

RSS no desenvolvimento de uma Central de Notícias

Darley Passarin¹, Parcilene Fernandes de Brito¹

¹Sistemas de Informação – Centro Universitário Luterano de Palmas (CEULP/ULBRA)
Palmas – TO – Brasil

darley@centralrss.com.br, pfb@ulbra-to.br

Resumo. *A tecnologia RSS possibilitou a distribuição em massa de conteúdos estruturados através da web, desta forma surgiu a relevância de se construir mecanismos que possibilitassem a junção dessas informações em um só local. A partir disso, aplicações podem ser desenvolvidas de forma a permitir o tratamento e a formulação de mecanismos de consultas funcionais, agregadas a interfaces eficientes de interação com o usuário. Esta é a premissa básica para o desenvolvimento deste trabalho, cujo objetivo é apresentar os estudos realizados para propiciar a utilização das fontes RSS disponíveis em alguns sites e, assim, desenvolver uma Central de Notícias (CN).*

1. Introdução

Tendo em vista a grande quantidade e diversidade de *sites* que trabalham com a divulgação de notícias, navegar de endereço a endereço é uma rotina comum e muitas vezes cansativa e nem sempre recompensada, no sentido de encontrar a notícia que realmente interessa. A partir disso, surgiu a idéia de se criar um padrão de divulgação de informações que atenuasse a busca, considerando, para isso, a organização da notícia e a forma de interação entre os *sites*. Este padrão foi criado e denominado RSS (*RDF Site Summary*, *Rich Site Summary* ou *Really Simple Syndication*) (HAMMERSLEY, 2003).

Com a possibilidade facilitada pelo RSS da divulgação de informações advindas de vários *sites*, observou-se a relevância do desenvolvimento de um sistema capaz de organizar as notícias por contexto e possibilitar várias formas de consulta. Com isso, tem-se a premissa básica do trabalho realizado nesse projeto, ou seja, a construção de uma Central de Notícias que agregue essas características, desenvolvida a partir da utilização das tecnologias e/ou padrões RSS e PHP.

A organização e obtenção de notícias são fatores decisivos no desenvolvimento da Central de Notícias. A obtenção de conteúdos (notícias) é feita através da leitura das informações contidas em várias fontes RSS de diferentes *sites*, possibilitando assim a criação de uma base de dados com informações sobre vários contextos. Para a organização das notícias foi adotado um esquema de categorização das informações, que possibilitou ao usuário uma busca mais precisa de notícias conforme o seu perfil.

2. RSS

O RSS é uma especificação para distribuição de conteúdo através da utilização da linguagem XML (*eXtensible Markup Language*) (HAMMERSLEY, 2003). Essa especificação surgiu com o intuito de padronizar a forma de distribuição de conteúdos estruturados existentes nos mais diferentes *sites* na internet. A distribuição do conteúdo estruturado, neste contexto, é representada através do termo “*feed*”. Um *feed* pode ser definido como qualquer informação importante disponível em um *site*, como, por exemplo, notícias, artigos, histórias, entre outras informações. Nesse projeto será abordado o trabalho de estruturação de conteúdo realizado através da utilização dos *feeds* de notícias.

Um *feed* é distribuído através de arquivos de extensão .xml, .rss ou .rdf, e referenciado através de uma URL (ex.: <http://static.userland.com/gems/backend/sampleRss.xml>). Sua distribuição também pode ser realizada através de um *script* que tenha embutido em si uma função de retorno a tais documentos (ex.: <http://www.centralrss.com.br/rss.php>). Vale ressaltar que esses documentos devem seguir as especificações definidas no padrão RSS.

A primeira versão do RSS (RSS 0.90) foi projetada por Dan Libby, através da empresa Netscape. Tal versão foi baseada no padrão RDF, cuja estrutura foi considerada inadequada para os usuários finais (HAMMERSLEY, 2003). Assim, surgiu um novo projeto de alteração do RSS de forma que possibilitasse a divulgação das mesmas informações num formato mais simples, baseado em XML, sem a utilização do modelo RDF. Esse projeto foi apresentado na versão RSS 0.91.

A partir da relevância de se adicionar novos elementos ao modelo RSS foram lançadas novas versões. Neste trabalho foi adotada a versão RSS 2.0, que além de ser a última versão desenvolvida, suporta as demais versões.

2.1 Estrutura RSS 2.0

A forma de estruturação dos documentos RSS segue a especificação da XML 1.0 (WINER, 2000), podendo possuir elementos, sub-elementos, atributos etc.. Tal estrutura é composta pelo elemento principal `<rss>` juntamente com o atributo *version*, que representa a versão do mesmo. Neste caso, representado da seguinte forma `<rss version="2.0">`. Pode ser visualizado através da Figura 1 o modelo representativo dos elementos utilizados nesta versão.

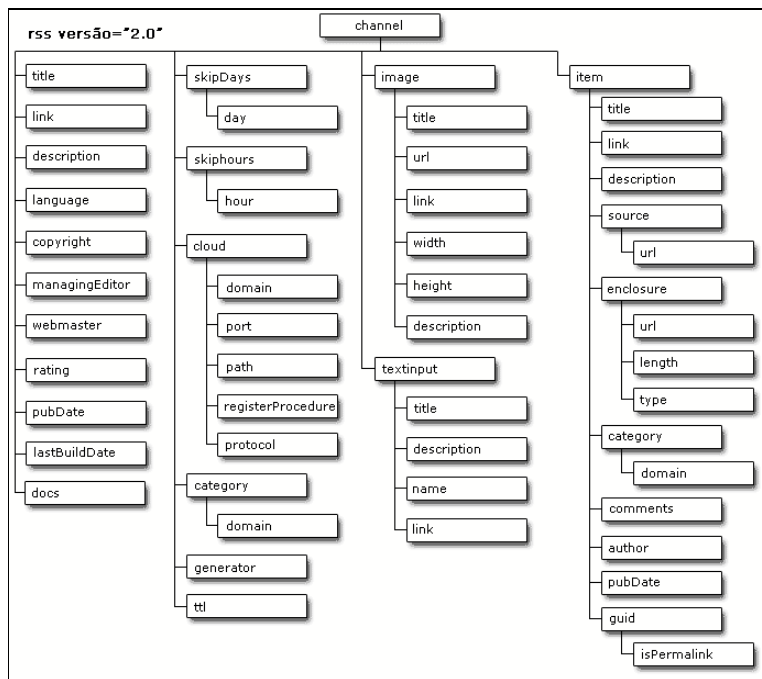


Figura 1 – Modelo representativo do RSS 2.0 (HAMMERSLEY, 2003).

Para cada *feed* RSS, há um sub-elemento `<channel>`, em que estão contidos todos os elementos que representam suas informações. Seus sub-elementos, em sua maioria, representam informações relacionadas ao *feed*, como título, descrição, linguagem e assim por diante.

Um *feed* poderá possuir vários sub-elementos `<item>`. Este é utilizado para representar as informações sobre o conteúdo (ex: notícias) do *feed*. Os seus sub-elementos descrevem informações como título do conteúdo, *link* de acesso e uma breve descrição ou resumo.

Para a estruturação de um documento RSS é necessário seguir as normas definidas pela XML 1.0 (WINER, 2000), pois um documento RSS é um documento XML que segue a especificação definida para o RSS 2.0. Como exemplo, para declarar informações sobre o título de uma determinada notícia é utilizada a expressão `<title>Título da notícia</title>`. Os demais elementos contidos na especificação do RSS 2.0 são declarados seguindo esta forma de estruturação.

Dentre todos os elementos pertinentes ao modelo RSS 2.0, os mais relevantes neste trabalho foram o elemento `<item>` e seus sub-elementos `<title>`, `<link>` e `<description>` por serem os responsáveis em descrever informações sobre as notícias.

3. Estrutura da Aplicação

A aplicação Central de Notícias tem como objetivo a obtenção de conteúdos provenientes de *feeds* RSS de diferentes fontes (*sites*), para o armazenamento em uma base de dados central. A partir disso, o sistema proporciona uma interface funcional ao usuário, possibilitando a ele pesquisas avançadas, definidas a partir de determinados contextos e da própria seqüência de buscas realizadas por ele.

Para melhor representação do funcionamento da aplicação, tem-se a Figura 2, que representa a estrutura, de forma geral, de todo o funcionamento da CN.

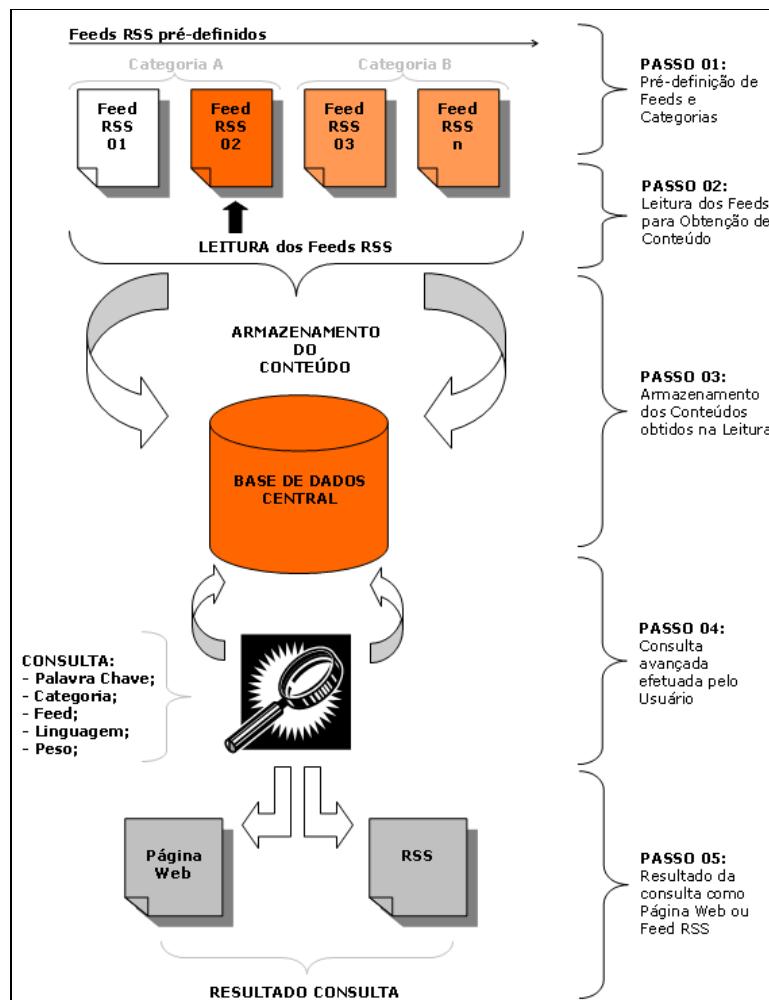


Figura 2 – Estrutura da Aplicação

Conforme a representação da Figura 2, o funcionamento da CN segue os seguintes passos:

1. É realizada a pré-definição dos *feeds* RSS e das suas categorias, ou seja, para que haja a obtenção das notícias dos *feeds* RSS é necessário que as informações

sobre os mesmos já estejam cadastradas, para que o sistema possa identificá-las e processá-las.

2. É obtido o conteúdo (que neste caso são notícias) através da leitura dos *feeds* RSS. Essa leitura é feita automaticamente em um determinado intervalo de tempo pré-definido no sistema, a fim de sempre manter a base de informações atualizadas;
3. É executado o processo de armazenamento das informações obtidas na leitura dos *feeds* RSS. Todas as informações são armazenadas em um banco de dados relacional (MySQL), denominado, neste caso, “Base de Dados Central”;
4. Está relacionado a interface com o usuário. Essa interface proporciona a execução de consultas sobre as informações armazenadas. Essas consultas podem ser refinadas através dos parâmetros informados pelo usuário, como Palavra Chave, Categoria, *feed* RSS e pela linguagem. Os *links* da notícia retornados ao usuário nas consultas são ordenados por Peso. O Peso é definido pelo sistema através da verificação de qual contexto de notícias o usuário mais acessa.
5. É possível observar que são disponibilizadas duas formas de visualização das informações resultantes da consulta efetuada, como uma página web ou através de um documento RSS.

3.1 Pré-definição e Categorização de Feeds RSS

É grande a diversidade e quantidade de notícias em toda a internet, mas grande parte dessas notícias já são sistematizadas em contextos, o que delimita sua abrangência a um dado domínio. Essa definição de assuntos é utilizada frequentemente em diversas mídias, não apenas na internet, para categorizar determinados assuntos, a fim de possibilitar ao usuário a disposição dos assuntos pertencentes a um dado perfil.

Seguindo esta linha, grande parte dos *feeds* RSS utilizados pela CN está sistematizada a partir de assuntos específicos. Através da Figura 3 pode ser visualizado um exemplo dessa sistematização, a partir das categorias pré-definidas, alguns *feeds* RSS retirados do *site* do Estadão (<http://www.estadao.com.br/ext/rss/>) são categorizados conforme o assunto equivalente ou relacionado às categorias do sistema.

Últimas Notícias	XML	3	Tecnologia	XML	9
Internacional	XML	8	Link	XML	9
Economia	XML	4	Especial & Multimídia	XML	6
Cidades	XML	1	Ciência e Meio Ambiente	XML	5
Arte e Lazer	XML	6	Educação	XML	2
Esportes	XML	7	Imóveis	XML	4
Futebol	XML	7	Autos	XML	9
Tênis	XML	7	Turismo	XML	6
Basquete	XML	7	Mídia Digital	XML	9
Automobilismo	XML	7	Convergência Digital	XML	9
Outros Esportes	XML	7	Sociedade Digital	XML	9
Consultor Jurídico	XML	4	Ponto Web	XML	9
Últimas Imagens	XML	3	Mundo sem fios	XML	9

1 - Brasil	6 - Entretenimento
2 - Ciência	7 - Esporte
3 - Diversas	8 - Mundo
4 - Economia	9 - Tecnologia
5 - Educação	

Figura 3 – Exemplo de Categorização.

A pré-definição e a categorização de *feeds* RSS são feitas de forma semi-automática, ou seja, as informações sobre os *feeds* RSS e suas respectivas categorias devem ser pré-definidas no sistema.

3.2 Gerenciamento de documentos RSS

Foram definidas classes responsáveis pelo processo de leitura e criação dos documentos RSS, implementadas utilizando a linguagem de programação PHP 5, juntamente com a interface DOM. Uma classe chamada de “RSS20” é responsável pela leitura das especificações RSS 0.91, 0.92 e 2.0, e pela criação de documentos RSS 2.0, sendo que esta versão foi utilizada como padrão pela CN.

A classe RSS20 possui dois métodos principais: “Ler(URL)” e “Gerar(ENCODING)”. Esses métodos são utilizados para a leitura e criação de documentos RSS, respectivamente. O método “Ler(URL)” recebe como parâmetro a URL do documento RSS e, posteriormente, utilizando métodos do DOM, é feita a leitura do documento e fixados os atributos utilizados pelo RSS 2.0.

Após a leitura do documento e/ou fixação dos valores dos atributos é utilizado o método “Gerar(ENCODING)”, que recebe como parâmetro a forma de codificação do documento (“UTF-8”ou “ISO-8859-1”), para gerar um novo documento RSS a partir dos valores. Esse documento poderá ser salvo em disco através do método “Salvar(CAMINHO)” ou apenas utilizar a *string* XML gerada através da função “getRSS()”. O método “getRSS()” é utilizado na CN para transformar os resultados obtidos nas consultas, em documentos RSS, sem que os mesmos sejam salvos em disco,

mantendo-os apenas na memória, fazendo que ao termino da consulta, essas informações sejam descartadas automaticamente (Ex: Figura 4 – Resultado em RSS).

3.3 Consultas

As consultas na CN têm a finalidade de possibilitar ao usuário uma busca mais precisa sobre as informações desejadas pelo mesmo, adotando para isso o emprego do estudo das interações dos usuários e a utilização de parâmetros importantes para o refinamento dos resultados.

No gerenciamento de Consultas é importante salientar a forma de disposição e ordenação dos resultados obtidos. Para garantir que os resultados fossem retornados de forma eficiente, foram trabalhadas duas formas de ordenação dos registros afetados na consulta, sendo elas a consulta “Geral” e a consulta específica por “Usuário”.

A consulta “Geral” é utilizada quando um usuário não obtém acesso à área interna do sistema (para isso é necessário login e senha), assim a disposição e ordenação dos resultados estão ligadas às interações dos usuários como um todo. As interações são tidas como eventos executados pelo usuário, como *cliques* e consultas, sendo assim contabilizadas através do incremento do *Peso*. O *Peso* é utilizado para definir os registros mais acessados e, conseqüentemente, utilizá-los para ordenar (forma decrescente, os maiores *Pesos* são os primeiros) as consultas. Na Figura 4 pode ser visualizada a forma de ordenação das notícias, que está relacionada a quantidade de acessos (neste caso declarado como *Peso*), seguindo a ordem decrescente em relação aos mesmos. Conforme os parâmetros recebidos, é montada uma SQL (*Structured Query Language*) para a execução da consulta no banco de dados, para que posteriormente seja gerado o resultado da consulta em uma interface para o usuário.



Figura 4 – Forma de Ordenação por Peso

A consulta específica por “Usuário” é utilizada quando o usuário obtém acesso à área interna do sistema, assim a disposição e ordenação dos resultados estão ligadas às suas interações. Para a montagem da SQL, o diferencial, em relação à consulta “Geral”, está no acréscimo das tabelas utilizadas para o controle do usuário e na forma de ordenação dos resultados. Esta forma de ordenação tenta obter uma maior precisão sobre as notícias mais atuais, sobre as notícias que pertencem ao contexto que o usuário mais interagiu e as notícias mais acessadas de modo geral.

Para a realização das consultas são necessários, como parâmetros, as Palavras-chave, o código da Categoria, o código do *Feed*, a linguagem, o código do Usuário, o valor do início da consulta e o valor da quantidade limite de resultados. As palavras-chave são utilizadas para identificar os termos específicos a serem procurados. O código do usuário é utilizado para definir se uma consulta será de modo geral ou específica a um determinado usuário, em relação ao atributo *Peso*. A partir destes parâmetros foi

possível montar uma busca mais eficaz sobre as informações contidas no banco de dados.

A visualização dos resultados das consultas pode ser feita através de uma página *web*, utilizada com padrão, ou através de um documento RSS. Na Figura 5 pode ser visualizado um resultado mostrado através de uma página *web* e através de um documento RSS.



Figura 5 – Telas de resultados de Consultas

4. Conclusões

O projeto Central de Notícias surgiu a partir da idéia de se desenvolver um mecanismo que buscasse notícias em diferentes *sites* de forma automática, para que posteriormente essas notícias pudessem ser disponibilizadas em outras aplicações. Desta forma qualquer *site* poderia utilizar esta funcionalidade para prover notícias sobre a sua área de forma automática e sem que houvesse a interferência do administrador do *site* neste processo.

O RSS foi fundamental para a formalização desta idéia, tornando possível a busca de notícias em sites que utilizam este padrão. Após o entendimento da relevância da utilização do RSS no projeto, foram estudadas formas de prover uma aplicação que possibilitasse a execução de consultas funcionais sobre os dados obtidos das fontes RSS. Assim, houve a necessidade da criação de categorias e da implementação de funcionalidades para o gerenciamento de usuários. Desta forma, o projeto guiou-se a

fim de criar uma ferramenta para uso diário de usuários e *sites* que necessitam manter-se atualizados sobre notícias variadas advindas de fontes diversas.

São previstas algumas melhorias para a Central de Notícias, a fim de torná-la mais eficiente e funcional para o usuário. Dentre estas melhorias estão: a criação de mecanismos para a busca de *feeds* RSS, através da utilização de motores de busca (Google, MSN etc); eliminar a pré-definição de categorias de *feeds* RSS e tornar este processo automático através da construção de um vocabulário de palavras-chave sobre um determinado contexto, obtidas, por exemplo, através da técnica de “Categorização” do “*Text Mining*”; buscar maneiras para melhorar a performance das consultas, adotando técnicas de indexação de conteúdos; desenvolver mecanismos para efetuar consultas mais precisas e eficientes sobre o contexto do usuário.

5. Referências Bibliográficas

- (HAMMERSLEY, 2003) HAMMERSLEY, Ben. **Content Syndication with RSS**. Sebastopol: O’Reilly, 2003.
- (MCGRATH, 1999) MCGRATH, Sean. **XML: Aplicações Práticas**. Rio de Janeiro: Campus, 1999.
- (SOARES, 2004) SOARES, Wallace. **PHP5: Conceitos, Programação e Integração com Banco de Dados**. São Paulo: Érica, 2004.
- (YARGER, 2000) YARGER, Randy Jay. Reese, George. King, Tim. **MySQL & mSQL**. Rio de Janeiro: Editora Ciência Moderna Ltda, 2000.
- (WINER, 2000) WINER, Dave. **RSS 0.91**. Junho de 2000. Disponível em: <<http://backend.userland.com/rss091>>. Acesso em: 25 de maio de 2005.