Ranqueamento de Notícias do Governo

Fabio Ferman[1], Matheus E. de Magalhães[1], Tiago S. da Silva[1], Luan B. Garrido[1], Miriam Chaves[2], Sérgio A. Rodrigues[1], Geraldo B. Xéxeo[1], Jano M. de Souza[1]

[1]Instituto Alberto Luiz Coimbra de Pós Graduação e Pesquisa de Engenharia – Universidade Federal do Rio de Janeiro (COPPE/UFRJ) – Cidade Universitária – RJ – Brasil

[2]Ministério do Planejamento, Orçamento e Gestão (SE/MP) – Esplanada dos Ministérios – Brasília/DF – Brasil

{fferman,emerick,tiagoss}@cos.ufrj.br, lbgarrido89@gmail.com, miriam.chaves@planejamento.gov.br,{sergio, xexeo, jano}@cos.ufrj.br

Abstract. *Follow the government news is an ordinary activity between managers, journalists and citizen in generals. Simply identify what is news of government can become a laborious process if the search engine is not the most appropriate. An even more challenging procedure is to establish a ranking between government news which are most relevant in a given spectrum. Into this scenario, this work aims to propose an algorithm that assigns a score to the government news that reflects its own importance. This score mainly uses attributes extracted from social networks of people or organizations profiles that publish government news. Besides showing the construction process of the ranking algorithm, this paper shows a case study with real news, the proposal and the validation results.*

Resumo. *Observar notícias de governo é uma atividade comum entre gestores, jornalistas e cidadãos em geral. O simples fato de identificar o que é uma notícia de governo pode se tornar um processo trabalhoso se o mecanismo de busca não for o mais adequado. Um procedimento ainda mais desafiador é estabelecer um ranqueamento entre quais notícias de governo são mais relevantes sob um determinado espectro. É neste cenário que encontra-se este trabalho, cujo objetivo é propor um algoritmo que atribua uma pontuação às notícias de governo e que reflita a importância de seu conteúdo. Esta pontuação utiliza principalmente atributos extraídos de redes sociais dos perfis de pessoas ou organizações que publicam notícias de governo. Além de apresentar o processo de construção do algoritmo de ranqueamento, este artigo mostra um estudo de caso com notícias reais, a validação da proposta e resultados obtidos.*

1. Introdução

No mundo atual a busca pelo conhecimento é eminente, as pessoas utilizam as mais diversas fontes de comunicação para expressar suas opiniões sobre os diversos cenários. Acompanhando essa tendência, os políticos brasileiros também têm expressado suas opiniões na *internet* não apenas em sítios privados e governamentais, mas também nas redes sociais. Essa nova vertente remete a uma situação em que os cidadãos estão cada

vez mais próximos e participantes dos processos de decisão de seus governantes, acompanhando as suas principais ações, divulgações e informações na grande rede.

Diversas alternativas podem ser utilizadas para o cidadão se interar do que está ocorrendo no mundo, sendo possível: ler jornais, acompanhar noticiários pela televisão ou na *Web*, entre outras formas. Com a sociedade em volta da era da informação digital, uma das principais maneiras de obter informações, é através de páginas na *Web* que cumprem o papel de divulgadores de notícias já existentes na *internet*. Estas disponibilizam informações criadas pelo próprio domínio, ou em alguns casos reúnem conteúdos de diversas fontes e o apresentam em sua página.

Devido à grande gama de notícias disponibilizadas pelos sítios, torna-se exaustiva a busca dos usuários por notícias específicas ou relevantes. Tendo em vista uma melhor utilização, é necessário que a página divulgadora faça uma pré-seleção destas notícias, possibilitando que os leitores encontrem as melhores e mais atuais, sem que estes tenham a necessidade de explorar seus respectivos conteúdos para encontrar informações mais interessantes.

No esforço de solucionar a problemática da enorme quantidade de notícias disponibilizadas em um só domínio e utilizando-se dessa nova era da busca de conhecimento em mídias sociais, o artigo apresenta uma solução de ranqueamento das notícias, no qual coloca as mais prioritárias e relevantes no topo, possibilitando o usuário encontrar o conteúdo que mais o agrade com uma maior facilidade.

Este artigo está organizado da seguinte forma: a seção 2 é apresentado um cenário mais específico e características deste ambiente que sugerem uma nova técnica para resolução do problema. A seção 3 mostra um caso de uso de aplicação do algoritmo. Na seção 4 são mostrados outros trabalhos com o mesmo foco e suas abordagens. A seção 5 mostra como é retirado conhecimento das redes sociais. Na seção 6 é mostrado como esse conhecimento é aplicado no ranqueamento das notícias. Na seção 7 é validado o algoritmo proposto e mostrado seus resultados. E por fim é mostrado as conclusões e trabalhos futuros na seção 8.

2. Motivação

Analisando o cenário de governo, existem diversos sítios eletrônicos que geram notícias, e, portanto, assim como encontrado nas notícias de assuntos diversos, a problemática do grande volume se apresenta.

Por estar se falando de notícias governamentais, é natural pensar que assuntos mais interessantes, são aqueles que os políticos e instituições governamentais estão discutindo no presente momento. No mundo atual das redes sociais, é fácil encontrar uma grande quantidade de perfis que seguem esta linha, no qual o conteúdo postado por estes, após ser processado, pode servir como um critério de ordenação de notícias, facilitando assim o usuário a chegar de forma mais rápida as notícias de seu interesse.

Existem já alguns sítios de notícias gerais, como o Google Notícias, que realiza um ranqueamento que estimula os seus usuários a lerem notícias em sua página, pois este consegue apresentar em sua primeira página as notícias de maior interesse. Dentre as técnicas que o Google utiliza para fazer o seu ranqueamento, estão: a utilização do número de cliques de usuários em cada notícia, estimativa gerada de uma medida de

interesse em determinado tópico, a sua data de publicação, informações geográficas das notícias, dentre outras [Sullivan, 2009].

Outros métodos de ranqueamento são encontrados na literatura, as principais técnicas e algoritmos associados ao seu desempenho e a qualidade dos resultados são: a análise do contexto gráfico do texto, diversas estatísticas de classificação, técnicas de aprendizagem de máquina, teorias de informações e o grau de entropia. Outras técnicas são utilizadas pelos grandes mecanismos de busca, atribuindo valores personalizados aos seus sistemas de ranqueamento, como critérios de localização geográfica, relatividade de interesses pessoais, linguagem do utilizador e histórico do browser [Elaysir and Anbananthen, 2012].

3. Caso de Uso: Noticias.Gov

O caso de uso a ser analisado será o do site do Noticias.Gov (http://www.noticias.gov.br) [Silva et. al., 2011]. Este sítio eletrônico é um sistema de busca de notícias do governo brasileiro, desenvolvido pela COPPE/UFRJ desde 2010 em parceria com o Ministério do Planejamento, Orçamento e Gestão.

O projeto é constituído por duas frentes de ação: um minerador e um portal *Web*. O minerador é responsável por visitar os diversos *sites* presentes em uma lista mantida no sistema, no qual procura pelas notícias para disponibilizar em seu portal. Além de notícias, o minerador também busca as publicações feitas por determinados políticos e instituições ligadas ao governo nas redes sociais (*Facebook* e *Twitter*). Junto com o conteúdo, também é recuperado para o caso dos *posts*: o número de curtidas, compartilhamentos e comentários; e para os *tweets*: o número de *retweets*. O portal *Web* disponibiliza todo o conteúdo minerado em diversos tipos de visualizações, que permite o usuário se inteirar dos acontecimentos do país e do governo.

O principal formato de disponibilização de notícias utiliza a ordenação por data de publicação como critério de ranqueamento. O portal também disponibiliza o ranqueamento por relevância, o qual somente se torna válido perante a busca do usuário. Nesta abordagem, a data de publicação é desconsiderada, o que faz com que notícias antigas possam estar entre as primeiras apresentadas.

4. Trabalhos Relacionados e Abordagem

Na literatura são apresentados outros trabalhos de ranqueamento de notícias os quais fazem uso de certas premissas. Em um primeiro trabalho, que utiliza apenas dos atributos presentes nas notícias para ranqueá-las, destaca-se a ideia dos autores de utilizarem a quantidade de notícias que abordam o mesmo tema como fator relevante no algoritmo [Xiaofeng, Chuanbo and Yunsheng, 2007]. Em um segundo trabalho, é feita a combinação de um atributo gerado através do retorno dado pelos usuários (comentários nas notícias) e de outros atributos (por exemplo a data de publicação, privilegiando notícias mais recentes) para se decidir como ordená-las [Kong et. al., 2012].

A abordagem a ser seguida neste trabalho, utiliza técnicas citadas anteriormente, e ainda conhecimentos que são retirados das mídias sociais (*Facebook* e *Twitter*) como o retorno dos usuários (número de curtidas, *retweets* e outros) e o conteúdo das publicações. Nas seções a seguir serão detalhados cada um desses aspectos.

5. Detecção dos Assuntos mais Relevantes

A primeira etapa a ser realizada é o cálculo da importância (a qual reflete a relevância do assunto) de cada publicação nas mídias sociais, no caso deste trabalho, publicações feitas no *Twitter* e no *Facebook* [Phelan, McCarthy and Smyth, 2009].

5.1. Cálculo das Importâncias das Publicações

Para se obter o valor da importância das publicações, são calculadas três pontuações: a pontuação interna (PI), pontuação externa (PE) e a pontuação do perfil publicador (PPP).

Para o cálculo da pontuação interna, as publicações dos últimos D dias são recuperadas e de acordo com sua data e hora de publicação são colocados nos seus respectivos grupos. Cada um dos G grupos existentes irá contemplar uma faixa horária tal que a composição de todos os grupos preencha todos os possíveis horários dos últimos D dias. A pontuação que cada publicação irá ganhar depende unicamente do grupo que esta foi colocada. Se a publicação está em um grupo que contemple uma faixa horária mais recente então ela ganha uma pontuação maior que outra publicação que está em um grupo que contemple uma faixa horária mais antiga.

Para calcular a segunda pontuação (PE), o primeiro passo é recuperar as demais publicações que sejam do mesmo tema (tenham uma maior similaridade entre si). Para o cálculo da similaridade, pode-se utilizar a biblioteca do Lucene [Mohd, 2011], que implementa esta funcionalidade, utilizando-se dentre diversas técnicas: a comparação do vetor de frequência das notícias e publicações e o cálculo do TF-IDF (este cálculo resulta em um número que tende a refletir a importância de cada palavra em relação ao documento que pertence, em meio a uma gama de documentos) [Zhang, Yoshida and Tang, 2011].

Para cada publicação relacionada recuperada, relacioná-la a algum dos G grupos, se possível (Figura 1). A pontuação de cada grupo é calculada da seguinte forma: soma-se um valor X caso a fonte da publicação relacionada seja igual à fonte da publicação em análise, e Y caso sejam diferentes; sendo Y>X para prevalecer os assuntos que foram publicados por diferentes perfis.

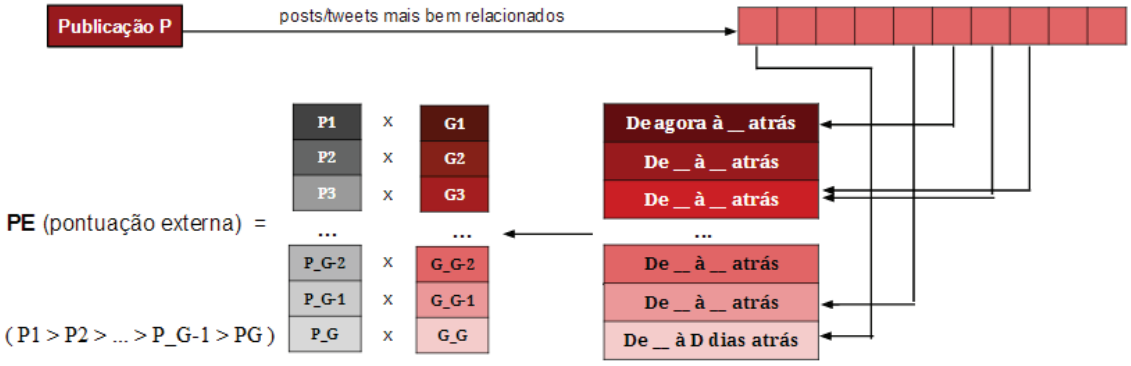

Figura 1. Pontos por grupo na pontuação externa.

A pontuação associada a cada grupo é multiplicado ainda por um valor fixo equivalente ao grupo, sendo este valor maior em grupos que se refiram às faixas horárias

mais recentes. A seguir, todos esses valores são somados, resultando na pontuação externa da publicação.

A última pontuação a ser calculada (PPP) faz as publicações serem também influenciadas por quem a postou. Assim, se a publicação foi feita por um perfil muito importante, por exemplo, do presidente da república, então pode-se dar preferência a todas as publicações deste perfil, uma vez que este tende a publicar conteúdos com maior chance de serem assuntos mais relevantes.

O cálculo final da importância (I) que cada publicação ganha ao final de uma rodada do algoritmo é dada pela fórmula:

1. Para os *tweets*:

$$I = PPP \times (PI \times (\text{número de } \textit{retweets}) + PE) \quad (1)$$

2. Para os *posts*:

$$I = PPP \times (PI \times (\text{número de curtidas} + \text{número de comentários} + \text{número de compartilhamentos}) + PE) \quad (2)$$

5.2. Normalização

Devido à possibilidade da ocorrência de valores de grandeza elevada e distâncias muito grandes entre as importâncias é necessária a sua normalização, uma vez que valores muito discrepantes podem não representar fielmente a realidade, afetando o resultado.

Para normalizar, deve-se partir da publicação com a menor pontuação (P_m) e ir seguindo até a publicação com maior pontuação ($P_m < ... < P_k < P_{k+1} < ... < P_M$), fazendo valer as seguintes regras:

- P_{k+1} menos P_k deve ser menor ou igual a R. Caso contrário, então P_{k+1} passa a valer P_k + R.;
- P_{k+1} menos P_k deve ser menor ou igual a P_k. Caso contrário, então P_{k+1} passa a valer P_k+P_k;

É importante destacar que não é possível fazer uma relação direta entre o retorno dado pelos leitores no *Facebook* e o retorno dado no *Twitter*, de modo que as normalizações devem ser feitas separadamente entre *posts* e *tweets*.

5.3. Conversão

Para facilitar a análise de importância entre as publicações à medida que o tempo passa, é feita uma transformação de variáveis, fazendo-se a importância (I) passar a ser uma data, chamada neste trabalho de data relativa (DR).

Para se fazer a conversão, determina-se qual será a faixa que a data relativa pode valer. Essa faixa tem seu valor máximo a exata data em que o algoritmo está executando e o valor mínimo igual à data em que o algoritmo está rodando menos os D dias determinados durante o cálculo da pontuação interna.

Aplicando a fórmula de conversão de variáveis (utilizada também em conversão de temperatura), e realizando os cálculos, tem-se que:

$$\text{Data relativa} = \frac{\text{Pontuação x (DAR - DARD)}}{I_{MAX}} + \text{DARD} \quad (3)$$

Com I_{MAX} sendo o valor da maior importância achado na última rodada do algoritmo, DAR representando a data em que o algoritmo foi rodado e DARD igual a data que o algoritmo foi rodada menos os D dias previamente estabelecidos.

5.4. Rodadas do Algoritmo de Detecção de *Hot Topics*

Os itens anteriores explicaram como em uma determinada rodada do algoritmo se calcula a importância das publicações. A freqüência com que o algoritmo é rodado determina de quanto em quanto tempo são atualizados os assuntos mais importantes. É interessante que se rode o algoritmo toda vez que uma nova remessa de publicações das redes sociais é carregada no sistema, uma vez que neste momento podem ser detectados novos tópicos importantes.

Nota-se que a cada rodada do algoritmo, uma determinada publicação tem seu valor de importância alterado, e cada vez mais o seu valor perde importância em relação às publicações mais novas, garantindo que a data de publicação continua sendo o principal critério de ordenação, porém não mais o único.

6. Ranqueamento das Notícias

O próximo passo é de fato realizar o ranqueamento das notícias. A cada vez que o algoritmo que atualiza o ranqueamento é executado, quatro passos são realizados: relacionar as publicações com as notícias, gratificar notícias com temas mais importantes, privilegiar fontes publicadoras e tratar as primeiras notícias do ranqueamento.

6.1. Relacionar as Notícias com as Publicações

Uma vez tendo as publicações das mídias sociais e suas importâncias, pode-se relacionar estas com as notícias, deixando a ordenação das notícias fortemente ligada ao que os políticos e instituições governamentais estão publicando e que o público esteja dando um bom retorno para o conteúdo.

Para cada notícia publicada até D dias atrás, encontra-se a publicação dentro de uma faixa de tempo de D_2 dias que tem a maior similaridade com o conteúdo da notícia. Assim, relaciona-se a toda notícia uma publicação [Guo et. al., 2012].

Assim como nas publicações, é criado para cada notícia o atributo da data relativa, na qual após rodado o algoritmo, o ranqueamento é dado pela simples ordenação deste novo atributo presente em cada notícia. Para atribuir um valor a data relativa da notícia (DRN), são utilizados outros dois valores, sendo estes a data de publicação da notícia (DPN) e a data relativa da publicação que foi anteriormente associada (DRP) à notícia. Assim, através da seguinte fórmula atribui-se uma data relativa a cada uma das notícias:

$$\text{DRN} = \frac{\text{A x DPN + B x DRP}}{\text{A + B}} \quad (4)$$

Sendo A e B valores escolhidos que dão mais peso a data de publicação da notícia ou a data relativa da publicação. Quanto maior o peso dado à data de publicação

da notícia, menor será a possibilidade de uma notícia mais antiga aparecer na frente de uma notícia mais recente.

No caso da notícia não ter uma publicação associada (ou seja, nenhuma postagem abordava o mesmo tema), então ou se retira a notícia ou então coloca como data relativa da notícia a data de publicação desta menos uma penalização P:

$$DRN = DPN - P \quad (5)$$

6.2. Gratificar as Notícias de Temas mais Freqüentes

As notícias são relacionadas umas com as outras (utilizando-se o Lucene, por exemplo), resultando para cada notícia a lista de notícias relacionadas a elas. A notícia com a maior quantidade de notícias relacionadas é selecionada por ter fortes indícios de se referir a um *hot topic*. A esta notícia é dado um acréscimo na data relativa de W, e às demais notícias é dado uma bonificação de Wr:

$$Wr = \frac{W \times NR}{NR_{MAX}} \quad (6)$$

Sendo NR o número de notícias relacionadas que a notícia que está tendo a sua bonificação calculada contém, e NR_{MAX} o máximo número de notícias relacionadas que uma determinada notícia dentro da faixa de horários de 0 à D dias atrás contém.

6.3. Tratar Notícias e suas Fontes Divulgadoras

Para se diferenciar notícias disponibilizadas por diferentes fontes de publicação, associa a cada fonte um valor de importância IF. A cada rodada do algoritmo é achado a média dos IFs, denominada pela letra X. Feito isso, para cada fonte é achado um valor Y que equivale à diferença entre o seu IF e a média X. Nota-se que caso o IF da fonte seja maior que a média, então Y terá valor positivo, caso o IF seja menor que a média, então Y terá valor negativo.

Para cada notícia, é acrescido no valor da sua data relativa um valor H, sendo H igual a Y vezes M minutos. Assim, as notícias ganham uma bonificação relativa à qualidade do seu conteúdo, sendo esta qualidade medida pela freqüência de notícias que abordam aquele assunto.

6.4. Tratar as Primeiras Notícias do Ranqueamento

Com o intuito de apresentar notícias de assuntos diferenciados na primeira página, essa tarefa consiste em analisar as notícias que atualmente nela estão sendo mostradas e verificar se existe alguma que está relacionada entre si, caso isso ocorra, desloca-se então a pior classificada para outra página, resgatando a próxima notícia presente em outra página para a primeira página. Com essa técnica, tende-se a ter na primeira página notícias que abordem diferentes temas, tentando chamar ao máximo a maior atenção da maior quantidade de leitores com a maior quantidade possível de assuntos.

6.5. Rodadas do Algoritmo de Ranqueamento de Notícias

Após as notícias serem ranqueadas, em primeira instância a ordem delas não se altera até que uma segunda rodada no ranqueamento ocorra. Todavia, sempre é possível que o ranqueamento que o algoritmo devolveu não seja o ideal, ou uma determinada notícia

começou a se destacar sem que esta tenha dado indícios anteriormente que deveria aparecer primeiro.

Para detectar que uma determinada notícia está sendo muito lida, pode-se contabilizar a quantidade de visualizações que ocorreram para ela. Assim, através do clique do leitor na notícia, pode-se inferir quais notícias estão sendo mais lidas. Para favorecer notícias muito lidas, pode-se atribuir a data relativa um acréscimo C por cliques que ela recebeu, permitindo mudanças no ranqueamento após a execução do algoritmo.

Assim como ocorre com o algoritmo de detecção de *hot topics* nas mídias sociais, este também é rodado de tempos em tempos. Quanto mais curto o intervalo de tempo que este algoritmo rodar, mais rápida será a atualização da ordenação. O ideal é que seja rodado esse algoritmo cada vez que uma nova remessa de notícias for carregada na base ou o primeiro algoritmo rodado, podendo assim inserir as novas notícias no ranqueamento ou atualizar com os novos assuntos mais comentados.

É importante notar que assim como ocorrido com as publicações, a cada rodada do algoritmo, uma notícia que esteja na faixa de 0 à D dias atrás tem o seu índice atualizado, e que após passados D dias, o seu índice permanecerá o mesmo exceto pelo acréscimo que a notícia pode ganhar por receber cliques. Assim como acontecido anteriormente, o algoritmo garante que notícias recentes tenham melhores índices que notícias mais antigas.

7. Valores Utilizados no Caso de Uso do Noticias.Gov e seus Resultados

Para realizar os testes da metodologia proposta anteriormente, foi necessário atribuir valores a todas as variáveis antes criadas. Para testar o algoritmo, utilizou-se o Noticias.Gov e os valores indicados nas sub-seções a seguir.

7.1. Os Grupos e Pontuações usados nas Pontuações Interna e Externa

Escolheu-se que seria aceitável que uma publicação de até três dias antes que uma terceira seja melhor ranqueada na escolha de um *hot topic*, embora essa situação raramente deva acontecer. Então por esse motivo, as faixas horárias deveriam variar até contemplar os últimos 3 dias (D igual a 3). Para a escolha dos grupos e suas faixas de tempo, é razoável pensar que deve-se diferenciar mais as faixas de tempo recentes do que faixas de tempo antigas, uma vez que para uma publicação de dois dias atrás é quase indiferente se esta foi publicada as oito da manhã ou dez da noite, porém se a publicação é de hoje, faz diferença se esta foi publicada agora ou há seis horas atrás. Por esse motivo, o tamanho da faixa de tempo de cada grupo cresce em uma ordem exponencial, resultando na criação de 7 grupos (G igual a 7), os quais receberam como pontuação interna e pesos a serem multiplicados no cálculo da pontuação externa os valores presentes na Tabela 1.

As pontuações internas e os pesos (P1, P2, ..., P_G) utilizados no cálculo da pontuação externa tiveram seus valores atribuídos a cada grupo seguindo uma distribuição exponencial, uma vez que é interessante beneficiar os mais recentes assuntos que estão sendo discutidos. Assuntos novos também são beneficiados recebendo uma maior pontuação, contrabalanceado sua desvantagem de não ter publicações mais antigas falando sobre o mesmo assunto.

Além dos pesos, para a pontuação externa também se precisa definir os valores X e Y, que definem respectivamente o valor somado quando a publicação que tem sua importância sendo calculada e a publicação relacionada possuem os mesmos ou diferentes perfis publicadores. Como já explicado, Y deve ser maior que X para favorecer assuntos que são publicados por diferentes pessoas. Assim, foram atribuídos os valores 1 e 3 para X e Y respectivamente.

Tabela 1. Valores utilizados para os grupos na validação do algoritmo

Grupo	Início da faixa	Fim da faixa	Pontuação Interna	Peso na pontuação externa
1	Instante atual	1h atrás	200	1000
2	1h atrás	3h atrás	100	50
3	3h atrás	6h atrás	50	25
4	6h atrás	12h atrás	25	13
5	12h atrás	1 dia atrás	12	7
6	1 dia atrás	2 dias atrás	4	3
7	2 dias atrás	3 dias atrás	2	1

Com o intuito de normalizar e deixar os valores mais próximos entre si na pontuação, escolheu-se experimentalmente R = 50.

7.2. Relacionar as Notícias com a Publicação mais Similar e Bonificações

O primeiro passo é definir o valor de D_2, que define dado uma notícia, o quão antiga pode ser a publicação que seja similar a ela. Seguindo o mesmo valor já utilizado, define-se que é razoável que uma publicação de até 3 dias antes esteja se referindo ao assunto que a notícia apresenta.

O próximo passo é atribuir os valores de A e B, que definem os pesos dados a data de publicação da notícia e a data relativa da publicação no momento do primeiro cálculo da data relativa da notícia. A fim de balancear e dar quase a mesma importância a ambos os atributos, foram atribuídos os valores 3 e 2 para A e B respectivamente. A escolha de valores muito próximos mas não iguais são para dar uma pequena preferência a data da notícia que é um valor certo e um peso menor para a data relativa da publicação que pode conter erros, uma vez que a publicação associada pode não ser a ideal. Para o caso da notícia não ter uma publicação associada, aplica-se uma penalização P de dois dias.

Para a bonificação de notícias com temas mais freqüentes, foi escolhido que W contribuiria com um valor de doze horas.

7.3. O que foi Proposto mas não é Validado

Alguns dos pontos citados na metodologia não foram utilizados para a validação dos resultados. Como não houve um trabalho de análise nas fontes das notícias e perfis das mídias sociais, foi atribuído o mesmo peso a todos estes. A melhora no ranqueamento por meio de cliques também não pôde ser validado devido ao tipo de validação proposta.

7.4. Validação e Resultados

Para validar a qualidade dos resultados obtidos no ranqueamento de notícias, utilizou-se o caso de uso do site Noticias.Gov para demonstração com os valores anteriormente descritos. Para esta analise, é feita a comparação na qualidade dos resultados apresentados na primeira página (que contém dez notícias) do portal de notícias.

A validação consiste na comparação do resultado ideal, com o resultado encontrado pelo algoritmo proposto e o resultado que atualmente existe no Noticias.Gov (ordenação por data de publicação). A forma escolhida para se determinar o que seria o resultado ideal foi: escolhidos alguns links dos principais sites de publicação de notícias (Tabela 2), se em algum destes aparecer uma determinada notícia que está entre as notícias divulgadas pelos Noticia.Gov, então esta notícia deveria estar na primeira página do site, por entender que estes principais sites divulgadores de notícias escolhem manualmente as notícias que merecem destaque e que terão uma maior quantidade de leitores.

Tabela 2. Links utilizados para realizar a validação do algoritmo proposto

Links do site da Globo		
http://g1.globo.com/	http://g1.globo.com/carros/	http://g1.globo.com/ciencia-e-saude/
http://g1.globo.com/concursos-e-emprego/	http://g1.globo.com/turismo-e-viagem/	http://g1.globo.com/economia/
http://g1.globo.com/educacao/	http://g1.globo.com/mundo/	http://g1.globo.com/musica/
http://globoesporte.globo.com/	http://g1.globo.com/natureza/	http://g1.globo.com/politica/
http://g1.globo.com/pop-arte/	http://g1.globo.com/tecnologia/	
Links do site do Terra		
http://www.terra.com.br/portal/	http://noticias.terra.com.br/	
Links do site da UOL		
http://www.uol.com.br/	http://noticias.uol.com.br/cotidiano/	http://noticias.uol.com.br/jornais/
http://noticias.uol.com.br/tabloide/	http://noticias.uol.com.br/	http://noticias.uol.com.br/internacional/
http://noticias.uol.com.br/politica/	http://tecnologia.uol.com.br/	http://noticias.uol.com.br/ciencia/
http://economia.uol.com.br/	http://noticias.uol.com.br/saude/	
Links do site do R7		
http://www.r7.com/	http://noticias.r7.com/distrito-federal	http://noticias.r7.com/empregos/
http://noticias.r7.com/tecnologia-e-ciencia/	http://noticias.r7.com/brasil	http://noticias.r7.com/economia
http://noticias.r7.com/internacional	http://noticias.r7.com/transito/	http://noticias.r7.com/cidades
http://noticias.r7.com/educacao/	http://noticias.r7.com/saude	http://www.triangulo.r7.com/capa/
Outros Links utilizados		
http://www.sbt.com.br/jornalismo/	http://www.hojeemdia.com.br/	http://www.correiodopovo.com.br/

O primeiro resultado encontrado (Tabela 3) faz um comparativo com o resultado encontrado pelo algoritmo proposto e o resultado que a ordenação atual (por data de publicação) oferece. Na primeira coluna de resultados, é apresentada a quantidade de notícias que obrigatoriamente deveriam aparecer entre as 10 notícias mostradas na primeira página que de fato apareceram (utilizando o algoritmo proposto); na coluna seguinte, é apresentada a quantidade de notícias que deveriam aparecer mas não

apareceram (utilizando o algoritmo proposto); e na última coluna, é apresentado a quantidade de notícias que deveriam aparecer e apareceram no caso da ordenação por data de publicação da notícia. Vale ressaltar que o resultado ideal de número de notícias é dado pela soma da primeira e segunda colunas de resultados.

Tabela 3. Resultado da validação do algoritmo proposto com relação à quantidade de notícias que deveriam aparecer e de fato apareceram

Dia da validação (realizada entre 23h e 00h)	Quantidade de notícias encontradas em cada caso		
	Notícias encontradas (algoritmo proposto)	*Notícias faltantes (algoritmo proposto)*	*Notícias encontradas (ordenação por data de publicação)*
08/12/2013	2	0	1
09/12/2013	1	0	0
10/12/2013	1	0	0
11/12/2013	3	2	1
12/12/2013	3	0	1
13/12/2013	1	0	0
Total	11	2	3

Através de cálculos chega-se ao resultado de 84% (11/13) de acerto no aparecimento das notícias pelo algoritmo proposto, além de um aumento em 61% (11/13 - 3/13) em relação à ordenação por data de publicação.

Uma segunda análise do resultado é quanto a sua recenticidade (Tabela 4), ou seja, analisar se as notícias mostradas na primeira página são atuais ou antigas. Para formalizar o teste, como a validação sempre foi feita quase na virada do dia, diz-se que a notícia é atual se aquela notícia é do próprio dia da análise, e diz-se que ela é antiga se a notícia é de algum dia anterior.

Tabela 4. Resultados apresentados com a recenticidade das notícias

Dia da validação	Recenticidade das notícias		
	Notícias Atuais	*Notícias Antigas*	*Valor de Recenticidade*
08/12/2013	7	3	70%
09/12/2013	9	1	90%
10/12/2013	8	2	80%
11/12/2013	9	1	90%
12/12/2013	10	0	100%
13/12/2013	8	2	80%
Total	51	9	85%

Tem-se assim que o resultado do ranqueamento proposto além de priorizar as notícias mais relevantes, também privilegia aquelas que são mais novas.

É importante ressaltar que mesmo identificando poucas das notícias de governo nos principais sites de divulgação de notícias, é razoável pensar que como as notícias em comum que deveriam aparecer na primeira página estão aparecendo, então as demais notícias que estão aparecendo a priori tendem a ser as mais relevantes para os usuários.

8. Conclusão

A tarefa de ordenar as notícias não é simples uma vez que não existe uma resposta certa. Utilizando-se do conhecimento disponibilizado nas mídias sociais, o trabalho propõe um algoritmo que procura reconhecer os assuntos que mais podem atrair os usuários.

Como trabalho futuro, é possível estender o algoritmo para que este também leve em consideração a consulta de um usuário sobre as notícias, além de permitir que juntamente com as notícias, apareçam as publicações (que são fundamentais no algoritmo de ranqueamento). Como possível melhora, o ranqueamento pode ser personalizado para cada usuário, utilizando do gosto pessoal de cada leitor para proporcionar uma melhor disponibilização das notícias.

Referências

Elaysir, A. M. H. and Anbananthen, K. S. M. (2012) "Focused Web Crawler", In: 2012 International Conference on Information and Knowledge Management (ICIKM 2012), Singapore, IPCSIT vol.45 (2012), p. 149-153.

Guo, J., Zhang, P., Jianlong, T. and Guo. L. (2012) "Mining Hot Topics from Twitter Streams", Procedia Computer Science, v. 9, p. 2008-2011.

Kong, L., Jiang, S., Yan, R., Xu, S. and Zhang, Y. (2012) "Ranking news events by influence decay and information fusion for media and users", In: CIKM'12, October 29–November 2, 2012, Maui, HI, USA, p. 1849-1853.

Mohd, M. (2011) "Development of Search Engines using Lucene: An Experience", Procedia-Social and Behavioral Sciences, v. 18, p. 282-286.

Phelan, O., McCarthy, K. and Smyth, B. (2009) "Using Twitter to recommend real-time topical news", In: Proceedings of the Third ACM Conference on Recommender Systems, New York, USA, p. 385-388.

Silva, T. S da, Chaves, M., Bretas, G., Peng, R., Rodrigues, S. A., Silva, R. T. and Souza, J. M. de (2011) "Um Ambiente Integrador de Notícias de Governo", In: VII Simpósio Brasileiro de Sistemas de Informação, Salvador, Bahia, Brasil, p. 373-383.

Sullivan, D. (2009) "Under The Hood: Google News & Ranking Stories". Disponível em: <http://searchengineland.com/google-news-ranking-stories-30424>. Acesso em: 01 fev. 2014.

Xiaofeng, L., Chuanbo and C. and Yunsheng, L. (2007) "Algorithm for Ranking News", In: Third International Conference on Semantics, Knowledge and Grid (2007), Shan Xi, p. 314-317.

Zhang, W., Yoshida, T. and Tang X. (2011) "A comparative study of TF*IDF, LSI and multi-words for text classification", Expert Systems with Applications: An International Journal, v. 38, n. 3, p. 2758-2765.