1. **Entrada de dados**

Tela do formulário

Dados do Editor

Dados imputados pelo usuário

Adição de estruturas em HTML

Dados de Fotos/Videos

Persistência na base de dados

Persistência no solr

A etapa de entrada de dados é o ponto de partida para o sistema de recomendação proposto, nesta fase, é apresentado ao editor uma interface web no formato de um formulário que solicita ao editor um conjunto de informações: Estas informações podem ser divididas em obrigatórias e não obrigatórias segundo a tabela 1.

As informações obrigatórias são necessárias para constituir uma notícia que pode ser publicada.

Table 1

|  |  |  |
| --- | --- | --- |
| Título | O título da notícia é um dado textual limitado em 255 caracteres alfa numéricos. | Obrigatória |
| Subtítulo | O subtítulo da notícia é um dado textual limitado em 1000 caracteres alfa numéricos. | Obrigatória |
| Corpo | O corpo da notícia é um dado textual com um limite de caracteres alfa numéricos grande. O Editor pode fazer uso de markup html na construção do conteúdo da notícia | Obrigatória |
| Editoria Principal | Dado que agrupa o conjunto de notícias. Esta informação é selecionada na interface através de um caixa de seleção. Este agrupamento é criado pelos editores e geralmente refletem seções do site na internet. Exemplo de Editoria para o techtudo: Jogos, Dicas e Tutoriais, Reviews | Obrigatória |
| Editorias Secundárias | Agrupamentos secundários para organização das notícias | Não Obrigatória |
| Entidades Associadas | Informações extras que associam as notícias | Não Obrigatória |

1. **Extrator de Atributos**

O módulo de extração de atributos é o responsável por recuperar e organizar as informações em estruturas de atributos conhecidos do sistema com base nos dados de entrada imputados pelo editor. Estes atributos permitirão ao sistema refinar as buscas por informações relacionadas uma vez que eles pode servir como pistas na identificação de como devemos relacionar o conteúdo com a matéria que esta sendo produzida.

As regras de extração são implementadas como funções que são executadas sobre os parâmetros de entrada.

O framework conta com a seguintes regras para extração de atributos:

* 1. **Dados em links**

Este extrator tem por objetivo extrair os termos marcados como link imputados pelo editor no momento da criação da matéria. Para realizar esta tarefa, o sistema faz uma análise do texto proveniente do campo “Corpo” que é o único campo capaz de aceitar a inserção de links.

A análise consiste em percorrer todo o texto a procura do mark-up <**a**> pois é este mark-up que define um link na especificação de **html**.

Ao final do processo é gerado um vetor de termos encontrados nos links. Os termos podem ser representados por uma única palavra ou um conjunto de palavras como mostra a figura abaixo:

Figure 1

Neste caso, “Fortaleza“ e “site do evento” serão armazenados no vetor de termos proveniente de links.

* 1. **Dados em Negrito e em Itálico**

Em uma análise semelhante ao extrator de links, o extrator de informações em negrito e em itálico também percorrerá todo o texto do campo Corpo. Porém agora, a procura dos mark-ups <**strong>** e <**em>** pois são estes mark-ups que definem respectivamente negrito e itálico na especificação de **html**.

Ao final do processo é gerado um vetor de termos encontrados em negrito e em itálico.

De acordo com a figura 1, serão selecionados os termos “Biblioteca Pública Dolor Barreira” e “Festival de Cinema e Cultura da Diversidade Sexual.”

* 1. **Dados em títulos de vídeos e fotos**

Este extrator difere dos anteriores apenas no que se refere ao padrão de procura, pois todo o texto do campo Corpo será percorrido porém agora, a procura de uma estrutura de mark-up que define o elemento título de vídeo e título de foto.

Na figura 1 o texto “**Quinta edição do festival For Rainbow começa nesta quinta-feira (27) (Foto: Divulgação**)” será selecionado.

* 1. **Extratação de entidades nomeadas**

Este extrator tem o objetivo de reconhecer entidades que aparecem no texto

Seleção de candidatos

Nesta etapa o objetivo é construir um entendimento acerca das informações recebidas do editor na etapa de entrada de informações. Para isso, é necessário um tratamento das informações com técnicas de mineração de textos para extrair o máximo de informações relevantes.

* Fase de limpeza:

O corpo da matéria é construído com texto livre o que permite aos editores a inclusão de mark-ups html para marcação do texto, com isso é comum termos palavras em negrito, itálico ou até mesmo elementos mais elaborados com tabelas e div’s que agrupam dados textuais. O extrator elimina os caracteres que representam marcadores html e realiza a conversão de normas de acentuação típicas da notação html para a tabela ascii. Assim temos:

Ex. <strong> OL&Aacute; MUNDO! </Strong>

É convertido para OLÁ MUNDO!

* Fase de geração do Bag of Words:

Com o texto livre de caracteres de marcação, o extrator realiza a eliminação de todos os caracteres de pontuação e realiza a quebra do texto em palavras utilizando como critério a presença de espaço em branco entre elas.

* Fase de geração do TF:

As palavras são então contadas e ordenadas em função de sua presença no texto, as palavras consideradas como stopwords e as palavras classificadas gramaticalmente como verbo são eliminadas. Assim podemos descobrir quais palavras são realmente relevantes para o texto.

Para remover as palavras chamadas stopwords, o extrator faz uso de um dicionário de palavras marcadas como stopwords. Este dicionário foi gerado com a recuperação das pala lavras com o menor TF-IDF que significa o grau de relevância da palavra para todo o corpus.

Para remover todas as palavras que gramaticalmente são classificadas como verbo, o extrator faz uso de um tagger gerado com base em uma acervo de artigos da folha de SP.

Após o tratamento do texto são também separados as entidades fornecidas na criação da matéria. Elas são armazenadas em uma array de entidades. Este mesmo procedimento é realizado para a informação de editoria.

1. **Seleção de Candidatos**

Nesta etapa é avaliado pesos e critérios para realização de consultas no corpus. Uma vez analisada a notícia e extraído os dados necessários poderemos estabelecer caminhos de procura de informações.

O corpus está armazenado em um servidor de busca por texto baseado em **lucene**, de tal forma que as consultas são realizadas por palavras chave e levam em consideração algumas regras de ordenação:

* As matérias são ordenadas por data de publicação das mais recentes para as mais antigas.
* Palavras chaves encontradas no campo “título” tem o dobro da relevância quando encontrada no corpo do texto.
* São observadas matérias que foram publicadas dentro de no máximo, 60 dias anteriores a data atual
* São selecionados um número fixo de matérias por consulta para permitir que as consultas sejam rápidas e padronizadas.

Embora tenhamos uma infinidade de heurísticas possíveis para seleção dos candidatos, optamos pela elaboração das seguintes heurísticas:

1. Mais da mesma entidade

Esta regra determina que o sistema selecione candidatos que falam da mesma entidade fornecida pelo usuário. As entidades são fornecidas pelo editor através de um sistema de reconhecimento de entidades nomeadas que extrai do texto as entidades citadas em suas devidas classes: Produtos, Lugares, Pessoas e Organizações. Deste modo podemos consultar a base de dados por conteúdos que fazem referencia a estas entidades.

1. Mais da mesma editoria

Esta regra determina que o sistema selecione candidatos que participam da mesma editoria fornecida pelo usuário, desta forma quando o editor esta produzindo um conteúdo na editoria jogos, são fornecidas outras matérias de jogos para o editor.

1. Mais das palavras mais relevantes

Esta regra determina que o sistema utilize a distribuição de frequência de termos e realize consultas por matérias que se relacionem com o termo mais relevante para o texto.

1. Mais das entidades similares a entidade relacionada

Esta regra determina que o sistema utilize a entidade fornecida para descobrir a superclasse da entidade de modo a oferecer conteúdos de entidades similares.

Como exemplo: Dada uma matéria que mencionam as vendas de IPad, o sistema também sugeriria, se existisse, matérias que falam de vendas de Tablets pois o IPad é um tipo de Tablet.

1. Mais das palavras grafadas com Bold

Esta regra determina que o sistema utilize as palavras grafadas com Bold para sugerir conteúdos ao editor.

1. Mais do assunto relacionado

Esta regra determina que o sistema utilize o assunto da matéria para sugerir outras matérias relacionadas ao mesmo assunto.

1. **Componente de Filtragem**

Nesta etapa os itens encontrados em cada uma das consultas realizadas, são ordenados em função do grau de similaridade com o item criado. A função de similaridade escolhida é o VSM.

Os itens são então agrupados em suas heurísticas e sugeridos para o Editor, de modo que ele possa escolher opções de heurísticas distintas.