um conjunto de regras formadas por conjuntos de ítens frequentes, e depois utilizam diversas estratégias para a seleção, ordenação, poda, predição e avaliação dessas regras, utilizadas para classificar um novo exemplo [Zaki 2000].

O problema dos classificadores associativos é que o custo de gerar todas as regras pode ser potencialmente proibitivo. Além disso, a grande maioria das regras pode ser desperdiçada, uma vez que muitas das regras geradas podem não ser aplicáveis durante a classificação. Porém, esses problemas podem ser resolvidos por um classificador associativo *lazy* [Veloso et al. 2006], que gera regras baseadas nos conjuntos de ítens que aparecem nos documentos (ou *tweets*) a serem classificados. Isso significa que podem-se gerar regras mais sofisticadas (por exemplo, de tamanho maior), uma vez que nem todas as regras possíveis serão geradas. A saída do algoritmo são regras de associação do tipo "se *corrupção* então *negativo*". O algoritmo utilizado para determinar a polaridade dos *tweets* foi o LAC (*Lazy Associative Classification*) [Veloso et al. 2006]. Para cada entidade, o LAC cria um conjunto de associações distintas entre termos que aparecem nos *tweets* e a conotação (positiva ou negativa) do termo quando relacionado a aquela entidade.

Em seguida, essas associações (i.e., regras) são empregadas em um processo de votação, onde cada regra $A \rightarrow c$ é interpretada como sendo um voto para a polaridade c. Os votos podem ter pesos diferentes, dependendo da força da associação entre A e c. Uma medida de associação entre A e c é a confiança da regra (a probabilidade condicional de que c seja a polaridade do *tweet*, dado que as palavras no conjunto A estão no *tweet* sendo analisado). Votos são somados e ponderados pelo peso correspondente, e no fim do processo, a polaridade com maior pontuação é a que será prevista pelo algoritmo. O algoritmo é muito eficiente, já que produz apenas regras que contenham as palavras no *tweet* sendo analisado.

3.2.1. Visualização

Como a intenção final do projeto é ter um portal que possa ser acessado pela comunidade em geral, a parte de visualização de dados é essencial para o sucesso do projeto. A maioria das metáforas visuais utilizadas foram implementadas utilizando o *Flare*[10] ou o *Open Flash*[11] e biblitecas e *plug-ins Javascript*. O *Flare* é uma biblioteca para criação de visualizações de dados interativas que roda sobre o *Adobe Flash Player*. Já o *Open Flash* é um projeto de código aberto para produção de gráficos, e constrói gráficos a partir de um arquivo de dados armazenado no servidor.

3.3. Modelo de Arquitetura de Sistema

O modelo de arquitetura de sistema utilizado atualmente é simples. Os dados estão armazenados em uma base de dados MySQL, e a coleta é centralizada. No momento,

[10]http://flare.prefuse.org

[11]http://teethgrinder.co.uk/open-flash-chart/

“Aecio”	“Serra”	“Lula”	“Arruda”	“Heloisa”	“Dilma”
“Ciro”	“Sarney”	“Cristovam”	“FHC”	“Alckmin”	“Marina”

Tabela 1. Vocabulário controlado utilizado para coleta

estamos construindo um novo modelo de arquitetura distribuído, que utilizará o HDFS[12] (*Hadoop Distributed File System*), além de termos um projeto de um armazém de dados para comportar o grande volume de dados coletado. Essa decisão de arquitetura deve criar um ambiente altamente tolerante a falhas e facilmente escalável.

4. O Observatório das Eleições

Neste artigo, descrevemos a instanciação do Observatório para o contexto das Eleições Presidenciais 2010. Nessa seção, listamos as fontes sendo coletadas, assim como as entidades sendo monitoradas. Além disso, descrevemos algumas funcionalidades do portal, disponível em http://observatorio.inweb.org.br.

Atualmente, estão sendo coletadas mais de 110 fontes, incluindo redes sociais como o Youtube, os principais portais brasileiros, além de jornais, revistas, *blogs* e o Twitter. A seleção de jornais coletados partiu de uma lista disponível no sítio da Associação Nacional de Jornais[13]. Já a lista de *blogs* baseia-se em um *ranking* criado em pesquisas acadêmicas nas áreas de ciências políticas.

Todas as coletas são realizadas a partir de um vocabulário controlado. As palavras utilizadas aparecerem na Tabela 1. Note que esse vocabulário não contêm nomes compostos pois, como mencioando anteriormente, a API do Twitter não aceita expressões (palavras separadas por espaço) como critério de filtragem. De qualquer forma, um segundo filtro é aplicado, onde expressões mais específicas, tais como “Heloisa Helena”ou “Marina Silva”são utilizadas como filtro.

Os dados estão sendo coletados desde 10 de dezembro de 2009. Até o dia 22 de março, 390.848 tweets e 194.487 *feeds*, provenientes de jornais, revistas e *blogs*, foram coletados utilizando o vocabulário controlado. Inicialmente, selecionamos seis entidades para monitorar: Aécio Neves, Ciro Gomes, Dilma Rousseff, José Serra, Luiz Inácio Lula da Silva, e Marina Silva.

Para cada entidade, calculamos o índice de visibilidade, conforme definido na Eq. 2. Treinamos também o algoritmo de identificação de polaridade, baseado no LAC, para cada uma delas. Porém, para treinar o algoritmo, necessitamos de um conjunto de exemplos manualmente rotulados pelo usuário. 2555 *tweets* referentes as seis entidades monitoradas foram classificados manualmente como positivos ou negativos por um conjunto de usuários, e utilizados como entrada para o classificador. Utilizando um método de validação cruzada de 10 partições, a taxa de acerto do algoritmo varia de 86% a 88%.

Tanto o indicador de visibilidade quanto o de polaridade podem ser visualizados utilizando diferentes períodos de tempo, variando de um dia a 3 meses. Note que, em todas as metáforas visuais, os dados prevenientes do Twitter são analisados separadamente dos dados provenientes de outras fontes. Isso acontece porque o número de *tweets* coleta-

[12]http://hadoop.apache.org/core/

[13]http://www.anj.org.br/associados

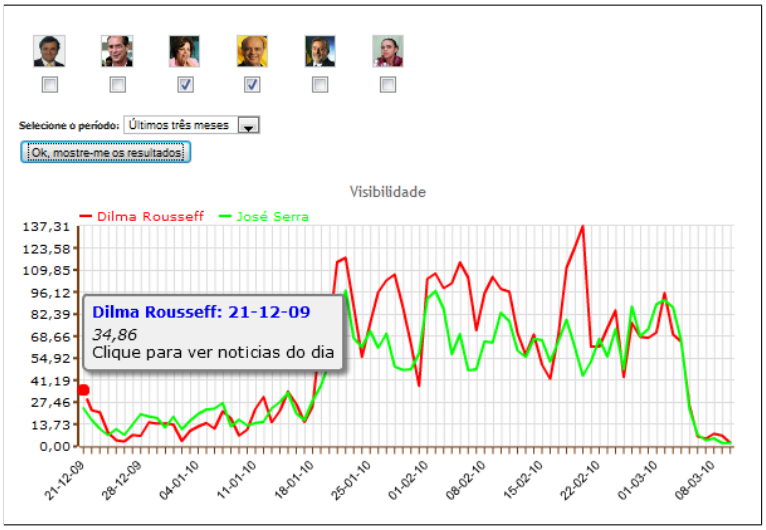

Figura 3. Visibilidade dos candidatos a Presidência

dos diariamente diferem em mais de uma ordem de grandeza do número de documentos coletados, e não seria justo reportar os resultados para essas duas fontes utilizando uma mesma métrica. Algumas das metáforas utilizadas podem ser observadas nas Figuras 3 e 4.

5. Resultados

A melhor forma de analisar os resultados obtidos pelo arcabouço descrito nesse artigo é acessar o portal do *Observatório da Web*, disponível em http://observatorio.inweb.org.br. Essa seção apresenta exemplos da visibilidade dos candidatos nas mídias exceto Twitter, e da polaridade de *tweets* descrita na Seção 3.2. A Fig. 3 mostra um gráfico de comparação do indicador de visibilidade para os pré-candidatos Dilma Rousseff e José Serra do dia 21 de dezembro de 2009 a 08 de março de 2010.

O pico máximo de visibilidade da pré-candidata Dilma Rousseff foi no dia 20 de fevereiro de 2010, quando seu índice de visibilidade chegou a 137.31. Já o pico de visibilidade de José Serra aconteceu no dia 22 de janeiro de 2009, quando o índice atingiu 97.12 pontos. Observe também que, ao clicar em qualquer ponto do gráfico, um *link* para um conjunto de notícias onde o político apareceu é exibido. Através desse conjunto, o usuário pode entender causas de picos ou baixas em termos de visibilidade.

Finalmente, note que a partir do dia 19 de janeiro de 2010, o valor do índice de visibilidade cresceu para os dois candidatos. Isso ocorreu porque, nesse dia, foram adicionadas a coleta novas fontes de dados. Considerando essas novas fontes, o número de ocorrências total dos candidatos nas mídias aumentou, causando então o aumento no valor do índice de visibilidade.

Já a Fig. 4 mostra a polaridade dos comentários no Twitter sobre o político Aécio Neves. Note que, em média, 54% dos tweets postados de 18 de dezembro de 2009 ao início de fevereiro de 2010 são positivos. O gráfico apresenta tanto o valor médio quanto o valor diário da polaridade do candidato, possibilitado a identificação de períodos em que se fala mais bem ou mal do candidato.

Figura 4. Visualização da Polaridade para Aécio Neves

6. Conclusões e trabalhos futuros

Esse artigo apresentou o *Observatório da Web*, um projeto desenvolvido com o intuito de criar indicadores e metáforas visuais que apresentem uma visão geral e em tempo real do que vem sendo comentado e disseminado na Web. O artigo descreveu o modelo conceitual criado e o arcabouço desenvolvido. Detalhamos como as fases de coleta, filtragem, extração, técnicas computacionais e visualização funcionam, contextualizando o projeto nas Eleições Presidenciáveis de 2010. Os resultados apresentados mostram como o Observatório da Web pode ser utilizado por cidadãos comuns para acompanhar o andamento das eleições na Web.

As técnicas computacionais descritas ainda estão em fase de aprimoramento, e muitas outras funcionalidades estão sendo desenvolvidas para inserção no portal. Entre elas, destacamos a propagação da informação na Web, a recomendação de outras fontes e tweets de acordo com os interesses dos usuários, a identificação da polaridade de notícias e *blogs*, e a associação de eventos com picos de visibilidade de alguns candidatos. Note também que, por ser um ambiente genérico, o Observatório pode ser modificado para monitorar outro contextos, tais como a Copa do Mundo ou as Olimpíadas.

Uma outra linha em que estamos trabalhando diz respeito a como medir a qualidade das informações disponíveis no portal. Pretendemos criar uma metodologia de avaliação de qualidade de cada uma das cinco fases que compõe o arcabouço, onde estamos interessados em medir a cobertura da coleta e filtragem, quantificar a acurácia dos nossos método de identificação de entidades e deduplicação, melhorar a eficácia das técnicas computacionais atualmente disponíveis no portal, além de analisar como o usuário interage com o portal.

7. Agradecimentos

Este trabalho é parcialmente financiado pelo INWeb - Instituto Nacional de Ciência e Tecnologia para Web, pelo projeto C5, pelo CNPq, FAPEMIG, e Santander.

Referências

Dodds, P. S. and Danforth, C. M. (2009). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*.

Jin, W., Ho, H. H., and Srihari, R. K. (2009). Opinionminer: a novel machine learning system for web opinion mining and extraction. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1195–1204, New York, NY, USA. ACM.

Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155, Morristown, NJ, USA. Association for Computational Linguistics.

Sarmento, L. and Oliveira, E. C. (2009). The design of optimism, an opinion mining system for portuguese politics. In *Encontro Português de Inteligência Artificial*.

Shannon, C. (1948). A mathematical theory of communication. Technical Report 3-4, Bell System Technical Journal.

Veloso, A., Meira Jr., W., and Zaki, M. J. (2006). Lazy associative classification. In *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*, pages 645–654, Washington, DC, USA. IEEE Computer Society.

Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE Trans. on Knowl. and Data Eng.*, 12(3):372–390.