Um Ambiente Integrador de Notícias de Governo

Tiago Santos Silva[1], Miriam Chaves [2], Giogonda Bretas[3], Ricardo Peng[3], Sergio Assis Rodrigues[1], Ricardo T. Silva[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 do Governo (SLTI/MP) – Esplanada dos Ministérios – Brasília/DF – Brasil

[3] Secretária de Comunicação Social da Presidência da República (SECOM-PR) – Esplanada dos Ministérios – Brasília/DF – Brasil

tiagoss@cos.ufrj.br, miriam.chaves@planejamento.gov.br, gioconda.bretas@planejamento.gov.br, ricardo.peng@planalto.gov.br, {sergio, rick, jano}@cos.ufrj.br

Abstract. *In the democratic state of Brazil, the growing desire for information transparency exposes Government and Society in complementary roles. Society is anxious for data increasingly free and democratic, on the other hand, Government has to ensure not only the transparency, but also consistency and reliability of information provided. In this context, this paper aims at presenting a System to Integrate Government News - an environment that allows searching of articles published by government agencies and also provides ways for the Government assess whether their communication actions comply with reporting criteria of appropriateness to public messages..*

Resumo. *No estado democrático brasileiro, o crescente desejo de transparência de informações expõe Governo e Sociedade a papéis complementares. Se por um lado a Sociedade anseia por dados cada vez mais democráticos e livres, o Governo necessita de meios que garantam não somente a transparência, mas também a coerência e idoneidade das informações disponibilizadas. Neste contexto, este artigo apresenta o Sistema Integrador de Notícias de Governo - um ambiente que permite a busca de notícias publicadas em órgãos públicos e ainda provê subsídios para o Governo avaliar se suas ações de comunicação obedeçam a critérios de sobriedade e adequação das mensagens ao público.*

1. Introdução

O Governo Federal, através das assessorias de imprensa dos vários órgãos que o constitui, disponibiliza informações de interesse a sociedade. A veracidade e da linguagem das noticias publicadas pelos órgãos são importantes para assegurar a imagem do governo. Atualmente a Secretaria de Comunicação Social da Presidência da República (SECOM-PR) conta com uma equipe que analisa diariamente o conteúdo destas notícias disponibilizadas ao público. .

Este artigo apresenta o desenvolvimento de um projeto proposto pelo Ministério do Planejamento *(MP),* juntamente com a Universidade Federal do Rio de Janeiro *(COPPE/UFRJ),* a *(SECOM) e o SERPRO,* para a informatização do ambiente de análise de notícias da SECOM.

O projeto contempla o desenvolvimento de dois sistemas, o Integrador de Notícias do Governo e o Portal de Notícias. O primeiro tem por objetivo prover uma interface para que outros sistemas tenham condições de acessar facilmente as notícias do governo em um formato estruturado. O segundo é um ambiente automatizado, desenvolvido em J2EE fazendo uso do MDA *(Model Driven Architecture)* [Stephen, Kendall, Axel, and Dirk 2002]**,** para a análise de notícias do governo destinado à SECOM que é rotineiramente atualizado por meio dos serviços disponíveis pelo Integrador de Notícias do Governo.

Este artigo apresenta na seção 2 a motivação para o desenvolvimento do Integrador de Notícias do Governo e do Portal de Notícias. Na seção 3 é discutida a arquitetura utilizada na concepção do Integrador de Notícias e como o sistema pode ser integrado com outros sistemas. Na seção 4 é realizada uma visão geral sobre a implementação do Integrador de Notícias do Governo e na seção 5 o Portal de Notícias é apresentado.

2. A SECOM e a Análise de Notícias do Governo

A Secom [SECOM, 2011] é responsável pela comunicação do Governo Federal, coordenando um sistema que interliga as assessorias dos ministérios, das empresas públicas e das demais entidades do Poder Executivo Federal. Ela atua para que as ações de comunicação obedeçam a critérios de sobriedade e transparência, eficiência e racionalidade na aplicação dos recursos, além de supervisionar a adequação das mensagens aos públicos. Também observa o respeito à diversidade étnica nacional e à regionalização no material de divulgação, avaliando os resultados.

Portanto é de responsabilidade da SECOM, analisar as notícias disponibilizadas ao público, pelo Governo Federal através das assessorias de imprensa dos vários órgãos que o constituem. Por exemplo, a análise permite localizar notícias que possam apresentar evidências de discriminação racial, religiosa, etc. O que pode afetar a imagem do governo.

Atualmente, a localização das notícias para a análise é realizada por analistas de forma manual. Um analista com posse de uma lista de sites acessa um por um procurando por notícias recentes. Esse processo, por ser manual, gera um custo de tempo elevado. Esse elevado custo de tempo, por sua vez, impõe um limite ao número de sítios que podem ser analisados pela SECOM, reduzindo a eficácia da análise geral.

Com o objetivo de reduzir os custos de tempo e ampliar a capacidade de análise de notícias o Ministério do Planejamento (MP) propôs um sistema que automatize a localização destas notícias e as disponibilize em um ambiente próprio para a SECOM, em que seja possível não só visualizar como também registrar informações adicionais a notícia.

Entretanto, essas notícias encontram-se descentralizadas e desestruturadas dificultando a sua recuperação automática. Por exemplo, no ano de 2009 estudos

realizados pelo *Projeto Censo Web .br* [CGI.br e NIC.br 2010] identificaram um total de 11.856 sítios sob o domínio .gov.br, sendo visitadas um total de 6.331.256 páginas no formato HTML (Figura1). Todas as notícias publicadas por entidades do governo estão espalhadas por esta vasta quantidade de sítios e páginas da web.

NÚMERO DE PÁGINAS HTML E SÍTIOS DA *WEB* - .GOV.BR		
NÚMERO DE SÍTIOS *WEB*	NÚMERO TOTAL DE PÁGINAS HTML DA *WEB*	NÚMERO MÉDIO DE PÁGINAS HTML POR SÍTIO
11.856	6.331.256	534,01

Figura 1. Número de Páginas HTML e Sítios da Web no domínio .GOV.BR [CGI.br e NIC.br 2010]

Uma possível maneira de estruturar as notícias dos sítios do governo é o uso de Feeds/RSS [Ben Hammersley 2005] contendo as notícias mais recentes. Entretanto, apesar de seu uso ter sido amplamente adotado na web, ainda não é utilizado pela maioria dos sítios do governo. Além disso, a grande maioria dos sítios do governo que apresentam RSS disponibiliza apenas o título, ou o título e um resumo da notícia e não o texto na íntegra.

Então, para solucionar a deficiência de estruturação e descentralização das notícias nos sítios do governo, o Ministério do Planejamento *(*SLTI/MP*)* juntamente com a Universidade Federal do Rio de Janeiro *(COPPE/UFRJ),* a Secretária de Comunicação Social da Presidência da República *(SECOM/PR) e o SERPRO* vislumbraram a idéia de conceber um Integrador de Notícias do Governo.

A contribuição do Integrador de Notícias do Governo para a comunidade, seja ela científica ou corporativa, é a de prover uma interface para que outros sistemas tenham condições de acessar facilmente as notícias do governo em um formato estruturado, sem ter que recorrer ao uso de mineradores de páginas, ao uso de crawlers para varrer a web e etc. Por exemplo, sistemas de análise de sensibilidade de notícias [Mostafa, Helmut e Mitsuru 2010] e sistemas de pesquisa sobre recuperação de informação podem ter acesso a uma base de notícias constantemente atualizada e de fácil leitura.

3. Integrador de Notícias do Governo

O objetivo do Integrador de Notícias do Governo é manter uma base centralizada de notícias do governo que possa ser facilmente acessada por outros sistemas. Ele atua como um Portal Web em que os leitores em vez de humanos são programas que consomem os metadados e as fontes são páginas da web.

Para isso o sistema deve localizar de forma automatizada as páginas de notícias publicadas em um determinado grupo de sítios do governo, extrair essas notícias, estruturá-las em metadados e disponibilizar mecanismos de recuperação para que sistemas possam consultar estes metadados. A (Figura 2) apresenta de forma resumida a arquitetura utilizada para a extração e disseminação das notícias.

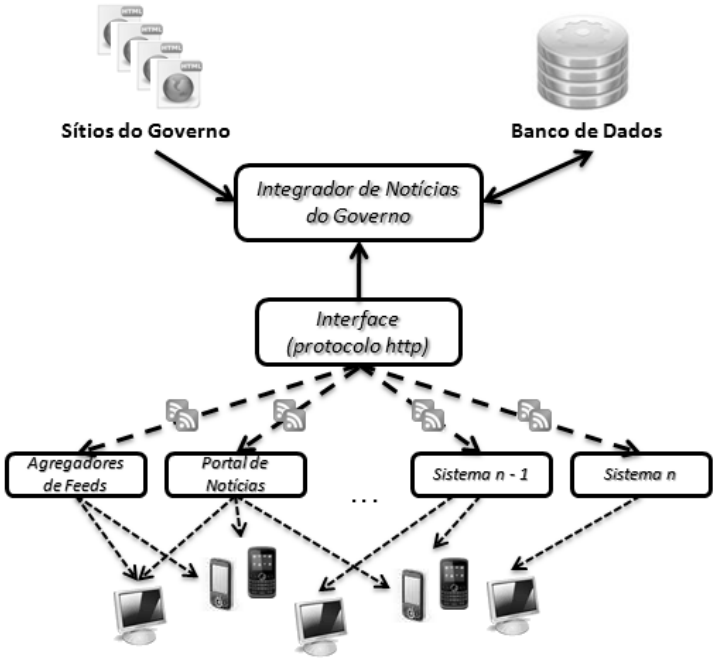

Figura 2. Arquitetura do Integrador de Notícias

Com o intuito de simplificar a comunicação entre os sistemas foi utilizado o protocolo HTTP para a troca de mensagens e o formato RSS para a representação das notícias. O formato RSS foi escolhido porque ele já tem sido largamente utilizado para o compartilhamento de notícias.

O uso do formato RSS apresenta algumas vantagens, como por exemplo, comunicação direta entre os agregadores de Feeds e o Integrador de Notícias, disponibilização de links com Feeds atualizados automaticamente para sítios do governo que ainda não possuem seus próprios Feeds e compatibilidade com os diversos sistemas e técnicas apresentados na literatura científica que usam como base Feeds, tais como os publicados em [Maria e Yiu-Kai 2008], [Mike, Rudy 2007] *e* [Mostafa, Helmut, e Mitsuru 2010].

A comunicação entre um sistema qualquer e o Integrador de Feeds é realizada por uma requisição GET ou POST. A reposta é um arquivo no formato RSS contendo as notícias que satisfazem os parâmetros escolhidos na requisição. O resultado apresenta o título, a descrição, o link original, a data de publicação e o órgão publicador das notícias. Os parâmetros para efetuar a requisição são:

1. **fonte:** Este parâmetro filtra as notícias por órgão. Uma lista de id_s deve ser repassada.
2. **busca:** Este parâmetro filtra as notícias como um campo de busca. O formato adotado é o utilizado pelo framework Lucene [Lucene]. Este framework foi utilizado para a implementação desta funcionalidade.
3. **região:** Este parâmetro filtra as notícias por região. Uma lista de id_s deve ser repassada. Esta funcionalidade ainda é básica, uma versão aprimorada está em desenvolvimento.

4. **step:** Este parâmetro define o número de notícias a serem recuperadas.
5. **page:** Este parâmetro é utilizado para paginação das notícias.

As informações sobre os órgãos e regiões necessários para a consulta estarão disponíveis em um arquivo RDF. Informações adicionais sobre a disponibilização deste serviço podem ser encontradas em [I3gov Planejamento] no link notícias.

4. Implementação do Integrador de Notícias do Governo

Este sistema é composto por dois módulos principais, o módulo responsável pela coleta e monitoramentos das páginas web do governo e o módulo responsável pela seleção, processamento e conversão da página web em uma representação estruturada.

O sistema monitora periodicamente as páginas publicadas nos sítios do governo e realiza extração de padrões a fim de classificar uma dada página como notícia ou não, em caso de ser classificada como notícia a página é submetida a um processamento com o objetivo de extrair de forma automatizada os metadados da noticia descoberta. Atualmente, o sistema monitora 62 sítios do governo. Entre estes sítios estão os ministérios, as secretarias e os conselhos do Governo Federal e alguns outros órgãos de interesse da SECOM.

4.1. Módulo Coletor

Atualmente a quantidade de informação disponível na web vem crescendo rapidamente. Por isso uma varredura completa pelas páginas web a fim de atualizar uma determinada base não é uma tarefa trivial e dependendo da taxa de atualização necessária da base, pode ser tornar inviável. O problema se agrava ainda mais quando o objetivo é coletar notícias, já que a publicação de novas notícias é realizada rapidamente. Em um intervalo de duas horas uma base contendo notícias previamente cadastradas da internet pode se tornar desatualizada.

Então, nesse trabalho houve a necessidade de desenvolver um módulo capaz de gerenciar a base de notícias a atualizando constantemente. Desta forma, foi desenvolvido o Módulo Coletor que é um Web Crawler desenvolvido especialmente para navegar pelos sítios do governo a procura de páginas com potencial de terem como conteúdo notícias.

Para reduzir o espaço de busca foram utilizadas heurísticas baseadas no comportamento humano quando este procura por notícias em sítios da web. As heurísticas adotadas consideram a localidade das páginas web nos seus respectivos sítios e a existência de termos demarcadores que indicam a possível existência de notícias na página. Estas heurísticas são baseadas em três hipóteses que são descritas abaixo:

1. *As notícias recentes de um sítio web estão nas vizinhanças da página principal.*
2. *As páginas contendo as notícia propriamente dita ou uma lista delas possuem textos que o identificam como notícias.*
3. *O sistema se atualizará periodicamente.*

Sítios da web que seguem as boas práticas na elaboração da navegação de suas páginas estão enquadrados na hipótese número 1. Com essa restrição imposta pela hipótese número 1 uma redução considerável do espaço de busca é realizada.

A hipótese número 2 baseia-se no fato de que humanos precisam identificar dentro dos sítios os espaços dedicados a notícias e para que isso seja possível o próprio sítio deve fornecer evidências disso, por exemplo, links com texto como "notícias", "destaque", etc. A hipótese número 2 é utilizada com o objetivo de pontuar determinada página [Eytan, Jaime, Susan, e Jonathan. 2009] como mais provável ou não a ser uma notícia. A fim de evitar falsos negativos essa hipótese é utilizada para indicar a direção da busca e não para a redução do espaço de busca a não ser que o tempo não seja suficiente para visitar todo o espaço.

A hipótese número 3 é utilizada apenas para fortalecer a hipótese número 1, pois considerando que o Coletor esteja periodicamente rodando apenas as notícias recentes são o foco da busca.

4.2. Módulo de Processamento

Foi visto em seções anteriores, que a maior parte das notícias publicadas nos sítios do governo está armazenada de forma desestruturada e mesmo alguns sítios que possuem RSS disponibilizam apenas títulos ou resumos e não o texto da notícia na íntegra.

As notícias disponíveis na web são constantemente visitadas por milhões de pessoas todos os dias. Cada pessoa que tem acesso a web e busca por novas notícias empenha de forma natural um conjunto de ações necessárias para obter acesso a elas. Por exemplo, um indivíduo ao acessar a página principal de um site, busca de forma intuitiva a área de notícias e após identificar essa área, procura por uma notícia de interesse, e então o indivíduo naturalmente consegue discernir o que é o título da notícia, o corpo da notícia e até mesmo a sua data de publicação.

De alguma forma o conhecimento necessário para discernir as partes integrantes da notícia encontra-se de forma tácita no indivíduo, então para que seja possível automatizar esse processo foi realizado um estudo a fim de documentar o processo intuitivo utilizado pelas pessoas na análise de uma notícia. A partir de um conjunto de premissa, extraídas do estudo acima, foi desenvolvido o Módulo de Processamento.

O Módulo de Processamento é responsável por ler uma página HTML e extrair a notícia desta e então fragmentá-la em metadados a fim de estruturá-la. Para realizar a atualização da base de dados o Módulo de Processamento e o Módulo Coletor são integrados. O Módulo Coletor transmite para o Módulo de Processamento as páginas encontradas, em seguida este módulo analisa cada página e armazena os metadados na base de dados caso a página seja considerada uma notícia.

O fluxo da (Figura 3) adotado para o Módulo de Processamento é dividido em quatro etapas, a primeira consiste em realizar um pré-processamento da página, a segunda consiste em filtrar as páginas relevantes para o sistema, a terceira consiste no processamento da notícia e a quarta e última consiste em atualizar a base de dados com as informações extraídas.

A etapa de pré-processamento consiste em fazer uma raspagem do código HTML e uma preparação do texto para a sua posterior mineração. Os métodos

implementados para a raspagem visam remover do código HTML áreas como menus, botões, anúncios e etc. Os métodos adotados para a raspagem foram adaptações de técnicas existentes na literatura como as encontradas em [Suhit, Gail, Peter, Michael e Justin 2005], [Lan, Bing e Xiaoli 2003] e [Lakshmish, Arun, Ling, e Fred 2003]. Para a preparação do texto foi utilizada a biblioteca HtmlUnit [HtmlUnit]. Uma das funcionalidades desta biblioteca é a de disponibilizar a partir de uma página web o conteúdo visível pelo usuário em um browser. O resultado da etapa de pré-processamento é o texto visível da página e a parte do código HTML resultante da eliminação dos ruídos de informação.

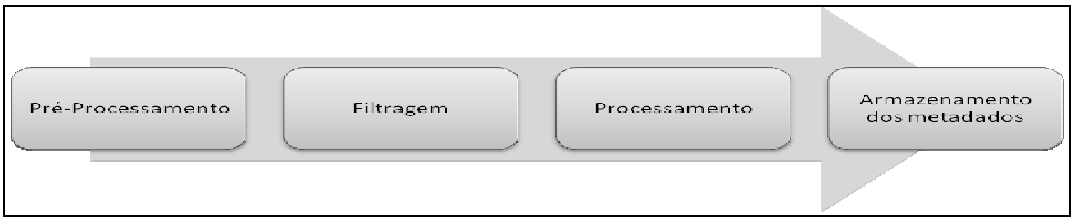

Figura 3. Fluxo do Módulo de Processamento

O Módulo Coletor consegue filtrar apenas partes das páginas que não são notícias a fim de diminuir o fluxo de páginas, não relevantes, que vão para o Módulo de Processamento. As demais páginas não relevantes são filtradas pelo Módulo de Processamento já que este dispõe de mecanismos de mineração da página mais robustos do que o Módulo Coletor.

Na etapa de filtragem das páginas é realizada uma análise sobre a estrutura do texto retornado pelo pré-processamento. Esse texto é submetido a uma série de verificações que checam se o texto atende as propriedades mínimas encontradas em textos de notícias, o qual algumas foram inspiradas em [Kjetil e Randi 2005]. Tais propriedades refletem a estrutura posicional do texto e as evidências léxicas de termos relacionados à área de notícias. Por exemplo, a densidade dos parágrafos, o número de parágrafos, presença de datas recentes, posição absoluta do texto na página e etc. As propriedades padrões para a checagem das páginas foram ajustadas após vários testes realizados sobre uma base de notícia preparada para este trabalho.

A etapa de filtragem retém um número considerável de páginas não relevantes para o sistema, entretanto algumas páginas não relevantes só poderão ser descartadas na etapa de processamento quando uma mineração mais apurada será realizada.

A etapa de processamento é responsável por selecionar a notícia dentro da página HTML e extrair os metadados relevantes para a representação da notícia em um formato RSS. Os metadados armazenados da notícia são o título, o corpo do texto, a data de publicação, a data de visitação da página pelo coletor, o órgão fonte e o link de publicação.

Para extrair a notícia e a separar em suas partes integrantes algumas heurísticas baseadas na estrutura posicional do texto são utilizadas. Segue abaixo as principais hipóteses que servem como base para as heurísticas utilizadas:

1. *Uma notícia é minimamente composta por um título, descrição e data de publicação.*

2. *O título de uma notícia está localizado próximo a descrição e sempre acima dela.*
3. *A data de publicação de uma notícia está localizada próximo ao título ou após a descrição.*
4. *A data de publicação de uma notícia é uma data dentro de um intervalo considerado aceitável para o escopo da busca.*
5. *As propriedades HTML da descrição da notícia são as mesmas para todas as palavras da notícia ou pelo menos para a maior parte delas. O mesmo é válido para o título e a data de publicação.*
6. *Um título e a descrição possuem tamanhos característicos, tendo um tamanho e densidade de palavras mínimas e máximas aceitáveis.*
7. *Sejam três parágrafos A, B e C dispostos em sequência. Se os parágrafos A e C pertencem à descrição da notícia então o parágrafo B também faz parte da descrição ou é uma legenda de uma imagem ou tabela.*
8. *Elementos como descrição, título e data de publicação tendem a ter propriedades semelhantes na estrutura de tags do código HTML bem como características de fontes semelhantes.*
9. *O uso de nomeações de variáveis, de classes e id_s nos códigos fontes das páginas HTML podem ser indícios de título, descrições e datas de publicação. Por exemplo, " A educação no Brasil ... "*
10. *A descrição da notícia é a área com maior densidade de palavras e linhas da página.*

Apesar de algumas das hipóteses serem consideradas como óbvias para seres humanos elas não são para o computador e por isso precisam ser descritas e implementadas. A (Figura 4) exemplifica de forma simplificada o uso das heurísticas para identificação automática dos metadados.

5. Portal de Notícias do Governo

O Portal de Notícias do Governo é uma iniciativa que envolve a cooperação da SLTI/MP, SECOM/PR, COPPE/UFRJ e SERPRO. O seu objetivo é atender as necessidades da SECOM mencionadas no capítulo 2 e prover ao público um ambiente intuitivo e robusto para a pesquisa de notícias publicadas nos sítios do governo. Este Portal foi desenvolvido em J2EE (Java 2 Enterprise Edition) fazendo uso do Padrão MDA (Model Driven Architecture) através do Framework do Ministério da Defesa e Ministério do Planejamento chamado MDArte [Roque, Filipe, Rodrigo e Geraldo 2011].

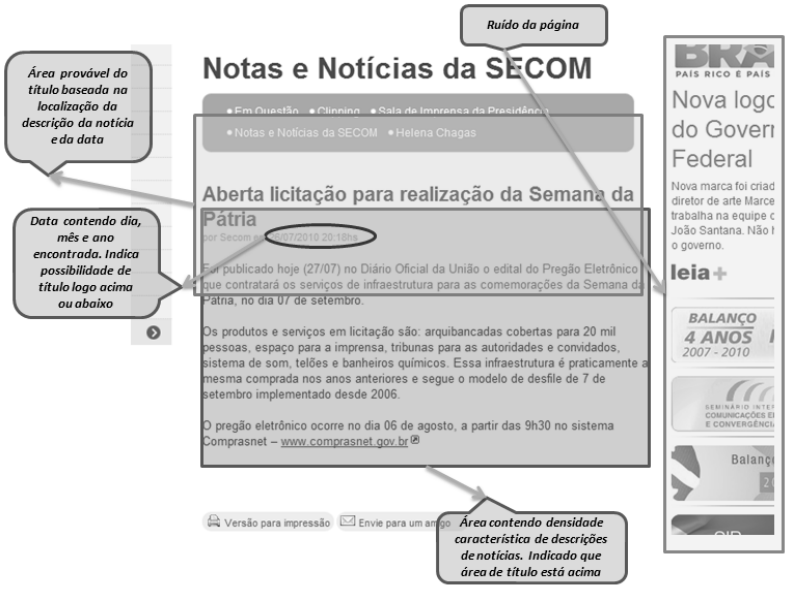

Figura 4. Exemplo simplificado do uso das heurísticas para a localização dos metadados

O Portal de Notícias do Governo é dividido em duas áreas, a pública e a restrita. A área pública (Figura 5) funciona como uma interface gráfica do Integrador de Notícias disponível a população. Por meio dessa interface usuários podem interagir com os dados obtidos pelo Integrador de Notícias do Governo. Por exemplo, usuários podem realizar buscas avançadas criando filtros por órgãos, regiões, datas e digitando consultas no campo de busca, os usuários também podem ter acesso a links contendo os Feeds de suas pesquisas. Os usuários podem adicionar estes Feeds aos agregadores de Feeds instalados em seus computadores, celulares, smartphones e ipads.

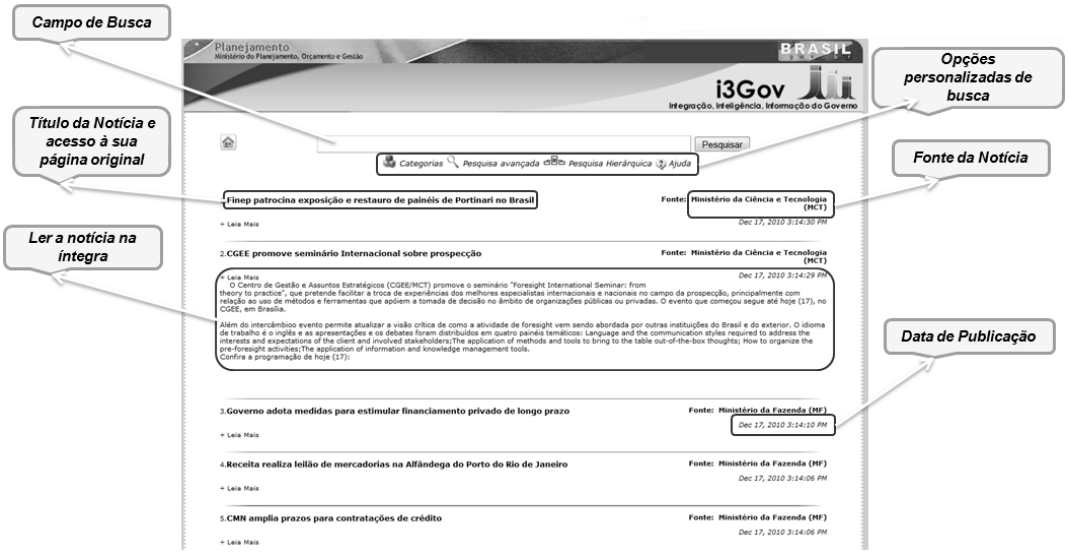

Figura 5. Área pública do Portal de Notícias

A área restrita do Portal é destinada a análise de notícias realizadas pela SECOM. A integração do Portal de Notícias com o Integrador de Notícias permite à SECOM manter um ambiente atualizado automaticamente com as notícias do governo. Isso permite aos analistas visualizarem as notícias publicadas recentemente sem ter que procurá-las pela Web. Além da exibição das notícias a área restrita do Portal oferece os mecanismos de busca disponíveis na versão pública, a geração de relatórios de interesse da SECOM e o armazenamento de dados vinculados as notícias conforme informado pelos analistas.

Conclusão

Este artigo apresentou detalhes sobre a informatização do ambiente utilizado pela SECOM. Os sistemas apresentados foram o Integrador de Notícias do Governo e o Portal de Notícias do Governo. A maior ênfase foi dada ao Integrador de Notícias do Governo já que este é o que mais contribui à comunidade científica e corporativa.

Como explanado nas seções anteriores, o Integrador de Notícias pode ser utilizado para que sistemas em geral, tenham acesso às notícias publicadas pelo governo de forma facilitada, mesmo que essas notícias não estejam originalmente armazenadas em Feeds ou em qualquer outro formato estruturado, o único requisito é que elas estejam disponíveis na web em formato HTML.

Como trabalho futuro, podemos citar a ampliação da quantidade de sites monitorados, a fim de cobrir o máximo de sítios possíveis sem perder a capacidade de rápida atualização. Outra prioridade é o estudo relacionado a inserção de ontologias para aprimorar a busca por região.

Referências

Stephen J. Mellor, Kendall Scott, Axel Uhl, e Dirk Weise. 2002. Model-Driven Architecture. In *Proceedings of the Workshops on Advances in Object-Oriented Information Systems* (OOIS '02), Jean-Michel Bruel e Zohra Bellahsene (Ed.). Springer-Verlag, London, UK, UK, 290-297.

SECOM. Secretária de Comunicação Social da Presidência da República: Secretária. Disponível em: <http://www.secom.gov.br/sobre-a-secom/a-secretaria>. Acesso em: 05 de fevereiro 2011.

CGI.br e NIC.br. Dimensões e características da Web brasileira: um estudo do .gov.br disponível em: <http://www.cgi.br/publicacoes/pesquisas/govbr/cgibr-nicbr-censoweb-govbr-2010.pdf>. Acesso em: 05 de fevereiro 2011.

Ben Hammersley. 2005. Developing Feeds with Rss and Atom (First ed.). O'Reilly.

Mostafa Al Masum Shaikh, Helmut Prendinger, and Mitsuru Ishizuka. 2010. Emotion Sensitive News Agent (ESNA): A system for user centric emotion sensing from the news. Web Intelli. and Agent Sys. 8, 4 (December 2010), 377-396.

Maria Soledad Pera and Yiu-Kai Ng. 2008. Utilizing phrase-similarity measures for detecting and clustering informative RSS news articles. *Integr. Comput.-Aided Eng.* 15, 4 (Dezembro 2008), 331-350.

Mike Thelwall , Rudy Prabowo, Identifying and characterizing public science-related fears from RSS feeds: Research Articles, Journal of the American Society for Information Science and Technology, v.58 n.3, p.379-390, February 2007.

Mostafa Al Masum Shaikh, Helmut Prendinger, and Mitsuru Ishizuka. 2010. Emotion Sensitive News Agent (ESNA): A system for user centric emotion sensing from the news. *Web Intelli. and Agent Sys.* 8, 4 (December 2010), 377-396.

Lucene. Apache Lucene - Overview: um estudo do .gov.br disponível em: < http://lucene.apache.org/>. Acesso em: 05 de fevereiro 2011.

I3gov Planejamento: disponível em < https://i3gov.planejamento.gov.br/ >. Acesso em: 05 de fevereiro 2011.

Suhit Gupta, Gail E. Kaiser, Peter Grimm, Michael F. Chiang, and Justin Starren. 2005. Automating Content Extraction of HTML Documents. *World Wide Web* 8, 2 (June 2005), 179-224.

Lan Yi, Bing Liu, and Xiaoli Li. 2003. Eliminating noisy information in Web pages for data mining. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '03). ACM, New York, NY, USA, 296-305.

Lakshmish Ramaswamy, Arun Iyengar, Ling Liu, and Fred Douglis. 2003. Techniques for efficient fragment detection in web pages. In *Proceedings of the twelfth international conference on Information and knowledge management* (CIKM '03). ACM, New York, NY, USA, 516-519.

HtmlUnit. Welcome to HtmlUnit. Disponível em: <http://htmlunit.sourceforge. net/>. Acesso em: 05 de fevereiro 2011.

Kjetil Norvag and Randi Oyri. 2005. News Item Extraction for Text Mining inWeb Newspapers. In *Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration* (WIRI '05). IEEE Computer Society, Washington, DC, USA, 195-204.

Portal do Software Público Brasileiro. MDArte: disponível em <http://www.softwarepublico.gov.br/ ver-comunidade?community_id=9022831>. Acesso em: 05 de fevereiro 2011.

Roque Elias Pinel, Filipe Braida do Carmo, Rodrigo Salvador Monteiro, e Geraldo Zimbrão. 2011. Improving tests infrastructure through a model-based approach. *SIGSOFT Softw. Eng. Notes* 36, 1 (January 2011), 1-5.

Eytan Adar, Jaime Teevan, Susan T. Dumais, and Jonathan L. Elsas. 2009. The web changes everything: understanding the dynamics of web content. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (WSDM '09), Ricardo Baeza-Yates, Paolo Boldi, Berthier Ribeiro-Neto, and B. Barla Cambazoglu (Eds.). ACM, New York, NY, USA, 282-291.