UMA FERRAMENTA PARA RECUPERAÇÃO DE TAGS DE BLOGS BASEADA EM MICROFORMATOS

A TOOL TO RECOVER TAGS OF BLOGS BASED ON MICROFORMATS

Célio Andrade Santana Júnior
Nilton Heck Santos
Steffane Ramires de Lima
Amanda Maria de Almeida Nunes

Resumo: O objetivo deste trabalho é apresentar a ferramenta Microgisbone que realiza a recuperação da informação de blogs na Internet utilizando como referência o padrão de microformato rel-tag. A ferramenta proposta tem como princípios (i) a adequação aos padrões de Big Data, (ii) utilizar uma arquitetura de serviços escalável para que seja possível utilizar computação em nuvem a partir de Infraestrutura como Serviços (IaaS) e (iii) que esteja disponível como um serviço de informação a comunidade. Para validar o funcionamento da ferramenta e escolher a arquitetura de dados foi realizado um experimento utilizando um cenário real, blogs hospedados pelo wordpress.com, onde as marcações (tags) foram coletadas por um período de 3 dias. Ao término deste experimento, onde foram coletadas cerca de 6,6 milhões de tags, foram desenvolvidos alguns serviços de informação, baseados nas marcações coletada. Foi observado que, de fato, um volume relativamente grande de informação foi recuperado de uma quantidade pequena de blogs e de um tipo de informação que é pequeno (marcações). Foi observado também padrão rel-tag dos microformatos tornam mais simples a identificação e recuperação das marcações nos blogs por máquinas se comparados com os mecanismos formais de web-semântica.

Palavras-chave: Micro Formatos. Blogs. Marcações. Produtos de Informação.

Abstract: This paper aims to present Microgisbone tool that performs the retrieval of information from Internet blogs using the standard rel-tag microformat. The proposed tool has the following principles: (i) the adequacy of the standards of Big Data, (ii) use an architecture for scalable services to be able to use cloud computing from Infrastructure as a Service (IaaS) and (iii) that is available as an information service to the community. To validate the tool efficacy and choose the data architecture we conducted an experiment using a real scenario, blogs hosted by wordpress.com, where the tags were collected for a period of 3 days. At the end of experiment, which were collected about 6.6 million tags, some information services, were developed. It was observed that, in fact, a relatively large volume of information was recovered from a small amount of blogs and a kind of information that is small (tags). It was also observed pattern of rel-tag microformat simplify the identification and retrieval of tags in blogs by machines compared to the formal mechanisms of semantic-web.

Keywords: Microformats. Blogs. Tags. Information Products.

1 INTRODUÇÃO

O rompimento da bolha das empresas ".com", no Outono de 2001, marcou um ponto de transformação da Internet que não ocorreu apenas influenciou no paradigma tecnológico vigente, mas principalmente, na forma como os negócio eram construídos na *web* (O'REILLY, 2007). Esta ruptura no modo de produzir e consumir conteúdo atraiu a atenção de profissionais e pesquisadores, levando à produção de ferramentas e pesquisas que

promoveram mudanças na Internet e transformaram naquilo em que conhecemos hoje. A percepção de uma evolução da *Web*, com base principalmente na maior interação entre usuários, levou ao surgimento do termo "*Web* 2.0", que representa a evolução dos produtos e serviços na Internet (O'REILLY, 2007).

Esta nova versão da *Web*, quando comparada à sua predecessora, destaca-se pela compreensão do ambiente digital como uma plataforma, isto pois, enquanto "a *Web* 1.0 limitava-se ao acesso e raras contribuição dos usuários" (MURUGESAN, 2007), enquanto a *Web* 2.0 apresenta um modelo completamente novo, focado nas relações entre os indivíduos e a sociedade digital. Por conta desta possibilidade de maior participação dos indivíduos, esta também pode ser chamada de "*Web* da Sabedoria", "*Web* Centrada nas Pessoas", "*Web* Participativa" ou "*Read/Write Web*" (MURUGESAN, 2007). De certo modo, se a Internet e a *Web* 1.0 poderiam ser consideradas "a rede mundial de computadores", então, a *Web* 2.0 pode receber a alcunha de "a rede mundial de pessoas". (CASSANO, 2008).

O surgimento da *Web* 2.0 transformou o perfil de seus usuários que passaram de passivos consumidores de conteúdo para produtores e consumidores de informação (PETERS; STOCK, 2007). Nessa transformação, Berners-Lee (2007) destaca o papel os *blogs*, uma vez que estes representam uma das ferramentas mais relevantes em colocar o usuário no controle de sua própria produção.

O próprio Bernes-Lee criou, em 1992, a página "*What is News*" que é considerado o primeiro *blog*, pelo menos da forma como o conhecemos hoje. Mas, apenas muito depois, em 1999, o termo "*blog*" foi criado por Peter Merholz derivado do termo "*web log*" que se tornou "*we blog*". A popularização dos *blogs* foi quase instantânea tanto que em 2002, um serviço de *blogs* chamado Blogger.com já apresentava mais de dez milhões de postagens em suas páginas. (FU, 2007).

Linpton (2002) afirma que os *blogs* caracterizam-se pela organização cronológica de suas publicações, dentro de uma estrutura onde, quase sempre, há somente uma página. Alexander (2006) destaca a diferença implícita na retórica dos *blogs*, quando comparada à sítios comuns na *Web* 1.0. A retórica mais pessoal, amplamente aplicada em *blogs*, culminou no surgimento de um público alvo específico, impelindo ao surgimento de práticas como o "*Blogrolling*", que se trata de uma listagem de endereços de outros *blogs* que o autor de um determinado *blog* recomenda e que formam uma espécie de rede de interesse. Outro fenômeno observado foi a popularização da utilização de folksonomias na organização do conteúdo digital, colocando o usuário no papel de arquiteto informacional.

O processo de utilização de folksonomias na indexação de conteúdo digital dentro da *Web* 2.0 recebe o nome de etiquetagem, ou *tagging* (LUND *et al.*, 2005; VANDER WAL, 2006) e consiste na atribuição livre de termos para representação de uma publicação. Deste modo, os termos atribuídos pelo autor (folksonomia estreita) ou pelos usuários (folksonomia ampla) em uma publicação, tem a finalidade de representar que assunto está sendo tratado naquele canal. A complexidade desta tarefa está expressas nos números, de acordo com estatísticas do Wordpress (2014), cerca de 40,5 milhões de novas publicações são realizadas todos os meses e mais de 409 milhões de pessoas leem publicações em *blogs* hospedados/construídos por esta ferramenta mensalmente.

O fenômeno da descentralização do processo de indexação trazido pela popularização da etiquetagem de recursos em *blogs*, que antes deveria ser feito, teoricamente, por um profissional da informação e é agora atribuído ao usuário, trouxe um grau considerável desorganização neste tipo de informação. Somado a isso o ambiente amplamente fluido e dinâmico dos *blogs* e grande o volume de informação, e de pessoas, que acessam esses serviços, tornou-se quase impossível recuperar conteúdo relevante nestes canais. Hewitt (2005) afirma que muitos destes *blogs* ainda se encontram inacessíveis para grande parte do público, particularmente para aqueles usuários que utilizam mecanismos de busca.

Esta dificuldade de encontrar conteúdo relevante na Internet a partir de sistemas de recuperação da informação (SRI), não foi percebido apenas nos *blogs*, mas em diversos contextos que começaram a se tornar cotidianos na vida dos usuários de Internet tais como informações pessoais, calendário, conexões sociais entre outros. Neste contexto surgiram os microformatos, criados por Tantek Çelik com o intuito de tornar os itens de dados existentes nas páginas da Internet reconhecíveis. A proposta é que esses elementos sejam passíveis de processamento automatizado por agentes de software e que também sejam diretamente legíveis por seres humanos (MENDEZ *et al*., 2008). Segundo Wharton (2005), a motivação para a criação dos microformatos foi atender necessidades, primeiramente, dos seres humanos e em segundo plano das máquinas.

A justificativa para este trabalho é que considerando as dificuldades encontradas pelos SRI no tratamento e processamento de informações não-estruturadas em *blogs* na *web*, em especial os meta-dados atribuídos pelos usuários a partir das etiquetas (*tags*), o presente trabalho tem como objetivo apresentar a ferramenta MicroGisbone, elaborada para obtenção automática de etiquetas em *blogs* a partir de marcações que utilizem o padrão de microformatos. A ferramenta será avaliada a partir de um experimento onde será recuperada

estas informações de *blogs* reais hospedados em serviços que suportem os microformatos. A partir dos dados coletados, serão desenvolvidos alguns serviços de informação para posterior uso.

Para uma melhor compreensão do contexto desta pesquisa, o presente trabalho encontra-se dividido em cinco seções e além desta Seção introdutória, a Seção 2 apresenta a fundamentação teórica que contempla os conceitos relacionados a esta pesquisa. A Seção 3 apresenta a ferramenta Microgisborne e seu processo de concepção. A Seção 4 apresenta o experimento. Por fim, na Seção 5 serão apresentadas as considerações finais.

2 LITERATURA PERTINENTE

Antes do início da construção da ferramenta se fez necessária uma revisão da literatura referente ao uso de etiquetas (*tags*) em *blogs* o uso de microformatos nestes ambientes. O resultado desta revisão é apresentado nas subseções a seguir.

2.1 Organização do Conteúdo em Blogs

Murugesan (2007) afirma que os *blogs* se tornaram populares pela sua forma simples e por possibilitarem que pessoas expressem seus pensamentos, ideias, sugestões e comentários na Internet. Rodzvilla (2002) argumenta que os *blogs* conceberam ao indivíduo a chance de satisfazer um desejo natural de se comunicar em sociedade. A consequência deste sucesso foi a explosão informacional que trouxe a tona a necessidade em organizar esta informação pelas empresas de marketing e monitoramento de mídias sociais (TREMAYNE, 2007).

No contexto dos *blogs*, a práticas tradicionais de organização da informação e do conhecimento não são aplicadas. Ao invés disso, são utilizados modelos de classificação natural, descentralizados, independentes de modelos pré-estabelecidos. A proposta dos *blogs* é, dar ao usuário a liberdade de organizar a sua produção de forma pessoal (RUSSEL, 2005). Dar ao usuário o controle sobre o processo organizacional não implica necessariamente em uma organização efetiva, ainda que este indivíduo possua, supostamente, um maior conhecimento semântico sobre a sua publicação (SOUZA; ALVARENGA, 2004).

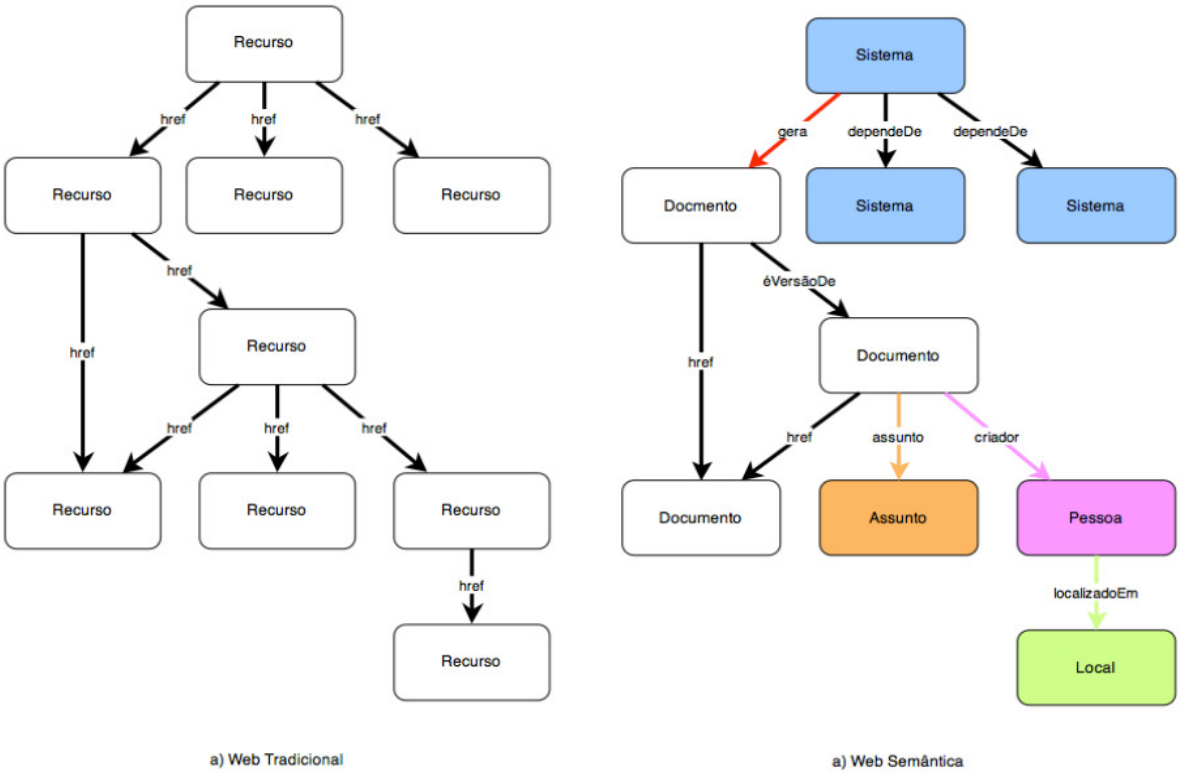
Cenários como este apresentam um grande desafio para que sistemas computacionais realizem o processo de mineração deste ambiente (*web mining*) uma vez que para isto se exige e existência de estabelecimentos semânticos no código fonte das páginas para auxiliar na atribuição de um significado àquilo que foi publicado (BAEZA-YATES; RIBEIRO NETO, 2013). Estes sistemas precisam deste auxilio para realizar a

identificação de conteúdo e uma das abordagens mais utilizadas para isso é a Web Semântica (KIM *et al*., 2009).

A Web Semântica permite que os dados contidos na Internet passem a ser compreendidos por computadores e esta é baseada na descrição dos recursos da Internet de forma compreensível pelos computadores. Esta descrição pode ser obtida através da anotação de meta-dados, agregando dados a outros pedaços de dados (OREN *et al*., 2006).

Deste modo, ao aplicar a Web Semântica, estes *blogs* são atendidos por um modelo mais inteligente de organização que prevê não apenas a organização interna, na forma de categorias ou marcações, mas também abre portas para modelos de extração externos que elevam o nível da organização para fora do ambiente do *blog*, apresentando um possível adjacente para novos modelos de organização, recuperação e representação da informação (WITTEN; FRANK; HALL, 2011). A FIGURA 1 apresenta um paralelo entre um conjunto de documentos organizados através de hiperlinks tradicionais (a) e outro utilizando os microformatos (b). Podemos perceber a dificuldade de prover significado a um elemento no modelo tradicional uma vez que não é possível determinar qualquer tipo de relação entre os recursos relacionados que só se utiliza do recurso de hiperlinks (*href*) que impossibilita a inferência informações.

FIGURA 1 – Comparativo dos modelos de organização Tradicional e de Web Semântica



Fonte: adaptado de Koivunen e Miller (2001)

2.2 Etiquetagem

Tradicionalmente o papel de catalogação e organização da informação em grandes unidades de informação e conhecimento, como as bibliotecas, é atribuição inquestionável de profissionais da informação. Contudo, estes profissionais, na maioria dos casos, lidam com ambiente e recursos padronizados que em nada se parece com o ambiente da Internet. Na rede, incontáveis recursos são produzidos diariamente e estes são criados nos mais diversos formatos (vídeo, texto, áudio, foto, etc.), e são organizados de formas completamente diferentes uns dos outros. Isto é um reflexo do aumento latente do conteúdo disponível, causado pela popularização do acesso e da produção de conteúdo e empregar profissionais de informação para catalogar e organizar o conteúdo da Web seria uma tarefa praticamente impossível (PRESSLEY, 2005).

Na Internet, muitos dos serviços disponibilizados adotaram um modelo de atribuição de palavras-chave que possibilita o usuário a atribuir etiquetas (*tags*) a recursos na *web* sem a necessidade de um vocabulário controlado (MARLOW *et al*., 2006). Este processo, denominado *Tagging*, propõe que usuários comum tornem-se o ator principal do processo de organização, consiste na associação de palavras-chave a um objeto ou item na Internet com o intuito de torna-los acessíveis de forma menos complicada para o usuário final (KIM *et al*., 2011).

Marlow *et al.* (2006) afirmam que:

> a utilização de tags no processo de organização apresenta potencial para melhorar o resultados de buscas, detectar spams, desenvolver sistemas de reputação e organização pessoal, enquanto possibilita novas possibilidades de comunicação e oportunidade de mineração de dados. Ainda, naquilo que tange sua concepção (MARLOW *et al*., 2006).

Kim *et al*. (2011) define que um modelo de marcação é composto de um conjunto de conceitos:

- "Objeto: o recurso que está sendo etiquetado. Por exemplo, um livro, uma foto, ou uma publicação de *blog*, etc.;
- *Tag*: a etiqueta associada ao recurso.
- Etiquetador: o agente – normalmente uma pessoa – que cria a ligação entre o objeto e a etiqueta.
- Fonte: o espaço onde o ato de etiquetar foi realizado. Por exemplo: Twitter.
- Polaridade: um voto contra ou à favor a declaração da etiqueta com o objetivo de solucionar problemas com *spams*.

A utilização de *Tagging* pode ser encontrada na web em duas vertentes: (a) através do uso colaborativo, ou amplo, e (b) através do uso restrito, ou estreito. No uso colaborativo, também conhecido como *social tagging*, há a interação social entre os

usuários de um sistema para a atribuição de etiquetas a um determinado recurso. Esta abordagem exige a existência de modelos de classificação que mensurem a relevância de uma etiqueta com base na quantidade de usuários que atribuem a mesma *tag* a um mesmo recurso, obtendo uma espécie de ranking (SINCLAIR; CARDEW-HALL, 2007). Este modelo de *Tagging* tem se tornado popular na Web em diversos serviços como por exemplo os *trend topics* do Twitter. (AMARAL; AQUINO, 2008) .

O conceito de folksonomia ampla de Varder-Wal (2013) é resultado do processo de etiquetagem a partir de um conjunto de palavras-chave que representam como um grupo de indivíduos classifica suas informações dentro de um contexto. Outro modelo possível de ser encontrado na *web* é aquele baseado no uso restrito, onde um único usuário, ou uma comunidade bastante reduzida, é responsável por atribuir etiquetas a um recurso, não havendo interação por parte dos outros usuários.

Guimarães (2008) afirma:

> Este modelo não possui a finalidade de ser interativo, mas representar como o autor do produto informacional classifica a sua produção, com a intenção de apresentar ao seu consumidor o conteúdo daquilo que se produziu, uma vez que o autor possui um conhecimento semântico profundo sobre aquilo que foi criado (GUIMARÃES, 2008).

2.3 Microformatos

Segundo Khare e Elik (2006)

> Para auxiliar na organização e recuperação da enorme quantidade de conteúdo que apresenta baixa semântica na web, especialmente aqueles advindos de blogs e ambientes de produção colaborativa, foram elaborados os microformatos sendo estes um conjunto de padrões que buscam "facilitar a descrição de pessoas, lugares, eventos e outras formas comuns de elementos semi-estruturados para linguagem humana" (KHARE; ELIK, 2006).

O conceito delimitado por Khare e Elik (2006) é fundamentado na descrição oferecida no próprio sítio da Microformatos: "Feito para humanos primeiro e máquinas em segundo". Os microformatos são um conjunto simples de formatos abertos de dados construída sobre a premissa de padrões existentes e já amplamente utilizados (Microformats, 2014).

A primeira menção aos microformatos ocorreu durante a conferência *South By Southwest* (SXSW) em 2004 e, desde então, "a utilização de microformatos vem ganhando adeptos rapidamente entre produtores de informação e desenvolvedores de serviços, grandes e pequenos, nos últimos dois ou três anos" (ALLSOP, 2007).

Para a utilização dos microformatos é necessária a existência de recursos semi-estruturados, tais como as linguagens de marcação *hipertexto mark-up language* (HTML) e

extensible mark-up language (XML), e através do uso de recursos já adotados, como o atributo *class*, são inseridos novos descritores contendo os microformatos responsáveis pelo valor semântico do documento (KHARE; ELIK, 2006). Um modelo de aplicação pode ser visualizado na FIGURA 2, onde há a descrição sintática de um evento acrescido do conteúdo semântico (em negrito).

FIGURA 2 – Exemplo de utilização do padrão hCalendar em microformatos para a definição de um evento em calendários

```
<div class="vevent">
<a class="url" href="http://conferences.oreillynet.com/pub/w/40/program.html">
http://conferences.oreillynet.com/pub/w/40/program.html
</a>
<span class="summary">Web 2.0 Conference</span>:
<abbr class="dtstart" title="2005-10-05">October 5</abbr>-
<abbr class="dtend" title="2005-10-07">7</abbr>,
at the <span class="location">Argent Hotel, San Francisco, CA</span>
</div>
```

Fonte: Os Autores

A grande vantagem dos microformatos para a organização e recuperação da informação na Internet reside, não em uma nova tecnologia, mas no reuso daquelas já utilizadas. Deste modo, facilitando a sua adoção. Neste sentido, Allsop (2007) argumenta à favor da importância dos microformatos para a Web Semântica: “os Microformatos são um conjunto de abordagens para resolver o importante problema da falta de produção de marcações com estruturas semânticas na *web* de hoje”.

De acordo com Kim *et al*. (2011) a semântica aplicada aos microformatos é menos poderosa quando comparada à capacidade de outros modelos tais como o *Resource Descritption Framework* (RDF). Mas, Khare e Elik (2006) sugerem que a utilização de um arquivo externo, como no caso do RDF, para atribuição de semântica em ambientes como a Blogosfera pode não funcionar por exigir um conhecimento específico do produtor da informação.

Khare (2006) destaca que existem quatro características que auxiliam na popularização de microformatos: (a) ser redutível, ou seja, ser utilizado somente quando há real necessidade, (b) ser reutilizável, não sendo necessária a definição de cada atributo sempre que este for utilizado, (c) ser reciclável, podendo ser utilizado em diversos recursos tais como *Feeds* sejam estes *Rich Site Summary* (RSS) ou *Atom* e *blogs*, e (d) ser representável e analisável, isto é, os meta-dados estão sempre visíveis, tanto para a máquina

quanto para o ser humano, embora sejam vistos de formas diferentes, proporcionando, por exemplo, análises automáticas do conteúdo.

Hoje, muitas são as classes, também conhecidas como vocabulários, criadas pela comunidade Microformats.org para atribuição de semântica na *web* com base nos microformatos, entre elas a hCalendar, hCard e rel-license, descritas a seguir (Microformats, 2014).

- **hCalendar**: Formato aberto de padronização para a descrição de eventos na web com base no padrão iCal utilizado em diversas aplicações de calendário. Este padrão permite que sistemas de busca na web sejam capazes de identificar um evento (no passado, no presente, ou no futuro) e converte-los de modo a serem automaticamente incorporados a ferramentas de calendário como, por exemplo, o Google Calendar.
- **hCard**: Este formato é utilizado para permitir que sistemas de busca sejam capazes de identificar pessoas ou empresas que estejam listadas em páginas na Internet, auxiliando na localização deste tipo de informação. Um serviço que utiliza este microformato é o "logar com Facebook" onde um usuário pode utilizar as suas informações contidas no Facebook e utilizá-las para se cadastrar em outros serviços.
- **rel-license**: Este padrão tem como objetivo delimitar espaços onde há a indicação do tipo de licença atribuída ao recurso. Se é livre, aberta, proprietária ou mesmo *copyleft*.
- **rel-tag**: Formato responsável por reconhecer as etiquetas em diversos serviços.

Outros vocabulários definidos são: rel-nofollow (comunicação com agentes), VoteLinks (concordância com links), XFN (relacionamentos entre pessoas), XMDP (metadados), XOXO (especificação de HTML), adr (endereços), geo (coordenadas geográficas), hAtom (informações semânticas), hAudio (conteúdo de áudio), hListing (listas abertas e distribuídas), hMedia (imagens, vídeo e áudio), hNews (informações semânticas em notícias), hProduct (informações padronizadas sobre produtos), hRecipe (receitas culinárias), hResume (currículos), hReview (revisões), rel-directory (indica que o destino de um hyperlink é um diretório), rel-enclosure (anexos), rel-home (indica um link para a página principal de um site), rel-payment (mecanismo de pagamento), robotsexclusion (informações para robôs de busca), xFolk (marcações sociais).

3 MICROGISBORNE

Com o objetivo elaborar um produto/serviço de informação que utilizasse a informação contida nas etiquetas criadas pelos usuários em blogs, os autores decidiram utilizar os padrões de microformatos para auxiliar nessa coleta baseado no vocabulário "*rel-tag*". Para tanto foi montado um experimento em um cenário real. Faz-se importante, ainda, ressaltar que a ausência de ferramentas para este tipo de problemática torna razoável a hipótese de pioneirismo desta pesquisa, principalmente devida sua relevância para a Organização e Recuperação da Informação.

Desta forma, para realização do experimento proposto nesta pesquisa, foi elaborada a ferramenta Microgisborne concebida como uma ferramenta acessível através da Web com o objetivo de processar e recuperar meta-dados, mais especificamente *tags*, de postagens de *blogs* que utilizem a classe de atributos *rel-tag* provenientes da semântica aplicada pelos microformatos. Neste sentido, a aplicação busca apresentar uma forma de se extrair este conhecimento através dos microformatos.

A ferramenta utiliza-se de algoritmos desenvolvidos em linguagem *Hypertext Preprocessor* (PHP), utilizando o framework Slim como base para efetuar a leitura dos microformatos dos documentos na Internet e recuperar para o usuário quais foram as *tags* atribuídas pelo produtor da informação para rotular uma postagem, tornando, assim, possível uma análise posterior do método de organização, ou mesmo a realização de inferências menos óbvias.

Os dados recuperados pela ferramenta são armazenado através do uso de bases de dados. Havia uma dúvida se a base de dados mais adequado seria a MySQL (livre e amplamente utilizada) ou a Google Big Query (livre e que suporta aplicações de *big data*). A intenção desta ferramenta é que ela seja aberta e acessível a outros desenvolvedores e para isso foi disponibilizada uma interface para aplicação de programas (API) para a utilização deste serviço de informação[85]. A arquitetura do sistema utiliza a arquitetura de serviços REST, também adequada para escalonamento de usuários se aplicado em infra-estrutura como serviço (IaaS), isso significa que se 1 ou 1.000.000 de usuários utilizarem o sistema de forma simultânea, o produto estará apto a realizar esse escalonamento de forma automática. A FIG. 3 apresenta as relações, métodos, e técnicas utilizadas para elaboração do serviço.

O modelo inicial o MicroGisborne foi delineado após grande dificuldade dos autores em encontrar outra ferramentas que tornassem possível a captura dos meta-dados utilizados

---

[85] http://docs.microgisborne.apiary.io/

no *blog* Brainsorm9.com.br, objetivando uma análise da organização informacional do *blog*, inicialmente quantitativa. As questões estabelecidas pela pesquisa giravam em torno da utilização de etiquetas no *blog* e eram as seguintes:

- Quantas *tags* os autores do *blog* utilizam em média para descrever seu conteúdo?
- Quais são as *tags* que mais se repetem?
- Os autores reutilizam *tags* ou tendem a criar novas *tags*?
- Qual o total de *tags* utilizadas pelos autores?

FIGURA 3 – Arquitetura do Microgisborne

Fonte: Os Autores

A ferramenta já está disponível no endereço www.microgisborne.com

4 EXPERIMENTO

O cenário escolhido para realização deste experimento foram os *blogs* presentes dentro das categorias "*News*[86]" e "*Music*[87]" do sítio da plataforma Wordpress. Estes blogs que utilizam a plataforma Wordpress foram escolhidos pois o suporte aos microformatos é uma configuração padrão em todos os blogs derivados deste serviço, assim, sendo garantida a consistência dos dados. O experimento foi realizado para avaliar três objetivos (i) se a utilização de microformatos era adequada para a recuperação desse tipo de informação, (ii)

86 http://en.wordpress.com/tag/news/
87 http://en.wordpress.com/tag/music/

saber qual base de dados seria mais adequada ao projeto, se o MySQL (relacional) ou Google Big Query (suporte a *big data*) e (iii) verificar se a arquitetura escolhida era escalável para múltiplos acessos.

Após a escolha do escopo a ser analisado, foi colocado o Microgisborne para executar o primeiro passo da recuperação das *tags* que é encontrar o endereço (URL) dos *blogs* a serem varridos. O *Wordpress* disponibiliza um arquivo .xml para cada categoria (*music* e *news*) onde se fez necessário buscar a marcação "*blog-url*" que apresentava a URL específica de um *blog*, e ao percorrer todo o arquivo xml, todos os *blogs* daquela categoria estariam mapeados. Esta primeira etapa finaliza quando todas as fontes de informação forem encontradas e os endereços dos *blogs* estejam armazenados nas base de dados (MySQL e *Google Big Query*).

O segundo passo foi verificar todas as postagens que existem no *blog*. Em boa parte dos serviços de *blog*s, uma nova postagem sempre é criada como uma nova página (nova URL), então era necessário encontrar os endereços destas páginas. Para isso, basta procurar pelo arquivo *sitemap.xml* que contem, como o nome sugere, o mapa do site e todos as suas "subpáginas" a partir do domínio original. Por exemplo, no *blog* "jonwizardnews.wordpress.com" era necessário procurar pelo endereço jonwizardnews.wordpress.com/sitemap.xml. A partir dai, e microgisborne deveria recuperar todo o conteúdo entre as marcações <loc></loc> que contem os endereços da subpágina. Esta segunda etapa finaliza quando todos as subpáginas forem localizadas e cadastradas nas bases de dados vinculadas as páginas principais.

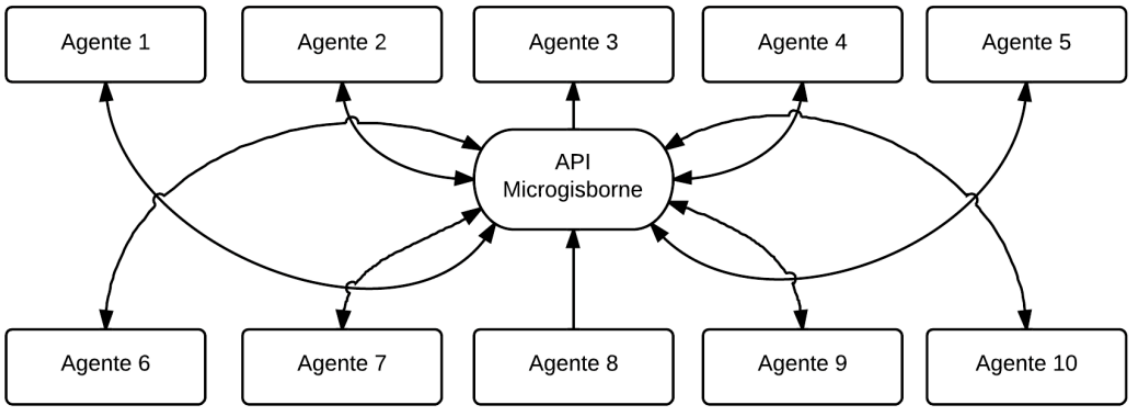
O terceiro passo seria "visitar" cada página e subpágina e ao olhar o código fonte procurar pelas marcações "rel-tag" do padrão de microformatos. Estas marcações representam as etiquetas realizadas pelos usuários em uma postagem específica. Então cada subpágina foi varrida pelo microgisbourne e cada *tag* foi vinculada a uma postagem "subpágina", e cada postagem vinculada a um *blog* (página).

Para este experimento a ferramenta foi configurada a Infraestrutura como serviço (IaaS) da amazon.com e foram criados 10 agentes onde cada uma simulava algum tipo de serviço do microgisborne. Os 10 agentes se alternavam, de forma aleatória, entre os três passos da atividade, para simular momentos em que fossem necessárias indexar mais informações ou simular que 10 usuários simultâneos estão buscando dados na ferramenta e assim o serviço não deveria estar realizando tarefas *back-end*. A FIGURA 4 a seguir apresenta este funcionamento.

Esta estrutura pode ser dimensionada para 1 único agente ou para 10.000 agentes simultâneos. A carga de trabalho dos agentes foram monitorados e estiveram em média, apenas 13% do tempo em estado de espera (*Idle*). Nas primeiras 6 horas, o número de instâncias, ou agentes funcionando, foram variados entre 3 e 30 para simular a escalabilidade do serviço.

Todo os dados processados foram colocados em dois sistemas gerenciadores de banco de dados (SGBD). O SGBD MySQL utilizou a modelagem relacional para a organização lógica dos dados, enquanto que no Google Big Query utilizou o formato JSON para modelar os dados. Em três dias de experimento, foram visitados 2.546 blogs que continham 764.429 páginas e foram coletadas 6.699.643 *tags*. Enquanto no MySQL estes dados ocuparam 418,72 Mb de dados, enquanto no Google Big Query se utilizou cerca 356,99 Mb. Entretanto, as consultas utilizando o Google Big Query são em média 56% mais rápidas do que no MySQL e a tendência é que c com a quantidade de dados aumentada (Big) é que este SGBD seja melhor.

FIGURA 4 – Funcionamento do Microgisborne



Fonte: Os Autores

5 CONSIDERAÇÕES FINAIS

O experimento demonstrou que a ferramenta foi bem sucedida em coletar etiquetas de *blogs* utilizando o vocabulário "*rel-tag*" de microformatos se mostrando uma solução simples para a recuperação de informações em serviços de Internet. A partir dessa possibilidade, novos produtos e serviços de informação, baseados em microformatos podem ser elaborados. Por exemplo, nuvens de *tags* das categorias *News* e *Music*, FIGURAS 5 e 6 respectivamente, foram criadas a partir dos dados coletados pela ferramenta e geradas pela ferramenta D34js.

FIGURA 5 – Nuvens de Tags dos Categoria *News*

Fonte: Os Autores

FIGURA 6 – Nuvens de *Tags* dos Categoria *Music*

Fonte: Os Autores

Atualmente a ferramenta já implementa os seguintes serviços:

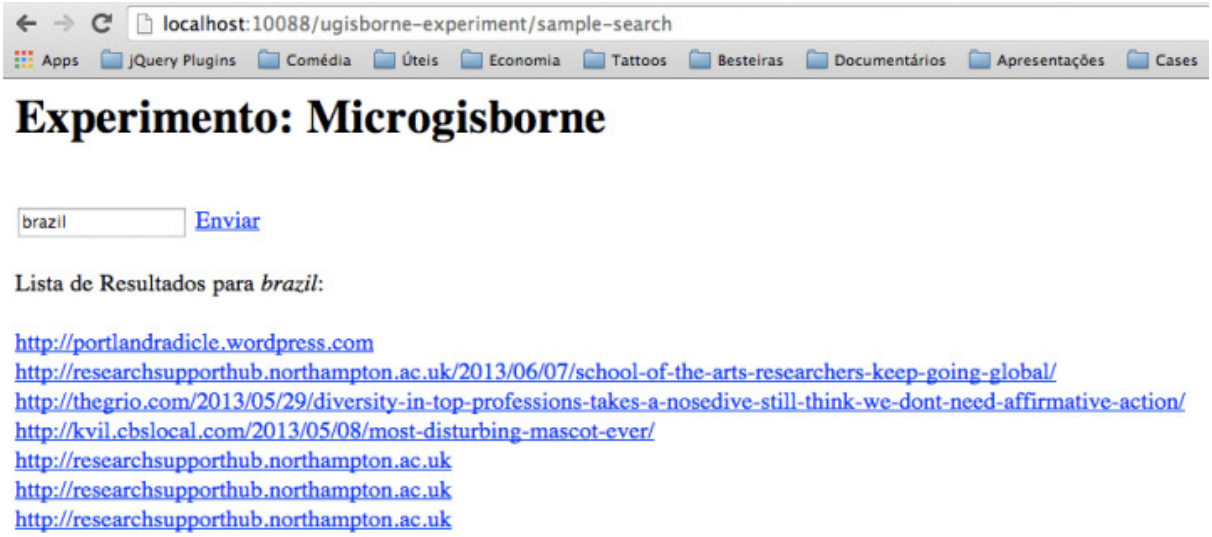
Tags de um *Blog*: Ao digitar a URL de qualquer *blog* que tenha suporte aos microformatos ele retorna a lista de *tags*, exemplo na FIGURA 7. A nuvem de *tags* por *blogs* ainda está em construção.

FIGURA 7 – Tags do blog vernasunhas.blogspot.com

Fonte: Os Autores

Lista de *Blogs* a Partir de uma *Tag*: Este serviço apresenta uma lista de *blogs* que contenham uma *tag* escolhida pelo usuário. Este ainda não está implementado no serviço *online*, mas a FIG. 8, mostra o resultado deste serviço no experimento.

FIGURA 8 – Lista de Sites a partir de uma Tag



Fonte: Os Autores

Tag Blogrolling: Este serviço apresenta uma lista de *blogs* que apresentam o mesmo conjunto de *tags* que um outro *blog* escolhidos pelo usuário. Por exemplo, um usuário pode entrar com o endereço http://vernasunhas.blogspot.com.br e o microgisborne sugere outros *blogs* que possuam o mesmo conjunto, ou o máximo, conjunto de *tags*. Este serviço ainda não está implementado devido.

Finalmente, sugerimos a reflexão de que os maiores desafios levantados quanto a evolução do microgisborne não provêm da tecnologia em si, mas sim da elaboração de fluxos da informação para as *tags* armazenadas nas bases de dados e transformar estes fluxos em produtos e serviços úteis a sociedade.

REFERÊNCIAS

ALEXANDER, B. Web 2.0: A new wave of innovation for teaching and learning? **Educause review**, v. 41, n. 2, p. 32-44, 2006.

ALLSOP, J. **Microformats:** Empowering Your Markup for Web 2.0. Nova York: Friends of, 2007. 368 p.

AMARAL, A.; AQUINO, M. Práticas de folksonomia e social tagging no Last. fm. In: Simpósio Brasileiro de Fatores Humanos e Sistemas Computacionais, 8., 2008, Porto Alegre. **Anais…,** Porto Alegre, Simpósio Brasileiro de Fatores Humanos e Sistemas Computacionais, 2008.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação**: Conceitos e Tecnologia das Máquinas de Busca. 2. ed. Porto Alegre: Bookman, 2013. 614 p.

CASSANO, R. **Somos tão Jovens**: Uma breve retrospective da Web. 2011. Disponível em: < http://www.techtudo.com.br/platb/internet/2011/06/09/somos-tao-jovens-uma-breve retrospectiva-da-web/> Acesso em: 13 Ago. 2013.

FU, B. Trust Management in Online Social Networks. 2007. 90 f. **Dissertação** (Mestrado em Ciência da Computação) - Departamento de Ciências da Computação. Trinity College, University of Dublin, Dublin. 2007.

GUIMARÃES, C. P. Tags: palavras-chave em blogs. In: Simpósio Hipertexto e Tecnologias na Educação. 2. 2008, Recife. **Anais…** Recife. Simpósio Hipertexto e Tecnologias na Educação, 2008. p. 1-22.

HEWITT, H. **Blog:** Understanding the information reformation that's changing your world. Nashville: Thomas Nelson Inc., 2005. 256 p.

KHARE, R; ELIK, T. Microformats: A Pragmatic Path to the Semantic Web. 2012. Disponível em: <http://commerce.net/wpcontent/uploads/2012/04/CN-TR-06-01.pdf>. Acesso em: 18 fev. 2014.

KHARE, R. Microformats: The next (small) thing on the Semantic Web**. Internet Computing**, v. 10, n. 1, p. 68-75, 2006.

KIM, H.; SCERRI, S.; PASSANT, A.; BRESLIN, J. Integrating Tagging into the Web of Data: Overview and Combination of Existing Tag Ontologies. **Journal Of Internet Technology,** Taiwan, v. 12, n. 4, p. 561-572, 2011.

KOIVUNEN, M.; MILLER, E. W3c semantic web activity. **Semantic Web Kick-Off in Finland**, p. 27-44, 2001.

LIPTON, R., **What is a Weblog?**. 2002. Disponível em: <http://radio.weblogs.com/ 0107019/stories/2002/02/12/whatIsAWeblog.html>. Acesso em: 13 Ago. 2013.

LUND, B.; HAMMOND, T; FLACK, M; HANNAY, T. Social bookmarking tools (II). **D-Lib magazine**. v. 11, n. 4, 2005.

MARLOW, C; NAAMAN, D; BOYD, D; DAVIS, M. Tagging paper, taxonomy. In: Conference on Hypertext and Hypermedia. 17. Nova York, 2006. **Anais…** Nova York, ACM, 2006.

MENDEZ, E; LÓPEZ, L. M.; SICHES, A.; BRAVO, A. G.. **DCMF**: DC & Microformats, a Good Marriage. In: International Conference on Dublin Core and Meta-Data Applications. Berlin, 2008. **Anais…** Berlin, DC-2008: International Conference on Dublin Core and Metadata Applications, 2008.

MICROFOMARTS. **What is Microformats**. 2014. Disponível em: <www.microformats.org>, Acesso em: 18 fev. 2014.

MURUGESAN, S. Understanding Web 2.0. **It Professional**, USA, v. 9, n. 4, p. 34-41, 2007.

O'REILLY, T. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. **Communications & Strategies**, v. 1, n. 65, p. 17, 2007.

OREN, E.; MÖLLER, K.; SCERRI, S.; HANDSCHUH, S. **What are semantic annotations**. Galway: Digital Enterprise Research Institute, National University of Ireland, 2006. 14 p.

PETERS, I.; STOCK, W. G. Folksonomy and information retrieval. **Journal of American Society for Information Science and Technology**, v. 44, n. 1, p. 1-28, 2007.

PRESSLEY, L. **Folksonomies, tags, and user-created metadata**: A truly emerging technological trend unpublished seminar on emerging technological trends in libraries. Winston-Salem : Reynolds Library - Wake Forest University, 2005. 22 p.

RODZVILLA, J. **We've Got Blog**: How Weblogs are Changing Our Culture. Cambridge: Basic Books, 2002. 176 p.

RUSSELL, T. **Contextual authority tagging**: cognitive authority through folksonomy. 2005. Disponível em: <http://www.terrellrussell.com/projects/contextualautho ritytagging/conanthtag200505.pdf>. Acesso em: 1 out. 2013.

SINCLAIR, J.; CARDEW-HALL, M. The folksonomy tag cloud: when is it useful?. **Journal of Information Science**, v. 34, n. 1, p. 15-29, 2008.

SOUZA, R.; ALVARENGA, L. A Web Semântica e suas contribuições para a ciência da informação. **Ciência da Informação**, Brasília, v. 33, n. 1, p 132-141, 2004.

TREMAYNE, Mark. Introduction: Examining the blog-media relationship. **Blogging, citizenship, and the future of media**, New York, p. 3-20, 2007.

VANDER WAL. Explaining and Showing Broad and Narrow Folksonomies. 2005. Disponível em: <http://www.vanderwal.net/random/entrysel.php?blog=1835>. Acesso em: 1 Ago. 2013.

WITTEN, H.; FRANK, E.; MARK, A. Hall. **Data Mining**: Practical machine learning tools and techniques. Morgan Kaufmann 2011. 560 p. (The Morgan Kaufmann Series in Data Management Systems – Series)

WHARTON. **What's the Next Big Thing on the Web? It May Be a Small, Simple Thing - Microformats**. 2005. Disponível em: <http://knowledge.wharton.upenn.edu/index.cfm?fa=printArticle&ID=1247> Acesso em: 1 ago. 2013

WORDPRESS. Stats -Wordpress.com. 2014. Disponível em: <http://en.wordpress.com/stats/>. Acesso em: 18 fev. 2014.