BLOGMINER: REPRESENTAÇÃO TEMPORAL DE ASSUNTOS ATRAVÉS DE MODELAGEM DE TÓPICOS

Júlia Ferreira de Almeida

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Geraldo Bonorino Xexéo

Rio de Janeiro
Setembro de 2012

BLOGMINER: REPRESENTAÇÃO TEMPORAL DE ASSUNTOS ATRAVÉS DE MODELAGEM DE TÓPICOS

Júlia Ferreira de Almeida

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

_____

Prof. Geraldo Bonorino Xexéo, D.Sc.

_____

Prof. Jano Moreira de Souza, Ph.D.

_____

Profa. Jonice de Oliveira Sampaio, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
SETEMBRO DE 2012

Almeida, Júlia Ferreira

BlogMiner: Representação temporal de assuntos através de modelagem de tópicos/ Júlia Ferreira de Almeida/ – Rio de Janeiro: UFRJ/COPPE, 2012.

XI, 96 p.: il.; 29,7 cm.

Orientador: Geraldo Bonorino Xexéo

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação, 2012.

Referências Bibliográficas: p. 90-93.

1. Modelagem de tópicos. 2. Recuperação da Informação. 3. *Collocation* 4. Similaridade de tópicos. I. Xexéo, Geraldo Bonorino. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III Título.

À minha família.

AGRADECIMENTOS

Agradeço a minha mãe que mesmo morando em Friburgo, Macaé, Cabo Frio, sempre fez com que parecesse estar ao meu lado, e me acalmando a cada crise. Agradeço ainda por ser essa mulher tão forte que não deixa se abater por problemas e sempre está disposta a ajudar aos outros. Agradeço também ao meu pai, que mesmo nem sempre concordando, me deu apoio em minhas decisões e esteve ao meu lado quando algumas deram errado. E por ter se tornado uma referência de profissional exemplar pra mim e pai zeloso, mesmo que nessa vida quase nômade de funcionário de banco. Ao meu irmão por colocar meus pés no chão e trazer estabilidade emocional quando necessário. E minha cunhadinha que já considero como da minha família.

Agradeço aos meus tios André e Eduardo por todo o suporte financeiro e emocional aqui no Rio, que sem eles nada disso seria possível. À Mônica e a Iaiá minhas companheiras de apartamentos pela imensa atenção e suporte nestes dez anos.

Agradeço aos meus queridos avós e minha bisa, por serem tão orgulhosos da neta, mesmo não conseguindo ao menos explicar o que ela faz. Aos meus tios e primos que mesmo longe sempre posso contar com eles.

Agradeço aos professores que me acompanharam durante toda minha jornada na UFRJ. Um agradecimento especial ao professor Xexéo, pela orientação, apoio e dedicação ao longo dos últimos dois anos, sem o qual este trabalho não teria sido possível. Agradeço ainda aos professores Jano e Jonice, por terem aceitado participar da minha banca de defesa de mestrado.

Agradeço aos meus amigos Paula, Aguas, Espirito, Sardinha, Joselito, Carlinha e João pelo apoio nos perrengues ao longo do mestrado, mesmo que muitas vezes para rir deles. Minhas amigas Hildi e Vanessa que me tiraram da realidade algumas vezes em nossas viagens pra Natal, Floripa,... e que me fizeram um bem imenso. Ao pessoal do laboratório que antes de serem colegas de trabalho, e com todas as briguinhas e ciúmes infantis, se tornaram grandes amigos e companheiros. Ao Pap por ter me dado um grande apoio no meu início de mestrado.

E ao meu amado Deus.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

BLOGMINER: REPRESENTAÇÃO TEMPORAL DE ASSUNTOS ATRAVÉS DE MODELAGEM DE TÓPICOS

Júlia Ferreira de Almeida

Setembro/2012

Orientador: Geraldo Bonorino Xexéo

Programa: Engenharia de Sistemas e Computação

Postagens em blogs estão proliferando e se tornando grandes influenciadores de opinião na web. Blogueiros postam sobre diversos assuntos, como análises de produtos, opiniões políticas e tendências tecnológicas. Com essa influência em expansão, monitorá-las de forma contínua, e extrair informações úteis sobre a "opinião pública" ganha grande importância. Blogs apresentam suas informações com uma dimensão temporal bem definida que não está presente em conteúdo web mais tradicional. Além disto, um post de blog pode desencadear novos posts pelo mesmo blogueiro ou por outros, levando a uma discussão na blogosfera. Estes fatores tornam informações em blogs e sua dinâmica, significativamente diferentes do conteúdo tradicional da web, ocasionando uma necessidade de tecnologias especializadas de pesquisa e análise sobre esses textos, diferentes das utilizadas hoje. Muitos dos trabalhos encontrados focam em análises de termos presentes nestes textos, mas poucos se focam em análises dos textos como um todo e no relacionamento com outros. Aqui procuramos viabilizar análises temporais sobre o conteúdo destas postagens e mostrar como algumas entidades podem influenciar a popularidade de outras. Propõe-se também que a ferramenta sirva com um agregador multifacetado de informações relevantes para uma determinada área e que não sofra grande interferência das fontes mais tradicionais de notícias.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

BLOGMINER: DYNAMIC ABOUTNESS REPRESENTATION BY TOPIC MODELLING

Júlia Ferreira de Almeida

September/2012

Advisors: Geraldo Bonorino Xexéo

Department: Computer Science Engineering

Blog posts are proliferating and are now great opinion leaders on the web. Blog authors post about various topics such as product reviews, political and technology trends. With their expanding influence, it is of great importance to monitor them and continuously extract useful information about the "public opinion". Blogs present their information with a well-defined temporal dimension that does not exist in any other traditional web content. Furthermore, a blog post is able to spark new posts, by the same author or others, leading to discussion over the blogosphere. These factors make information on blogs and their dynamics significantly different from traditional web content, and thus rises the need for specialized technologies, different from those used today, in order to research and analyze these texts. Many of the studies are focused on analysis of terms present in these texts, but few focus on analysis of texts as a whole and in their relationship with other texts. This work seeks viable temporal analysis on the content of these posts and shows how some entities can influence the popularity of others. It is also proposed a tool that serves a multifaceted aggregator of information relevant to a particular area and does not suffer major interference from more traditional news sources.

ÍNDICE

LISTAGEM DE FIGURAS

Capítulo 1 Introdução

1.1 Motivação

A adoção maciça de mídia social criou novas formas dos indivíduos expressarem suas opiniões on-line. Em 2007 existiam mais de 50 milhões de blogs, e cerca de cem mil novos blogs eram criados todos os dias [5]. Hoje esse número é maior, em torno de 450 milhões, mas cresce bem mais lentamente, em torno de 40 mil a cada dia [32].

Blogueiros[1] postam sobre diversos assuntos, incluindo suas vidas pessoais, análises de produtos, opiniões políticas, tendências tecnológicas, experiências de turismo, eventos esportivos e indústria do entretenimento.

Sem dúvida, o blog é um fenômeno social. Este fenômeno vai persistir e crescer, assim como nossas vidas tornam-se mais fortemente dependentes das tecnologias da Internet. Dado o crescimento exponencial da quantidade de blogs, surge uma possibilidade interessante de monitorá-los de forma contínua, e extrair informações úteis sobre a "opinião pública" em uma variedade de assuntos.

Com essa explosão da comunicação e publicação na Internet, processar com sucesso textos relativamente curtos, informais e que levam em consideração a dimensão tempo, como mensagens de fóruns e *chats*, *feeds* e notícias de blogs, análises de produtos, resumos de filmes e livros, se torna cada vez mais relevante na área de recuperação da informação.

Encontrar textos que falem sobre o mesmo assunto, classificar e agrupá-los, levando sempre em consideração as datas de postagem, acarreta novos desafios. Ao contrário de documentos “normais”, estes segmentos de texto são mais ruidosos, menos focados em assuntos específicos, e muito menores, ou seja, eles são formados por uma reduzida quantidade de palavras contidas em algumas frases. Por causa do comprimento muitas vezes curto, eles não proporcionam uma coocorrência suficiente de palavras ou compartilhamento de contexto para uma boa medida de similaridade. Portanto, métodos de aprendizagem de máquina de tarefas textuais geralmente não alcançam o desempenho desejado devido à “escassez” de dados. [2]

[1] Designação dada a quem escreve em um blog

Ao levar em consideração o momento das postagens dos blogs, podemos tentar detectar períodos em que a popularidade de um termo ou evento específico aumenta drasticamente, marcando assim "estouros". Podemos também tentar descobrir assuntos interessantes em intervalos de tempo específicos, assim como entender como eles se desenvolveram durante o tempo. Outra informação interessante pode ser a obtenção de correlações de palavras-chave, ranking de blogueiros e blogueiros influentes e assim por diante.

A identificação dos assuntos subjacentes é essencial para selecionar e estabelecer o estado da arte de áreas de pesquisa e empreendimentos de negócios que seriam atrativas, por exemplo. [1]

Muitos dos trabalhos encontrados focam em análises de termos presentes em documentos dentro da Blogosfera, também chamados de *tags*, mas poucos se focam em análises dos textos como um todo e como eles se relacionam com outros.

Além destas análises procura-se também agregar o conteúdo coletado de vários blogs diferentes, para que a ferramenta sirva com um agregador multifacetado de informações relevantes para uma determinada área e, que não sofra grande interferência das fontes mais tradicionais de notícias, tais como grandes portais.

1.2 Problema

Por serem disponibilizados em fluxos ao longo do tempo, documentos eletrônicos como as postagens/comentários em blogs, possuem conteúdo com uma ordem temporal forte. Considerar a informação sobre o tempo é essencial para entender melhor os assuntos subjacentes e rastrear a evolução e propagação destes dentro de seus domínios. Por exemplo, o domínio de blogs sobre política ou tecnologia.

Além disso, a literatura trabalha mais com blocos fixos de tempo onde grandes coleções de texto são coletadas de forma total antes de qualquer análise, sendo que seria mais interessante e real para aplicações: analisar, resumir e categorizar o fluxo de textos em fatias de tempo dinâmicas como se fossem de certa forma em tempo real. Por exemplo, como notícias chegam em fluxos, organizá-las como *threads* de postagens relevantes é mais eficiente e conveniente. Áreas como TDT (*topic detection and tracking*) já trabalham com

fatias de tempo mais dinâmicas, mas apresentam um custo de tempo de modelagem de tópicos bastante elevado. [27]

Com a avalanche de fluxos de postagens vindos de toda a Web, são necessárias novas formas de processar documentos que facilitem a extração automática de informação útil. Uma abordagem que tenta entender os aspectos chave de um documento ou um conjunto de documentos é analisar os eventos presentes nestes documentos e automaticamente encontrar "cenários de eventos" relacionados. Chamamos "cenário de evento" um conjunto de eventos que podem interagir uns com os outros em situações específicas. Por exemplo, a prisão de um traficante de drogas. O cenário do evento descreve a prisão do criminoso. A acusação de um crime e a captura do criminoso seguida pelo seu interrogatório são eventos típicos que acontecem em um cenário de crime. [3]

Informação em blogs tem uma dimensão temporal bem definida que não está presente em conteúdo web mais tradicional. Além disso, um post de blog pode desencadear novos posts pelo mesmo blogueiro ou por outros levando a uma discussão na blogosfera. Esses fatores tornam informações em blogs e sua dinâmica, significativamente diferentes do conteúdo tradicional da web, e, portanto, há uma necessidade de tecnologia especializada de pesquisa e análise sobre estes textos. [5]

Apesar de existirem inúmeros sites de busca especializando-se na Blogosfera, os resultados retornados ainda não são muito interessantes, como o próprio Google[2], pois não permitem análises em cima dos dados, apenas filtros. Além de se focarem em termos e não nos assuntos em si e seu aspecto temporal. Assuntos estes, que nada mais são do que as "histórias" contidas nas postagens.

1.3 Objetivo do Trabalho

Dado que conteúdos gerados por usuários de mídia social, em nosso caso postagens em blogs, estão proliferando e se tornando grandes influenciadores de opinião na web, propomos um modelo que recupera a informação contida nesses textos, e permite análises sobre essas informações. E através desta informação coletada na forma de assuntos, pode-se

[2] http://www.google.com/blogsearch?hl=en

conduzir análises em grande escala dos dados em mídias sociais, no caso em blogs para rastrear as entidades e histórias de interesse nas postagens.

O modelo proposto busca facilitar a exploração interativa de conteúdo, permitindo que os usuários descubram assuntos interessantes ou surpreendentes. Assuntos de interesse podem ser explorados de diversas maneiras, como obtendo conteúdo relacionado e examinando a suas evoluções temporais. [6]

Resumindo, este trabalho tem como objetivo permitir ao usuário analisar a dinâmica dos assuntos contidos nas postagens em blogs, levando em conta a forte influência temporal e como algumas entidades podem influenciar a popularidade de outras.

Toda essa análise poderá servir de estudo na realização de prospecções tecnológicas, análise de tendências de mercado, auxilio aos cientistas sociais na compreensão de processos eleitorais, fonte de informações condensadas de uma determinada área, etc.

O trabalho apresenta um modelo conceitual e um protótipo de uma ferramenta de análise dos assuntos contidos na Blogosfera, que automaticamente captura padrões temáticos e identifica assuntos "quentes" nos fluxos de textos e suas mudanças ao longo do tempo.

A abordagem proposta permite que a estrutura de modelagem de tópicos- representação concreta dos assuntos- aja de forma que permita análises sobre textos coletados em datas não muito anteriores a atual, construindo gradativamente um modelo atualizado, com uma mistura de assuntos por documento e de palavras por assunto, quando um novo documento (ou um conjunto de documentos) aparece.

O trabalho apresenta também análises sobre os termos mais relevantes de cada documento. É possível detectar períodos de *bursts*[3] através de gráficos dia-a-dia de popularidade destes termos (assim como ver os termos correlacionados ao buscado no período de tempo pré-selecionado) e efetuar análises comparativas de popularidade entre termos.

A ideia é atualizar incrementalmente a distribuição mais atual dos assuntos, de acordo com a informação inferida do novo fluxo de dados, sem a necessidade de acessar dados mais antigos. A dinâmica da abordagem proposta também provê um meio eficiente de rastrear os assuntos ao longo de tempo e detectar termos emergentes. [1]

[3] Períodos de tempo em que há picos positivos de popularidade

Utilizamos modelos de tópico (assuntos) que exploram as correlações entre as palavras e os temas latentes em fluxos de documentos. Modelos de Tópico podem extrair uma estrutura bastante interpretável e útil, sem qualquer "entendimento" explícito da língua pelo computador.

Apresentamos um modelo de tópicos correlacionados, que modela a correlação entre os assuntos contidos na coleção, e permite a construção de gráficos de tópicos e "navegadores" de documentos que possibilitam ao usuário navegar na coleção de uma maneira mais guiada. [4]

O ponto de partida deste trabalho é uma limitação percebida em modelos de tópicos como LDA: eles não conseguem modelar diretamente correlação entre os tópicos. Na maioria das coleções de textos, é natural esperar que os temas presentes sejam altamente correlacionados. Na Ciência, por exemplo, um artigo sobre genética pode ser provavelmente também sobre saúde e doença, mas improvável de ser sobre astronomia.

Para o modelo LDA, a limitação citada acima decorre das suposições de independência implícitas na distribuição Dirichlet, relativas às proporções dos tópicos. Em uma Dirichlet, os componentes do vetor de proporções são quase independentes, o que leva a suposição de que um tópico não é correlacionado com a presença de outro. [4]

Com isso, antes de desenvolver quaisquer modelagens é preciso encontrar formas de correlacionar os assuntos (tópicos), para até mesmo sabermos se eles se tratam do mesmo (só com pequenas diferenças) ou não.

1.4 Metodologia de Pesquisa

O trabalho pretende seguir as etapas da concepção atual do método científico moderno, segundo o qual uma investigação alcança seus objetivos de forma científica quando cumpre ou se propõe a cumprir as seguintes etapas [9]:

- Descoberta do problema;
- Colocação precisa do problema;
- Procura de conhecimentos ou instrumentos relevantes ao problema;
- Tentativa de solução do problema com o auxílio dos meios identificados;
- Produção de novos dados empíricos;
- Obtenção da solução;

- Investigação das consequências da solução obtida;
- Prova (comprovação) da solução;
- Correção das hipóteses, teorias, procedimentos ou dados empregados na obtenção da solução incorreta.

Em conformidade com esse passo, a descoberta do problema ocorreu através de certa necessidade em analisar melhor as postagens em blogs. Tal necessidade fez surgir um questionamento de como se desenvolver um conjunto de métodos que resultem em uma ferramenta que permita minerar de modo temporal as postagens em uma determinada área de blogs de forma completa e flexível. O problema a ser solucionado é propor um conjunto de métodos de recuperação da informação voltados especificamente para blogs, com o objetivo de relacionar e agrupar suas postagens satisfatoriamente.

Durante a procura de conhecimentos ou instrumentos relevantes ao problema, foi necessário um estudo sobre os formatos de disponibilização de notícias em blogs, mecanismos analíticos, medidas de similaridade e diversos aspectos relacionados a sistema de recuperação da informação.

A produção de dados empíricos foi realizada através da utilização e avaliação da ferramenta desenvolvida, responsável por coletar, processar, agrupar e exibir os assuntos provenientes de diversos blogs.

Através da execução e utilização da ferramenta, encontramos a solução, e investigamos e comprovamos suas consequências através de experimentos que visavam avaliá-la.

A correção de alguns procedimentos empregados foi realizada com o objetivo de deixarmos a ferramenta de acordo com as necessidades que foram encontradas durante o experimento.

1.5 Organização do Texto

O Capítulo 1 deste trabalho corresponde a presente introdução.

O Capítulo 2 apresenta uma visão geral do que são blogs e porque a análise deles é uma área interessante, além de exemplos de ferramentas da área.

No Capítulo 3 são apresentadas as características mais relevantes da área de recuperação da informação, tendo como ponto de partida uma análise detalhada das principais técnicas e algoritmos utilizados no desenvolvimento da ferramenta proposta.

O Capítulo 4 da dissertação descreve algumas aplicações de LDA ou análises de termos em blogs. Para isso, trabalhos relacionados são detalhados em comparação à proposta apresentada nesta pesquisa.

O Capítulo 5 descreve o modelo proposto por este trabalho.

O Capítulo 6 apresenta uma descrição geral da ferramenta proposta de análises de blogs *BlogMiner*.

O Capítulo 7 detalha o funcionamento dos módulos pertencentes à arquitetura do sistema.

No Capítulo 8 descrevemos os experimentos realizados, juntamente com os resultados retornados, de forma a mostrar que a ferramenta desenvolvida satisfaz os requisitos propostos.

O Capítulo 9 apresenta a conclusão desta pesquisa, ressaltando suas contribuições e sugerindo futuras melhorias.

Por fim o Capítulo 10 lista as referências bibliográficas.

Capítulo 2 Análise de Blogs

2.1 Blogs

Blogs são sites pessoais, geralmente desenvolvidos por ferramentas específicas, que possuem certas características próprias tais como:

- o arranjo cronológico das informações, apresentando a última atualização na parte mais superior da página;
- a vasta utilização da hipertextualidade, para referenciar tanto outros blogs e sites como os materiais utilizados como referência para discutir determinado assunto;
- a atualização frequente, e
- a possibilidade de interatividade.

A facilidade de publicação de materiais na web acarreta um confronto com uma vasta gama de interesses particulares, condensados em blogs com diversos tipos de conteúdos, distintos entre si. O artigo [36] cita uma análise estrutural sobre o assunto, em que se categorizam os blogs em: diários, publicações, literários, clippings e publicações mistas.

Blogs começaram a se tornar populares depois da criação do primeiro software de *blogging* em 1999 e se tornaram grandes influenciadores de opinião após os comentários dos blogueiros no episódio do "11 de setembro" e da guerra do Iraque em 2003.

Assim como outros documentos da web, os blogs podem ser multimodais ou puramente textuais. Um aspecto em que eles se diferenciam da mídia tradicional é a opção que os blogueiros têm de permitir que os leitores comentem sobre suas postagens no blog, o que pode dar origem a trocas comunicativas entre blogueiros e comentadores dentro de um único blog, o que funde a fronteira entre páginas estáticas HTML e fóruns de discussão interativos.

Na mídia tradicional já é possível comentar também sobre as notícias publicadas, mas a relevância dada a esses comentários ainda é menor em comparação aos comentários em blogs. Um dos fatores levados em conta para um blog ser descrito como "popular" é a quantidade de comentários por postagem que ele possui -existem grupos de pessoas especializados em comentar blogs-, o que ainda não é tanto o caso da mídia tradicional.[46]

Devido à sua natureza temporal e acessível, os blogs originaram um poderoso fenômeno social, com as discussões em blog muitas vezes influenciando os meios de comunicação e a opinião

pública, além da indústria de marketing. Blogs possuem estrutura de "comunidade" e aspectos de dinâmica temporal, o que os torna um domínio de estudo mais rico do que páginas estáticas da Web.

2.2 Análise

Blogueiros, pessoas que escrevem blogs, usam de suas características singulares para expressar livremente suas opiniões e emoções, tornando os blogs cada vez mais populares. Uma análise destas expressões pessoais poderiam criar oportunidades para os governos e empresas, por exemplo, compreenderem o público de uma maneira que anteriormente era caro ou mesmo indisponível.

Apesar da blogosfera conter muita informação útil, os dados são ruidosos pelo fato dos blogs não serem estruturados e cobrirem uma vasta variedade de assuntos. Para minerar as informações valiosas é preciso ferramentas especializadas nesta área.

Ao analisar a expressão de opiniões dos blogueiros através da análise de blogs, comerciantes, por exemplo, podem entender melhor seus clientes, ou usuários comuns podem saber mais sobre o que estão dizendo sobre determinados produtos, empresas ou questões políticas. Entretanto, dado o grande número de blogs existentes, monitorar e analisar manualmente este grande número de dados é um trabalho intenso e extremamente demorado se realizado por seres humanos.

Intuitivamente, a primeira coisa que pensamos é em utilizar técnicas de mineração de texto para análises de blogs, mas dados os inúmeros desafios, não é aconselhável usar diretamente estas técnicas. Um dos desafios é o fato de blogueiros falarem sobre diversos assuntos em uma mesma postagem, com isso possivelmente apenas um parágrafo poderia ser relacionado ao interesse de algum usuário – por exemplo, um produto sendo analisado.

Além do que foi citado anteriormente, com o crescente número de blogs interessantes, os usuários comuns estão cada vez mais usando os chamados agregadores de notícias como um ponto único de leitura do que mais lhes chamam a atenção.

Apesar de já serem muito úteis sendo pontos únicos de acesso a várias informações de interesse da pessoa, eles ainda têm deficiência no campo de Recuperação da Informação. Em geral, os usuários não podem, nativamente, fazer nenhum tipo de análise a partir dos documentos coletados pelo agregador, como por exemplo, quais assuntos são novos ou qual a curva de popularidade de algum outro.

Assim como para o usuário comum é interessante saber as *Hot Trends* dos blogs presentes em seu agregador, elas podem ser uma mina rica de dados para marketing online que tenta detectar algum desvio de curiosidade do público ao longo do tempo.

2.3 Ferramentas

A seguir, apresentaremos ferramentas, de certa forma, comerciais que foram consideradas interessantes e relevantes para o nosso trabalho durante nossas pesquisas. Estas ferramentas apresentam uma visão do que está sendo feito nas áreas que tentamos atingir, dentro do contexto de notícias: avaliação de impacto e agregação de notícias.

2.3.1 Avaliadores de Impacto-Buzz

Estas ferramentas mostram como um dado termo se comporta ao longo de um período de tempo, pré-determinado pelo usuário. Apresentam as seguintes características:

- Apresentam as notícias relacionadas aos picos apresentados na curva;
- Podem comparar com outros termos;
- Em geral não filtram as buscas somente por blogs;
- Apresentam apenas o gráfico relacionado à popularidade sem nenhuma funcionalidade a mais;

Algumas ferramentas são muito interessantes, mas não focam em análises sobre os assuntos contidos nos documentos (o que poderia enriquecer as informações retornadas) somente sobre os termos. Termos estes que podem ser definidos como palavras presentes nos documentos e que serão mais detalhados em capítulos posteriores. Em geral essas ferramentas são voltadas para a análise de "marcas" e não sobre notícias.

Os motores de busca sempre tiveram visão perspicaz quando se trata de saber o que está quente e o que não está, mas a maioria dos usuários da web ainda não se deu conta de que corporações como o Google e Yahoo! criaram aberturas para esses dados valiosos que qualquer um pode manusear, através de ferramentas que possibilitam "análises" sobre estes dados.

Analisamos a seguir a abordagem de busca sobre dados que são considerados "tendência"(Trends) de dois dos maiores "influenciadores" da Internet: Yahoo! e Google.[39]

2.3.1.1 Google Trends

O Google Trends[4] é uma ferramenta que permite o usuário ver por debaixo da superfície de milhares de buscas diárias ao Google. Ao submeter termos para a pesquisa, a ferramenta irá retornar quantas pesquisas sobre esse termo foram feitas (em relação às buscas totais do Google) ao longo de um período determinado de tempo.

As notícias relacionadas aos picos de volume de pesquisa são colocadas à direita do gráfico, enquanto outro gráfico abaixo do principal apresenta o número de vezes que o termo apareceu em notícias durante o mesmo período. Esta ferramenta do Google permite também que o usuário se aprofunde pelas regiões e cidades que possui interesse em uma determinada pesquisa.

A Figura 1 apresenta um exemplo de busca no Google Trends. Foram comparados os termos "Hillary Clinton" e "Barack Obama" durante o ano de 2007. É possível através da Figura 1 percebermos também uma pequena falha em que aparecem notícias relacionadas ao pico de popularidade de janeiro de 2008, sendo que a busca foi filtrada por ano, no caso 2007.

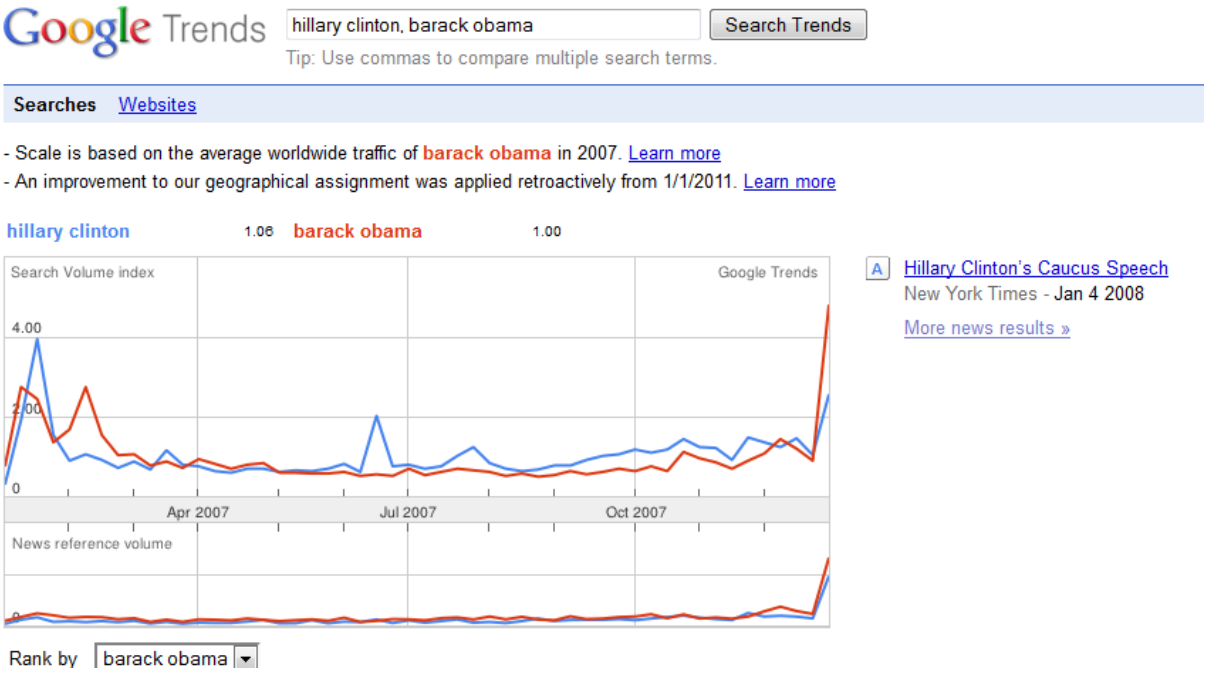


Figura 1: Google Trends

Os dados do Trends são atualizados a cada hora. A opção "Hot Trends"[5](Figura 2) apresenta ao usuário os termos que tiveram picos de interesse no dia em que ele está interessado. Ela apresenta

[4] http://www.google.com/trends/

[5] http://www.google.com/trends/hottrends

também os posts em blogs, sites relacionados e notícias do dia que contêm o termo que ajudam a entender o motivo da onda de popularidade. A cada termo é dada uma classificação de calor e alguns dados sobre o momento em que houve maior interesse no termo e onde as buscas ocorriam em maior quantidade.

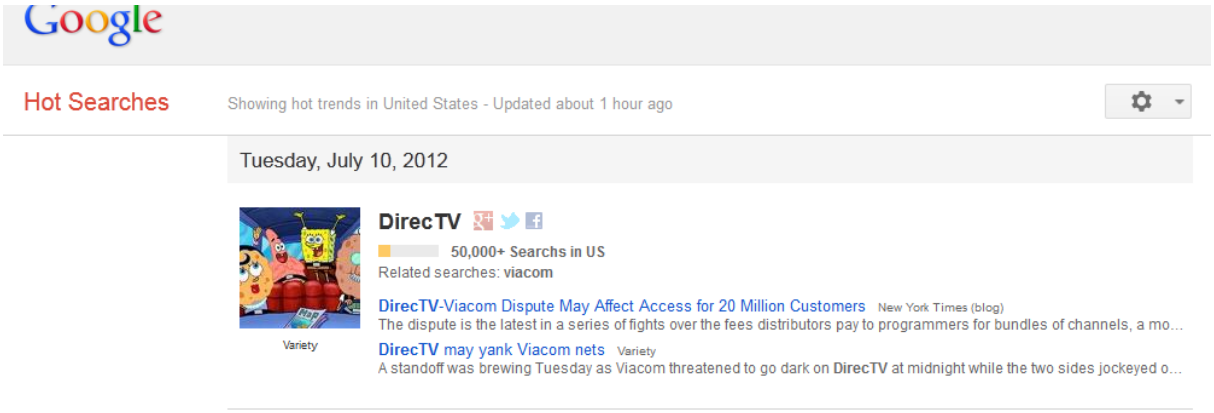


Figura 2: Hot Trend

2.3.1.2 Yahoo! Buzz Index

Assim como o Google, o Yahoo! possui sua ferramenta de tendências que é definida como: "a pontuação do assunto do Buzz é a porcentagem de usuários do Yahoo! procurando por este termo em um determinado dia, multiplicado por uma constante para tornar o número mais fácil de ler. Os líderes semanais são os assuntos com maior média de pontuação de Buzz para uma determinada semana" .

Segundo [39], o Buzz Index[6] é considerado menos *nerd* e interativo do que a ferramenta Google Trends e apresenta um marcador estilo "mercado de ações" dos assuntos que mais impulsionaram e abalaram as pesquisas do dia. Ele é habilmente dividido em categorias, de modo que o usuário pode acompanhar de maneira online a popularidade de atores, filmes ou programas de TV, entre outros.

O Yahoo! Buzz Index editoria seus serviços através de um Blog e de RSS com suas atualizações. Assim como o Google eles tentaram também regionalizar seus dados, com lista de *buzz* canadenses e franceses. E desde que o Yahoo! abriu um pouco do seu conteúdo de *buzz* através do RSS, alguns *mashups* interessantes estão começando a aparecer. A Figura 3 apresenta a interface do Yahoo! Buzz.

[6] http://buzzlog.yahoo.com/overall/

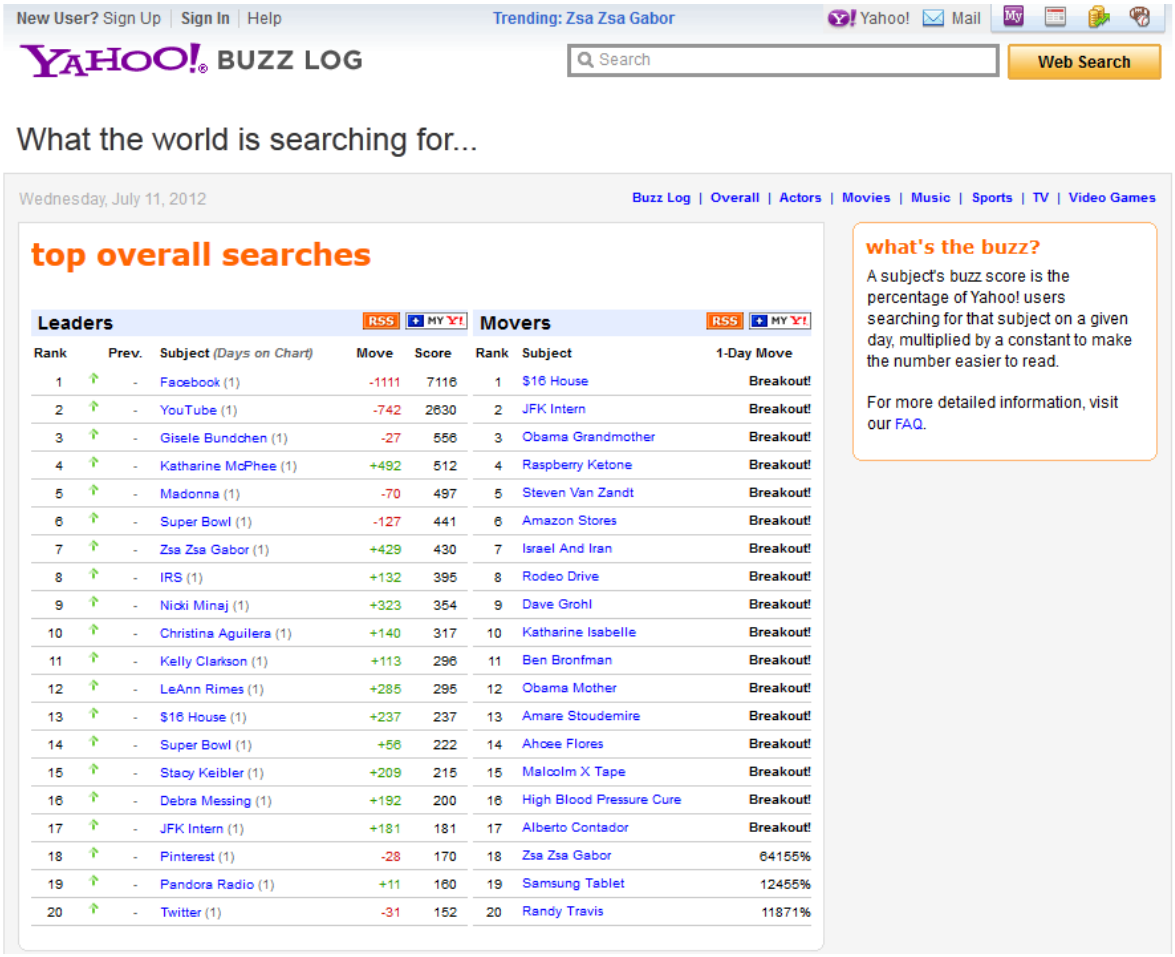


Figura 3: Yahoo Buzz

2.3.1.3 Buzz Trends-Nielsen

Cada vez mais, as pessoas vêm se afastando das mídias tradicionais como a televisão, rádio e jornais e se aproximando das mídias sociais, criando e compartilhando seus próprios conteúdos através de sites como Youtube, Facebook, Blogs e Twitter. Mas para os proprietários destas, relativamente novas, mídias existe um grande problema: Como fazer dinheiro a partir desta popularidade?

O enigma para a mídia social é que se você tentar comercializar o espaço gerado pelo usuário ele perde o encanto para o usuário que quer estar no controle. Mas para empresas de pesquisa, não existe enigma apenas um bom nicho a ser explorado. Uma empresa que vem trabalhando neste nicho é a Nielsen BuzzMetrics[7], que têm como objetivo medir o buzz nas mídias sociais como os blogs e fóruns e depois empacotar estes dados para as empresas cliente.

[7] http://buzz-trends.com/tag/nielsen-ratings/

As áreas de marketing de empresas como a Toyota, Sony e Coca-cola utilizam o BuzzMetrics para saber o que as pessoas em redes sociais estão dizendo sobre suas marcas e produtos. Conseguindo medir e nutrir este *buzz*, os marqueteiros esperam transformar o *buzz* positivo em possíveis vendas.

A Figura 4 apresenta um exemplo de utilização do BuzzMetrics.

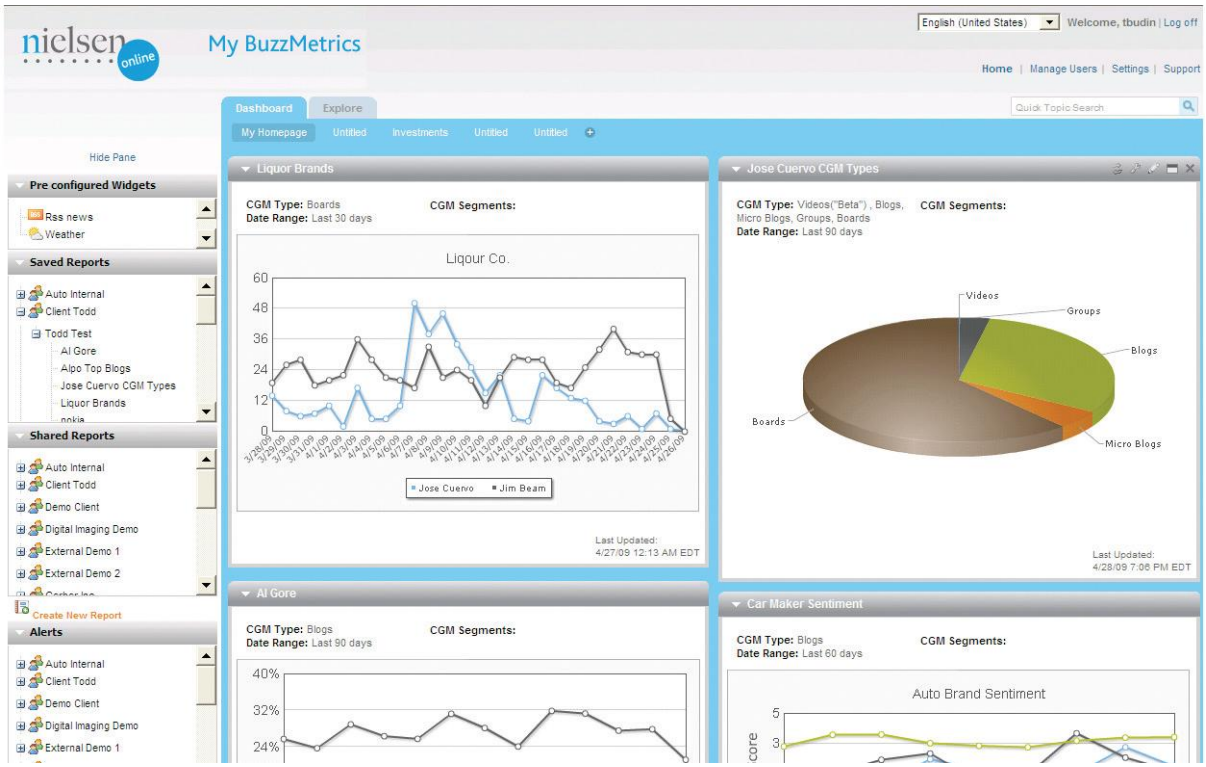


Figura 4: BuzzMetrics

2.3.2 Agregadores de Notícias

Com a mudança nos meios de comunicação, agregar notícias se tornou essencial em quase todas as organizações que trabalham com notícias. Leitores encontram um crescente e abundante volume e fontes de notícias.

Outros meios de comunicação profissionais estão acelerando sua produção. Seria interessante aproveitar-se disto, as organizações olharem além de suas próprias redações e dar ao leitor um resumo mais abrangente. A organização que percebe isto se torna a "primeira parada" e a mais frequente dos leitores. Adicionar a agregação pode retornar informações com maior profundidade e um público maior para as notícias originais.

Segundo [41], alguns itens têm que ser discutidos para se desenhar uma estratégia inteligente de agregação e fazer do site um ponto central de informações:

- Agregação automatizada ou manual;
- Como dar "poder" aos usuários;
- Escolher o que agregar;
- "Linkar" as notícias ou resumi-las;
- Como decidir entre múltiplas fontes de notícias;
- Escolher a frequência de postagem dos itens agregados;
- Como dar "poder" às suas fontes de notícias.

Há uma grande variedade de aplicativos e serviços que permitem que o usuário adicione apenas seus blogs favoritos e acompanhe seus *feeds*, mas existem alguns que, além disto, auxiliam o usuário a descobrir notícias novas e interessantes. A seguir falaremos um pouco mais sobre alguns deles.

2.3.2.1 Google News/ Reader

Google News[8] (Figura 5) e Reader[9] ainda são, provavelmente, os serviços de agregação de notícias mais populares da web, quando se considera os dois juntos. Google Reader é um leitor robusto de *feeds* e que permite que o usuário: adicione quantos *feeds* quiser, organizá-los e ler tudo ou assuntos/*feeds* selecionados ou apenas o que o usuário ainda não viu desde a última atualização. O Google News é um serviço de notícias que agrega milhares de blogs, jornais, agências de notícias e revistas on-line e apresenta as tendências e os conteúdos que são mais interessantes ao usuário quase que instantaneamente.

[8] http://news.google.com/

[9] www.google.com.br/reader/

Figura 5: Google News

2.3.2.2 Newsola

A Internet está sempre buscando novas maneiras de mostrar suas notícias. Sejam novas maneiras de exibir os feeds do Google Notícias do usuário ou alguma forma inteiramente nova de assinar os feeds RSS, há sempre algo novo surgindo a cada dia. Alguns deles utilizam muitos gráficos, e outros se focam em acabar com os extras e construir uma forma simples. Newsola combina a simplicidade com artifícios visuais para fazer um leitor de notícias "diferente".

Newsola divide seus feeds em seis categorias. O usuário pode ver notícias do Mundo, Nacional, Showbiz, Esportes, Tecnologia e Finanças. Ele possui uma tela com as manchetes, codificadas através de cores por categoria para que o usuário possa encontrar o que deseja. O usuário também pode optar por mostrar assuntos de todas as categorias, ou somente as que acha interessante. Os assuntos mais relevantes são exibidos em textos maiores, e as menos, com menores. Além dos itens anteriores o usuário também pode escolher seu país no menu superior. A Figura 6 apresenta a interface do Newsola.[42]

Apesar de "diferente", segundo uma pesquisa informal feita durante este trabalho, a maioria dos usuários ainda prefere formas que apresentam a notícia com mais imagens e menos texto.

Figura 6: Newsola

2.3.2.3 10x10

Segundo seu próprio site, 10x10 ("10 por 10")[10] é uma exploração interativa das palavras e imagens que definem o momento atual. O resultado está em uma mudança constante, por ser atualizado de hora em hora, mas sempre dando uma noção do mundo. A cada hora, 10x10 recolhe as 100 palavras e imagens que foram mais importantes em uma escala global, e as apresentam na forma de imagens, que servem para encapsular o dado momento no tempo. Ao longo de dias, meses e anos, 10x10 armazena estas informações feitas por hora que, servem para dar um panorama do desenrolar das notícias.

A cada hora é apresentada uma tela composta por 100 quadros diferentes, cada uma das quais contendo a imagem de um único momento no tempo. Ao clicar em um determinado quadro, o usuário visualiza um pouco mais a fundo a história que está por trás da imagem. Desta forma, o usuário pode se aprofundar ou ter uma visão macro das notícias e a compreensão de tanto as histórias individuais como a forma pelas quais eles se relacionam.

A Figura 7 apresenta a interface do 10x10.

[10] http://tenbyten.org/10x10.html

Figura 7: 10x10

2.4 Conclusões

A seguir apresentamos uma tabela comparativa entre as ferramentas citadas neste capítulo e o modelo de ferramenta proposto neste trabalho (BlogMiner), sendo o que foi considerado interessante enfatizado em lilás e o comum em salmon. Foram também divididas em subgrupos as com o foco em Buzz e os Agregadores de Notícias.

Após analisar as ferramentas citadas anteriormente podemos dizer que uma das principais contribuições deste trabalho é apresentar em um só lugar várias das características mais interessantes presentes nas outras ferramentas, levando em consideração a dimensão tempo e os assuntos em si (na maioria das vezes as ferramentas só trabalham com termos), além de ter bases matemáticas mais aprofundadas, aumentando indiretamente a confiabilidade dos dados.

		Buzz			Agregadores		
	G.Trends	Y.Buzz	Nielsen	G.News/Reader	Newsola	10x10	BlogMiner
Gráfico temporal de termos	x		x				x
Gráfico temporal de assuntos							
Assuntos relacionados	x					x	x
Termos correlacionados						x	x
Filtra geograficamente	x				x		
Ferramenta analítica			x				x
Confiabilidade das informações	x			x	x		x
Forma inovadora de ver as notícias					x	x	x
Foco no visual			x		x	x	x
Várias fontes de notícias	x			x	x	x	x
Boa usabilidade	x			x	x		x
Agregam notícias							
Foco em blogs							

Capítulo 3 Revisão da Literatura

Apresentamos neste capítulo referências da literatura sobre os principais assuntos presentes neste trabalho. Entre eles podemos destacar a recuperação da informação, que serve como base para o processo de modelagem que gera os tópicos automaticamente (como o LDA), medidas de similaridade entre os tópicos gerados, que visam determinar a relação entre tópicos presentes em períodos de tempo diferentes, com o objetivo de analisar se trata-se do mesmo tópico com apenas algumas variações; analisa-se também o grau de *Collocation*[11] entre termos que aparecem em um certo período de tempo escolhido pelo usuário e o grau de covariância entre tópicos gerados; e formas de coletar e armazenar documentos XML(dos posts de blog coletados) para a posterior análise dos dados.

Por fim, temos a Análise Formal de Conceitos (FCA) que auxilia na compreensão de como os assuntos variam ao longo do tempo, de acordo com os termos contidos nestes.

3.1 Recuperação da Informação (RI)

Recuperação da Informação pode ter um leque variado de significados dependendo da área e do contexto utilizado. [8] define RI como um artifício para encontrar documentos que satisfaçam certa necessidade de informação dentro de grandes coleções.

Nos anos 1990, estudos mostraram que a maioria das pessoas preferia saber sobre informações por outras pessoas ao invés de sistemas de recuperação de informação. Apesar de que, nessa época, a maioria das pessoas também preferia usar agentes "humanos" para reservar as suas viagens, por exemplo.

No entanto, durante as últimas décadas, a elevada otimização da área de Recuperação de Informação tem levado os motores de busca da web para novos níveis de qualidade onde a maioria das pessoas está satisfeita a maior parte do tempo, e pesquisas na web se tornaram uma fonte padrão e muitas vezes preferida para encontrar informação. Por exemplo, já em 2004 estudos mostravam que 92% dos usuários diziam que a Internet era um bom lugar para obter informações todos os dias.

Para a surpresa de muitos, o campo de Recuperação da Informação deixou de ser uma disciplina principalmente acadêmica para ser o acesso à informação preferido pela maioria das pessoas.

[11] Quando dois termos aparecem com uma distância reduzida entre eles, em um mesmo documento.

Recuperação da Informação não começou com a web. Em resposta a dificuldades de prover acesso a informações, este campo evoluiu para a criação de princípios para a busca de várias formas de conteúdo. O domínio começou com publicações científicas e registros de bibliotecas, mas logo se espalhou para outras formas de conteúdo, especialmente os dos profissionais da informação, tais como jornalistas, advogados e médicos. Grande parte da pesquisa científica sobre recuperação da informação tem ocorrido nestes contextos, e muito dela também trabalha com o acesso a informações não estruturadas em vários domínios empresariais e governamentais.

Grandes inovações científicas, avanços da engenharia e o grande declínio do preço do hardware de computador, por exemplo, conspiraram para termos os grandes motores de busca atuais, que são capazes de fornecer resultados de alta qualidade dentro de tempos de resposta de milissegundos para centenas de milhões de buscas por dia em bilhões de páginas da web. [7]

O nosso trabalho está em grande parte relacionado com a recuperação de documentos e com as informações que podem ser obtidas indiretamente destes. Com isso, a Recuperação da Informação assume um papel de suma relevância, pois ela serve como base para a descoberta de tópicos presentes no Corpus.

3.2 Medidas de Similaridade

Similaridade é um conceito fundamental e amplamente utilizado. Muitos métodos de similaridades têm sido propostos, tais como o coeficiente de Dice [11 apud 21], coeficiente por cosseno [11 apud 21], baseados em medições de distância [11 apud 22], modelo de recurso de contraste [11 apud 23], entre outros.

Esse conceito fundamental pode ser definido como a semelhança entre A e B, quanto mais coisas comuns eles compartilham mais parecidos eles são. Assim como também podemos dizer que a Similaridade está relacionada com as diferenças entre eles, quanto mais diferenças eles têm menos parecidos eles são. A semelhança máxima entre A e B é alcançada quando estes são idênticos, não importando quantas coisas comuns eles compartilham. [11]

3.2.1 Correlação

A informação na Blogosfera é altamente dinâmica por natureza. Ao longo da evolução dos tópicos, palavras-chave tendem a se alinhar para formar histórias, e quando os tópicos recuam, esses agrupamentos de palavras-chave tendem a se dissolver. Esta formação e dissolução de aglomerados de palavras-chave é capturada por este trabalho sob a forma de correlações. Tais palavras-chave

podem ser utilizadas para auxiliar na compreensão do contexto em que o termo buscado apareceu durante o período de tempo selecionado pelo usuário.

A grosso modo, as palavras-chave citadas acima são as que coocorrem mais frequentemente com os termos buscados. Correlações não são estáticas, elas podem, e geralmente variam de acordo com o intervalo temporal especificado na consulta. Essas correlações podem ser utilizadas para entender melhor a razão de *Bursts*, "estouros", de alguns termos. [10]

Correlações podem ser selecionadas pela frequência, pela média e variância da distância entre a palavra foco e a correlacionada, por testes de hipóteses, etc. Aqui optamos por trabalhar com técnicas que auxiliem na busca pelos melhores resultados, assuntos ou termos que pertencem a um mesmo evento, dentro de um período de tempo pré-determinado pelo usuário.

Certamente o modo mais simples de encontrar correlações em uma coleção de documentos é contando. Se duas palavras ocorrem muito juntas, então isso é evidência que elas possuem uma função especial que não é explicada simplesmente pela função que resulta da combinação entre elas.

Mas somente isto não é suficiente para o nosso caso, dado que só com essa informação podem surgir muitos falsos positivos, pois estamos em busca de termos que se relacionam durante um intervalo de tempo dentro de um contexto específico e não palavras compostas, por exemplo. Logo, analisar a distância média entre certos termos dentro do corpus (dos termos que aparecem dentro da faixa de frequência desejada e de outros que aparecem dentro dos mesmos documentos) se torna bastante interessante e necessário para enriquecer nossos resultados, assim como a probabilidade condicional entre eles.

Em nosso trabalho queremos descobrir quando duas palavras coocorrem bem mais que não só pelo acaso, o que no nosso caso determina termos que são relacionados. Avaliar se algo é ou não um acontecimento ao acaso é um problema clássico em estatística. É geralmente expresso em termos de Hipóteses. Nós utilizamos uma hipótese nula H0 que não há associação entre as palavras para além de ocorrências ao acaso, calculamos a probabilidade p de que o evento poderia ocorrer se H0 fosse verdade, e depois rejeitar H0 se p for muito baixo (normalmente é abaixo de um nível de significância de $p < 0.05$, 0.01, 0.005, ou 0.001) e manter H0 enquanto for possível de outra forma[12].

[12] Significância a um nível de 0.05 é a evidência mais fraca que é normalmente aceita na ciências experimentais.

É importante notar que este é um modo de análise de dados, onde olhamos para duas coisas ao mesmo tempo. Como antes, nós estamos procurando por padrões particulares nos dados. Mas estamos também levando em conta a quantidade de dados que nós já vimos. Mesmo que exista um padrão notável, iremos descontar isso se não vimos dados suficientes para ter certeza de que não poderia ser devido ao acaso.

Para aplicarmos a metodologia do teste de hipóteses, primeiro precisamos formular uma hipótese nula que diz o que deve ser verdade, se duas palavras não formam uma colocação. Para tal uma combinação livre de duas palavras, vamos supor que cada uma das palavras w1 e w2 são geradas de forma completamente independente da outra, e assim a chance de aparecem juntas é dada simplesmente por:

$$P(w_1w_2) = P(w_1)P(w_2)$$

Segundo o modelo, a probabilidade de coocorrência é o produto das probabilidades das palavras individualmente.

3.2.2 Similaridade por Cosseno

Segundo [13], dado um conjunto grande de itens (objetos) e dados de observação sobre a coocorrência destes itens, a análise de associação está preocupada com a identificação de subconjuntos que sejam fortemente relacionados. A análise associativa tornou-se um dos problemas centrais no campo da mineração de dados e desempenha um papel importante em muitos outros domínios de aplicação. Por exemplo, a análise de associação pode encontrar padrões que são úteis para promoção de vendas, gestão de prateleira, e gerenciamento de inventário.

Embora vários métodos escaláveis têm sido desenvolvidos para minerar padrões frequentes em análises de associação, o framework tradicional de apoio e confiança tem mostrado as suas limitações em descobrir relacionamentos interessantes.

Para enfrentar este desafio crítico, correlações estatísticas ou medidas de similaridade têm sido exploradas para a mineração de padrões associativos, como $\chi 2$, e a similaridade do cosseno. Entretanto, a maior parte destas medidas é utilizada apenas para pós-avaliação por não serem "amigáveis" computacionalmente.

A Similaridade por Cosseno mantém simetria, desigualdade triangular, invariância-nula [13 apud 24], e propriedades de suporte cruzado [13 apud 25]. Além disso, este estilo de similaridade é muito simples e tem um significado real, isto é, que mede a diferença de ângulo de dois vetores. Isto

faz com que a semelhança de coseno seja particularmente útil para medir a proximidade em um espaço altamente dimensional.

O cosseno é nulo-invariante e, portanto, é uma boa medida para minerar relacionamentos interessantes em bases de dados transacionais.

A Similaridade por Cosseno é a medida de similaridade entre dois vetores de n dimensões. Cada objeto é representado por um vetor e o cosseno da medida do ângulo entre eles representa o grau de similaridade. O valor da Similaridade por Cosseno, para vetores não negativos, sempre varia de [0,1], onde 1 indica uma combinação perfeita dos dois vetores (idênticos) e 0 o completo oposto.

$$\cos(\theta) = \frac{u \cdot v}{\|u\| \|v\|}$$

Neste trabalho utilizamos esse tipo de medida de similaridade para medir a similaridade entre dois tópicos de fluxos de textos diferentes para verificar o quão são semelhantes. Dentro de um limite pré-definido, que será mais bem explicado em capítulos posteriores, eles podem ser considerados sobre o mesmo assunto. Medimos a similaridade entre os vetores de probabilidades (probabilidade de um determinado conjunto de termos, os mesmos para os dois vetores estarem naquele tópico) dos tópicos presentes em certo período de tempo e do anterior a ele, por exemplo.

3.3 RSS

RSS (Really Simple Syndication, ou Rich Site Summary ou RDF Site Summary) é uma sintaxe que pertence a Web 2.0 para agregar conteúdo. Usuários podem usar o RSS para serem alertados de notícias relevantes, posts novos de blogs, podcasts, e etc. Pelo fato de RSS enviar "feeds" para um site agregador (por exemplo, o Google Reader), o usuário não precisa ser bombardeado com e-mails, e podendo limitar o período de tempo para alertas (por exemplo, as últimas duas semanas). [14]

Os documentos resultantes da agregação do conteúdo podem conter tanto um resumo como o conteúdo completo. Eles apresentam também informações sobre a data de publicação e do emissor do conteúdo.

Os feeds são escritos em XML, sendo que atualmente existem três especificações importantes para a criação desses arquivos:

- RSS 1.0;
- RSS 2.0;

- Atom.

Sendo que o formato RSS 2.0 é o mais utilizado atualmente.

RSS é muito utilizado pelos blogueiros, com ele um usuário pode assinar certos blogs ou palavras-chave e depois receber todos os itens relevantes em um único lugar. O usuário pode utilizar apenas o RSS ou outros agregadores, para compartilhar as últimas manchetes ou os textos completos sem precisar monitorar periodicamente atualizações.

Nós coletamos durante um ano, RSS de blogs em inglês para um dos experimentos. Escolhemos trabalhar com eles por gerarem *feeds* em XML com conteúdo bastante estruturado e sem muito "lixo" como é o caso de quando coletamos informações sobre sites comuns através de crawlers, o que nos ajuda a ganhar tempo e precisão ao realizarmos a Extração, Transformação e Carregamento dos dados para o banco de dados relacional.

3.3.1 YQL

YQL (Yahoo! Query Language) é uma linguagem estilo MySQL para uma API Yahoo! que trata todas as formas de dados on-line como tabelas. Os programadores podem acessar e "misturar" dados a partir de praticamente qualquer Web Service, *feed* RSS, HTML, ou mesmo XML estático e planilhas.

Utilizamos YQL para unir em um só documento *feeds* de diversas fontes em um mesmo período de tempo, assim como só transpassando para eles os itens de interesse contidos no RSS.

3.4 Agregadores de Conteúdo

Segundo citação em [43], "A agregação de conteúdo (content syndication) pode ser definida como uma forma de localização controlada de um mesmo conteúdo em múltiplos destinos na internet. Geralmente, ela se refere à disponibilização de *feeds* web de uma determinada página com o objetivo de fornecer a outras pessoas um resumo ou atualização do conteúdo desses sites (por exemplo, as últimas notícias de um jornal web".

RSSs podem conter tanto um resumo como o conteúdo completo de um documento. Diferentemente das páginas HTML, que apresentam seu conteúdo em qualquer navegador web, o conteúdo dos *feeds* são apresentados apenas através dos chamados agregadores. Um dos agregadores de conteúdo mais populares é o Google Reader[13], mas existem muitos outros utilizados.

[13] www.google.com.br/reader

Eles servem como um ponto central em que o usuário pode visitar para saber as últimas informações sobre blogs de seu interesse e, por conseguinte áreas que o interessa, como por exemplo, Humor ou Tecnologia.

3.5 Modelagem Probabilística de Tópicos

Segundo [1], a modelagem de tópicos probabilística é uma abordagem relativamente nova que está sendo aplicada com sucesso na exploração e previsão de estruturas subjacentes em dados discretos, como textos, por exemplo. Um modelo de tópico, como a indexação semântica latente probabilística (PLSI) proposta por [26 apud 1], é um modelo estatístico gerador que relaciona documentos e palavras através de variáveis latentes que representam os tópicos [27 apud 1].

Ao considerar um documento como uma mistura de temas, o modelo é capaz de gerar as palavras em um documento, dado o pequeno conjunto de variáveis latentes (ou assuntos). Este processo de inversão, isto é, encaixando o modelo gerador para os dados observados (palavras em documentos), corresponde a inferir as variáveis latentes e, portanto, aprender as distribuições de tópicos subjacentes.

A seguir, abordamos o modelo gerador de tópicos escolhido para este trabalho: LDA. Este modelo foi escolhido, dentre os vários modelos existentes, por ser altamente modular e, por isso, facilmente estendido.

3.5.1 LDA(Latent Dirichlet Allocation)

LDA é uma forma de encontrar tópicos automaticamente na coleção desejada. Ele representa documentos como misturas de tópicos-conjuntos de palavras com probabilidades específicas. Assume-se que os documentos são produzidos da seguinte maneira:

Quando se escreve cada documento:

- Decide-se o número de palavras N que o documento terá, de acordo com a distribuição de Poisson;
- Escolhe-se uma mistura de tópicos para o documento (de acordo com uma distribuição de Dirichlet sobre um conjunto pré-determinado de tópicos K). Por exemplo, supondo que temos dois tópicos, podemos dizer o documento consiste de 1/3 do tópico 1 e 2/3 do tópico 2;
- Gera-se cada palavra w_i no documento:

- Primeiro escolhendo um tópico (de acordo com a distribuição multinomial que foi gerada acima, por exemplo, podemos escolher o tema 1 com 1/3 de probabilidade e o tema 2 com probabilidade 2/3);
- Usando o tópico para gerar a própria palavra (de acordo com a distribuição multinomial do tópico). Por exemplo, se foi selecionado o tema 1, podemos gerar a palavra "x" com probabilidade de 30%, "y" com probabilidade de 15%, e assim por diante;

Assumindo este modelo gerador para uma coleção de documentos, LDA então tenta voltar atrás dos documentos para encontrar um conjunto de tópicos que são prováveis de ter gerado a coleção.

3.5.1.1 Aprendizagem

Suponha um conjunto de documentos. Escolhe-se um número fixo de tópicos K para serem descobertos, e opta-se por LDA para aprender a representação por tópicos de cada documento e as palavras associadas a estes. Mas como se faz isso? A forma utilizada neste trabalho (conhecida como amostragem de Gibbs colapsado[14]) é a seguinte:

- Atribua aleatoriamente a cada palavra de cada documento um dos K tópicos;
- Observe que esta atribuição aleatória já lhe fornece ambas as representações: dos tópicos de todos os documentos e das distribuições de palavras de todos os tópicos (embora não muito boas);
- Então, para aperfeiçoá-los, para cada documento d :
  - Passe por cada palavra w em d:
    - E para cada tópico t, calcule duas coisas: 1) p (tópico t | documento d) = a proporção de palavras no documento d que estão atribuídas atualmente ao tópico t, e 2) p(palavra w | tópico t) = proporção de atribuições para o tópico t em relação a todos os documentos em que aparece esta palavra w. Reatribuir w a um novo tópico, onde escolhemos um tópico t com probabilidade p (tópico t | documento d) * p (palavra w | tópico t) (de acordo com nosso modelo gerador, isto é, essencialmente, a probabilidade de que o tópico t

[14] Collapsed Gibbs Sampler é um método estatístico bastante utilizado nesta área

gerou a palavra w, por isso faz sentido calcular um novo modelo do atual tópico da palavra com essa probabilidade).

- Depois de repetir o passo anterior um número grande de vezes, finalmente chega-se a um estado mais ou menos estável, onde suas atribuições são muito boas. Assim, utilizar estas atribuições para estimar as misturas de tópicos de cada documento (contando a proporção de palavras atribuídas a cada tópico dentro desse documento) e as palavras associadas a cada tópico (contando a proporção de palavras atribuídas a cada tópico global).

A Figura 8 apresenta uma representação gráfica do LDA.

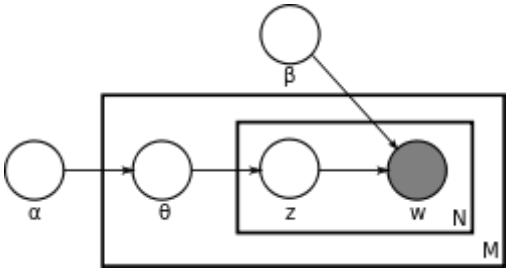


Figura 8: Representação gráfica do LDA

3.6 Análise Formal de Conceitos (FCA)

Segundo [20], FCA é um método bastante utilizado para a análise de dados, deriva relações implícitas entre objetos descritos por meio de um conjunto de atributos, por um lado e esses atributos, por outro. Os dados são estruturados em unidades que são abstrações formais de conceitos do pensamento humano, permitindo a interpretação significativamente compreensível (Ganter & Wille, 1999). Assim, FCA pode ser visto como uma técnica de agrupamento conceitual assim como também proporciona descrições intensionais para os conceitos abstratos ou unidades de dados que produz. A noção central para a FCA é a de um contexto formal.

[19]Considera a Análise Formal de Conceitos um ramo da teoria matemática *reticulada* que fornece meios para identificar grupos significativos de objetos que compartilham atributos comuns, assim como fornece um modelo teórico para analisar as hierarquias desses agrupamentos. [19 apud 28]

O principal objetivo do FCA é definir um conceito como uma unidade de duas partes: *extensão* e *intensão*. Extensão de um conceito abrange todos os objetos que pertencem ao conceito, enquanto a intenção compreende todos atributos compartilhados por todos os objetos sob consideração.

A fim de aplicar FCA, é necessário o contexto formal ou a tabela de incidência dos objetos e de seus respectivos atributos. O contexto formal consiste de um conjunto de objetos O, um conjunto de atributos A, e uma relação binária $R \subseteq S \times A$ entre objetos e atributos, indicando que atributos cada objeto possui. Formalmente, pode ser definida como $C = (A,S,R)$. A partir do contexto formal, FCA gera um conjunto de conceitos, onde cada conceito é uma coleção máxima de objetos que possuem atributos comuns. Mais formalmente, um conceito é um par de conjuntos (X, Y) de tal forma que:

$X= \{o \in O \mid \forall a \in Y: (o,a) \in R\}$

$Y= \{a \in A \mid \forall o \in X: (o,a) \in R\}$, onde

X é considerado como sendo a *extensão* do conceito e Y é *intenção* do conceito. Este conjunto de conceitos é chamado uma ordem parcial completa, onde alguns conceitos são super ou subconceitos em relação aos outros. O conjunto de todos os conceitos constitui um conceito *reticulado*. [19]

Capítulo 4 Trabalhos Relacionados

Nesta seção descrevemos trabalhos estreitamente relacionados a esta pesquisa, com o foco mais acadêmico do que os citados anteriormente. Entre eles podemos citar os que utilizam LDA e trabalham com a dimensão tempo e tópicos correlacionados, fatores muito relevantes para este trabalho; o Blogscope[15], uma ferramenta acadêmica de análise de keywords vindas de milhares de blogs coletados, que serviu como ideia para algumas análises sobre os dados coletados e como um apoio no algoritmo de detecção de Bursts; e o Grapevine16, pertencente ao mesmo grupo do Blogscope, que permite descobrir histórias interessantes, dentre os assuntos mais discutidos na blogosfera ou durante intervalos de tempo.Por fim, falamos de trabalhos que também utilizaram Análise Formal de Conceito(FCA).

4.1 Latent Dirichlet Allocation (LDA)

O modelo LDA, foi introduzido pela primeira vez por [29 apud 2], e é considerado um modelo probabilístico gerador que pode ser utilizado para estimar observações multinomiais por aprendizado não supervisionado. A intuição por trás do LDA é encontrar a estrutura latente de "tópicos" ou "conceitos" em um corpus de texto. [30 apud 2] mostrou empiricamente que a coocorrência (tanto direta quanto indireta) de termos em documentos de texto pode ser usada para recuperar esta estrutura latente de tópicos. [2]

4.1.1 Dimensão Tempo

O artigo [15] fala de como tópicos evoluem ao longo do tempo em grandes coleções de documentos, o que auxiliou no desenvolvimento da ideia de agregar à ferramenta desenvolvida neste trabalho, um modelo gráfico que mostre a dinâmica dos tópicos- como um assunto se torna outro/outros com o passar do tempo, quando ele tem picos de popularidade ,etc. A Figura 1, retirada do artigo citado acima, mostra um exemplo de evolução de tópico ao longo de várias décadas.

[15] O site http://www.blogscope.net/ foi descontinuado ao longo do desenvolvimento deste trabalho

[16] O site http://www.onthegrapevine.ca/ também foi descontinuado

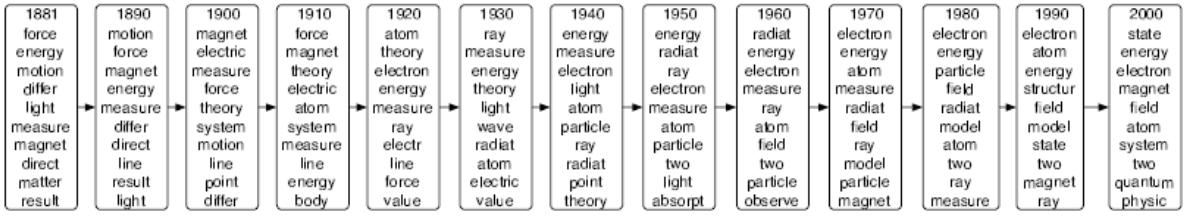


Figura 9: As dez principais palavras da distribuição posterior inferida ao longo de dez anos [15]

Um ponto negativo em relação a esse trabalho foi que não foi citado explicitamente como sabiam que todos estes grupos se tratavam necessariamente do mesmo tópico, simplesmente afirmava-se isso.

4.1.2 Tópicos Correlacionados

Modelos de Tópicos tais como o LDA, podem ser muito úteis, utilizando-os como ferramentas de análise estatística de coleções de documentos e outros dados discretos. O modelo LDA assume que as palavras em cada documento surgem a partir de uma mistura de tópicos, sendo cada um dos quais uma distribuição ao longo do vocabulário. Uma limitação desse modelo é a incapacidade de modelar correlações entre tópicos, conforme citado no Capítulo 1. Esta limitação deriva do uso da distribuição Dirichlet para modelar a variabilidade entre as proporções de tópicos. Em [4], desenvolve-se um modelo de tópicos correlacionados (CTM) onde as proporções dos tópicos apresentam correlações através da distribuição Normal.

O artigo [4] serviu como uma ideia base para o cálculo de similaridade entre tópicos relacionados. A partir dele surgiu a ideia de se criar um modelo gráfico mostrando esta relação entre tópicos, através da covariância entre eles. O algoritmo em si deste artigo não foi utilizado. Usa-se em [4] Distância de Hellinger que em nosso caso, após testes, não se mostrou muito interessante, por isso optamos pela Similaridade do Cosseno, que apresentou textos de fato similares.

4.1.3 Modelos de Tópicos

Segundo [18], cientistas precisam de novas ferramentas para explorar e navegar por grandes coleções de literatura acadêmica. Graças a organizações como a JSTOR, que digitalizam e indexam arquivos físicos originais de muitas revistas, os cientistas modernos podem fazer buscas por bibliotecas digitais que abrangem centenas de anos. Um cientista, ao se confrontar com o acesso a milhões de artigos de sua área, pode não ficar satisfeito com pesquisas simples. Usar tais coleções efetivamente requer uma interação com elas de uma forma mais estruturada: encontrar artigos semelhantes aos de interesse, e explorar a coleção através dos temas subjacentes presentes nela.

Para desenvolver as ferramentas necessárias para explorar e navegar pelas modernas bibliotecas digitais, precisa-se de métodos automatizados de organização, gerenciamento e entrega de seus conteúdos.

Em [18] descreve-se modelos de tópicos para descobrir a estrutura semântica subjacente de uma coleção de documentos com base em uma análise Bayesiana hierárquica. Modelos de tópicos foram aplicados a vários tipos de documentos, incluindo e-mail, papers, e Journals. Ao descobrir padrões de uso de palavras e documentos conectados que apresentam padrões semelhantes, modelos de tópicos surgiram como uma nova e poderosa técnica para encontrar estruturas interessantes em uma coleção não estruturada.

A ferramenta[19] apresentada por [18] permite organizar automaticamente arquivos eletrônicos para facilitar a navegação e análise eficiente. Tendo como exemplo o arquivo do JSTOR sobre a revista Science. Este exemplo pode ser visto na Figura 20.

A ferramenta citada acima serviu de ideia para possíveis análises a serem feitas sobre os assuntos encontrados através do uso do LDA.

Figura 10: Modelo navegável estimado a partir da revista "Science"

4.2 Blogscope

BlogScope, é um sistema acadêmico que foi desenvolvido pela Universidade de Toronto no Canadá com o objetivo de gerar análises sobre a Blogosfera, mas que atualmente foi descontinuado e se tornou um novo produto comercial. Seu foco principal era extrair dados que auxiliassem na análise e descoberta de informações de forma interativa. A ferramenta BlogScope rastreava cerca de nove milhões de blogs, indexando mais de 65 milhões de posts em seu banco de dados. Entre suas

[19] http://www.cs.cmu.edu/~lemur/science/

características estavam a detecção de *Bursts* de palavras-chave, identificação de palavras-chave correlatas, navegação espacial pelos posts dos Blogs, apoio para a detecção de palavras-chave "quentes" ao longo do eixo temporal, etc.[5]

O Blogscope, que esteve em funcionamento por quase todo o tempo de pesquisa para este trabalho, auxiliou na criação da nossa ferramenta através de ideias para a criação de nossos modelos gráficos relacionados a termos, do algoritmo para o cálculo de "bursts" e de noções do seriam os termos realmente "quentes". Toda a análise feita no Blogsscope é relacionada aos termos presentes nos documentos e nunca aos assuntos em si.A Figura 5 apresenta como era a tela inicial do Blogscope.

4.2.1 Dimensão Tempo

Segundo [6], a análise de posts de blogs ainda é um ramo da área de Recuperação da Informação pouco explorado quando se leva em conta a forte dimensão temporal presente neles. Em geral, só se leva em conta as *Tags* presentes nestes posts.

4.2.2 Detecção de Bursts

O BlogScope inteligentemente identifica e indica possíveis eventos interessantes presentes na curva de popularidade,eventos os quais referenciados como "bursts". O conceito de "burst" utilizado por essa ferramenta é relacionada ao conceito de popularidade inesperada de uma palavra-chave dentro de uma janela temporal.

"Bursts" desempenham um papel central na análise e navegação pelos blogs utilizando o BlogScope , pois eles identificam intervalos temporais para se focar e detalhar, refinando a busca. Eles podem ser classificados em dois tipos principais: antecipados e imprevistos. A popularidade para "bursts" antecipados aumenta de forma constante, atinge um máximo e depois recua da mesma maneira. Um lançamento de um filme ou a Copa do Mundo se enquadra nesta categoria. Ao contrário de "bursts" antecipados, a popularidade de "bursts" imprevistos aumenta inesperadamente. O furacão Katrina ou a morte de Abu Musab al-Zarqawi se enquadra nesta categoria. [6]

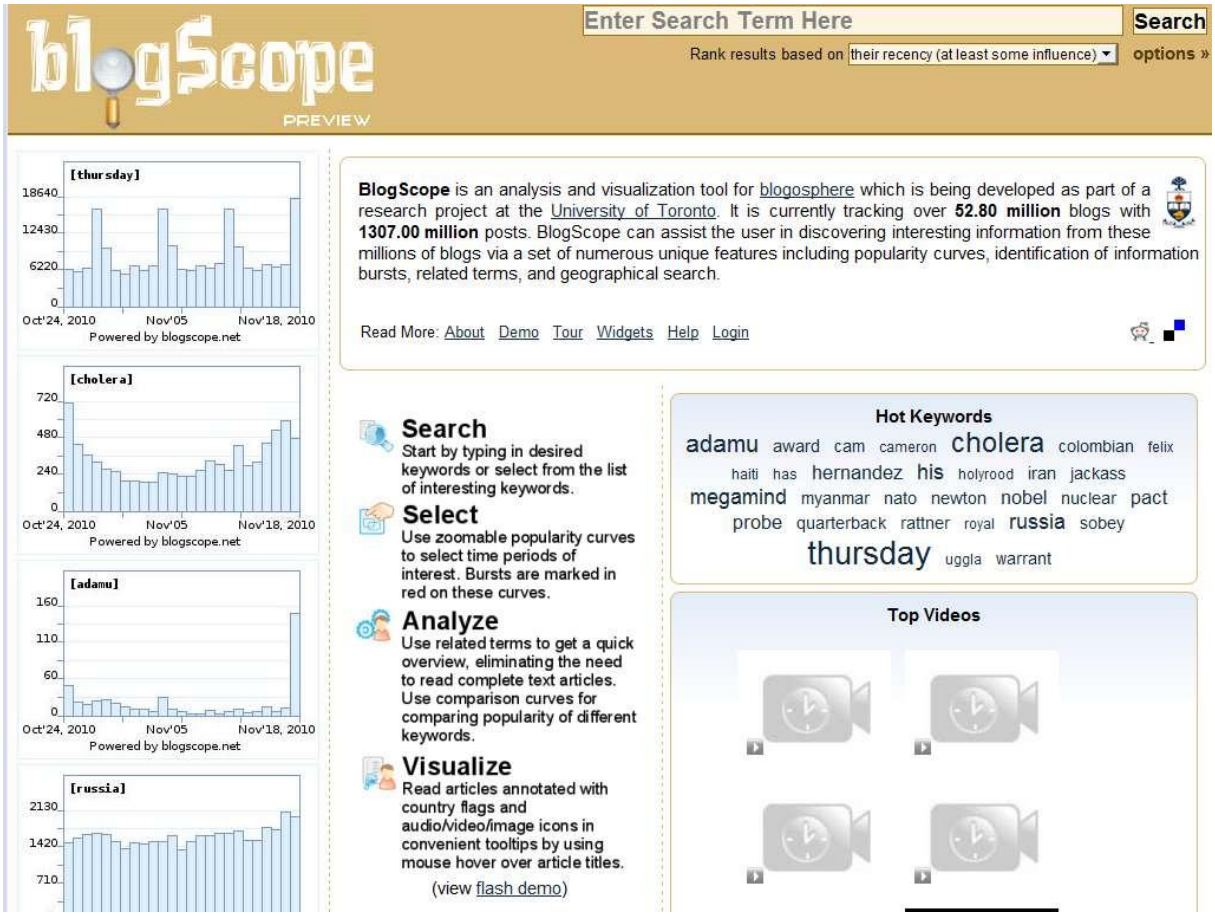


Figura 11: Tela inicial da ferramenta BlogScope [17]

Os trabalhos que referenciam o projeto BlogScope colaboraram para uma maior clareza sobre tipos de análises interessantes a serem feitas nos posts dos blogs, apesar de serem feitas a partir de termos somente e não assuntos. Em nosso caso apesar de também analisarmos termos, temos um foco principal mais abrangente incluindo também as análises sobre as histórias presentes nesse nicho de documentos.

4.3 Grapevine

Segundo [16], o objetivo o qual o site Grapevine se propõe é minerar informações e fornecer "insights", capturando tendências populares à medida que elas surgem. Por exemplo, ele permite aos usuários descobrirem assuntos interessantes que estão sendo muito falados na blogosfera ou descobrirem assuntos que são/foram de interesse para um grupo demográfico específico, e / ou durante um intervalo de tempo específico.

Algumas das funcionalidades que o Grapevine possui são:

- Descobrir assuntos e termos de interesse popular (figura 6, itens 2-4);

- Apresentar blogs relacionados, notícias, vídeos, etc. (figura 6, item 8);
- Identificar o impacto de certas histórias em diferentes grupos demográficos (figura 6, item 1);
- Entender a evolução temporal das histórias desejadas, através da ferramenta de "tendências de popularidade";
- Navegar entre histórias relacionadas, através da "nuvem de entidades" (figura 6, item 7);
- Inspecionar o impacto de um tópico particular ou um tipo de tópico (figura 6, item 5-6);

A ferramenta acadêmica Grapevine foi desenvolvida pelo mesmo grupo de pesquisadores do Blogscope, citado anteriormente. Ela serviu também como uma ideia base para as análises feitas por esta dissertação, não levando em conta seus algoritmos, pois pouco conteúdo foi achado sobre como foi desenvolvida esta ferramenta. O site Grapevine não estava mais sendo atualizado durante o desenvolvimento desta dissertação, mas para períodos mais antigos podia-se ver a maioria das funcionalidades.

Outra coisa que foi bastante interessante para o nosso projeto citada em [16], foi como nos dias de hoje em que se leva muito em conta as mídias sociais, é importante saber sobre o que as pessoas estão falando nestes meios. Além de saber o que as pessoas estão falando é muito válido saber dentre as histórias faladas, quais são mais interessantes discutidas, e quais são os textos que ajudaram a montá-las.

Uma diferença marcante entre os dois trabalhos citados é que o BlogScope é mais focado em análises de palavras-chave em contrapartida ao Grapevine que faz análises em um nível conceitual mais alto, focado em entidades e eventos do mundo real. Já em relação ao nosso trabalho, levamos em conta para gerar os assuntos presentes nos documentos os textos como um todo e não só entidades, o que enriquece as histórias.

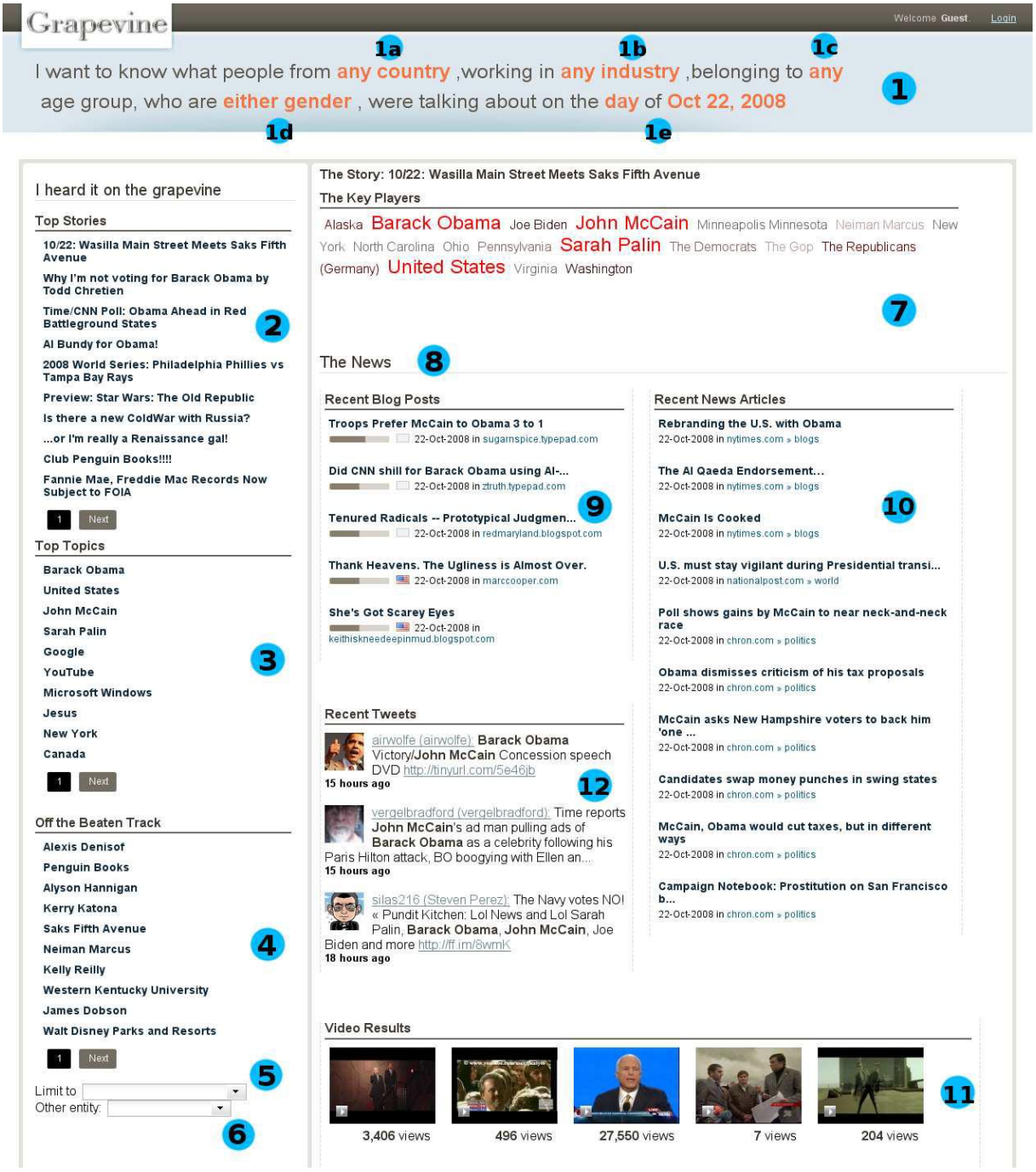


Figura 12: Tela inicial da ferramenta Grapevine [16]

4.4 Observatório da web

Este projeto brasileiro foi desenvolvido com o objetivo de monitorar, em tempo real, fatos importantes, eventos e entidades nas várias mídias e pelos vários usuários da Web. Auxilia na criação de panoramas de assuntos relevantes sob o ponto de vista das informações e das opiniões que circulavam na Web e nas redes sociais online, incluindo jornais, revistas, portais e o Twitter.

A partir da identificação de entidades nos textos coletados, a ferramenta possibilita gerar produtos de análise e visualização. Um exemplo de um destes produtos é apresentado na Figura 13.

Figura 13 Observatório da web

Antes da extração propriamente dita, segundo[48], esta ferramenta executa um pré-processamento dos textos coletados, incluindo a padronização da codificação dos caracteres, a eliminação de código HTML, cabeçalhos e anúncios de páginas coletadas através de *feeds*, e métodos tradicionais de pré-processamento de textos: remoção de *stop words*[20] e *stemming*[21].

A identificação de entidades presentes nos textos é feita através de uma ferramenta que utiliza técnicas de processamento de linguagem natural para identificar referências a entidades (pessoas, organizações, locais, etc.) em texto livre. Apos a fase de identificação, segue-se uma fase de

[20] palavras de pouco valor informacional como artigos, preposições e conjunções

[21] consiste na extração dos radicais das palavras do texto

desambiguação de entidades. Para isso, um método de classificação foi utilizado para aprender a associar entidades a determinados contextos.

4.5 FCA

Existe uma bibliografia online sobre o FCA[22] que contém links para bibliografias mantidas pelos grupos de pesquisa da área e conferências.

Em [33] é detalhado o que é a Análise de Conceitos Formais e apresentados alguns exemplos de utilização, que foram bastante interessantes para se ter uma noção melhor da área, assim como áreas em que ele pode ser utilizado com os trabalhos relacionados existentes. Duas áreas bastante interessantes para este trabalho foram a de Recuperação da Informação e a de usar o FCA como uma ferramenta de representação e descoberta de conhecimento. As figuras 18 e 19 apresentam exemplos citados em [33].

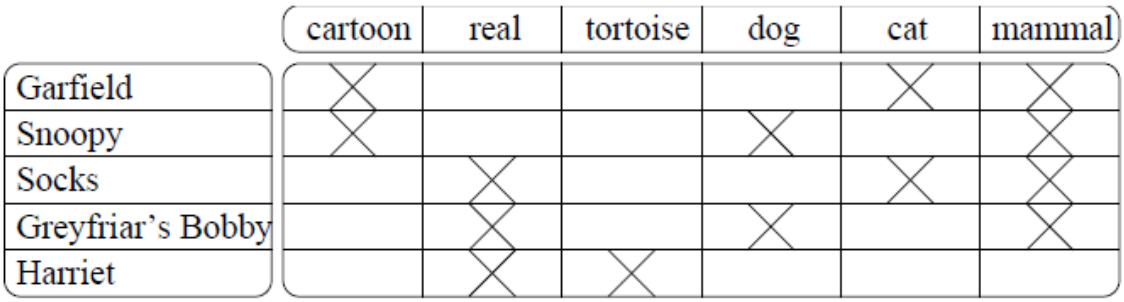


Figura 14: Um contexto formal de "animais famosos" [33]

[34] apresenta uma técnica semiautomática que reconstrói o mapeamento das *features* que são acionadas pelo usuário e exibe um comportamento observável, como um mapa mental facilitando os usuários a compreender o que o sistema em que ele está trabalhando possui.

[35] apresenta um *survey* sobre trabalhos que utilizam Análise de Conceitos Formais em engenharia de software. A figura 20 apresenta um exemplo de utilização na área de Engenharia de Software.

4.5 Conclusão

[22] http://www.fcahome.org.uk/fca.html

A seguir apresentamos uma tabela comparativa entre os trabalhos e áreas relacionados e a ferramenta proposta. Tendo estes trabalhos, um cunho mais acadêmico do que os apresentados no Capítulo 2.

	Metodologias		Ferramentas		
	LDA	FCA	BlogScope	Grapevine	BlogMiner
Dimensão temporal	x		x	x	x
Assuntos relacionados	x				x
Detecção de *bursts*			x		x
Foco em blogs			x		x
Agregam notícias				x	x
Explora o lado visual		x	x	x	x
Forma inovadora de ver as notícias					x
Filtra geograficamente			x	x	
Confiabilidade das informações	x	x	x	x	x
Descoberta de assuntos populares	x				x
Descoberta de termos populares		x	x	x	x
Termos correlacionados ao assunto(que ajudam a compreende-lo)			x	x	x
Ferramenta analítica		x	x	x	x

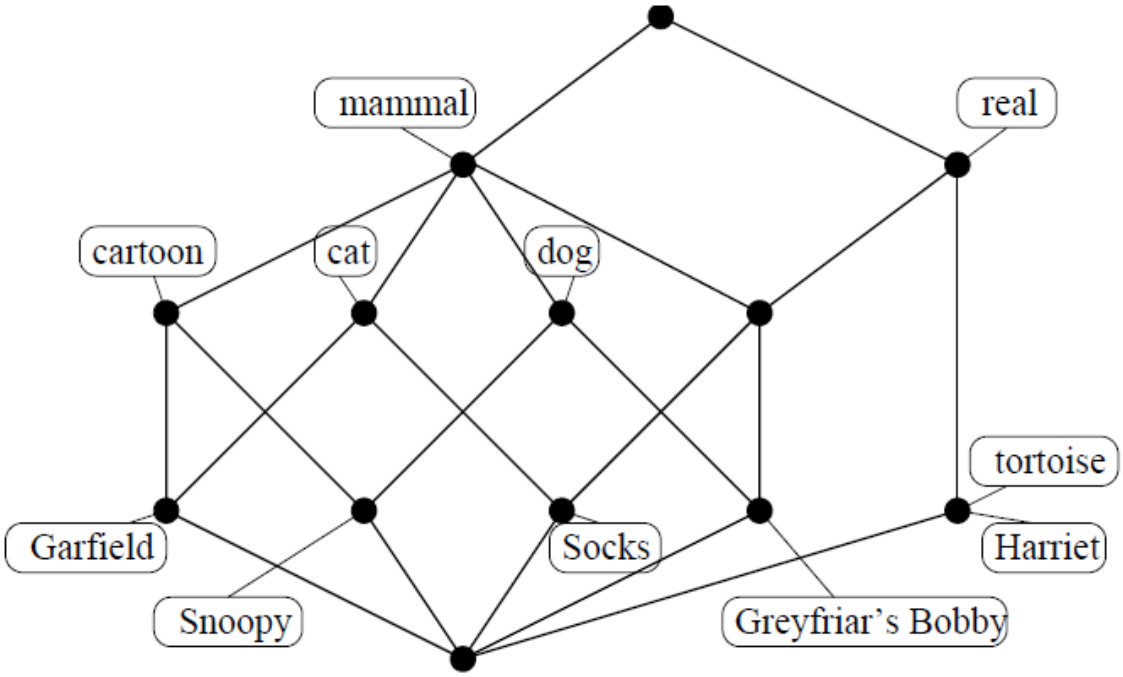


Figura 15: Um conceito *lattice* para o contexto formal da Figura 18 [33]

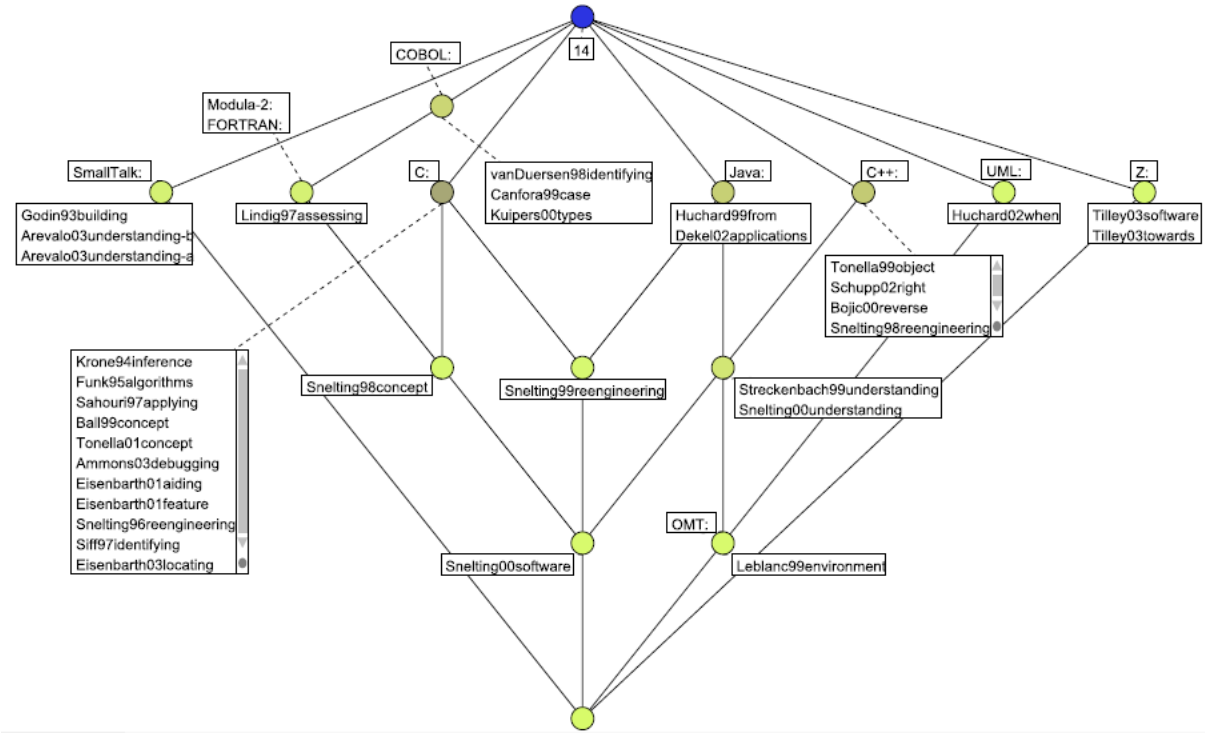


Figura 16: FCA utilizado na área de Engenharia de Software[35]

Capítulo 5 Tópicos

A seguir serão apresentadas algumas definições sobre tópicos, bem como destacar as propriedades relacionadas a eles. Também vamos abordar algumas aplicações suportadas por eles e destacar seus pontos positivos e negativos. E finalmente iremos apresentar o modelo proposto que possibilitará posteriores análises temporais dos tópicos. Modelo que é um dos objetivos a ser alcançado por este trabalho e que permitirá continuar a conduzir a construção da ferramenta proposta.

5.1 Definição

Nesta seção são fornecidas algumas definições relacionadas aos tópicos e alguns conceitos relacionados a eles, com o objetivo de promover um entendimento uniforme que irá permitir que posteriormente se possa construir um framework formal para a construção deles. Também serão discutidas algumas questões relacionadas à diferença entre os tópicos gerados por seres humanos e computadores.

Neste trabalho é adotada a definição proposta por [44]: "Formalmente, um tópico é uma distribuição de probabilidades sobre os termos de um vocabulário. Informalmente, um tópico representa um tema semanticamente subjacente". Sendo o significado da palavra "tema", neste trabalho, considerado similar ao da palavra "assunto", usada com maior frequência.

Utilizamos os tópicos para representar os assuntos contidos na Blogosfera

Usamos a linguagem de "coleções de textos" ao longo do trabalho, referindo-se a entidades como "palavras", "documentos" e "corpora". Isto é útil na medida em que ajuda a guiar a intuição, especialmente quando introduzimos variáveis latentes que visam capturar noções abstratas, como tópicos.[23]

Podemos definir formalmente as seguintes entidades, que compõem um tópico:

- Um *termo* é a unidade básica de dados discretos, definido para ser um item de um vocabulário indexado por {1,...,V}.
- Um *documento* é uma sequência de N termos denotada por $W = \{W_1, W_2, ..., W_n)$, onde W_n é o Nésimo termo na sequência.
- Um *corpus* é uma coleção de M documentos denotada por $D = \{D_1, D_2, ..., D_m\}$.

Tópicos são compostos por um conjunto de termos que juntos indicam a qual assunto este tópico está relacionado. Por exemplo, um tópico pode se referir a um evento que ocorreu ou irá ocorrer, como um lançamento de um novo produto.

Os **termos** mais relevantes são relacionados a um ou mais assuntos presentes na Blogosfera. A descrição do que são termos relevantes será abordado mais a frente.

Cada **termo** é uma representação concreta, na forma de palavra, de um conceito relacionado ao assunto o qual ele está ligado. Conceitos podem incluir ideias de propósito, descrição, composição entre outros.

Um determinado conceito pode ser representado por um ou mais **termos**: por causa de palavras compostas ou por mais de uma palavra ter o mesmo "significado", blogueiros podem utilizar palavras diferentes ao longo de suas postagens para representar um mesmo conceito.

Um campo semântico é o conjunto de conceitos conectados a um assunto foco, de tal forma que é independente dos blogueiros que criaram as postagens em que estão contidos os termos "originais" e, compreensível para outras pessoas que tenham o entendimento destes conceitos. Como resultado, um tópico é uma representação gráfica e concreta de um campo semântico de um assunto.

A Figura 17 apresenta uma representação visual de um campo semântico.

Figura 17: Campo Semântico

O processo de criação de um tópico pode ser resumido nas seguintes etapas iniciais:

1. Compreensão do assunto foco e dos conceitos que podem ser relacionados a ele (Figura 18).

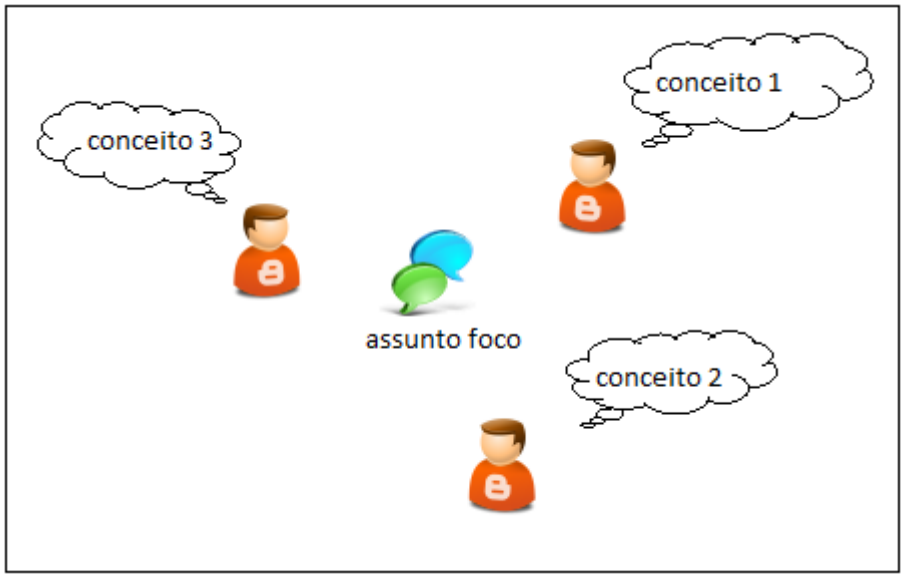


Figura 18: Compreensão do assunto foco

2. Captura do campo semântico ao redor deste assunto foco (Figura 19).

Figura 19 Capturando campo semântico

3. Transformação do campo semântico em um tópico (Figura 20).

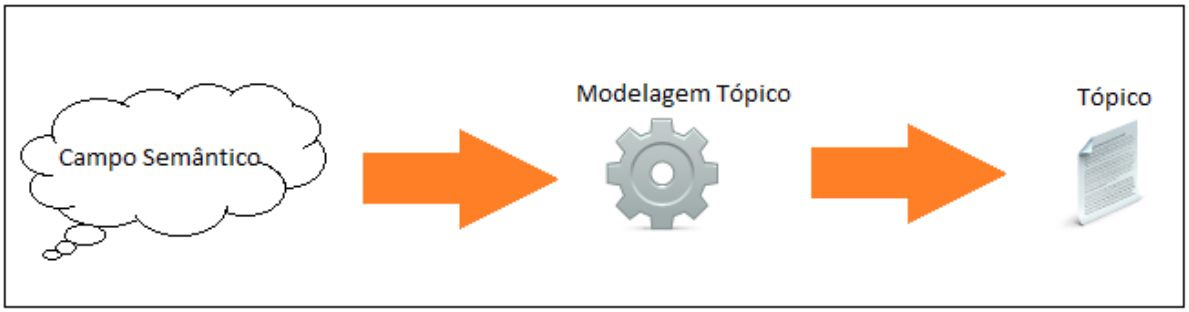


Figura 20: Processo de construção do Tópico

4. Interpretação do tópico pelo usuário final (Figura 21).

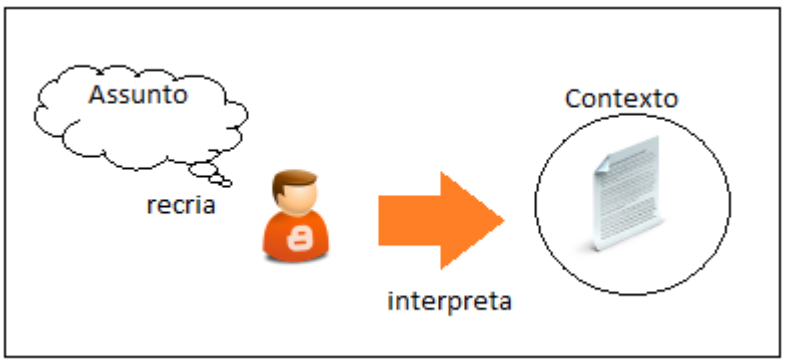


Figura 21: Interpretação do Tópico

A quarta etapa deste processo de criação é a "recriação" do assunto através da interpretação do usuário final que tenta entendê-lo como um possível tópico dentro de um contexto. Ou seja, o usuário final interpreta o tópico e tenta recriar em sua mente o assunto que deu origem a este.

5.2 Proposta de modelo formal para *Tópicos*

Tópicos fornecem um resumo ou uma visão semântica dos conceitos mais importantes que representam um assunto qualquer.

Seres humanos constroem este campo semântico associando os conceitos que são percebidos através da compreensão do assunto e dos termos que representam estes conceitos, isto por meio de sua interpretação particular. A partir dos assuntos contidos nas postagens, estes termos podem ser encontrados em seus conteúdos ou inferidos através do entendimento do conteúdo destas postagens.

Em geral, as pessoas podem facilmente identificar os termos que representam os conceitos, desde que elas tenham um razoável conhecimento da linguagem utilizada e do mundo. Elas também podem descrever conceitos através de sentenças. No entanto, sistemas automáticos não possuem este conhecimento, portanto, eles devem inferir o conhecimento exclusivamente através das palavras (termos) que compõem o corpus ou alguma outra informação textual.

Também devemos destacar o fato da relação entre termos e conceitos não ser bijetora. Uma palavra pode representar mais de um conceito, fato que é conhecido como polissemia.

O termo "apple", por exemplo, pode representar uma fruta , uma empresa ou uma cidade, Nova York que também é conhecida como Big Apple. Como outro exemplo, podemos citar o termo "icecream", que pode se referir tanto ao alimento quanto a uma determinada versão do sistema operacional Android. Por outro lado, um conceito pode ser representado por vários termos, que podem ser palavras tanto simples como compostas, tal como "president", "chief of state", "head of state" e "premier", o que é conhecido como sinonímia.

Além disto, pode ocorrer de não existir termo que solitariamente represente um conceito, como nos casos de, “Google Plus” e “Cloud Computing”. E por fim, também existem palavras, que não chegam a ser consideradas termos, conhecidas como *stop words*, que não carregam qualquer significado, tais como, preposições, artigos e conjunções, mas que ocorrem em grande frequência nas postagens. No trabalho proposto, também podem ser consideradas *stop words* palavras referentes às datas das postagens, como “monday” e “november”. Também é possível a existência de conceitos que não podem ser representado por palavras, mas isto é um caso muito raro. No entanto, com menor raridade são encontrados conceitos que necessitam de textos complexos para que possamos compreendê-los.

Considerando estas limitações, uma solução automática para geração de tópicos, semelhante a que é proposta neste trabalho, somente pode ser efetiva quando é desenvolvida como uma aproximação do comportamento humano. Para possibilitar esta aproximação, deve-se considerar que é necessário criar um modelo que permita a descrição dos campos semânticos do corpus ambos quando analisados sob o ponto de vista humano e quando analisado sob o “ponto de vista” de um computador.

Nesta seção, é iniciada a criação desde o início de uma definição conceitual e formal de tópicos que irá permitir desenvolver um método abstrato para construí-los, levando em conta a forte influência temporal.

5.3 Considerações iniciais

Neste trabalho é apresentado um modelo para geração de tópicos a partir de fluxos de postagens em blogs, tentando atingir ao máximo a aproximação do modelo de geração de tópicos realizados pelos humanos. Para se criar um tópico, primeiramente, as pessoas devem criar conceitos em suas mentes. Estes conceitos são pensamentos abstratos que nem sempre podem ser descritos em palavras.

Portanto, para adotar este processo deve-se primeiramente, decidir como representar conceitos em um computador.

5.3.1 Definições iniciais

Esta seção começa apresentando alguns conceitos iniciais que serão úteis durante a definição do modelo.

Para começar é apresentada a definição de assuntos e contextos. A motivação para estas definições está no fato de que, uma das propostas deste trabalho se baseia na construção de um modelo para geração de tópicos que descrevam assuntos presentes em um determinado contexto.

As definições apresentadas a seguir seguem a ideia do modelo proposto por [45] onde é definido *Tag Clouds* e não tópicos, mas que possui muitas definições com pensamentos semelhantes as que iremos apresentar.

5.3.1.1 Assunto e contexto

Neste trabalho, um assunto é qualquer objeto identificável que pode ser descrito, ao menos parcialmente, por um conjunto de termos. Sendo estes conjuntos de termos, os **tópicos** aqui discutidos.

Estes termos agem como representações dos conceitos que residem na mente das pessoas e que foram de certa forma materializados, através das postagens nos blogs. Podendo ser aplicados aos assuntos segundo alguma razão. Um assunto é representado por uma letra *r*, possivelmente indexada. No nosso caso podemos dizer que os assuntos *r* são os assuntos discutidos na Blogosfera.

Segundo [45] "Um **contexto**, denotado pela palavra *w*, é um conjunto de assuntos que podem ser analisados como um todo. O contexto que contém todos os assuntos será denotado por *W*, por causa de "web". Portanto, *w* não é um elemento de *W*, mas um subconjunto dele." Contextos podem ser abstratos, como quando são definidos por uma única palavra como "tecnologia", "política americana", ou muito mais objetivo, tal como, quando é definido como "as postagens na Blogosfera sobre a eleição presidencial americana de 2008". No caso deste trabalho, podemos dizer que o contexto *w* são conjuntos de postagens de diversos blogs em áreas específicas da Blogosfera em períodos de tempo previamente selecionados.

Figura 22: Modelo UML para assuntos e contextos

5.3.1.2 Conjunto par atributo

Seguindo as definições de par atributo apresentadas em [45] podemos dizer que quando definida, uma função de atribuição de domínio representa os tipos de valores que podem ser atribuídos a um atributo.

Com isto, consideramos que nossos conjuntos *A*, *V* e a função f_{ad} estão definidos. A seguir mostramos um exemplo com possíveis valores para os conjuntos e para a função relacionados aos termos, baseado em um cenário plausível a este trabalho:

$$A=\{relevância,nome\}$$

$$V=\{Valores,Strings\}$$

$$Valores=\{0,1...1\}$$

$$Strings=\{\text{“}google\text{”},\text{”}apple\text{”}\}$$

$$f_{ad}=\{(relevância,Valores),(nome,Strings)\}$$

Um **par atributo** é um par ordenado (a_i,v_{ij}):

$$(a_i,v_{ij}) \in A \times V_i \text{ onde } f_{ad}(a_i)=V_i .$$

A definição de par atributo foi criada para permitir a seleção dinâmica de atributos que podem ser aplicados a um objeto. Desta forma, posteriormente, não será obrigatório definir previamente quais atributos podem ser usados para descrever um objeto, isto é, sua classe, como na teoria orientada a objetos. Os principais objetos contidos no modelo proposto neste trabalho serão apresentados a seguir.

5.3.2 Conceitos e Campos Semânticos

Nesta seção começamos a construir o conceito de tópicos abstratos variantes no tempo que poderia ser aplicado a grande maioria dos fluxos de postagens, dentro do domínio da Blogosfera. Para começar, é formalizado o conceito de campo semântico. Para isto, deve-se supor a existência não só de um conjunto de assuntos, mas também de um conjunto de conceitos, denotado por *C*. Conceitos podem ser extremamente abstratos como, por exemplo, no caso em que estamos falando sobre quando pessoas formam conceitos, ou bem mais concreto, como no caso da representação de conceitos em estrutura de dados.

Dado um assunto *r*, de um contexto *w*, e um conjunto de conceitos *C*, um **campo semântico** para *r* é um conjunto *CS(r)* de conceitos:

$$CS(r) = \{c_i \mid c_i \in C \wedge aplica(c_i,r)\}$$

onde, *aplica(c,r)* é um predicado lógico que representa o fato que um determinado conceito pode ser utilizado para descrever, de alguma forma, um assunto ou ser uma propriedade deste assunto.

Portanto, um campo semântico é um conjunto de conceitos que, de algum modo, podem ser relacionados a um assunto com o objetivo de construir algum entendimento sobre ele. Em alguns casos estaremos interessados em descrever o campo semântico de um assunto sob um contexto específico, e para representar isto será utilizado $CS_w(r)$.

Vale destacar que apesar de ser possível gerar um campo semântico a partir de um assunto, é muito mais compreensível considerar que para gerar este campo semântico é necessário analisar não apenas o assunto, mas também o contexto no qual ele está inserido.

5.3.3 Geração de Tópicos Abstratos a partir dos termos

Um conjunto de termos, os **tópicos,** assume o papel de uma representação concreta de um campo semântico. Também, não existe diferença de um campo de termos criado por humanos ou computadores. Ambos são conjuntos de símbolos concretos. Um campo semântico pode induzir diferentes conjuntos de termos de acordo com os símbolos (palavras) disponíveis, a função *representa* escolhida e o período de tempo abrangido.

Um **gerador de conjunto de tópicos** é uma função f_{gct} que dado um contexto *w*, um fluxo de documentos específico b, $b \subset w$, gera um conjunto de tópicos *CT(b)* o qual representa o grupo de tópicos que podem ser considerados, sob alguma razão, relacionados a *b* no contexto *w*. Sendo Tp o total de tópicos e $CR_w(b)$ o conjunto de assuntos presentes no fluxo de documentos *b* no contexto *w*.

$$f_{gct} : W \times \wp(W) \rightarrow Tp$$

$$f_{gct}(b,w) = \{tp_i \mid \exists\ r_i \in CR_w(b) \wedge representa_w(tp_j,r_j)\} = CT_w(b)$$

Um **tópico** *TCA(r)* é um conjunto de tuplas

$$TCA(r) = \{(t_j,m_i)\},$$

onde t_i é um termo qualquer presente no vocabulário formado , e m_i é representa os atributos deste termo.

A partir das definições expostas acima, temos agora o vocabulário para discutir como, dado um conjunto de documentos, podemos gerar dinamicamente um conjunto de tópicos. Os conceitos descritos como conjuntos e funções podem ser vistos na figura 4 como um modelo UML.

5.3.4 Modelo abstrato de análise de tópicos

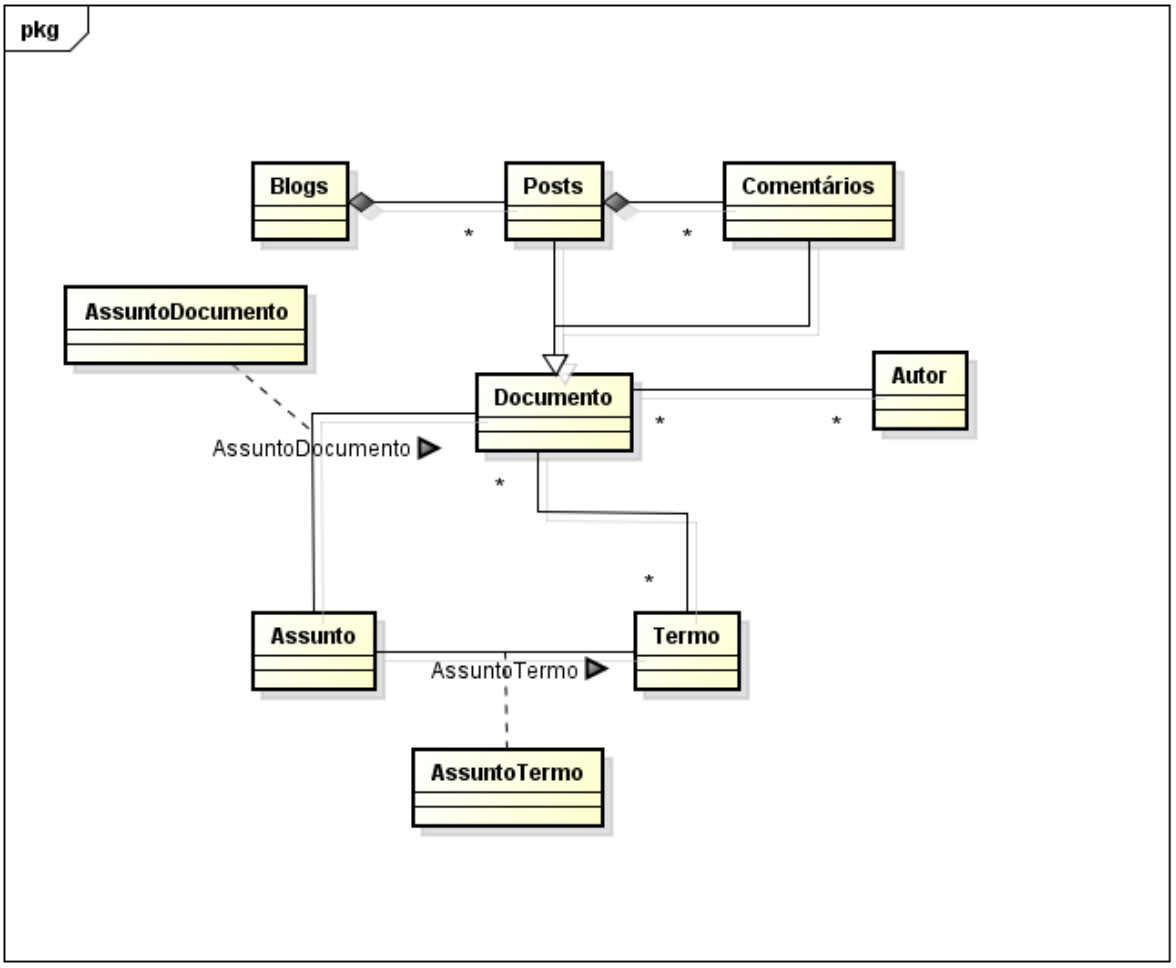


Figura 23: Modelo proposto

Um documento, conforme dito anteriormente, pode ser definido como todo texto relacionado ao conteúdo do blog e que contenha um selo temporal agregado. As postagens dos blogueiros e os possíveis comentários relacionados a elas podem ser consideradas documentos dentro do contexto deste trabalho.

Todo documento deve estar ligado a algum outro objeto que o identifica. Todas as postagens tem que estar ligadas a algum blog raiz, podendo ser replicadas por outros blogs, assim como todo comentário deve estar ligado a alguma postagem específica.

Cada documento possui um ou mais autores, sendo que em nosso trabalho não levamos obrigatoriamente em questão a autoria, o que pode ser um possível trabalho futuro. Levamos em conta apenas os nomes dos blogs.

O assunto pode ser descrito como a parte abstrata- onde a definição total está somente na cabeça das pessoas que leem e escrevem sobre ele- do que é abordado em uma parte ou na totalidade de todas as postagens.

Dentro do nosso contexto, um assunto é representado concretamente através de um conjunto de termos relevantes oriundo de parte dos documentos coletados. Este conjunto de termos é chamado de **tópico**, e cada um dos termos que o compõe possui um atributo temporal que indica o momento no tempo em que ele apareceu**.** Termos que aparecem em vários momentos diferentes dentro de um mesmo tópico são considerados mais frequentes e auxiliam mais na compreensão do tópico.

Os termos, além de aparecer mais de uma vez dentro de um mesmo tópico, podem fazer parte de vários tópicos ao mesmo tempo. Assim como um documento pode abordar mais de um tópico. A Figura 11 apresenta um exemplo de documento que fala sobre tópicos diferentes, com diferentes proporções, e os termos mais representativos de alguns destes tópicos, com suas respectivas probabilidades.

A partir do modelo apresentado acima é possível criar um *framework* que possibilite o desenvolvimento de uma ferramenta de análises de blogs.

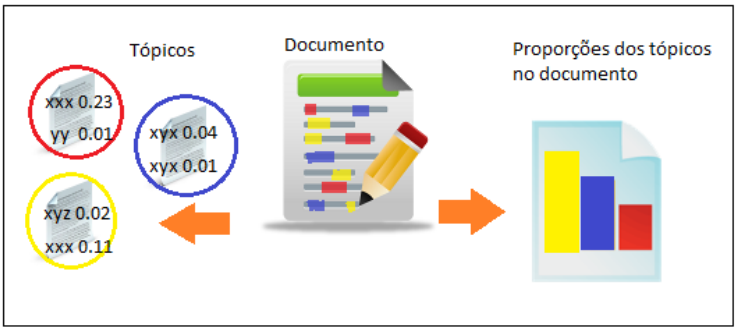


Figura 24: Proporções dos tópicos em um documento

5.3.5 Criação de um tópico

A seguir é apresentado o algoritmo desenvolvido para a modelagem dinâmica de tópicos a partir dos fluxos de textos que chegam ao longo do tempo.

Algoritmo Modelagem dinâmica dos tópicos

Entrada: Instância C_t de um dado corpus C para um período de tempo T+1, hiperparametros α,β, número de tópicos K e número de iterações

[1] *Agregar a instância C_t a C:*

a. $C=C_t + \sum_{t=1}^{T} C_t$

[2] *Gerar os tópicos, através do modelador de tópicos LDA, para $C_{t+1;}$*

a. ***for*** *k=1 to K* ***do***

i. Instanciar Y como sendo o conjunto de termos do tópico k;

ii. Instanciar W_k como sendo o conjunto de termos relevantes de k;

iii. ***for*** *y=1 to Y* ***do***

1. *Filtrar pelos termos mais relevantes de k, através de um threshold z referente a sua popularidade:*
2. *F(y,z) {true,false};*
3. ***If*** *F(y,z)=true* ***then***
    a. $W_k=W_k + y$
4. ***end if***

iv. ***end for***

v. *Inserir o tópico k e seus termos relevantes contidos em W_k ,no banco de dados*

b. ***end for***

Saída: K tópicos modelados para um período de tempo T+1

A próxima imagem apresenta o algoritmo que identifica a similaridade entre os tópicos, permitindo identificar quando se tratam de tópicos similares ou o mesmo tópico com apenas algumas mudanças oriundas do novo período de tempo.

Algoritmo Similaridade entre tópicos

Entrada: Os K tópicos e seus respectivos termos relevantes, da instância C_t do corpus, última instância inserida no banco de dados; os Y tópicos previamente inseridos no banco de dados; thresholds TS_h, TS_l;

[1] *Comparar os termos de k com todos os outros pré-existentes*

a. ***for*** *k=1 to K* ***do***

i. Instanciar W_k como sendo o vetor de termos do tópico k;

ii. ***for*** *y=1 to Y* ***do***

1. Instanciar W_y como sendo o vetor de termos do tópico y;
2. Comparar o tópico k com o tópico y, através da semelhança de seus vetores de termos, usando a função de similaridade por cosseno

a. *Sim(* W_k , W_y *) {float};*

b. ***if*** *Sim(* W_k , W_y *)>=*TS_h ***then***

i. Os tópicos k e y falam sobre o mesmo assunto

c. ***end if***

d. ***else if*** *Sim(* W_k , W_y *)<*TS_h ***and*** *Sim(* W_k , W_y *)>=*TS_l ***then***

i. Os tópicos k e y são considerados relacionados

e. ***end if***

3. Armazenar no banco de dados o valor de *Sim(* W_k , W_y *)*

iii. ***end for***

b. ***end for***

Saída: Relação de similaridade entre os novos tópicos e os inseridos anteriormente

5.3.6 Tópicos propostos

Nesta seção são apresentados os modelos propostos neste trabalho: ***Tópicos interessantes de um determinado período de tempo*** e ***Dinâmica temporal de um determinado tópico***.

5.3.6.1 Modelo de tópicos interessantes de um determinado período de tempo

Nesta seção são exibidas as definições formais para o conceito de ***Tópico Interessante de um Determinado Período de Tempo***, *TIDT(w,pt)*, que é apresentado neste trabalho. Seguindo o esquema das seções anteriores, é pretendido fornecer um modelo que possa ser usado por um processo para gerá-los.

Primeiramente, deve-se definir um *TIDT(w,pt)*, como instâncias de um conjunto dos tópicos mais discutidos na Blogsofera para um período pt de tempo. Tópicos estes, que resumem os conceitos presentes no conteúdo das postagens que formam o contexto *w*.

Para começar, é apresentada a definição de **Tópico de União do Conjunto**, *TCUC(w,pt)*, o qual é obtido a partir da união de todos os tópicos encontrados, referentes as instâncias no período de tempo pt, dos assuntos *r* que pertencem ao contexto *w*, denominados ITP(*r*,*pt*).

$$TCUC(w,pt) = \bigcup_{r \in w} ITP(r, pt)$$

O tópico de união do conjunto representa todos os assuntos existentes em um determinado contexto, em um determinado período de tempo. No entanto, não é necessariamente uma boa representação, uma vez que ela representa todos os conceitos possíveis de todos os assuntos, mas não o resumo da coleção de assuntos como um todo. E o nosso interesse é prover tópicos que possam resumir a coleção de assuntos de um contexto qualquer. Portanto, é necessário funções que permitam determinar se um elemento deste tópico de união do conjunto pertencerá ou não aos tópicos que desejamos criar.

Uma **função de relevância** é uma função $f_s(t,tp)$, que dado um termo *t* sabe decidir se ele deve ou não aparecer em algum dos tópicos interessantes de um determinado período de tempo.

$$f_s(t,tp) \rightarrow \{true, false\}$$

Sendo assim, a definição da *TIDT (w,tp)* é estabelecida como:

$$\text{TIDT}(w,tp) = \{t \mid (t \in TCUC(w)) \wedge (f_s(t,tp) = true)\}$$

Portanto, resumidamente, estes tópicos são formados por *termos* que estão relacionados a pelo menos a algum dos assuntos do contexto e, ao mesmo tempo, atendem às condições especificadas por $f_s(t,tp)$.Esta função deve ser utilizada com mais de um filtro: com filtro temporal para selecionar somente os termos naquele período de tempo(tp) e com algum filtro de *threshold* para selecionar os termos que irão pertencer aos tópicos em questão(t).

Desta forma, nos tópicos apresentados nesta seção, irão constar os termos mais importantes para a representação dos conceitos existentes em uma coleção de assuntos de acordo com algum critério.

5.3.6.2 Modelo para Tópicos variantes no tempo

Agora definimos formalmente o conceito de ***Tópicos Variantes no Tempo***, *TVT(r,w)*. Seguindo o esquema das seções anteriores, pretende-se fornecer um modelo que possa ser utilizado por um processo para gerá-los.

De acordo com a proposta do trabalho também estamos interessados em como representar as diferenças de um assunto específico ao longo do período de tempo em que ele "existe". O conceito de diferença é usado no sentido de realçar, destacar ou evidenciar as características ou conceitos particulares de um recurso que mudam durante sua vida dentro do contexto em que está envolvido.

Portanto, um *TVT(r,w)* é definido como uma representação dos conceitos presentes em um assunto específico *r*, e que de alguma forma, se destacam ou se diferenciam dos demais conceitos presentes durante a vida de *r*, dentro do mesmo contexto *w*.

O *TVT(r,w)* é a união de todas as instâncias variantes no tempo, de te=0 até te=tp do assunto *r* dentro do contexto *w*

$$TVT(r,w) = \bigcup_{te=tp}^{te=0} TVT(r,w,te)$$

Para começar, é apresentada a **função de continuidade** que é uma função $f_d(t,r,w)$, que irá fornecer uma condição para definir se o termo *t* deverá ser utilizado para representar um conceito que existe no assunto *r* durante a maior parte do tempo tp que está sendo analisado, no caso se tornando um termo "constante" .

$$f_d(t,r,w) = \{true, false\}$$

Desta forma, também podemos definir um tópico variante no tempo, *TVT(r,w)* como:

$$TVT(r,w) = \{\ t \mid (\ t \in \mathrm{TIDT}(w) \vee t \in ITP(r)) \vee fd(t,r,w) = true\ \}$$

Sendo assim, a principal função de tópico variante no tempo é representar características individuais e importantes de um determinado assunto e por consequência auxiliar os usuários nas análises dos assuntos desejados. Filtrar pelos termos que aparecem na maior parte do tempo auxilia na elucidação do que de fato se trata o assunto.

Assim como o usuário tem a possibilidade de saber se o recurso é pontual ou já vem sendo discutido por um bom tempo, assunto constante.

Capítulo 6 BlogMiner

Nesse capítulo propomos uma ferramenta de análise de blogs capaz de concentrar em um ponto único, através de um agregador de notícias, notícias de várias fontes diferentes sobre campos de interesse do usuário e que permite análises gráficas temporais dos principais assuntos e termos contidos nas postagens coletadas.

Proposta de Ferramenta

A ferramenta BlogMiner tem como objetivo, auxiliar o usuário comum na compreensão do que está em evidência na Blogosfera, apresentando o que é tendência e o porque, através de termos relacionados e gráficos de popularidade. Esses usuários podem ser tanto pessoas que desejam saber mais sobre áreas de seu interesse, como moda ou política, ou pequenos empresários que desejam ter um panorama mais amplo de sua área de atuação, mas não possuem recursos financeiros para ferramentas altamente sofisticadas.

O paradigma de análises em que o BlogMiner se propõe a cobrir é composto de quatro pilares:

- Identificar “o que” é interessante;
- “Quando” foi interessante;
- “Quem” interage com “o que” é interessante;
- “Por que” é interessante.

Com a utilização da ferramenta, pesquisadores, como os de ciências humanas e sociais, podem ganhar tempo e robustez valiosos em seus trabalhos, muitas vezes extremamente braçais.

Assim como o público feminino tem a oportunidade de compreender melhor, de uma forma mais simples, unificada e sob vários pontos de vista, o motivo e quais peças do vestuário e marcas que estão em maior evidência. É possível descobrir também se o dado objeto de pesquisa já está “na moda” há muito tempo ou não, fator levado bastante em conta por esse tipo de usuário, pois se já está em evidência há bastante tempo o desejo de consumo diminui drasticamente.

Ela é capaz de fornecer ao usuário uma visão da evolução no tempo dos assuntos, por meio de diagramas, gráficos e representações textuais. Essas visões são construídas por meio de análises baseadas no modelo definido no capítulo anterior.

O usuário tem a possibilidade de compreender as diferenças que surgem com o passar do tempo: como o conjunto de termos que forma um tópico (representação física de um assunto) se modifica entre períodos de tempo ou um determinado termo se comporta ao longo do tempo. Possui também a possibilidade de ver comparativamente como termos relacionados se comportam ao longo de um mesmo período de tempo.

O trabalho proposto é composto de módulos capazes de encontrar automaticamente tópicos dentro dos fluxos de textos coletados, tópicos e termos correlacionados, tópicos e termos constantes ou quentes, picos de popularidade e o grau de relação entre os principais termos.

A seguir descreveremos detalhadamente o funcionamento, especificação e arquitetura da ferramenta.

6.1 Visão Geral

Partindo do modelo conceitual apresentado na Figura 23, a ferramenta proposta de análise temporal de postagens em blogs, BlogMiner, permite destilar os textos através dos termos mais relevantes encontrados ou através de assuntos subjacentes, assim como entender a dinâmica dos assuntos ao longo do tempo através de conceitos formais e similaridade entre eles. A seguir a detalharemos melhor.

Ao selecionarmos um determinado termo (ou mais de um, para efeito de comparação) e um período de tempo, o sistema apresenta um gráfico dia-a-dia da frequência e seus períodos de *bursts* (períodos em que teve uma frequência mais alta do que da sua distribuição normal, conforme visto no Capítulo 3), podendo ser selecionadas subfatias de tempo para se ter uma visão mais macro dos valores assim como os termos e documentos correlacionados ao buscado naquele período. Facilitando a compreensão do ambiente em que o assunto/termo está inserido.

Podemos também saber os assuntos mais falados em períodos de tempo variável (através da técnica de modelagem de tópicos LDA, descrita no Capítulo 3), os termos de maior relevância para estes e documentos que os compõem (os posts dos blogs), assim como eles se desenvolvem ao longo do tempo.

Outra capacidade da ferramenta é a partir de um determinado período de tempo gerar os termos mais populares, com seus devidos graus de popularidade e "idade" (se é um termo sempre presente ou novo) e quando o usuário selecionar um deles redirecioná-lo para sua curva de popularidade citada anteriormente. O mesmo é feito para os assuntos, o que torna a ferramenta mais

poderosa, pois em geral as ferramentas se focam em termos somente, por ser de mais fácil desenvolvimento.

A BlogMiner é considerada temporal, pois todas as análises apresentadas levam em conta a data de postagem dos posts e como isso pode interferir nos resultados encontrados.

A Figura 25 ilustra a curva de popularidade de um termo X em um período de tempo previamente selecionado pelo usuário, sendo os pontos em verde, dias considerados *bursts*. O ponto verde evidenciado no gráfico, no mês 09, poderia representar o início de um novo produto da empresa X, por exemplo. Já a Figura 26 ilustra a comparação da popularidade de dois termos relacionados, dentro do período de tempo escolhido, por exemplo, os dois principais candidatos à eleição presidencial.

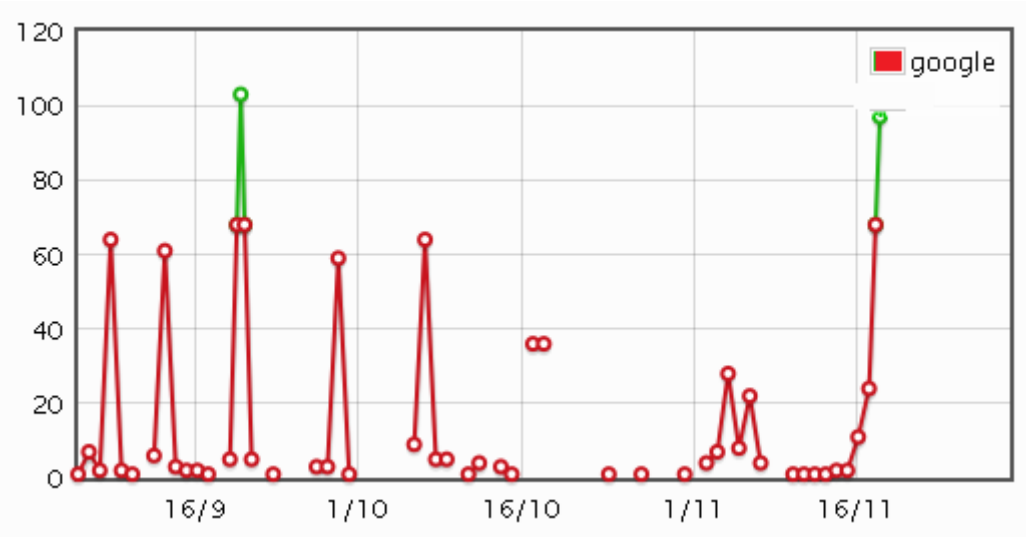


Figura 25: Busca de termos

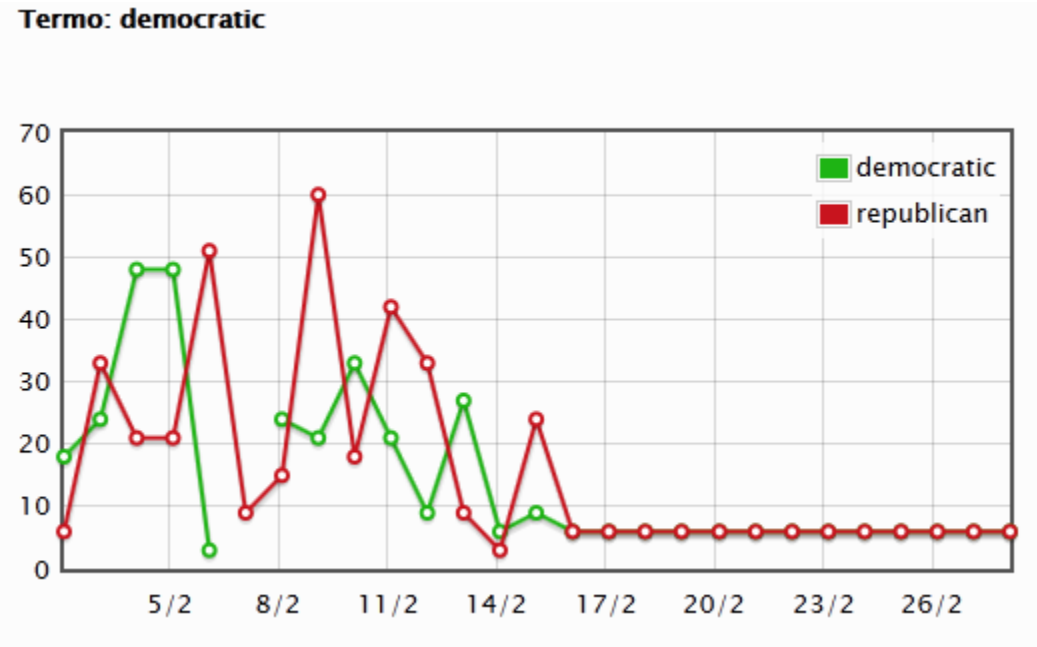


Figura 26: Comparação de popularidades

A Figura 27 apresenta os assuntos mais falados em um determinado período de tempo previamente escolhido pelo usuário. Os tamanhos dos retângulos representam a popularidade dos assuntos listados, ou seja, quanto maior, mais popular é o assunto. Já os termos dentro dos retângulos nada mais são do que os termos mais relevantes de cada assunto. Buscando aprimorar os resultados, os termos foram indexados através de seus radicais, fazendo com que as mesmas palavras conjugadas diferentemente não fossem contabilizadas mais de uma vez, poluindo o resultado.

Percebemos, através de pesquisas informais, o quanto a utilização de imagens é importante para prender a atenção do usuário no agregador de notícias, por isso ao passar o mouse sobre o assunto é adicionado uma legenda com uma imagem representativa do assunto.

A Figura 28 ilustra um fragmento da tela derivada da apresentada na Figura 27. Na tela da Figura 27 o usuário pode clicar em um dos retângulos e ler os documentos que falam sobre aquele assunto no período em questão, documentos estes listados na tela da Figura 28, com suas respectivas datas de postagem e paginados para evitar a poluição de conteúdo na tela.

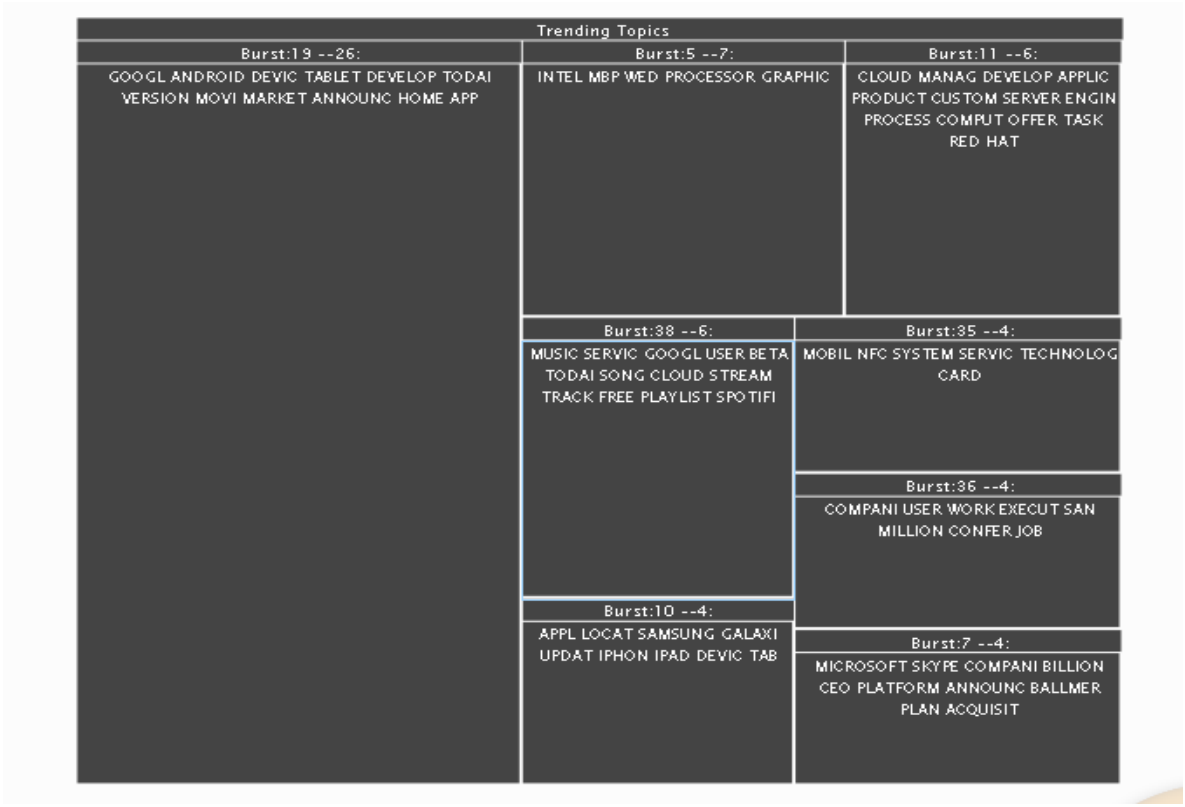
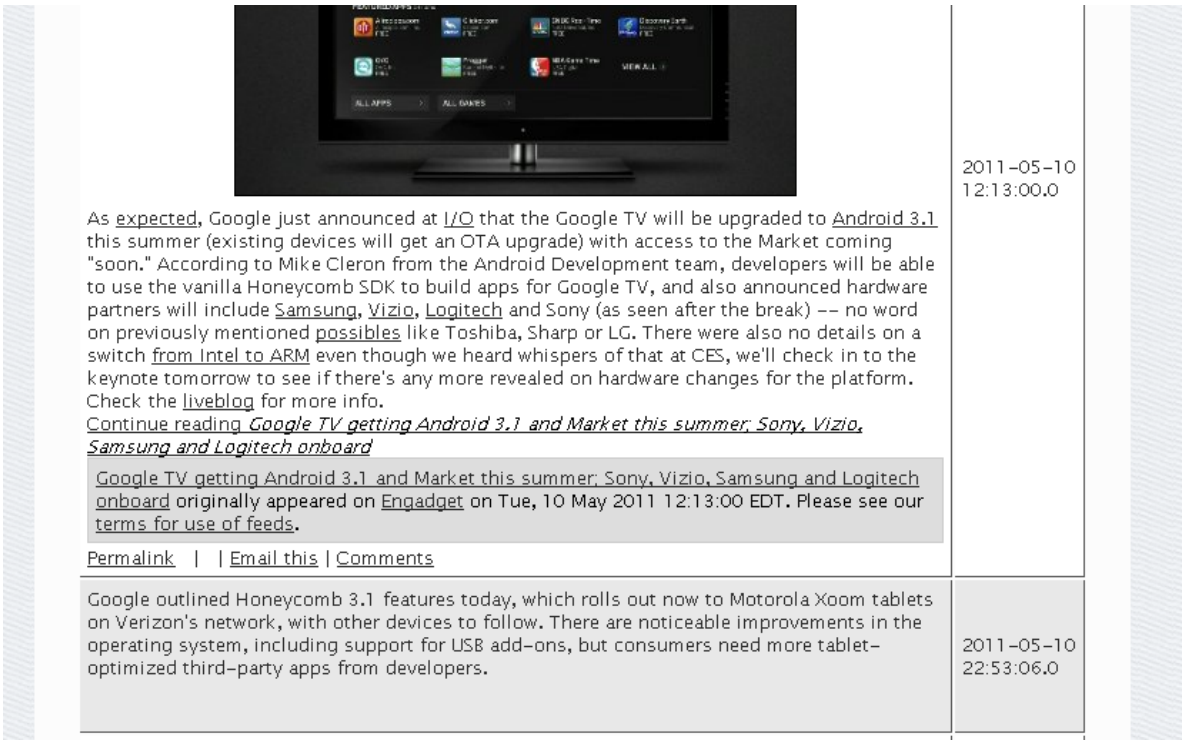


Figura 27: Detalhamento de um grupo de assuntos



2011-05-10 12:13:00.0

As expected, Google just announced at I/O that the Google TV will be upgraded to Android 3.1 this summer (existing devices will get an OTA upgrade) with access to the Market coming "soon." According to Mike Cleron from the Android Development team, developers will be able to use the vanilla Honeycomb SDK to build apps for Google TV, and also announced hardware partners will include Samsung, Vizio, Logitech and Sony (as seen after the break) -- no word on previously mentioned possibles like Toshiba, Sharp or LG. There were also no details on a switch from Intel to ARM even though we heard whispers of that at CES, we'll check in to the keynote tomorrow to see if there's any more revealed on hardware changes for the platform. Check the liveblog for more info.

Continue reading *Google TV getting Android 3.1 and Market this summer; Sony, Vizio, Samsung and Logitech onboard*

Google TV getting Android 3.1 and Market this summer; Sony, Vizio, Samsung and Logitech onboard originally appeared on Engadget on Tue, 10 May 2011 12:13:00 EDT. Please see our terms for use of feeds.

Permalink | | Email this | Comments

Google outlined Honeycomb 3.1 features today, which rolls out now to Motorola Xoom tablets on Verizon's network, with other devices to follow. There are noticeable improvements in the operating system, including support for USB add-ons, but consumers need more tablet-optimized third-party apps from developers.

2011-05-10 22:53:06.0

Figura 28: Contextualização de um assunto

A dinâmica de um assunto, como ele se desenvolve ao longo de um período de tempo é apresentada na tela da Figura 29. A Figura 29 mostra uma Análise Formal de Conceitos (FCA, conforme visto no Capítulo), sendo os objetos (nós): uma representação do período de tempo analisado, por exemplo, meses e os atributos: os termos de maior relevância para o assunto. No caso da imagem, o grafo mostra os meses em que os termos apareceram e a relação entre eles. No exemplo real da Figura 29 podemos dizer que os termos que melhor representam o assunto discutido são “Mobile” e “Android”.

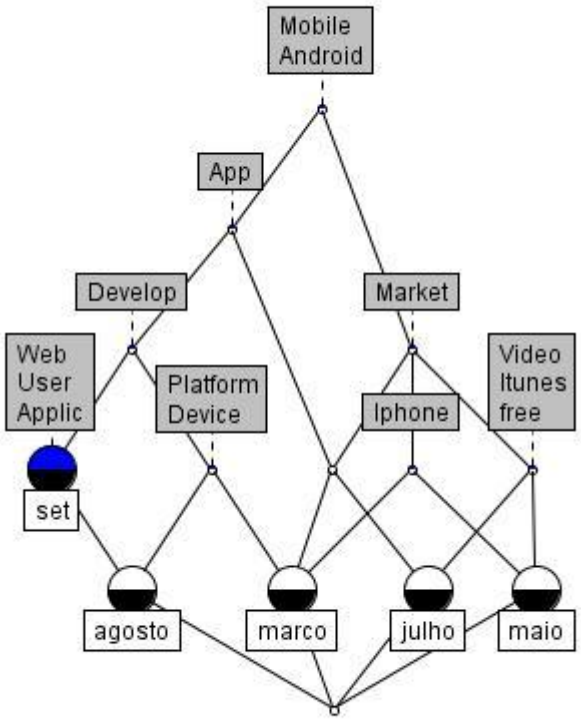


Figura 29: Dinâmica de um assunto

6.2 Definição dos Requisitos

A Figura 31, como dito anteriormente, apresenta a arquitetura proposta.

Os requisitos funcionais da ferramenta proposta, tendo como base os conceitos apresentados nos capítulos anteriores são:

- RF1- O sistema deve permitir que o usuário escolha a janela de tempo do termo a ser analisado.

- RF2- O sistema deve gerar um gráfico de popularidade do termo selecionado.
- RF3- O sistema deve permitir ao usuário ter uma visão mais macro de uma determinada "fatia" da curva.
- RF4- O sistema deve informar os pontos de *burst* na curva.
- RF5- O sistema deve informar os termos correlacionados ao buscado, na "fatia" selecionada do gráfico gerado.
- RF6- O sistema deve permitir ao usuário gerar curvas comparativas de popularidade entre dois termos.
- RF7- O sistema deve apresentar a listagem dos termos mais populares de um determinado período de tempo.
- RF8- O sistema deve informar o quão "quente" é um determinado termo da listagem.
- RF9- O sistema deve permitir a visualização dos documentos relacionados a uma determinada "fatia" da curva de popularidade do termo.
- RF10- O sistema deve permitir que o usuário escolha a "fatia" de tempo de visualização de assuntos.
- RF11- O sistema deve possibilitar a visualização dos assuntos mais relevantes do período previamente selecionado.
- RF12- O sistema deve informar o grau de popularidade dos assuntos listados.
- RF13- O sistema deve apresentar os documentos relacionados a um determinado assunto selecionado pelo usuário.
- RF14- O sistema deve gerar um gráfico de como um assunto se desenvolve ao longo de sua existência.

6.2.1 Análises supridas pela ferramenta proposta

A seguir serão listadas algumas das análises que a ferramenta se propõe a cobrir:

- **Curva de Popularidade:**
    - Analisar o que ocorreu em determinado ponto desta curva, através de um *drill down*[23] ou de termos relacionados;
    - Descobrir os pontos da curva considerados *bursts*[24];
    - Períodos de tempo com granularidades mais flexíveis do que em outras ferramentas;
    - Analisar assuntos e não somente termos, como a maioria dos trabalhos relacionados;
    - Tentar compreender o motivo de um determinado assunto ou termo estar entre os mais populares (apresentar a curva quando selecionar um determinado termo ou assunto);
    - Comparar dois termos para verificar se a popularidade de um pode estar afetando do outro, assim como se os dois se comportam de maneira semelhante ao longo do período de tempo.
- **Grafo de termos ou assuntos:**
    - Analisar o grau de relação entre os termos/assuntos mais populares em um determinado período.
- **Listagens:**
    - Listagem dos termos/eventos mais populares (popularidade relativa) de um determinado período;
        a. Definir o quão "quente" eles são. Se aquilo está no auge da moda ou já está na moda há um bom tempo;
    - Analisar quais assuntos de uma determinada área são os mais populares de um determinado período de tempo;
    - Analisar se o assunto é um assunto novo ou constante.

6.2.2 Casos de Uso

No Anexo I são apresentados alguns dos principais casos de uso que foram desenvolvidos para implementar os requisitos listados anteriormente.

[23] Aumentar o nível de detalhamento dos dados.

[24] *Outliers* positivos

6.2.3 Arquitetura do Sistema

A Figura 31 mostra a arquitetura proposta. Ela é formada por um banco de dados e duas grandes camadas que contêm uma série de componentes que, atuando em conjunto, compõem todas as funcionalidades do mecanismo proposto. Nas próximas seções serão descritos, com maiores detalhes, os módulos que formam a arquitetura proposta.

6.2.3.1 Camada de Aplicação

A Camada de Aplicação envolve todos os módulos responsáveis por consultar (*Keyword Search* e *TopicFinder*) e visualizar os termos/assuntos (*TopicFlowAnalyser*, *KeywordTrending* e *PostList*), ou seja, ela corresponde à parte do sistema que interage diretamente com os usuários.

6.2.3.2 Camada de Extração e Processamento

A Camada de Extração e Processamento envolve todos os módulos responsáveis por obter (*Blog Collector*), indexar (*Post Indexer*) e relacionar (*Topics Connector*) as postagens, ou seja, ela prepara o banco de dados para a Camada de Aplicação.

Outra característica de grande importância desta camada é o fato dela atuar de forma independente (*background*) da Camada de Aplicação. Como os processos da Camada de Extração e Processamento são contínuos, com essa abordagem, evitamos que o seu processamento impacte na utilização do sistema através da Camada de Aplicação.

6.2.3.3 Banco de Dados

Uma característica comum a todos os módulos deste trabalho é atuar sobre o mesmo banco de dados, que neste contexto, funciona como um repositório central de documentos (postagens, comentários, entre outros) devidamente processados, indexados e relacionados. A Figura 30 apresenta o modelo deste banco de dados.

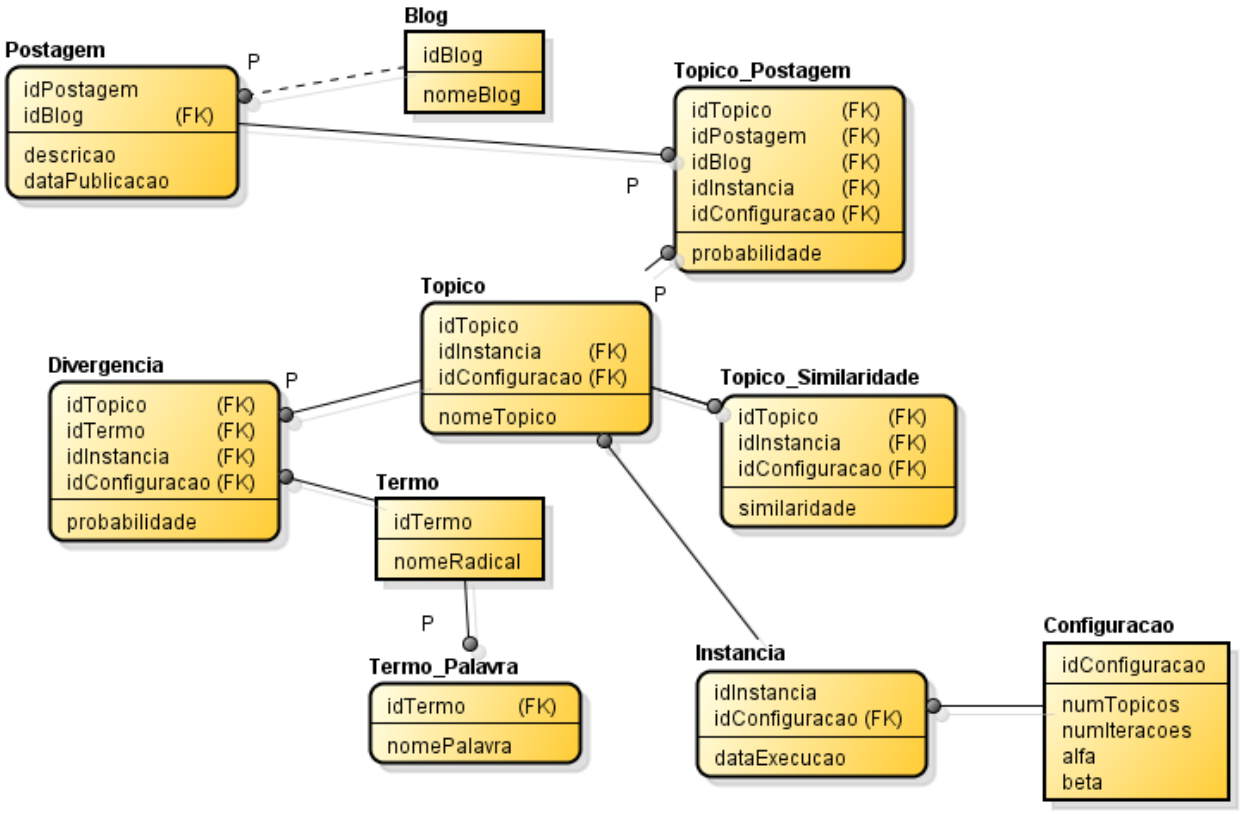


Figura 30: Modelo de dados

O modelo do banco de dados apresentado na Figura 30 é composto pelas seguintes tabelas:

- **TOPICO_POSTAGEM:** armazena as informações de que documentos estão ligados a quais tópicos;
- **DIVERGENCIA:** registra a probabilidade de um dado termo estar presente em um dado tópico;
- **TOPICO:** contém a descrição dos tópicos em si: sua identificação, a instância do corpus em que ele pertence e um nome que o representa;
- **TERMO:** contém o radical das palavras indexadas pela ferramenta;
- **BLOG:** armazena informações sobre os blogs coletados;
- **TERMO_PALAVRA:** contém as palavras que estão ligadas a um determinado radical;
- **INSTANCIA:** guarda quando foi executada uma determinada instancia do corpus;
- **POSTAGEM:** armazena os dados relevantes sobre cada documento coletado;
- **TOPICO_SIMILARIDADE:** registra o valor da similaridade entre dois tópicos.

- **CONFIGURACAO:** armazena os parâmetros de configuração do modelo de tópicos.

6.2.3.4 Tecnologias Utilizadas

A implementação protótipo da ferramenta BlogMiner (ela será mais bem detalhada no próximo capítulo) foi desenvolvida em Java como um sistema Web, tendo como base o framework *Struts*. A base de dados, derivada da modelagem apresentada na seção anterior, foi implementada utilizando o sistema de gerenciamento de banco de dados MySQL 5. As seguintes bibliotecas e ferramentas externas suportam diretamente as atividades da metodologia:

- JGIBBLDA[25]: implementação Java que utiliza a técnica de amostragem Gibbs para inferir e estimar parâmetros;
- JIT[26]: biblioteca gráfica Javascript;
- Flot[27]: biblioteca gráfica Javascript ;

[25] http://sourceforge.net/projects/jgibblda

[26] http://thejit.org/demos/

[27] http://code.google.com/p/flot/

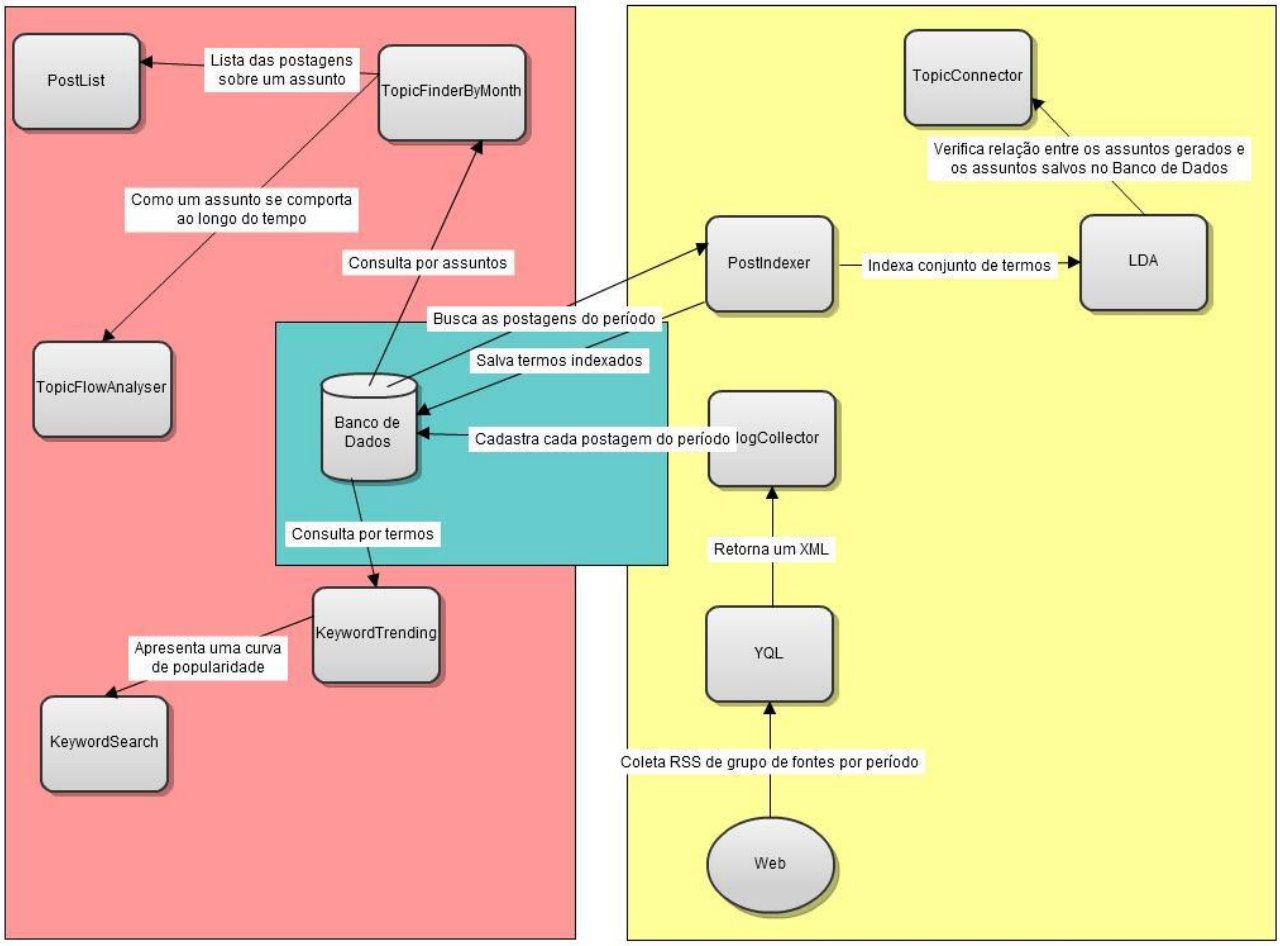


Figura 31: Arquitetura proposta

Capítulo 7 Implementação Protótipo do BlogMiner

Este capítulo descreve em detalhes os módulos da implementação protótipo da ferramenta BlogMiner, proposta neste trabalho.

Inicialmente serão descritos os módulos referentes à camada de Extração e Processamento, responsável por obter, executar o pré-processamento, inserir no Banco de Dados e relacionar as postagens dos blogs. Estes módulos são os seguintes: ***Blog Collector, Post Indexer*** e ***TopicConnector.***

Em seguida serão vistos os módulos pertencentes à camada de Aplicação, responsável por executar as análises sobre os textos ou termos dos blogs, assim como apresentar relacionamentos entre eles. Estes módulos são os seguintes: ***KeywordSearch, TopicFinder, TopicFlowAnalyser, PostList*** e ***KeywordTrending***.

7.1 Blog Collector

Seu objetivo é formar um grande repositório de postagens em blogs. Ele é responsável por obter, fazer o pré-processamento e salvar no banco de dados as postagens extraídas dos documentos previamente coletados. Documentos estes, no formato XML, formados por *RSSs* coletados através de consultas a determinados blogs através do YQL (ver capítulo 3 e Figura 32).

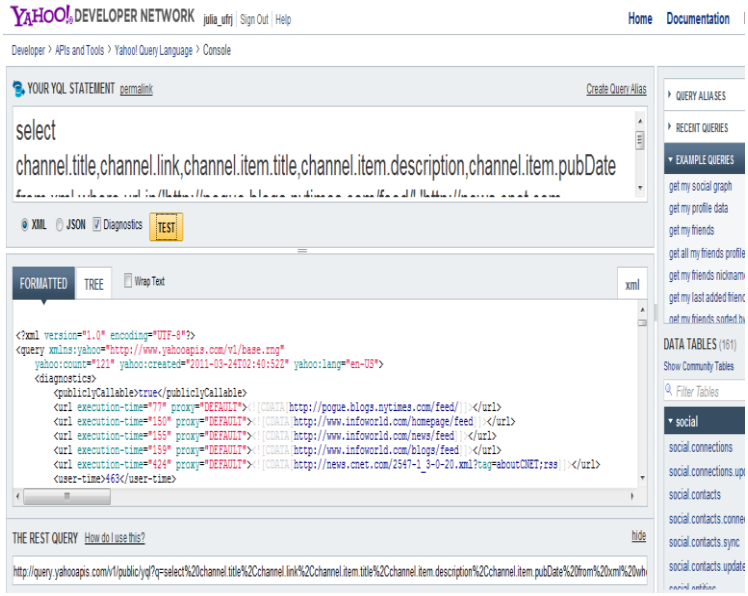


Figura 32: Console do YQL

A seguir descrevemos cada atividade contida neste módulo:

1. Definir o intervalo de tempo: Para cada base de dados coletamos em intervalos de tempo diferentes, de acordo com uma análise intuitiva da média de publicações dos blogs selecionados. No caso da área de tecnologia, coletamos uma vez por semana;
2. Definir a lista de blogs a serem coletados: Criar uma lista de URLs válidas para posteriormente serem consultados através do *webservice* YQL;
3. Consultar a lista de URLs no YQL: Verificar as URLs dos blogs selecionados e adicionarmos na consulta ao YQL, juntamente com o intervalo de tempo;
4. Extrair informações do canal de RSS: Filtrar a consulta descrita acima para que retorne somente os campos desejados dos RSSs coletados, no caso: a data de publicação, o nome do blog e o campo de descrição;
5. Realizar download dos documentos resultantes: Realiza o download dos arquivos XML resultantes da consulta citada anteriormente. Estes arquivos contêm informações de todos os blogs coletados com os campos filtrados no passo anterior;
6. Realizar inserção: Realiza uma inserção. Essa atividade será expandida e detalhada posteriormente;
7. Disparar o *Post Indexer:* Dispara o módulo *Post Indexer*.

A seguir detalhamos cada uma das atividades presentes na expansão da atividade "Realizar Inserção":

1. Lê o arquivo XML correspondente a um dos subconjuntos de RSSs coletados;
2. Obtém o conjunto de feeds;
3. Retorna um item da lista;
4. Obtém a lista de postagens do feed;
5. Retorna um item da lista de postagens;
6. Trata o item. Essa atividade será expandida e explicada em detalhes posteriormente;
7. Verifica se ainda existem itens (postagens daquele feed específico) não processados na lista;
8. Verifica se ainda existem feeds não processados na lista.

A seguir detalhamos cada uma das atividades presentes na expansão da atividade "Tratar Item":

1. Salva no banco de dados o texto da postagem coletada e a sua data de publicação. Para extrair a postagem, foi utilizado um algoritmo, que através do uso de expressões regulares, retira "ruídos" do documento, como *tags* HTML e eventuais subtextos que não estejam ligados a notícia em si;
2. Verifica se a fonte do item (o nome do blog) já foi cadastrada. Em caso positivo, segue com o fluxo de atividades, senão, insere a nova fonte;
3. Cadastra o item no banco de dados.

O Blog Collector não possui uma interface gráfica. Ele é executado diretamente pelo prompt de comando, e recebe como entrada do usuário apenas os arquivos a serem processados. Nesse prompt, a única informação exibida é uma mensagem informando que a inserção no banco de dados foi bem sucedida.

7.2 Post Indexer

O *Post Indexer* é o módulo responsável por indexar as postagens.

7.2.1 LDA

Este módulo é responsável por encontrar os assuntos falados pelos documentos coletados (no caso as postagens nos blogs) através do algoritmo LDA (Latent Dirichlet Allocation) que automaticamente aloca palavras em assuntos e documentos em uma mistura de assuntos, conforme detalhado no Capítulo 3. Este algoritmo já foi aplicado, por muitas vezes, com sucesso em diversos modelos científicos, conforme dito em [27].

Para desenvolver este módulo utilizamos o JGibbLDA[28] ,com algumas modificações para se adaptar ao contexto do BlogMiner (a criação de interface, por exemplo). O JGibbLDA é uma implementação Java de LDA que utiliza a técnica de amostragem de Gibbs[47] para estimar e inferir parâmetros.

A seguir detalhamos cada uma das atividades presentes neste módulo:

1. Define o período de datas de publicação dos textos para a modelagem de tópicos;

[28] http://jgibblda.sourceforge.net/

2. Define o número de tópicos e o número máximo de termos/documento a serem inseridos no banco de dados (além deste filtro os termos devem ter uma probabilidade maior que o valor pré-estabelecido de estar relacionado a determinado tópico), conforme detalhado no capítulo 5;
3. Define o número de iterações do algoritmo. Quanto maior a base maior é o número de iterações que são necessárias para que haja uma conversão mais interessante dos tópicos (em geral este número fica próximo de 2000);
4. Executa o algoritmo de modelagem de tópicos;
5. Retorna o status da execução do algoritmo. Como é um algoritmo custoso, pode demorar bastante tempo, acarretando alguns erros derivados deste problema e também de leitura e escrita em arquivos. Com isso, pode ser necessário a reexecução do processo.

7.3 TopicConnector

Calcula através de algoritmos de similaridade de vetores (no nosso caso, vetores com as palavras que melhor representam os tópicos), a semelhança entre os tópicos encontrados no fluxo de dados mais recente e os já existentes na base de dados.

É através do grau de similaridade deste módulo que o sistema calcula se o "novo" tópico encontrado se trata de um novo assunto ou de algum assunto já existente na base. Este módulo é baseado no algoritmo detalhado no Capítulo 5.

7.4 KeywordSearch

Este módulo é responsável por gerar a distribuição no tempo do termo selecionado durante o intervalo de tempo escolhido. Esta distribuição é calculada da seguinte forma:

1. Seleciona todos os documentos, presentes no intervalo de tempo informado, que possuem aquele termo, agrupados pelo dia;
2. Executa o pré-processamento dos documentos retornados;
3. Indexa os documentos pré-processados, armazenando também os termos contido nele e suas respectivas frequências relativas (com as ***stop words*** já removidas);
4. Calcula o valor mínimo para que a frequência relativa a um dado dia seja considerada um ***burst*** (conforme detalhado no Capítulo 3);

A partir dos valores encontrados gera-se um gráfico (conforme apresentado no Capítulo 6) mostrando as frequências dia a dia do termo e caso exista ***bursts*** eles são marcados no gráfico com uma cor diferente.

O módulo *Keyword Search* possibilita também a comparação da distribuição de dois termos, como na análise da popularidade de dois modelos de celulares. A única diferença para o modelo citado acima é que agora temos duas curvas e não mais uma, com isso executamos o processo duas vezes, uma para cada termo.

Como existe a opção de gerar o gráfico para um período de tempo bastante grande, fica praticamente impossível saber os dias exatamente em que ocorreram ***bursts*** por exemplo. Pensando nisso, este trabalho sugere a opção do usuário escolher trechos do gráfico para se ter uma visão macro (ou seja, gera-se um novo gráfico com os mesmos valores só que somente para aquela fatia de tempo) .

7.4.1 Termos Correlacionados

Além dos itens citados acima este módulo é responsável por calcular quais termos estão correlacionados ao buscado dentro do período de tempo desejado (após a filtragem inicial é possível selecionar fatias de tempo dentro do gráfico gerado).

Calcula-se a correlação (conforme detalhado no Capítulo 3), com os termos indexados ao gerar a distribuição e apresenta-se na tela somente os que estão acima de um limiar mínimo.

7.5 TopicFinder

O módulo TopicFinder é responsável por retornar os assuntos mais falados no período de tempo selecionado pelo usuário. Neste trabalho, os assuntos são representados “fisicamente” pelos tópicos.

O primeiro passo para retornar os tópicos é consultar o banco de dados para saber os tópicos encontrados pelo módulo *Full-Text Indexer.* Apesar de cada documento, em geral, falar sobre mais de um assunto/tópico, é armazenada apenas a relação com o tópico mais falado, como se ele falasse daquele tópico em específico, e sua porcentagem (quanto aquele documento fala sobre aquele tópico). Isto facilita encontrar os tópicos falados naquele período de tempo.

Para calcular o número de documentos que “falaram” sobre aquele tópico escolhemos apenas os que estão relacionados em uma porcentagem maior que 40% com aquele tópico, e após isto é

feito o somatório do número de documentos. Para referenciar os tópicos, escolhemos os termos com maior peso dentro destes.

A interface do módulo apresenta graficamente, depois de selecionado o período de tempo, os principais tópicos, sendo que quanto mais falados mais espaço na tela eles possuem, e o usuário tem a opção através desta interface de escolher um tópico desejado e visualizar as postagens referentes àquele tópico naquele intervalo de tempo, estes tópicos são ordenados por data de postagem e paginados para uma melhor visualização.

7.6 TopicFlow Analyser

Este módulo é responsável por entender como é a dinâmica de um assunto/tópico ao longo de períodos de tempo. Como um tópico pode se tornar outro ou como ele pode apresentar subtópicos completamente distintos ao passar do tempo, por exemplo.

O problema inicial encontrado na construção deste módulo foi de como saber que dois documentos de períodos de tempo distintos falam sobre o mesmo tópico (testes apontaram que os resultados encontrados pela modelagem de tópicos eram mais interessantes com períodos menores de tempo, como um mês, por exemplo). Muitas publicações falam de como tópicos se comportam com o tempo, mas praticamente nenhum deixa claro, como sabem que se trata do mesmo tópico.

Assim que um novo bloco de postagens é inserido no banco de dados e são modelados os tópicos, calculamos a similaridade entre os vetores com os termos mais relevantes dos novos tópicos e os vetores dos pré-existentes. Essa similaridade é calculada através da Similaridade do Cosseno, detalhada no Capítulo 3. São considerados sobre o mesmo tópico/assunto os documentos que possuírem vetores com similaridade maior que um determinado *threshold*, nas bases que foram utilizadas como teste este valor era igual a 60% (valor escolhido após ser testado em experimentos).

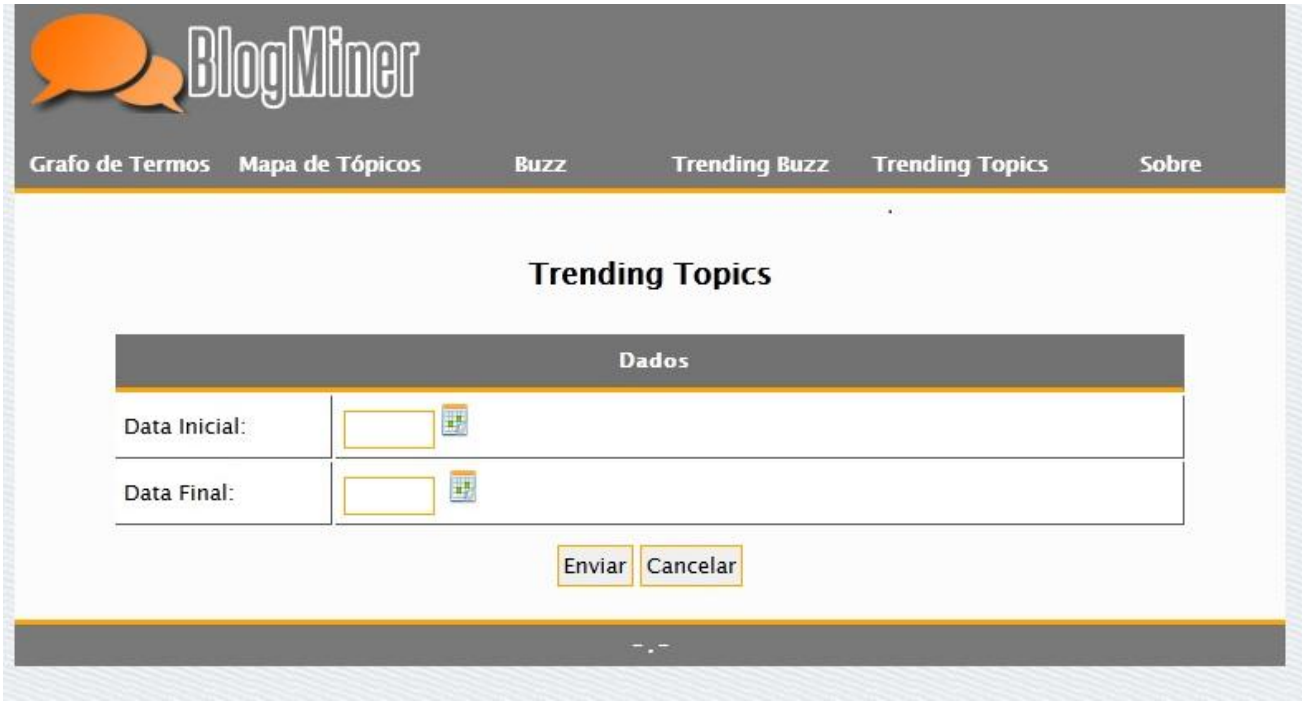
Após executar os cálculos anteriores, fazemos a FCA (Análise dos Conceitos Formais) de um assunto selecionado pelo usuário através da interface. Análise esta que auxilia na compreensão de que termo representa o assunto, que termos são derivados de outros, quais nunca estão ligados e etc.

7.7 PostList

Este módulo é responsável por recuperar e listar os documentos que falam sobre um determinado tópico com uma probabilidade acima de um determinado *threshold*. T*hreshold* este que possui o mesmo valor do citado na seção anterior .

7.8 KeywordTrending

Este módulo calcula a popularidade dos termos presentes em um determinado período de tempo, através de sua frequência relativa (média dos dias em que a frequência é maior que zero) e o quão "quente" ele está (analisando se é um termo novo dentro da base de dados ou se já está há algum tempo).



Capítulo 8 Exemplos de Uso

8.1 Visão Geral

Os experimentos propostos consistem em analisar o *BlogMiner* através da utilização de duas bases de dados distintas, uma sobre política americana e outra sobre tecnologia. Estas bases foram escolhidas por serem de fácil compreensão, ou seja, grande parte das pessoas ao lerem sabem de que se trata o assunto falado.

8.2 Política Americana

Com o desenvolvimento da web, novos modelos de mídia surgiram e ainda vem surgindo, como os blogs. Este modelo de mídia é recoberto de possibilidades, principalmente em relação à disponibilidade de informação para um mercado global, em tempo real, e à interação entre autores e leitores.

Os blogs, com suas publicações dinâmicas e gratuitas, têm demonstrado a necessidade de reformulação de conceitos e técnicas jornalísticas tradicionais quando o assunto é web.

É o que nota-se com o aumentos dos chamados blogs jornalísticos, alternativas de jornalismo na web, que utilizam uma linguagem pessoal, direta e que propicia a discussão pública no meio virtual.[36]

Este experimento pretende averiguar, através da análise de notícias online, sobre as últimas eleições presidenciais americana, a possibilidade de tratarmos os blogs como ferramentas jornalísticas na web e o BlogMiner como um novo meio de visualizar e compreender melhor estas notícias.

8.2.1 Corpus

Este corpus foi encontrado na web[29], no site do Instituto de *Language Technologies* da Universidade Carnegie Mellon. Ele foi coletado para utilização em pesquisas acadêmicas e possui dados de cinco blogs políticos americanos dos anos de 2007 e 2008.

Segundo [31] foram coletadas: postagens e comentários de blogs focados em política Americana no período de novembro de 2007 e outubro de 2008, período que antecedeu as eleições presidências de 2008. As discussões nestes blogs focam na política americana, e em muitos deles aparecem: os candidatos Democrata e Republicano, especulações sobre os resultados em vários

[29] http://www.ark.cs.cmu.edu/blog-data/

estados, e vários aspectos de política doméstica (mais comumente discutida) e internacional. Os sites foram selecionados de uma forma que abrangessem várias tendências políticas. De todos os blogs coletados, foram escolhidos apenas cinco que "acumularam" um maior número de postagens durante este período: Carpetbagger (CB)[30], Daily Kos(DK)[31], Matthew Yglesias (MY)[32], Red State (RS)[33] e Right Wing News (RWN)[34]. CB e MY deixaram de ser blogs independentes em agosto de 2008.

Foram descartados os blogs que não receberam comentários durante seis dias e os que possuíam muito poucas palavras. Todas as postagens foram representadas como texto apenas (foram ignorados, por exemplo, imagens e hiperlinks). Para padronizar os textos, removeu-se as 670 *stop words* mais usadas, símbolos não alfabéticos, incluindo pontuações, etc. Foram descartadas também termos com a frequência muito baixa.

8.2.2 Utilização da Ferramenta

8.2.2.1 Termos

Na Figura 33 apresentamos o gráfico comparativo dos termos "Democratic" e "Republican", que representam os dois principais partidos americanos (Republicano e Democrata) entre os meses de junho e setembro de 2008, meses que antecederam as eleições presidenciais. Podemos perceber uma correlação entre eles e que o Democrata, partido vitorioso, foi bem mais falado que o Republicano.

Na Figura 34 apresentamos a listagem com os termos mais relevantes durante o mesmo período de tempo e suas respectivas "popularidades" (conforme detalhado anteriormente). A esses termos populares damos o nome de *buzz*, conforme dito anteriormente. Podemos perceber que os nomes dos candidatos dos principais partidos aparecem no topo da lista, com o de Obama em primeiro lugar (candidato do partido Democrata, citado acima).

Ainda na Figura 34, podemos perceber certa duplicação dos resultados encontrados pela ferramenta: Obama e Barak Obama, por exemplo, se tratam da mesma pessoa e as duas variações

[30] http://www.thecarpetbaggerreport.com

[31] http://www.dailykos.com

[32] http://www.redstate.com

[33] http://www.rightwingnews.com

[34] CB washingtonmonthly.com e MY http://yglesias.thinkprogress.org

aparecem na listagem. Isto pode ser solucionado através do uso de entidades nomeadas, por exemplo, que foge ao escopo deste trabalho, podendo ser considerado como um trabalho futuro.

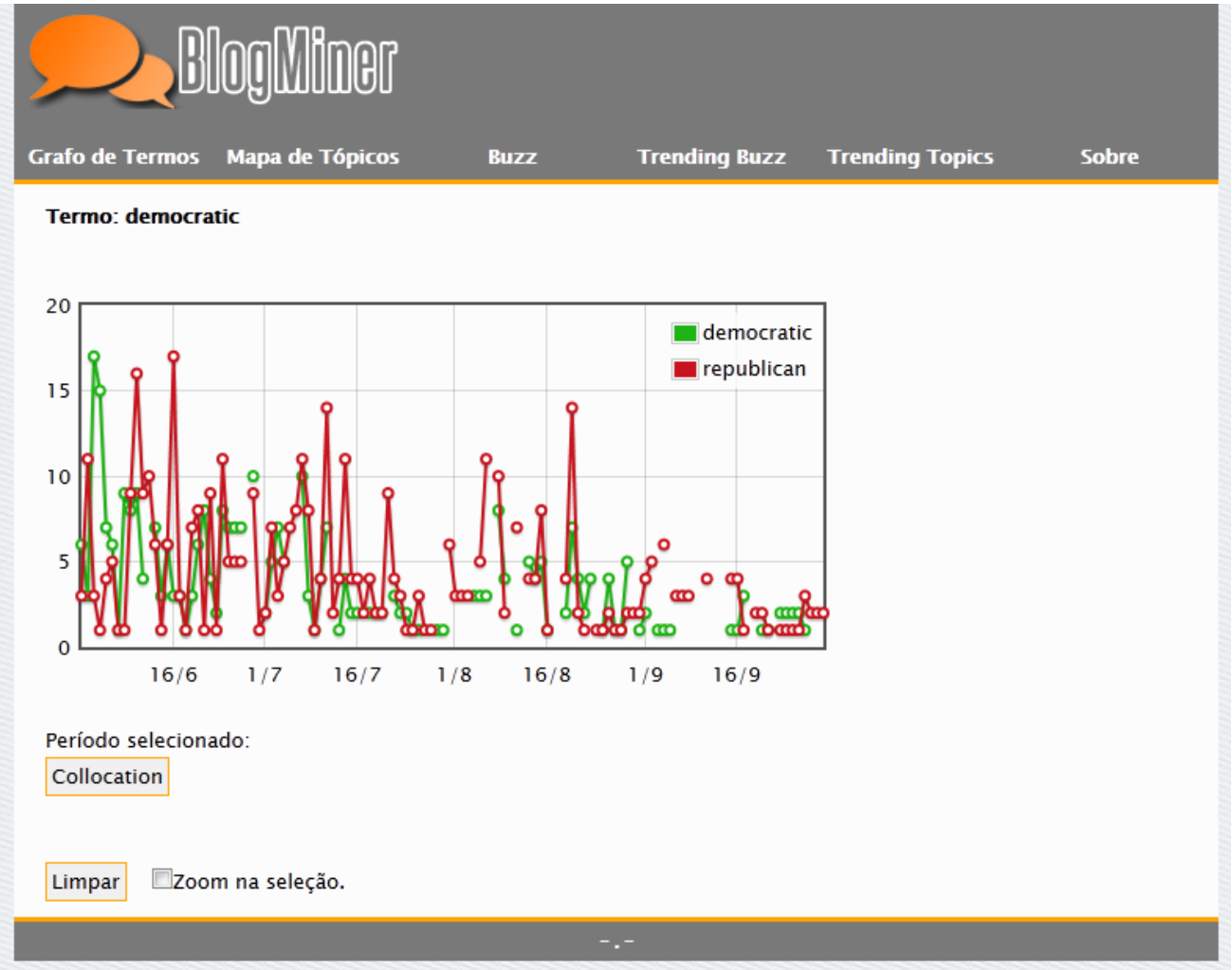


Figura 33: Buzz

Já na Figura 35 analisa-se o termo "Iraq", um dos termos mais discutidos no período de junho a setembro de 2008, conforme a Figura 34, para o mesmo período de tempo. Podemos ver através do gráfico um pico de popularidade no início de junho e ao selecionarmos a fatia do gráfico que abrange este período de tempo podemos ver os termos correlacionados a ele, que auxiliam na compreensão do motivo do pico. Caso o usuário se interesse mais em saber o motivo do pico, ele pode, através de outra funcionalidade presente no sistema, ver os documentos que falam sobre o assunto. No caso o módulo TopicFinder, que seria o item de menu "Mapa de Tópicos".

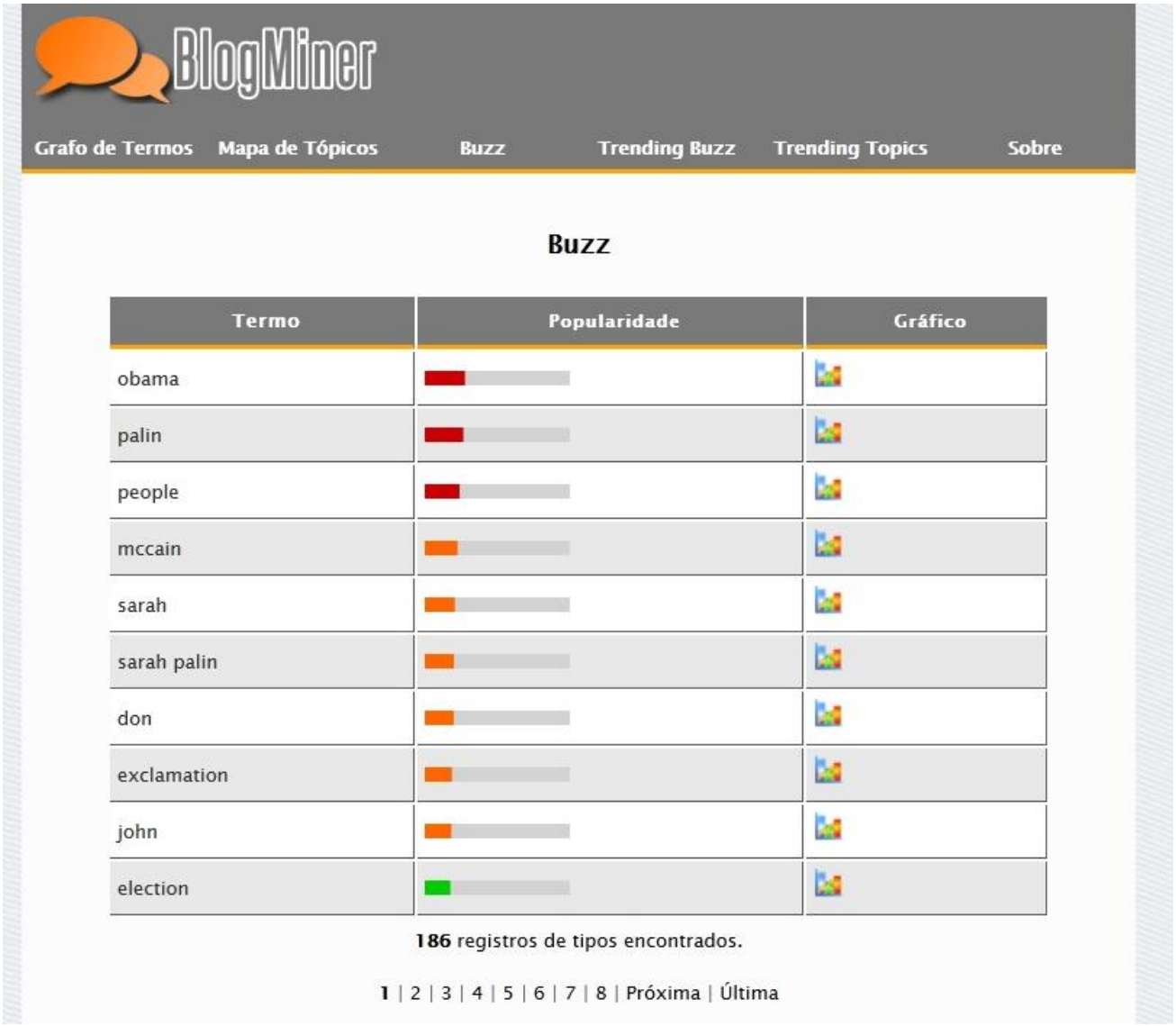


Figura 34: Principais termos

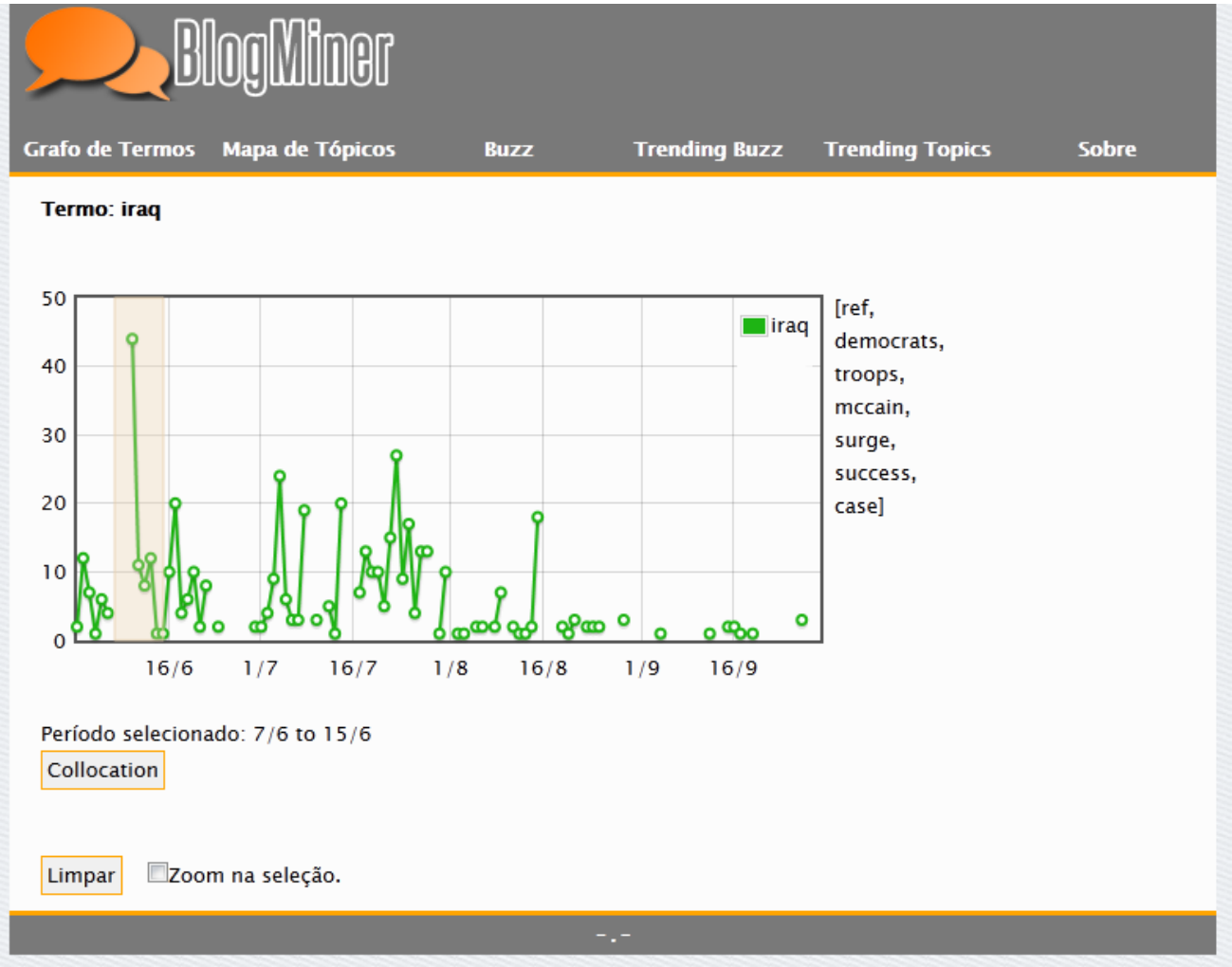


Figura 35: Curva do termo "Iraq"

8.2.2.2 Assuntos

Na Figura 36 podemos ver como se comporta um dos assuntos mais comentados entre os meses de dezembro de 2007 e fevereiro de 2008 (após isso ele não aparece mais). A ferramenta possibilita ao usuário, saber também os documentos que compõem este assunto, para compreendê-lo melhor. A partir deste conceito formal podemos tirar várias conclusões de quais termos derivam de quais outros, por exemplo. Através da ferramenta também pudemos constatar q este assunto decaiu sua frequência 50% ao mês.

Uma das possibilidades da ferramenta é saber os assuntos mais falados em um determinado período de tempo escolhido pelo usuário, além de ver os documentos que os compõem (clicando em um dos assuntos). A Figura 37 apresenta os principais assuntos encontrados no mês de janeiro de 2008, representados através de tópicos (a imagem mostra os principais termos-radicais presentes em cada tópico com seus tamanhos-áreas dos retângulos- diferindo de acordo com suas popularidades).

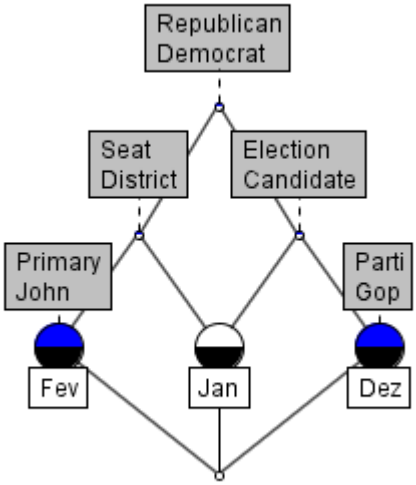


Figura 36: Conceito Formal

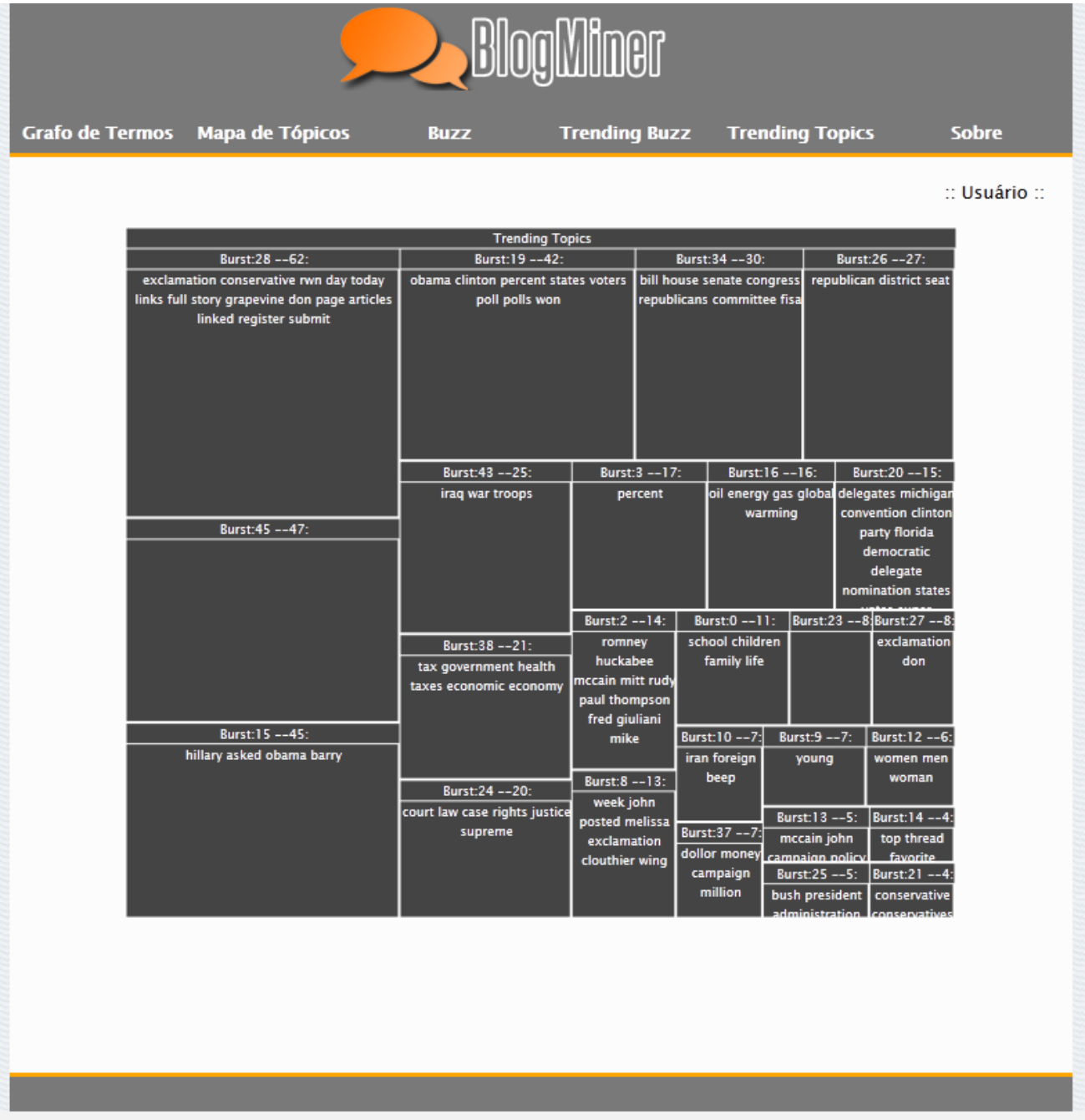


Figura 37: Mapeamento de assuntos

8.3 Tecnologia

Neste estudo de caso trabalhamos com blogs da área de tecnologia coletados ao longo do ano de 2011.

8.3.1 Corpus

Foram coletados RSS no período de fevereiro a novembro de 2011 dos seguintes blogs:

- http://pogue.blogs.nytimes.com/feed;
- http://news.cnet.com/2547-1_3-0-20.xml?tag=aboutCNET;RSS;
- http://www.infoworld.com/news/feed;
- http://www.infoworld.com/homepage/feed;
- http://www.infoworld.com/blogs/feed;
- http://feeds.ziffdavisenterprise.com/RSS/tech.xml;
- http://rss.news.yahoo.com/rss/techblog;
- http://www.bbc.co.uk/blogs/thereporters/maggieshiels/rss.xml;
- http://feeds.nytimes.com/nyt/rss/start-ups;
- http://bits.blogs.nytimes.com/feed/;
- http://feeds.nytimes.com/nyt/rss/Technology;
- http://feeds.nytimes.com/nyt/rss/business-computing;
- http://feeds.nytimes.com/nyt/rss/companies;
- http://feeds.nytimes.com/nyt/rss/internet;
- http://gadgetwise.blogs.nytimes.com/feed/;
- http://rss.businessweek.com/bw_rss/techbeat';
- http://feeds.technologyreview.com/technology_review_top_stories;
- http://feeds.reedbusiness.co.uk/e102d436-2872-4a80-b393-723ab272b4fe/CW360/Computer-Weekly-All-Content.xml ;
- http://feeds.feedburner.com/technologyevangelist/bkxI;
- http://feeds.abcnews.com/abcnews/technologyheadlines;
- http://feeds.feedburner.com/readwriteweb;
- http://feeds.feedburner.com/CloveTechnologysBlog?format=xml;
- http://feeds.feedburner.com/TechCrunch;
- http://feeds.feedburner.com/readwriteweb;

- http://www.engadget.com/rss.xml .

Foram escolhidos alguns dos blogs listados como mais relevantes na área de tecnologia-2011 em rankings na web, como do Technorati[35]. Foram armazenadas no banco de dados apenas informações, dos XMLs, julgadas na época como relevantes ao nosso experimento: a data de publicação, a descrição da notícia e o blog que a publicou.

As postagens foram armazenadas sem nenhum pré-processamento (isto foi feito posteriormente), pois poderia ser interessante armazenas as imagens contidas nelas, por exemplo.

Um problema encontrado ao tratar a base de dados foi que alguns blogs de um mesmo veículo de informação, por exemplo o New York Times, utilizavam a mesma postagem. Alguns filtros tiveram que ser criados para cobrir esta falha.

8.3.2 Utilização da Ferramenta

8.3.2.1 Termos

Na Figura 38 apresentamos o gráfico comparativo dos termos “IOS” e “Android”, durante os meses de setembro e novembro de 2011. Percebemos pela curva que Android foi consideravelmente mais falado neste período.

A Figura 39 apresenta os termos com maior frequência relativa no período coletado no experimento, podemos ver que a empresa Google e seu Sistema Operacional móvel Android foram os termos mais falados durante o ano em toda a área de tecnologia. Com isso, poderíamos tirar como uma conclusão inicial que a área móvel vem dominando os assuntos de tecnologia.

[35] http://technorati.com/blogs/directory/technology/

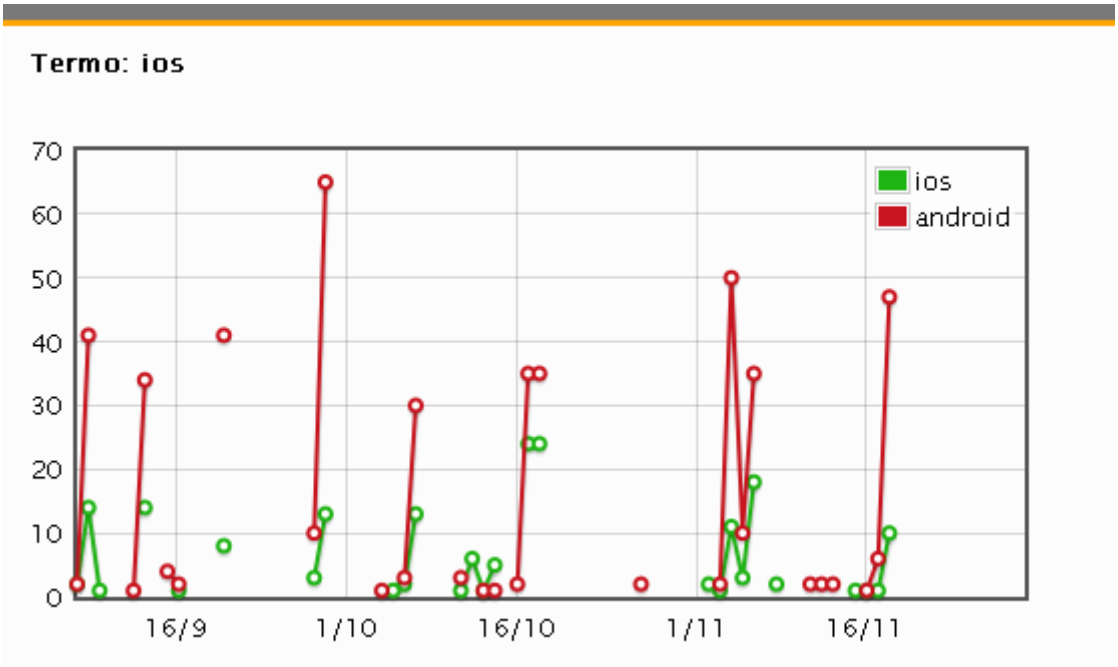


Figura 38: Análise comparativa dos termos IOS e Android

Figura 39: Termos mais falados durante o ano de 2011

Na Figura 40 temos a curva de popularidade do termo mais comentado do ano, no caso Google, no período entre setembro e novembro de 2011. A parte em verde significa quando ocorreram *bursts* (detalhados em capítulos anteriores).

Analisando os documentos da data do primeiro *burst* destacado em vermelho, vimos que se tratava do surgimento do Google Plus, o que pode ser averiguado através de buscas na web sobre o assunto.

Ao analisar este caso verificamos que apesar da ferramenta encontrar termos compostos praticamente não encontrou nenhuma referência ao Google Plus, pois analisando os documentos da maioria das referências, ele é tratado como "Google +", sendo que o "+" é retirado no pré-processamento, acarretando este problema.

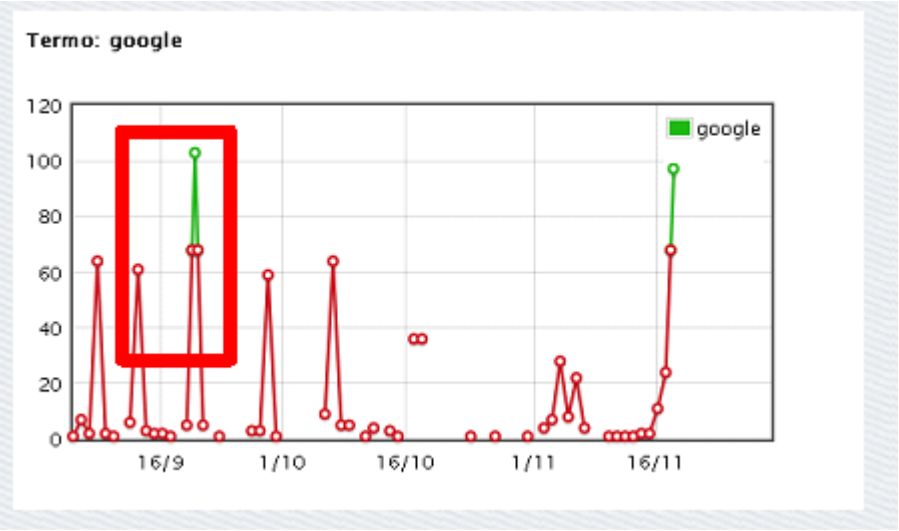


Figura 40: Curva de popularidade do termo Google

Já a apresenta os termos relacionados ao termo Google no período de 05 a 25 de novembro. Período que cobre o outro *burst* apresentado na curva da Figura 40.

8.3.2.2 Assuntos

A Figura 42 apresenta os assuntos mais discutidos do mês de novembro de 2011, sendo o evidenciado em vermelho o mais falado. Os assuntos são referenciados através de uma lista com os termos que melhor os representam, sendo estes termos apresentados através de seus radicais (neste estudo de caso em especial).

A Figura 43 apresenta os documentos que falam sobre o assunto evidenciado na Figura 42 e suas respectivas datas. No caso da listagem, não foram excluídas as imagens contidas nos textos, para facilitar a compreensão do assunto pelo usuário.

Dado que os tópicos são gerados de forma automática, algumas vezes são recuperados documentos que não são sobre exatamente o mesmo assunto que os demais, mas pela opinião de alguns usuários teste (em torno de dez testadores) isso não impactou na qualidade do resultado recuperado, pois a grande maioria falava sobre a mesma coisa.

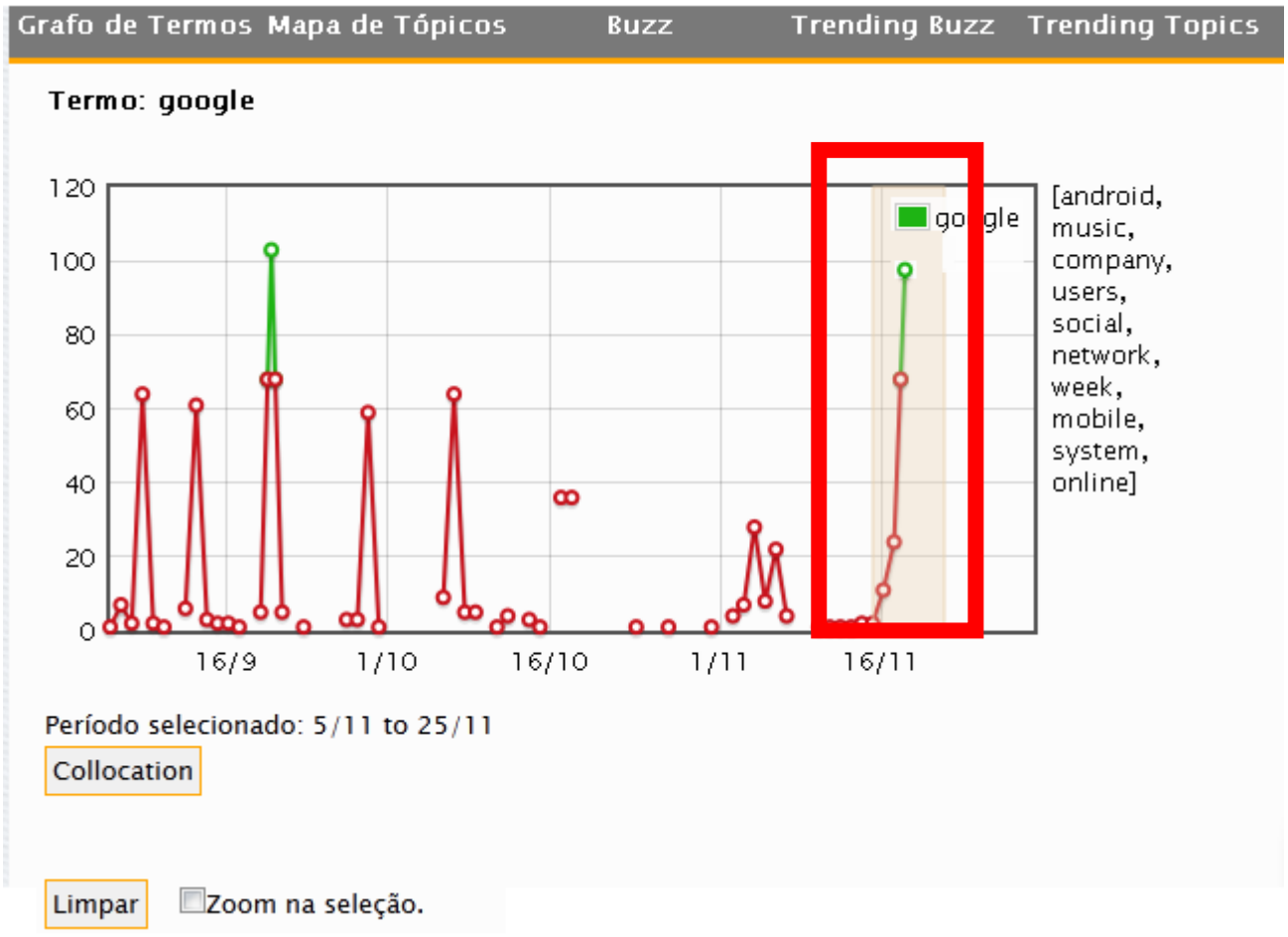


Figura 41 Termos relacionados ao termo Google

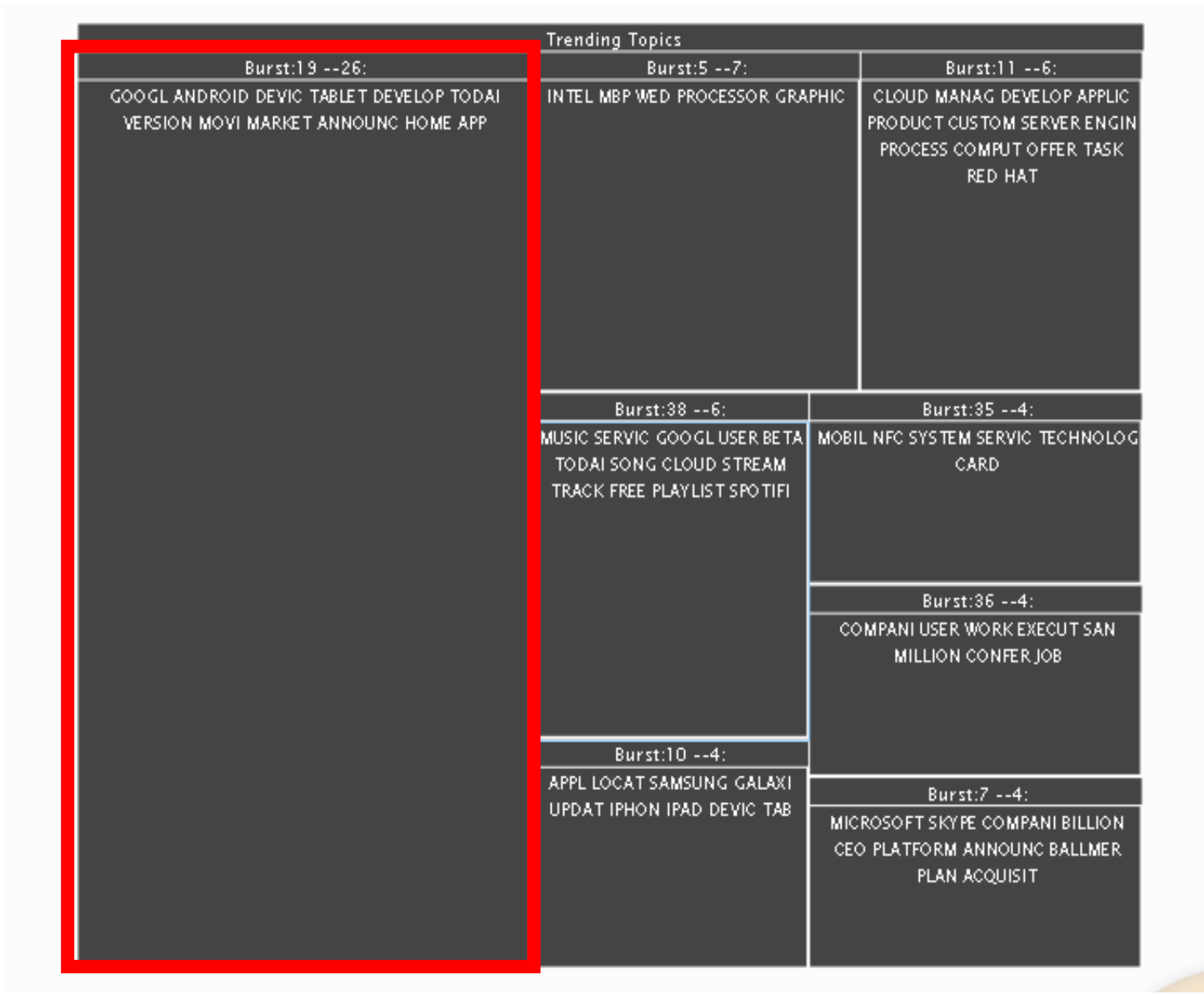
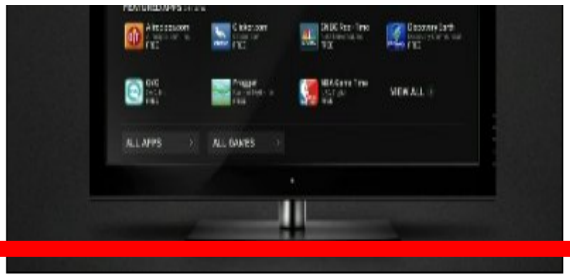


Figura 42: Mapa de tópicos



As expected, Google just announced at I/O that the Google TV will be upgraded to Android 3.1 this summer (existing devices will get an OTA upgrade) with access to the Market coming "soon." According to Mike Cleron from the Android Development team, developers will be able to use the vanilla Honeycomb SDK to build apps for Google TV, and also announced hardware partners will include Samsung, Vizio, Logitech and Sony (as seen after the break) -- no word on previously mentioned possibles like Toshiba, Sharp or LG. There were also no details on a switch from Intel to ARM even though we heard whispers of that at CES, we'll check in to the keynote tomorrow to see if there's any more revealed on hardware changes for the platform. Check the liveblog for more info. Continue reading *Google TV getting Android 3.1 and Market this summer; Sony, Vizio, Samsung and Logitech onboard* Google TV getting Android 3.1 and Market this summer; Sony, Vizio, Samsung and Logitech onboard originally appeared on Engadget on Tue, 10 May 2011 12:13:00 EDT. Please see our terms for use of feeds. Permalink | | Email this | Comments	2011-05-10 12:13:00.0
Google outlined Honeycomb 3.1 features today, which rolls out now to Motorola Xoom tablets on Verizon's network, with other devices to follow. There are noticeable improvements in the operating system, including support for USB add-ons, but consumers need more tablet-optimized third-party apps from developers.	2011-05-10 22:53:06.0

Figura 43: Listagem das postagens sobre o assunto evidenciado na Figura 41

Capítulo 9 Conclusão e Trabalhos Futuros

Neste trabalho, é apresentada uma abordagem alternativa às abordagens tradicionais de compreensão do que está acontecendo na Blogosfera ou em sites de notícias. Em geral, são feitas análises a partir de **termos**, palavras presentes nos documentos, mas em nosso caso fazemos análises com estes termos e com os assuntos em si, presentes nos blogs coletados, onde inferimos relações temporais, levando em conta diversos fatores como a similaridade dos documentos, suas datas de publicação, etc.

Foi escolhida a área de blogs pelas suas características, de certa forma, peculiares que servem principalmente para a criação de opinião. A criação autônoma, a liberdade de edição, a gratuidade e os comentários são aliciantes para quem quer liberdade e compartilhamento de suas opiniões.

Opinião que assume diversas formas: sobre política, sobre música, sobre mídias ou outros assuntos quaisquer.

Blogs além de gerarem opiniões dos blogueiros, geram também, de forma intrínseca, opiniões em seus leitores que estão interessados no tema exposto ou naquele blog.

Além de geradores de opinião, blogs podem ser descritos como fontes de notícias, ou até mesmo de “furos” jornalísticos. [46]

9.1 Contribuições

De acordo com os objetivos propostos, podemos destacar as seguintes contribuições deste trabalho:

1. Foi proposta uma arquitetura para apoiar o mecanismo desenvolvido de análise de postagens em blogs;

2. Foi desenvolvido um módulo (RSS Torrent) capaz de popular um repositório com postagens obtidas a partir de uma lista de *feeds* RSS;

3. Foi desenvolvido um módulo (Keyword Search) para analisar a curva de popularidade de termos presentes nos documentos coletados, considerando fatores como tempo, ”*bursts*” e termos relacionados;

4. Foi desenvolvido um módulo (TopicFinder) que permite determinar os assuntos mais falados em um determinado período de tempo e os documentos que os compõem;

5. Foi proposto um algoritmo (TopicFlow Analyser) que mapeia os caminhos que um assunto pode seguir, ou seja, mapeia essa dinâmica ao longo do tempo;

6. O mecanismo proposto (BlogMiner) foi avaliado por meio de experimentos que analisaram seu comportamento em relação a uma base de dados de blogs sobre política americana e uma sobre tecnologia.

9.2 Trabalhos Futuros

Grandes coleções de documentos estão disponíveis na Web e amplamente acessadas por diversas comunidades. Como exemplos notáveis, podemos citar os artigos acadêmicos que estão sendo publicados cada vez mais em formato eletrônico, e arquivos históricos que estão sendo digitalizados e se tornando mais acessíveis. Como os dados são em grande parte não estruturados e incluem milhares de artigos que abrangem séculos de trabalhos acadêmicos, análise automatizada é essencial. O desenvolvimento de novas ferramentas para navegar, pesquisar e que permita o uso produtivo de tais arquivos é, portanto, um desafio tecnológico importante, e oferece novas oportunidades para modelagem estatística, o que seria outra área interessante que a ferramenta poderia ser utilizada. [4]

Como trabalhos futuros, podemos destacar a realização de uma análise mais aprofundada da popularidade dos assuntos encontrados, similar à análise feita com termos.

- Melhorias na interface para se tornar mais dinâmica;

- Conseguir prever os próximos assuntos que serão quentes, ideia inicial deste trabalho;

- Interface para facilitar a coleta dos blogs;

- Prospecção Tecnológica;

- Adicionar ontologia aos vocabulários dos corpi.

Dessa forma, fica claro que o BlogMiner é apenas o início de uma pesquisa que, envolvendo causa e tempo, consegue melhorar ainda mais a recuperação da informação em blogs.

Capítulo 10 Referências Bibliográficas

[1]AlSumait, L., Barbara, D., Domeniconi, C., (2008), "On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking". In: *Eighth IEEE International Conference on Data Mining, 2008. ICDM '08Eighth IEEE International Conference on Data Mining, 2008. ICDM '08*, p. 3-12

[2]Phan, X.-H., Nguyen, L.-M., Horiguchi, S., (2008), "Learning to classify short and sparse text & web with hidden topics from large-scale data collections". In: *Proceedings of the 17th international conference on World Wide Web*, p. 91–100, New York, NY, USA.

[3]Bejan, C.A.,(2008), "Unsupervised discovery of event scenarios from texts". In: *The Florida Artificial Intelligence Research Society*, Coconut Grove, FL

[4]Blei, D. M., Lafferty, J. D., (2007), "A Correlated Topic Model of Science", *The Annals of Applied Statistics*, v. 1, n. 1 (jun.), p. 17-35.

[5]Bansal, N., Koudas, N., (2007), "Searching the blogosphere". In: 10th International Workshop on the Web and Databases (WebDB 2007)

[6]Platakis, M., Kotsakos, D., Gunopulos, D.,(2008), "Discovering Hot Topics in the Blogosphere". In Proc. of the 2nd Panhellenic Scientific Student Conference on Informatics, Related Technologies and Applications EUREKA 2008, pp. 122--132.

[7] R. Baeza-Yates e B. Ribeiro-Neto, 2011, *Modern Information Retrieval: The Concepts and Technology Behind Search*. Addison Wesley.

[8]Manning, C. D., Schuetze, H., (1999), *Foundations of Statistical Natural Language Processing*. 1 ed. The MIT Press.

[9]Marconi, M., Lakatos, E., (2004), *Metodologia Científica*, 4ed., São Paulo, Atlas.

[10]Bansal, N., Koudas, N., (2007), "BlogScope: a system for online analysis of high volume text streams". In: Proceedings of the 33rd international conference on Very large data bases, p. 1410–1413

[11]Lin, D., (1998), "An information-theoretic definition of similarity". In: *Proceedings of the 15th international conference on machine learning*, p. 296–304

[12]Griffiths, T. L., (2004), "Finding scientific topics", *Proceedings of the National Academy of Sciences*, v. 101, n. suppl_1 (jan.), p. 5228-5235.

[13]Zhu, S., Wu, J., Xiong, H., Xia, G., (2011), "Scaling up top-K cosine similarity search", *Data Knowl. Eng.*, v. 70, n. 1 (jan.), p. 60–83.

[14]McLean, R., Richards, B. H., & Wardman, J.I. (2007). "The effect of Web 2.0 on the future of medical practice and education: Darwikinian evolution of folksonomic revolution?" *Medical Journal of Australia* ,187 (3), 174-177.

[15]Blei, D. M., Lafferty, J. D., (2006), "Dynamic topic models". In: *Proceedings of the 23rd international conference on Machine learning*, p. 113–120, New York, NY, USA.

[16]Angel, A., Koudas, N., Sarkas, N., Srivastava, D., (2009), "What's on the grapevine?". In: *Proceedings of the 35th SIGMOD international conference on Management of data*, p. 1047–1050, New York, NY, USA.

[17]BlogScope. Disponível em: http://vbeta.pl/2010/11/23/nie-tylko-google-najlepsze-alternatywne-wyszukiwarki-w-sieci-top-13. Acesso em: 15 jul 2012.

[18]Blei, D.,Lafferty, J. (2009). "Topic models". In *Text Mining: Theory and Applications*. Taylor and Francis, London, UK.

[19]Poshyvanyk, D., Marcus, A., (2007), "Combining Formal Concept Analysis with Information Retrieval for Concept Location in Source Code". In: *Program Comprehension, 2007. ICPC '07. 15th IEEE International Conference on*, p. 37 -48

[20]Cimiano, P., Hotho, A., Staab, S., (2005), "Learning concept hierarchies from text corpora using formal concept analysis", *J. Artif. Int. Res.*, v. 24, n. 1 (ago.), p. 305–339.

[21]Baeza-Yates, R. A., (1992), "Information retrieval", In: Frakes, W. B., Baeza-Yates, R. [orgs.] (eds)Upper Saddle River, NJ, USA: Prentice-Hall, Inc., p. 13–27.

[22]Lee, J. H., Kim, M. H., Lee, Y. J., (1993), "Information Retrieval based on conceptual distance in is-a hierarchies", *Journal of Documentation*, v. 49, n. 2 (dez.), p. 188-207.

[23]Tversky, A. (1977). "Features of similarity".*Psychological Review*, 84:327–352.

[24]Omiecinski, E. R., (2003), "Alternative interest measures for mining associations in databases", *IEEE Transactions on Knowledge and Data Engineering*, v. 15, n. 1 (fev.), p. 57 - 69.

[25]Xiong, H., Tan, P.-N., Kumar, V., (2006), "Hyperclique pattern discovery", *Data Mining and Knowledge Discovery*, v. 13, n. 2, p. 219-242.

[26]Hofmann, T., (1999), "Probabilistic latent semantic indexing". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, p. 50–57, New York, NY, USA.

[27]Griffiths, T. L., & Steyvers, M. (2004)," Finding scientific topics. *Proceedings of the National Academy of Science,101*

[28]Ganter, B. (1999). *Formal concept analysis* . Berlin, Heidelberg: Springer.

[29]Blei, D. M., Ng, A. Y., Jordan, M. I., (2003), "Latent dirichlet allocation", *J. Mach. Learn. Res.*, v. 3 (mar.), p. 993–1022.

[30]Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R., (1990), "Indexing by latent semantic analysis". *Journal of the American Society for Info. Science* 41, 391-407.

[31]Yano, T., Cohen, W. W., Smith, N. A., (2009), "Predicting response to political blog posts with topic models". In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p. 477–485, Stroudsburg, PA, USA.

[32]BlogPulse.The Nielsen Company. Disponível em: http://www.blogpulse.com. Acesso em: 17 fev 2011.

[33]Priss U.,(2006), "Formal Concept Analysis in information science". *Annual Review of Information Science and Technology (ARIST)* 40, in press.

[34]Eisenbarth, T., Koschke, R., Simon, D., (2003), "Locating features in source code", *IEEE Transactions on Software Engineering*, v. 29, n. 3 (mar.), p. 210 - 224.

[35]Tilley, T., Cole, R., Becker, P., Eklund, P., (2005), "A Survey of Formal Concept Analysis Support for Software Engineering Activities", In: Ganter, B., Stumme, G., Wille, R. [orgs.] (eds), *Formal Concept Analysis*, , chapter 3626, Springer Berlin / Heidelberg, p. 250-271.

[36]Storch, L. ,(2007), “Weblogs como ferramentas jornalísticas através da análise das notícias online “. *Celacom.*

[37][1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, e Y. Yang, 1998, Topic Detection and Tracking Pilot Study Final Report, *Computer Science Department* (fev.)

[38]Chau, M., Lam, P., Shiu, B., Xu, J., Cao, J., (2009), "A Blog Mining Framework", *IT Professional*, v. 11, n. 1 (fev.), p. 36 -41.

[39]Yahoo! Buzzlist versus Google Trends. *ReadWriteWeb*. Disponível em: http://www.readwriteweb.com/archives/yahoo_buzzlist_versus_google_trends.php. Acesso em: 2 set 2012.

[40]Nielsen BuzzMetrics Tries to Measure Buzz in Social Media. *MediaShift*. Disponível em: http://www.pbs.org/mediashift/2007/01/nielsen-buzzmetrics-tries-to-measure-buzz-in-social-media003.html. Acesso em: 2 set 2012.

[41]The seven steps to a successful aggregation strategy for your news organization.*Poynter*. Disponível em: http://www.poynter.org/how-tos/digital-strategies/137285/the-seven-steps-to-a-successful-aggregation-strategy-for-your-news-organization/ Acesso em: 2 set 2012

[42] Newsola: A Cool, Visual Way To Read Google News. *MakeUseOf Directory*. Disponível em: http://www.makeuseof.com/dir/newsola-cool-visual-find-happening-world/. Acesso em: 2 set 2012.

[43]Rodrigues, T., (2011), WHYSEARCH: Um mecanismo causal e temporal de encadernamento de notícias. Dissertação de Mestrado, UFRJ.

[44]Blei, D. M., McAuliffe, J. D., (2010), "Supervised Topic Models", *arXiv:1003.0783* (mar.)

[45]Xexéo G., Morgado F., Fiuza P., 2009a. “Automatically Generated Tag Clouds”. *In Proceedings of the XXIV Simpósio Brasileiro de Banco de Dados, SBBD 2009*, pp. 136-150, Fortaleza, October, 2009.

[46] J. Simão, 2010, Relação entre os Blogs e Webjornalismo, *Revista Prisma.Com*, v. 0, n. 3 (out.)

[47] G. Casella e E.I. George, 1992, Explaining the Gibbs Sampler, *The American Statistician*, v. 46, n. 3, p. 167-174.

[48] BENEVETUDO, Fabrício; ALMEIDA, Jussara; SILVA, Altigran. Explorando Redes Sociais Online: Da Coleta e Análise de Grandes Bases de Dados às Aplicações. *XXIX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, Campo Grande, 2011.

ANEXO I

Caso de Uso	Keyword Search
Objetivo	Gerar um gráfico de tempo x frequência de um determinado termo.
Atores	Usuário do sistema
Pré-condições	Notícias previamente coletadas
Pós-condições	Curva(s) de popularidade plotada
Cenário Principal (Fluxo Principal)	1. Usuário seleciona o período de meses e o termo a serem analisados 2. Sistema gera a curva com a frequência do termo no período selecionado, realçando as datas que apresentaram picos positivos de popularidade
Fluxo Alternativo	1. Se o usuário selecionar mais de um termo a. Sistema gera uma curva para cada termo dado um mesmo período de tempo 2. Se o usuário selecionar uma determinada região do gráfico a. Sistema gera uma curva detalhando mais o período de tempo b. Se o usuário também selecionar a opção de Collocation i. Sistema apresenta os termos relacionados ao buscado naquele subperíodo de tempo
Exceções	

Caso de Uso	TopicFinder
Objetivo	Apresentar os assuntos mais falados em um determinado período de tempo.
Atores	Usuário do sistema
Pré-condições	Notícias previamente coletadas
Pós-condições	Os assuntos mais falados no mês selecionado representados graficamente
Cenário Principal (Fluxo Principal)	1. Usuário seleciona o período de tempo a ser analisado 2. Sistema gera os assuntos mais populares naquele período de tempo, através de um modelo gráfico em que enfatiza os mais falados entre os apresentados e apresenta os termos mais relacionados com aquele assunto
Fluxo Alternativo	1. Se o usuário selecionar um dos assuntos a. Sistema apresenta uma nova tela com a listagem dos documentos com maior relacionamento com o assunto selecionado
Exceções	

Caso de Uso	Hot Terms
Objetivo	Apresentar os termos mais populares em um determinado período de tempo.
Atores	Usuário do sistema
Pré-condições	Notícias previamente coletadas
Pós-condições	Listagem com os termos mais populares no período de tempo desejado
Cenário Principal (Fluxo Principal)	1. Usuário seleciona o período de tempo a ser analisado 2. Sistema gera os termos mais populares
Fluxo Alternativo	-
Exceções	