Sistema para extração de informações de artigos científicos - IESYSTEM

Rafael Geraldeli Rossi
Solange Oliveira Rezende
Alneu de Andrade Lopes

Universidade de São Paulo – USP
Instituto de Ciências Matemáticas e de Computação – ICMC
Laboratório de Inteligência Artificial – LABIC
Av. Trabalhador São-carlense, 400 – Centro
Caixa Postal: 668 – CEP: 13560-970 – São Carlos – SP – Brasil

[ragero, solange, alneu]@icmc.usp.br

Resumo. *Neste relatório técnico é apresentada a ferramenta IEsystem, que extrai metadados de coleções de artigos científicos. Esta ferramenta é capaz de realizar a extração de metadados mesmo quando os artigos científicos são provenientes de diferentes fontes ou escritos em diferentes línguas. O processo de extração de metadados pauta-se em modelos, que descrevem posições relativas de conteúdos ou indicadores de conteúdos a serem extraídos. Um conjunto de ferramentas de pré-processamento e de criação/edição de modelos também é disponibilizado. Este relatório apresenta detalhes sobre as telas, módulos, funções, e uso da ferramenta.*
Palavras-chave: *Mineração de Textos, Extração de Informações.*

Relatórios Técnicos do ICMC
Junho de 2010

Sumário

1. Introdução

Atualmente, o número de artigos científicos disponíveis em formato digital na rede mundial de computadores têm aumentado incessantemente. A necessidade para localizar o artigo científico desejado ou os artigos mais relevantes torna-se mais complexa devido a esse grande volume de documentos, além de tornar mais complexa as tarefas de gerenciamento, organização e extração de conhecimento útil. Para auxiliar o usuário nessas tarefas, pode-se utilizar a Mineração de Textos para organizar as coleções de artigos científicos, extrair conhecimento, e apresentar os resultados em representações gráficas que auxiliem a exploração e o entendimento dos documentos da base (Lopes et al., 2007).

Porém, devido a enorme quantidade de atributos gerados por uma coleção de documentos textuais, pode ser de grande utilidade usar apenas algumas partes do texto, por meio da Extração de Informações, o que pode aumentar a compreensibilidade do usuário no resultado final do processo. A Extração de Informações preocupa-se em localizar padrões específicos de dados e por meio disso extrair informação estruturada de dados não estruturados (Nahm e Mooney, 2000).

Dado isso, neste relatório técnico é apresentada e detalhada a ferramenta `IESystem`, que foi desenvolvida para extrair medatadados de artigos científicos, como título, autoria, resumo, conclusões, e referências. Vale ressaltar que a ferramenta é capaz de extrair metadados mesmo em coleções de artigos científicos com diferentes línguas e formatos.

O restante deste relatório técnico está dividido da seguinte maneira: na Seção 2 são apresentadas as características da ferramenta, suas telas e a interação entre os módulos do `IESystem`; na Seção 3 é apresentada a biblioteca de classes, juntamente com o diagrama de classes e as explicações de cada classe que compõe a ferramenta; na Seção 4 são apresentadas as utilidades da ferramenta desenvolvida; na Seção 5 é ilustrado o uso da ferramenta; e finalmente na Seção 6 são apresentadas as considerações finais.

2. A ferramenta IESystem

A ferramenta IESystem foi desenvolvida para extrair os elementos dos artigos científicos escritos em qualquer língua suportada por caracteres da tabela ASCII estendida.

2.1. Visão Geral

Uma das principais características do IESystem é extrair os elementos dos artigos de acordo com um modelo fornecido pelo usuário. Um modelo é uma estrutura que possui nome, descrição, elementos dos artigos científicos, as outras possíveis escritas para os elementos dos artigos científicos, e quais desses elementos serão gravados no resultado.

Um exemplo dos elementos de um modelo são: `Título`, `Autoria`, `Resumo`, `Palavras-Chave`, `Corpo`, e `Referências`. O texto entre Título e Autoria é gravado no arquivo resultante do processo entre as *tags* `<Título>` e `</Título>`, em seguida extrai o texto entre Autoria e Resumo e coloca o conteúdo extraído entre as *tags* `<Autoria>` e `</Autoria>`, e assim por diante.

Note que, provavelmente, um artigo científico não possui um descritor de seção `Corpo` abaixo do resumo, e sim `Introdução`, `Motivação`, entre outros. Quando isso ocorre, `Introdução` e `Motivação` podem ser inseridos no dicionário[1] como significados de `Corpo`. Neste caso, o extrator extrai, por exemplo, o conteúdo entre `Resumo` e `Introdução` e marca o início desse conteúdo extraído com a *tags* `<Resumo>` e o final desse conteúdo com a *tag* `</Resumo>`, o conteúdo entre `Introdução` e `Referências` é marcado da mesma forma com as *tags* `<Corpo>` e `</Corpo>`. Note que mesmo tendo encontrado `Introdução`, o arquivo gerado conterá as *tags* dos elementos a que se referem no dicionário. Vale ressaltar que os elementos de um modelo e as variações desses elementos devem ser fornecidos pelo usuário. Entretanto, a ferramenta apresenta recursos para auxiliar o usuário nestes quesitos.

A ferramenta foi desenvolvida utilizando as linguagens de programação Java e Perl. A interface com o usuário, o gerenciamento dos modelos, da base de nomes de autores, e dos resultados da extração, foram desenvolvidos em Java, devido a sua portabilidade e facilidade de desenvolvimento de interfaces gráficas. Estas funcionalidades desenvolvidas em Java serão chamadas de *Módulo de Gerenciamento* no decorrer deste relatório. O parser, para separar os elementos dos artigos científicos, a geração dos arquivos contendo os resultados do processo de extração

[1] Contando com o fato de que um descritor de seção de um artigo científico pode ser escrito de maneiras diferentes, como `Referências` e `Bibliografia`, ou `Conclusão`, `Conclusões` e `Trabalhos Futuros`, a ferramenta propicia um dicionário de significados para que, por exemplo, `Conclusão`, `Conclusões` e `Conclusões e Trabalhos Futuros` signifiquem apenas `Conclusão`.

de metadados dos artigos científicos, e detectores de cabeçalho e rodapé foram desenvolvidos em Perl, devido ao seu grande suporte e facilidade ao se trabalhar com expressões regulares. Estas funcionalidades desenvolvidas em Perl serão chamadas de *Módulo de Extração* no decorrer deste trabalho. Na Figura 1 são mostrados as principais funcionalidades dos módulos desenvolvidos e a direção do fluxo de informação entre eles. Na Figura 2 são mostrados as informações que são trocadas entre os módulos durante o processo de extração de metadados.

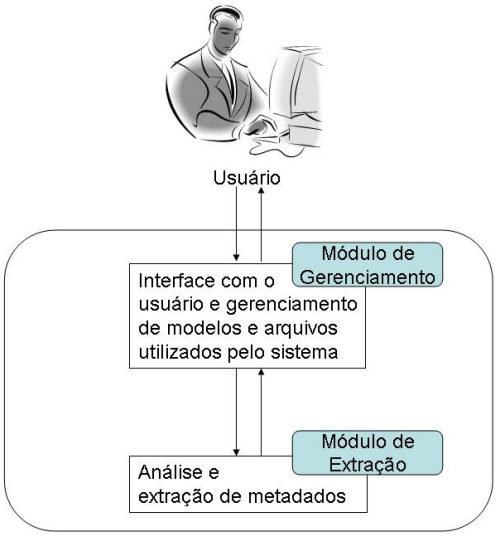

Figura 1. Módulos da ferramenta IESystem.

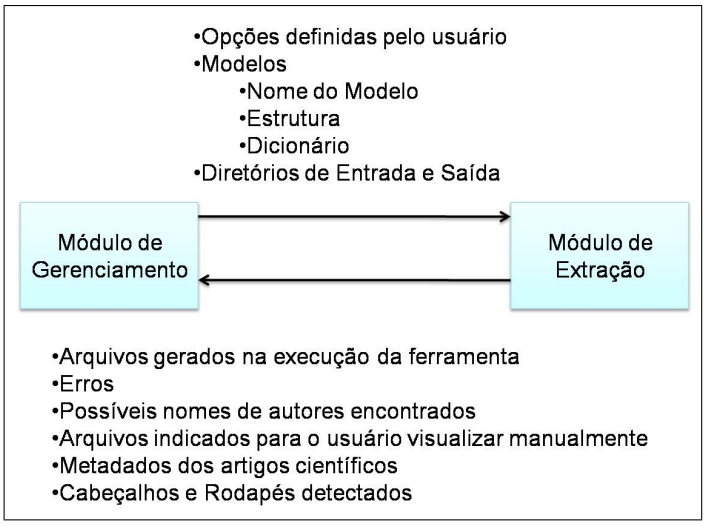

Figura 2. IESystem: Interação entre os módulos da ferramenta IESystem.

2.2. Telas do IESYSTEM

Nesta seção do relatório são apresentadas as telas da ferramenta IESystem, suas funcionalidades, e como utilizá-las.

2.2.1. Tela Principal

Na tela principal, apresentada na Figura 3, é possível acessar todas as opções da ferramenta e acompanhar o processo de extração de metadados. Esta tela permite ao usuário:

- Criar, editar, carregar e deletar os modelos que serão utilizados no processo de extração de metadados;
- Definir os diretórios de entrada (que contém os arquivos a serem analisados) e saída (que conterão os arquivos resultantes do processo de extação);
- Escolher opções para o formato dos arquivos gerados;
- Indicar os elementos que serão gravados no(s) arquivo(s) resultantes do processo de extração de metadados;
- Habilitar/Desabilitar a remoção de *stopwords*.[2];
- Acessar as opções e a ajuda da ferramenta;
- Visualizar o progresso do processo de extração de metadados.

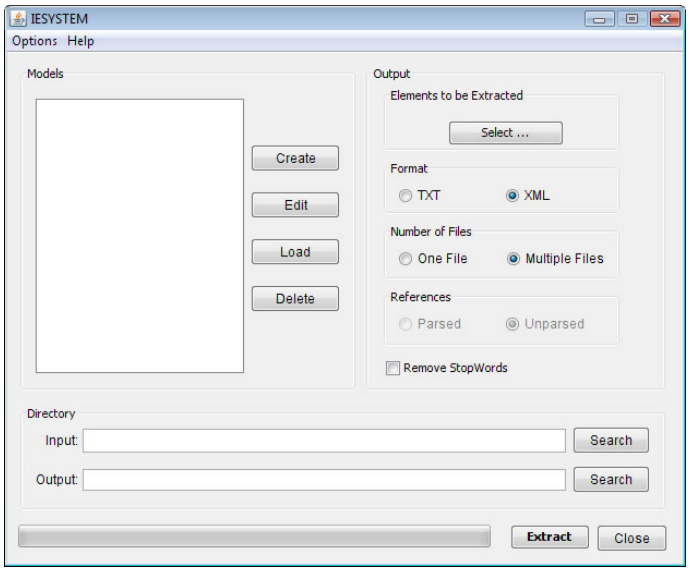

Figura 3. IESystem: Tela inicial da ferramenta IESystem.

[2]Palavras que não são úteis para o processo de Mineração de Textos

2.2.2. Criação/Edição de Modelos

Os modelos são fundamentais para a extração de metadados dos artigos científicos. Eles contêm toda a informação necessária para o processo de extração: nome do modelo, descrição, lista de elementos dos artigos científicos, possíveis significados desses elementos, e quais elementos serão gravados nos arquivos resultantes do processo de extração.

A tela de criação/edição de modelos, apresentada na Figura 4, disponibiliza os seguintes itens para o usuário:

- *Nome*: é usado para identificar o modelo na lista de modelos a serem utilizados no processo de extração de metadados;
- *Descrição*: utilizada para lembrar o usuário do por que, ou para que, aquele modelo é utilizado;
- *Elementos*: são os elementos a serem extraídos da base de artigos científicos. Há dois elementos que foram identificados como padrão: *Título* e *Autoria*. A partir destes, os elementos dos artigos científicos podem variar. O usuário deve então definir esses elementos **na mesma ordem** em que eles ocorrem nos artigos científicos, pois a ferramenta considera que os conteúdos dos elementos aparecem na mesma ordem em que foram definidos no modelo;
- *Dicionário de Elementos*: adição de significados ou outras formas que os elementos do modelo podem estar escritos nos artigos científicos.

Para editar os elementos de um modelo, a tela apresentada na Figura 5 é disponibilizada ao usuário. O usuário pode utilizar os botões *Include* e *Remove* para incluir e remover respectivamente os elementos de um modelo, e os botões *Up* e *Down* para ajustar a ordem.

Para editar os significados dos elementos de um modelo, há um painel no canto inferior esquerdo da tela de criação e edição de modelos, apresentado na Figura 4, que exibe os elementos dos modelos juntamente com seus significados. Há duas maneiras para editá-los: clicar no elemento que deseja adicionar os significados e depois clicar no botão `Edit`, ou clicar no botão `Get Section Descriptors`.

Caso a opção `Edit` seja escolhida, a tela apresentada na Figura 6 será disponibilizada ao usuário. Nesta tela, o usuário pode incluir e remover o significado dos elementos dos artigos científicos por meio dos botões `Include` e `Remove`, respectivamente.

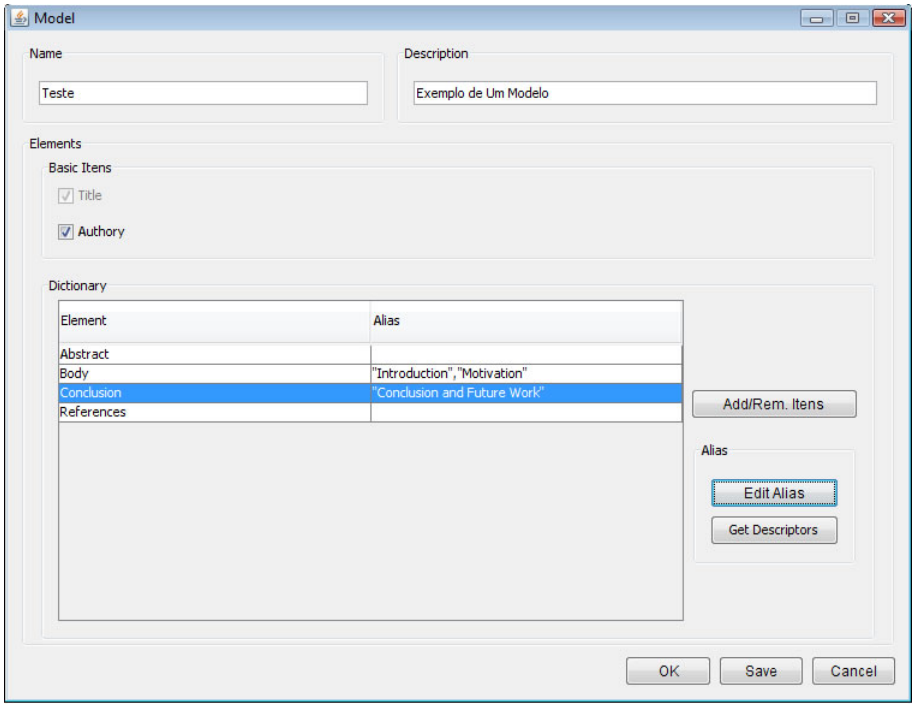

Figura 4. IESystem: Tela de criação/edição de modelos.

Caso o usuário opte pela opção `Get Descriptors`, a tela apresentada na Figura 7 é apresentada ao usuário. Para que o usuário obtenha os descritores de seções, é necessário definir um diretório onde se encontram os artigos científicos. A ferramenta IESystem automaticamente converte os arquivos no formato *pdf* para o formato *txt* e aplica os filtros para remoção de acentos (para que as expressões regulares possam funcionar adequadamente), e junção de palavras (para que não apareçam tópicos com o caractere "-" caso estes contenham alguma palavra separada em duas linhas).

2.2.3. Base de Nomes de Autores

Esta opção da ferramenta, apresentada na Figura 8, disponibiliza uma base de nomes de autores, que é utilizada pela ferramenta para extrair a autoria de um artigo científico. Esta opção encontra-se no menu `Options`. Nesta tela, o usuário pode incluir ou excluir nomes de autores, e realizar buscas nomes dos mesmos.

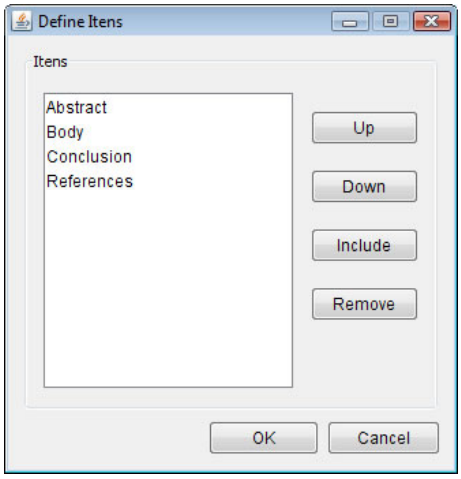

Figura 5. IESystem: Tela de edição de elementos de um modelo.

Figura 6. IESystem: Tela de edição de itens do dicionário.

2.2.4. Detecção de Nomes de Autores

Esta tela da ferramenta, apresentada na Figura 9, exibe os possíveis nomes de autores encontrados durante o processo de extração de metadados, exibidos na lista que se encontra ao lado esquerdo da janela. O sistema só irá cadastrar os nomes que constarem na lista que se encontra ao lado direito da tela. Para gerenciar a lista de nomes encontrados e a lista de nomes que serão cadastrados, a ferramenta disponibiliza ao usuário botões para adicinar um nome por vez (>), todos os nomes de uma vez (»), remover um nome por vez (<) e remover todos os nomes de uma vez («).

Sempre que é feita a inclusão de um nome após ter sido realizado o processo de extração de metadados, o sistema emite um aviso alertando o usuário para refazer o processo de extração agora com o nomes inseridos na base.

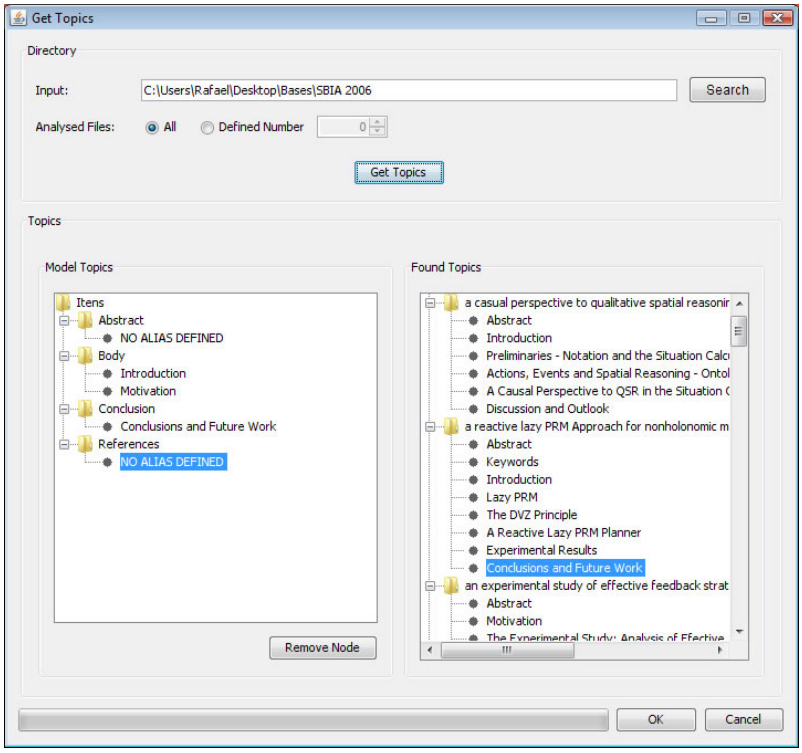

Figura 7. IESystem: Tela de detecção de descritores de seções.

2.2.5. Conversor e Filtros

Esta opção da ferramenta, apresentada nas Figuras 10 e 11, encontra-se no menu `Options`. O usuário pode configurar o diretório de destino dos arquivos convertidos (formato *pdf* para o formato *txt*) e filtrados, bem como as opções de filtro. Vale ressaltar que a IESystem aplica filtros tanto nos arquivos no formato *pdf* quanto nos arquivos já no formato *txt*. As opções de filtros da ferramenta são:

- *Remoção de números*: para o processo de Mineração de Textos pode ser desnecessário a presença de números nos documentos base textual.
- *Conversão para letras minúsculas*: algumas ferramentas podem fazer distinção entre letras maiúsculas e minúsculas. Esta opção serve para fazer uma padronização para que palavras escritas em caixas diferentes sejam reconhecidas igualmente.
- *Remoção de caracteres não alfa-numéricos*: remoção de caracteres dos textos que podem não ser necessários para o processo de Mineração de Textos.
- *Remoção de pontuação*: remoção de caracteres de pontuação que podem não ser necessários para o processo de Mineração de Textos.

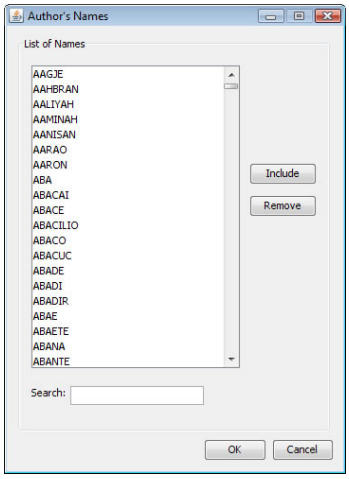

Figura 8. IESystem: Tela para inclusão, remoção e busca de nomes gravados na base de nomes de autores.

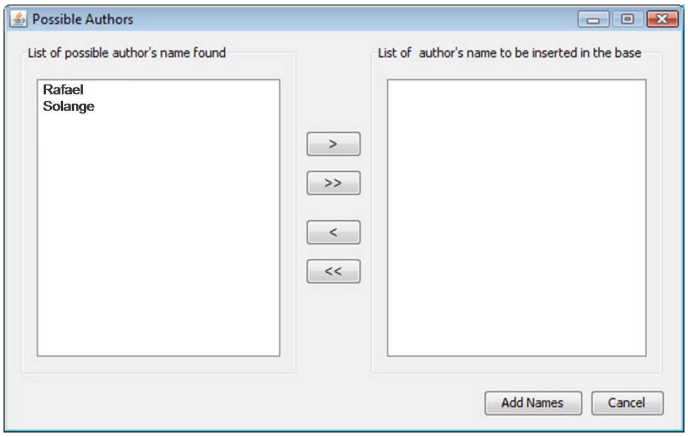

Figura 9. IESystem: Tela para adição de nomes de autores encontrados duranto o processo de extração de metadados.

- *Mínimo de caracteres por linha*: no processo de conversão de *pdf* para *txt*, o conversor separa cada item de uma fórmula matemática, ou números sobrescritos, e os colocam em uma única linha. Esta opção serve para minimizar a ocorrência de caracteres de fórmulas matemáticas que podem não ser necessários para o processo de Mineração de Textos.

  Vale ressaltar que por padrão a ferramenta IESystem aplica a remoção de acentos automaticamente, pois, a ferramenta utiliza expressões regulares para a extração de metadados, e estas não se comportam bem em *strings* que contém caraceres com acento.

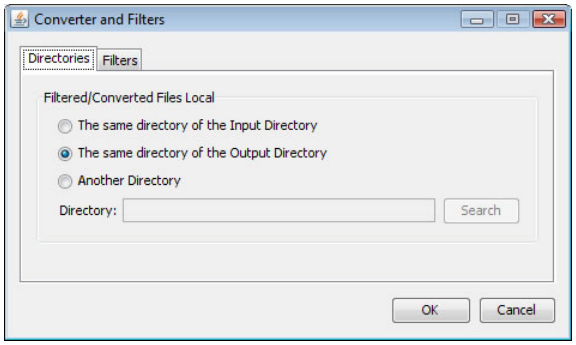

Figura 10. IESystem: Tela de configuração do diretório dos arquivos convertidos e filtrados.

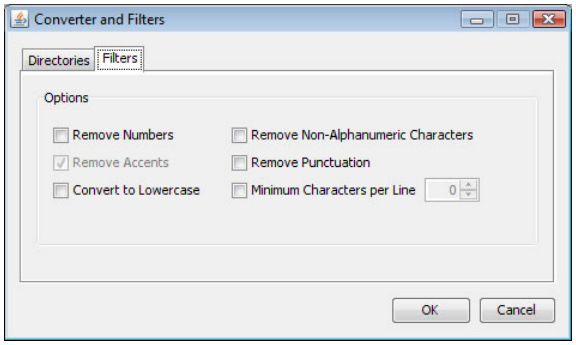

Figura 11. IESystem: Tela de configuração dos filtros.

2.2.6. Editor de *Stopwords* de Domínio

Esta opção da ferramenta, apresentada na Figura 12, encontra-se no menu `Options`. O usuário pode definir palavras, frases, cabeçalhos, rodapés, e marcas de conferências para serem eliminadas dos metadados extraídos dos artigos científicos. O usuário também pode definir se a remoção das *stopwords* serão aplicadas a todos os arquivos da coleção de artigos científicos, ou a um arquivo específico. A ferramenta disponibiliza as seguintes opções de *stopwords*:

- *Words:* remoção de palavras simples. A ferramenta separa as palavras com espaço ou vírgula e as tratam como palavras simples. São removidas todas as ocorrências dessas palavras nos arquivos especificados;
- *String:* remoção de cadeias de caracteres. São removidas todas as ocorrências exatas das cadeias de caracteres dos arquivos especificados;
- *Header and Footer:* remoção de cabeçalhos e rodapés presentes nos artigos científicos. São removidas cadeias de caracteres que se repetem segundo um padrão nos artigos científicos;

- *Brand Conference:* remoção de marcas de conferência[3]. São removidas cadeias de caracteres dos artigos científicos. A diferença nesta situação é que, ao invez de procurar várias ocorrências da cadeia de caracteres nos arquivos especificados, procura-se por apenas uma ocorrência, uma vez que dificilmente uma marca de conferência aparece mais de uma vez em um artigo científico.

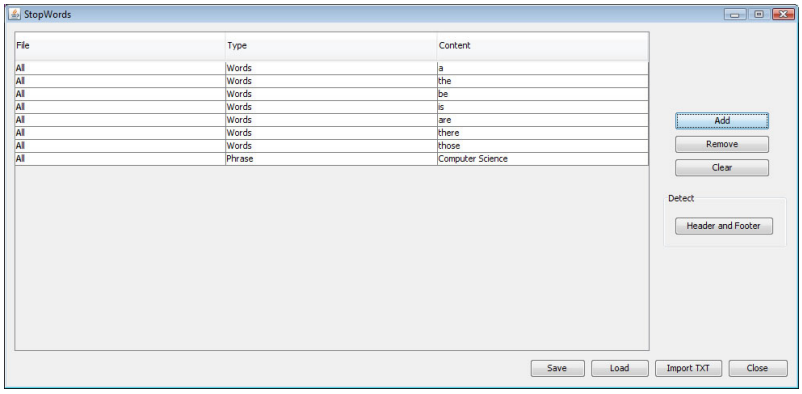

Figura 12. IESystem: Tela de edição de *stopwords*.

Para criar um arquivo contendo as *stopwords* que serão utilizadas no processo de extração de metadados, o usuário pode adicionar manualmente as palavras utilizando o botão `Add`, ou utilizando o detector de cabeçalhos e rodapés, por meio do botão `Header and Footer`, ou ainda por meio de um arquivo texto, onde cada linha do arquivo corresponde a uma *stopword*.

Na Figura 13 é apresentada a tela para a adição manual de *stopwords*. Nesta tela, o usuário pode definir o tipo de *stopword* (*word*, *string*, *header and footer*, *brand conference*), e especificar qual arquivo serão removidas as *stopwords* (se um único arquivos ou todos os arquivos da coleção).

Caso o usuário possua um arquivo texto contendo as *stopwords*, este pode ser carregado pela ferramenta através do botão `Load`. Este arquivo deve estar no formato *txt* e cada linha deste arquivo deve conter uma *stopword*.

Para remover os cabeçalhos e rodapés dos artigos científicos, o usuário pode optar por utilizar o detector de cabeçalhos e rodapés, por meio do botão `Header and Footer`. Ao acessar esta opção, a tela apre-

[3]Neste trabalho são consideradas marcas de conferência as *strings* contidas nos artigos científicos referente aos nomes de conferência e informações sobre a publicação.

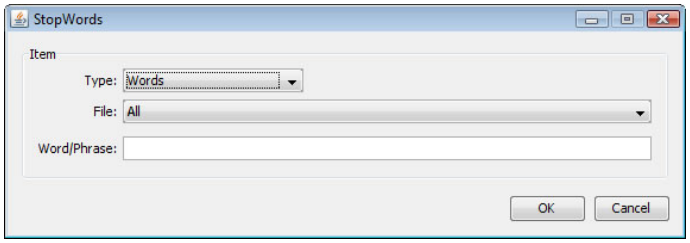

Figura 13. IESystem: Tela para adição manual de *stopwords*.

sentada na Figura 14 será exibida ao usuário. Nesta tela, o usuário deve definir o diretório onde se encontram os artigos científicos que serão submetidos ao processo de extração de metadados e clicar no botão `Detect Header and Footer`. Após realizar a detecção de cabeçalhos e rodapés dos artigos científicos, a ferramenta exibe no canto inferior esquerdo da tela apresentada na Figura 14, os arquivos analisados e os cabeçalhos e rodapés detectados nesses arquivos. O usuário pode então selecionar o cabeçalho/rodapé de um arquivo desejado e clicar no botão >, para que este cabeçalho/rodapé seja inserido na tabela localizada no canto inferior direito da tela apresentada na Figura 14, que corresponde aos cabeçalhos e rodapés que serão inseridos na lista de *stopwords*. Caso o usuário deseje inserir todos os cabeçalhos e rodapés detectados, basta clicar no botão >>. Se o usuário desejar remover um cabeçalho/rodapé específico, ou remover todos, basta clicar nos botões < e << respectivamente.

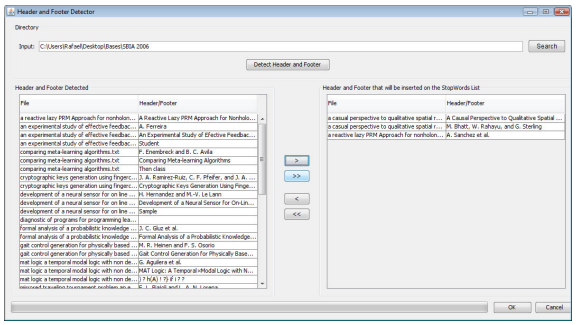

Figura 14. IESystem: Tela para detecção de cabeçalhos e rodapés de artigos científicos.

2.2.7. Ajuda

A ajuda da ferramenta consiste de uma página *html* contendo uma explicação dos conceitos, funcionamentos e informações sobre as versões da ferramenta. O IESystem utiliza o navegador padrão do sistema operacional para exibir o conteúdo da ajuda. Este conteúdo contém os seguintes

tópicos:

- *About IESystem*: uma breve descrição da ferramenta, características técnicas e os autores das versões da ferramenta;
- *Models*: uma explicação sobre os modelos e seus componentes;
- *Configurations*: explicação das opções encontradas na tela de configurações da ferramenta;
- *Author's Name*: explicação sobre a utilidade da base e nomes de autores e como modificá-la;
- *Output Types*: explicação dos tipos de arquivos que podem ser gerados pelo processo de extração de metadados dos artigos científicos;
- *How it works?*: explicação dos passos e dos arquivos gerados durante todo o processo de extração de metadados dos artigos científicos;
- *Result*: explicação dos itens que aparecem na tela de resultados.

2.3. Arquivos de Saída Gerados pelo IESystem

O IESystem gera arquivos para seu próprio controle e também para os resultados obtidos no processo de extração de metadados.

Para iniciar o processo de extração de metadados, o IESystem gera o arquivo `IESFiles.txt`, que contém o caminho de todos os artigos científicos que serão submetidos ao processo de extração de metadados. Uma vez gerado esse arquivo, a ferramenta submete o modelo mais específico[4] inserido pelo usuário, e tenta extrair os metadados dos arquivos contidos no arquivo `IESFiles.txt` de acordo com esse modelo. Se o arquivo apresenta os elementos descritos no modelo, este é retirado do arquivo `IESFiles.txt`, caso contrário, permanece no mesmo. Para os endereços de arquivos restantes no `IESFiles.txt`, a ferramenta submete o segundo modelo mais específico, e repete novamente o processo de extração de metadados. Novamente, os artigos científicos que não apresentaram os elementos descritos no modelo permanecem no arquivo `IESFiles.txt`. Esse procedimento é realizado até que todos os modelos inseridos na ferramenta sejam submetidos. Ao final do processo restarão no arquivo `IESFiles.txt` somente os arquivos que não possuem os elementos de nenhum dos modelos fornecidos pelo usuário, ou seja, os arquivos que não foram extraídos os metadados.

[4] O modelo mais específico é o que possui uma maior quantidade de elementos.

Durante o processo são criados mais 3 arquivos: `IESLog.txt`, `IESTakeLook.txt` e `IESPossibleAuthors.txt`. O arquivo `IESLog.txt` contém os erros de cada arquivo submetido ao processo de extração de metadados para cada modelo fornecido pelo usuário. Esse arquivo tem a seguinte estrutura:

```
<Article>
<File>Caminho do arquivo</File>
<Model>Nome do modelo</Model>
<Error>Descrição do erro encontrado</Error>
⋮
<Error>Descrição do erro encontrado</Error>
</Article>
<Article>
⋮
</Article>
```

O arquivo `IESTakeLook.txt` contém uma lista de arquivos recomendados para que o usuário visualize manualmente. Essas indicações são obtidas pelo Módulo de Extração, que é o núcleo do extrator. Neste módulo há uma série de expressões regulares desenvolvidas baseadas em um estudo de diversos formatos de artigos científicos. Caso a ferramenta não consiga extrair os elementos por meio de suas expressões regulares mais específicas, a ferramenta faz uso de expressões regulares mais abrangentes. Por exemplo, a primeira expressão regular utilizada pela ferramenta, tenta casar o texto sendo analisado à uma string que contém digitos e pontos, ou numerais romanos precedendo uma quebra de linha, uma sequência de caracteres referentes a um descritor de seção informado pelo usuário, uma sequência de caracteres qualquer, até encontrar novamente uma sequencia de digitos e pontos ou numerais romando que precedem uma quebra de linha e que em sequencia apresente uma sequencia caracteres referente a um descritor de seção também informado pelo usuário. Caso o sistema não consiga casar a expressão regular com o texto analisado, o sistema tenta casar outras expressões regulares menos específicas, como uma sequencia de caracteres informado pelo usuário precedendo uma quebra de linha, uma sequência de caracteres qualquer, e uma outra sequecia de caracteres

correspondente a um descritor de seção informado pelo usuário. Alguns exemplos de expressões regulares utilizadas pela ferramenta IESystem são apresentados no Apêndice A.

Quando o extrator utiliza expressões regulares mais abrangentes para extrair os metadados dos artigos científicos, o arquivo que está sendo extraído é recomendado para que o usuário olhe manualmente. O conteúdo do arquivo `IESTakeLook.txt` consiste de um caminho completo para o arquivo que está sendo indicado.

O arquivo `IESPossibleAuthors.txt` contém a lista de possíveis nomes de autores encontrados durante o processo de extração que não constam na base de nomes da ferramenta. Esses possíveis nomes são indicados se há uma *string* encontrada entre o nome localizado na base de nomes de autores e o título.

Quanto aos arquivos que conterão os metadados extraídos, o usuário tem duas opções:

- `One File`: todos os metadados extraídos serão gravados em um único arquivo, cujo nome será composto pelo nome do modelo seguido de `_One_XML_Out.xml`, e terá o seguinte formato:

  ```
  <?xml version="1.0"encoding="ISO-8859-1"?>
  <Database>

  <Article>
  <file_name>
  Nome do arquivo usado no processo de extração
  </file_name>
  <Title>
  Título do artigo científico
  </Title>
  <Authorship>
  Autores e informações dos autores
  </Authorship>
  <Primeiro item do modelo>
  Conteúdo do primeiro item do modelo
  </Primeiro item do modelo>
  ⋮
  ```

<Enésimo item do modelo>
Conteúdo do enésimo item do modelo
</Enésimo item do modelo>
</Article>
<Article>
⋮
</Article>

</Database>

- `Multiple Files`: para cada artigo científico analisado e que tenha seus elementos extraídos sem erro, será criado um arquivo de metadados correspondente. O nome dos arquivos gerados serão compostos pelo nome do modelo ao qual o extrator conseguiu extrair os elementos, seguido de um "_", e seguido do nome do arquivo original sem extensão. Os arquivos gerados por esta opção terão o seguinte formato:

<?xml version="1.0"encoding="ISO-8859-1"?>
<Article>
<Title>
Título do artigo científico
</Title>
<Authorship> Autores e informações dos autores
</Authorship> <Primeiro item do modelo>
Conteúdo do primeiro item do modelo
</Primeiro item do modelo>
⋮
<Enésimo item do modelo>
Conteúdo do enésimo item do modelo
</Enésimo item do modelo>
</Article>

3. Descrição dos Módulos da ferramenta IESystem

Conforme mencionado anteriormente, a ferramenta IESystem foi desenvolvida utilizando as linguagens Java e Perl. Na primeira linguagem, foi

desenvolvida a interface com o usuário, o gerenciamento dos modelos, e o gerenciamento dos arquivos produzidos pelo processo de extração. Este conjunto de funcionalidades desenvolvido em Java é denominado Módulo de Gerenciamento. Na segunda linguagem, foi desenvolvida o *parser* para extrair os metadados dos artigos científicos, o detector de descritores de seções, e o detector de cabeçalhos e rodapés. Este conjunto de funcionalidades é denominado Módulo de Extração.

A seguir serão descritos o Módulo de Gerenciamento e o Módulo de Extração.

3.1. Módulo de Gerenciamento

O módulo de gerenciamento provê a interface com o usuário, o gerenciamento dos modelos, da base de nomes de autores, e dos resultados da extração. Na Figura 15 é apresentado o diagrama de classes do Módulo de Gerenciamento. Este módulo foi desenvolvido utilizando a linguagem Java devido sua portabilidade, fácil desenvolvimento de interface gráfica e comunicação com outras linguagens de programação.

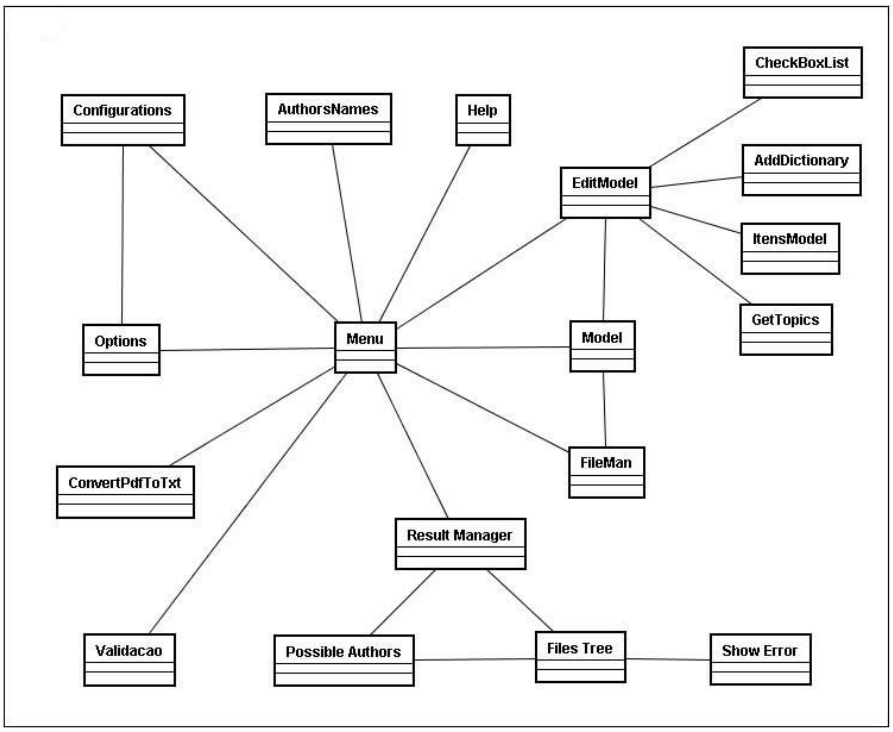

Figura 15. Diagrama de Classes do Módulo de Gerenciamento.

A seguir serão descritas as classes presentes no Módulo de Gerenciamento:

- `AddDictionary`: A classe `AddDictionary` estende a classe javax.swing.JFrame e é responsável por adicionar itens ao dicionário referente ao modelo que foi passado para esta classe.
- `AuthorsNames`: A classe `AuthorsNames` estende a classe javax.swing.JFrame e é responsável por gerenciar a base de nomes de autores. Esta base está armazenada no arquivo `namebaseies.dat` no diretório corrente da aplicação.
- `CheckBoxList`: A classe `CheckBoxList` estende a classe javax.swing.JList e é responsável por criar as caixas de seleção ao lado da lista dos metadados que serão gravados durante o processo de extração.
- `Configurations`: A classe `Configurations` estende a classe javax.swing.JFrame e é responsável por exibir as opções da ferramenta e associá-las ao objeto `Options`.
- `ConvertPdfToTxT`: A classe `ConvertPdfToTxT` é responsável por converter os arquivos que estão no formato *pdf* localizado no diretório de entrada, para o diretório de destino definido em um objeto que instancia `Options`.
- `EditModel`: A classe `EditModel` estende a classe javax.swing.JFrame e é responsável por associar os itens disponibilizados na tela com um objeto que instancia a classe `Model`. Esta classe também permite salvar o objeto que instancia a classe `Modelo` que está sendo editado.
- `FileMan`: A classe `FileMan` provê métodos estáticos para o gerenciamento de Modelos em disco e telas para escolher diretórios, abrir e salvar modelos.
- `FilesTree`: A classe `FilesTree` estende a classe javax.swing.JFrame e é responsável por exibir os resultados do processo de extração de metadados.
- `FilesDescriptors`: A classe `FilesTree` estende a classe javax.swing.JFrame e é responsável por exibir uma árvore contendo os descritores de seções detectados dos artigos científicos, e permitir que o usuário edite o modelo corrente facilmente clicando e arrastando os descritores.
- `Help`: A classe `Help` é responsável por abrir o navegar padrão do dispositivo que estiver rodando a ferramenta e carregar o arquivo `IEShelp2.html`.
- `ItensModel`: A classe `ItensModel` estende a classe

javax.swing.JFrame e é responsável por gerenciar os elementos que o usuário deseja extrair ao modelo corrente.

- `Menu`: A classe `Menu` estende a classe javax.swing.JFrame e é responsável por exibir a tela inicial ao usuário, possibilitar que este acesse as opções do sistema, gerenciar os modelos e preparar os arquivos para o processo de extração.
- `Model`: A classe model implementa a interface `Serializable` para possibilitar a gravação do objeto em um arquivo. Esta classe contém os elementos que serão usados para extrair os metadados dos artigos científicos.
- `Options`: A classe `Options` contém as atributos para armazenar as opções selecionadas pelo usuário referentes ao destino dos arquivos convertidos/filtrados e as opções de filtros.
- `PossibleAuthors`: A classe `PossibleAuthors` estende a classe javax.swing.JFrame e é responsável por exibir os possíveis nomes de autores que foram encontrados durante o processo de extração e gerenciar a inclusão desses nomes na base de nomes de autores.
- `ResultManager`: A classe `ResultManager` é responsável por gerenciar os resultados obtidos no processo de extração, instanciando objetos da classe `PossibleAuthors` e `FilesTree`.
- `ShowError`: A classe `ShowError` e responsável por exibir os erros dos artigos contidos no nó `Errors` na árvore de resultados.

3.2. Módulo de Extração

O Módulo de Extração é constituído por nove *scripts*: `HFDetector.pl`, `HFDetectorL.pl`, `HFDetectorR.pl`, `parser.pl`, `SWRemoverC.pl`, `SWRemoverHF.pl`, `SWRemoverP.pl`, `SWRemoverW.pl`, e `topics.pl`.

Os *scripts* `HFDetector.pl`, `HFDetectorL.pl`, e `HFDetectorR.pl` são responsáveis por detectar os cabeçalhos e rodapés dos artigos científicos. Esses *scripts* não recebem nenhum parâmetro, pois analisam os arquivos contidos na pasta `temp` da ferramenta. Os *scripts* `SWRemoverC.pl`, `SWRemoverHF.pl`, `SWRemoverP.pl`, e `SWRemoverW.pl` são responsáveis pela remoção das *stopwords*. Esses scripts recebem como parâmetro: o nome do arquivo, as *stopwords* e o diretório de entrada, caso o usuário opte por passar a stopword em todos os arquivos.

O *script* `parser.pl` é responsável por extrair os elementos dos artigos científicos, gravá-los de acordo com a preferência do usuário e fazer o gerenciamento de erros. Os parâmetros passados para esse *script* são:

- *Elementos a serem extraídos:* elementos que se desejam extrair dos textos, excluindo `Título` e `Autoria`. Aqui há duas possibilidade de sintaxe para estes itens. Caso a opção de usar o dicionário esteja habilitada, a sintaxe é a seguinte:
  `[s|n]»[Elemento 1](Significado 1|...|Significado L) ...`
  `[s|n]»[Elemento N](Significado 1|...|Significado M)`
  na qual `[s|n]` significa gravar aquele metadado extraído deve ser gravado no arquivo ou não (`s` para extrair e `n` para não extrair.)
  Ex: `s»Body(Introduction|Motivation|Body).`
  Caso a opção de utilizar o dicionário não esteja habilitada a sintaxe fixa mais simples: `[s|n]»[Item]`;
  Ex: `s»Abstract`
- *Diretório de entrada*: caminho completo do diretório de entrada;
- *Diretório de saída*: caminho completo do diretório de saída;
- *Número de arquivos gerados:* pode ser `only-one-xml`, para um único arquivo contendo o resultado da extração para todos os arquivos de entrada, e `multiple-xml-files`, para um arquivo contendo o resultado da extração para cada arquivo de entrada;
- *Gravar o título*: há duas possibilidades: `yt`, para gravar o título extraído nos arquivos resultantes da extração, e `nt`, para não gravar o título;
- *Gravar a autoria*: há duas possibilidades: `ya`, para gravar autoria extraída nos arquivos resultantes da extração, e `nt`, para não gravar a autoria;
- *Usar dicionário*: há duas possibilidades: `s` indica para o *script* que os itens informados ao sistema contém outros significados, e `n` para informar ao sistema que os itens informados ao sistema não contém uma lista de significados;
- *Nome do modelo*: este nome será usado para compor o nome dos arquivos resultantes do processo de extração.

Exemplo completo da linha de comando para executar o *script* `parser.pl`:

> perl parser.pl s>>Resumo#(Abstract|Resumo) s>>Body(Introduction|Introducao|Body) s>>Conclusao#(Conclusoes e Trabalhos Futuros|Conclusoes e Recomendacoes|Consideracoes Finais|Conclusoes|Conclusao) s>>Referencias#(Referencias|References) C:\Users\Rafael\Desktop\Testes\Teste1 C:\Users\Rafael\Desktop\Testes\Teste2 multiple-xml-files yt ya s Exemplo xml

O *script* `topics.pl` visa extrair os tópicos da base de artigos científicos para auxiliar o usuário no processo de construção de modelos. Esse *script* não recebe nenhum parâmetro, e analisa os arquivos contidos dentro da pasta `temp` do diretório corrente da aplicação, que é criado automaticamente pelo `IESystem` e contém os artigos científicos convertidos e filtrados.

4. Utilidades do IESystem

Os metadados extraídos dos artigos científicos podem ser utilizados para diversas finalidades, a saber:

- *Utilizar somente algumas partes em determinados processos*: pode-se extrair características relevantes dos artigos e suas relações de similaridade usando *part-of-speech (POS) tagging* (Álvarez, 2007) na parte das referências bibliográficas dos artigos. Um outro exemplo seria utilizar somente os atributos do `Título` e `Abstract` para a descrição de coleções documentos;
- *Busca*: pode-se realizar busca somente em determinados elementos do texto;
- *Dar peso aos atributos*: durante o processo de Mineração de Textos na metodologia descrita por Moura et al. (2009), pode-se multiplicar os valores de cada atributo na **matriz atributo-valor** gerada no processo de acordo com a parte do texto que este atributo aparece para aquele documento;
- *Eliminar atributos*: pode-se excluir elementos da matriz atributo-valor de acordo com o local que esse atributo aparece no texto;
- *Diferentes Linguas:* algumas ferramentas podem não funcionar corretamente se trabalharem com textos que tenham línguas diferentes em seu conteúdo. Em muitos artigos escritos em português é comum aparecer os elementos `Abstract` e `Resumo`. Com auxílio do IESystem, por exemplo, o elemento `Abstract` poderia ser facilmente eliminado da coleção de artigos científicos;
- *Agrupamento e Classificação:* ao invés de utilizar todos os atributos de toda a base de artigos para o agrupamento/classificação de documentos, pode-se usar apenas os atributos de alguns elementos essenciais como `PalavrasChave` e `Abstract`.

5. Exemplo de Uso

Para ilustrar o uso do sistema foram selecionados 20 arquivos do evento `SBIA 2006` (Simpósio Brasileiro de Inteligência Artificial) em for-

mato *pdf* e escritos na língua inglesa. Foi criada uma pasta chamada `IESystem-Entrada` e os arquivos foram copiados para esta pasta, como ilustrado na Figura 16. Para o diretório de saída foi criado um diretório chamado `IESystem-Saida`.

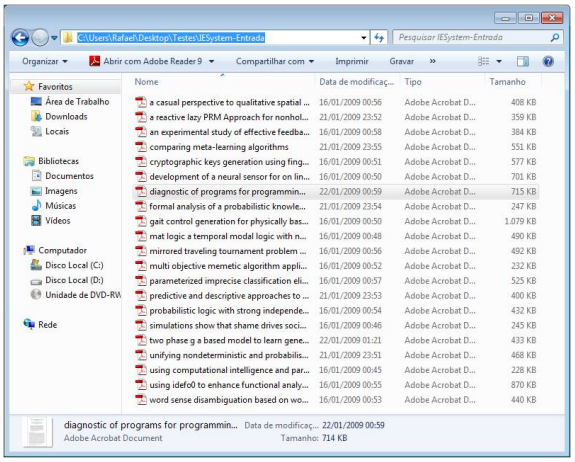

Figura 16. Diretório criado para armazenar os artigos científicos para o exemplo de uso.

Para este exemplo de uso, foi criado um modelo chamado `Exemplo`. Na Figura 17 são apresentados os passos para criar o modelo. Foram escolhidos os seguintes elementos para serem extraídos dos artigos científicos: `Title`, `Authory`, `Abstract`, `Body`, `Conclusion`, e `References`. Para criar um modelo com esses elementos, basta acionar o botão `Create` (Passo 1), que se encontra na tela principal, digitar o nome do modelo e sua descrição nas caixas de texto rotuladas por `Name` e `Description` respectivamente (Passo 2), selecionar os itens `Title` e `Authory` na caixa rotulada por `Basic Itens` (Passo 3), acionar o botão `Add/Rem. Elements`, no painel `Define Elements` e inserir os elementos `Abstract`, `Body`, `Conclusion` e `References` (Passo 4).

Pode-se inserir significados aos elementos definidos no passo anterior de duas maneiras: manualmente, ou por meio do detector de descritores de seções. Na Figura 18 são mostrados os passos para inserir os significados dos elementos manualmente. Primeiramente, deve-se selecionar o elemento que se deseja inserir um significado e acionar o botão `Edit Alias` (Passo 1). Após isso, o usuário deve inserir ou remover os significados do elemento por meio dos botões `Include` e `Remove`, respectivamente (Passos 2 e 3).

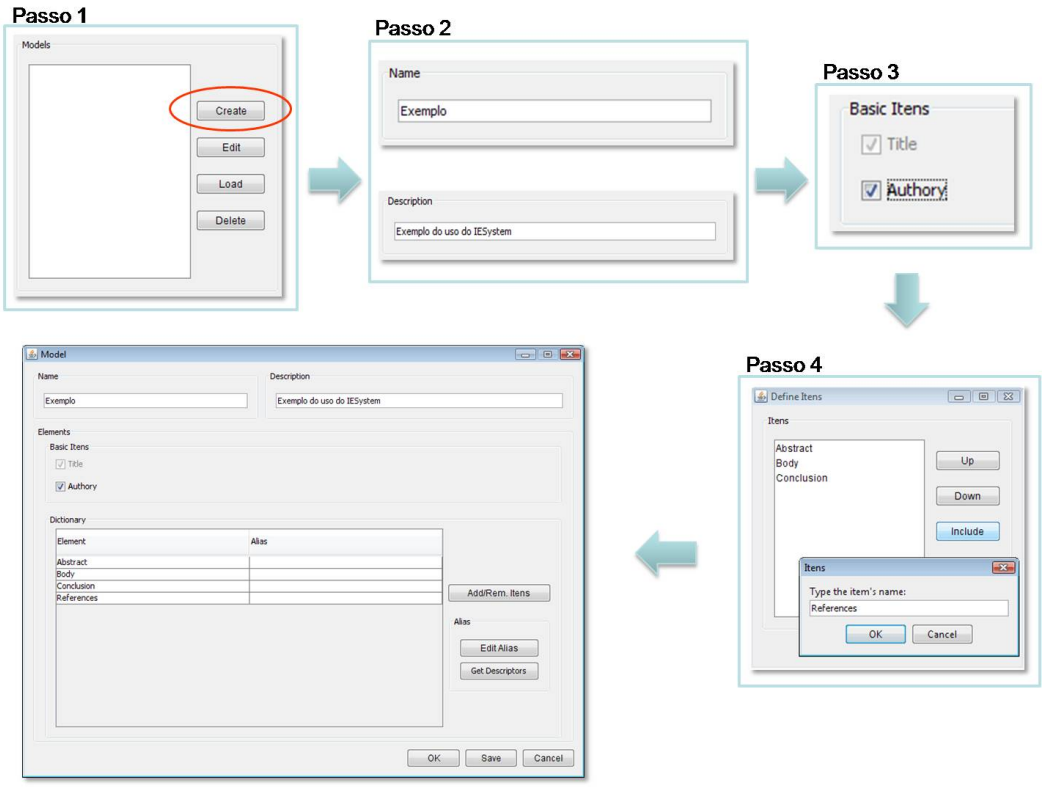

Figura 17. Passos para inserir elementos no modelo.

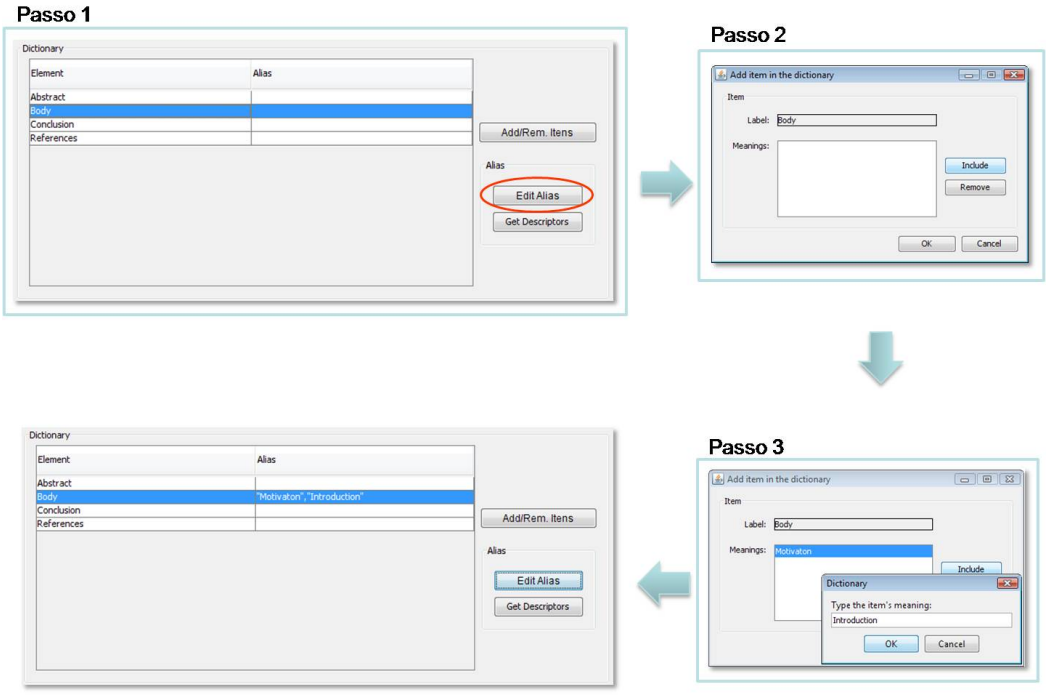

Figura 18. Passos para inserir significados aos elementos do modelo manualmente.

Na Figura 19 são mostrados os passos para inserir os significados dos elementos por meio do do detector de descritores de seções. Para isso, basta acionar o botão `Get Descriptors` (Passo 1), definir o diretório onde se encontram os artigos científicos, e acionar o botão `Detect`. Os artigos científicos presentes no diretório informado pelo usuário são analisados e os descritores de seções detectados são exibidos em um formato de árvore, na qual os ramos são os artigos científicos, e as folhas são os descritores de seções que foram detectados nos respectivos artigos científicos. Após detectados os descritores de seções, basta arrastá-los aos seus respectivos elementos, que se encontram em formado de árvore no canto inferior esquerdo da tela (Passo 2).

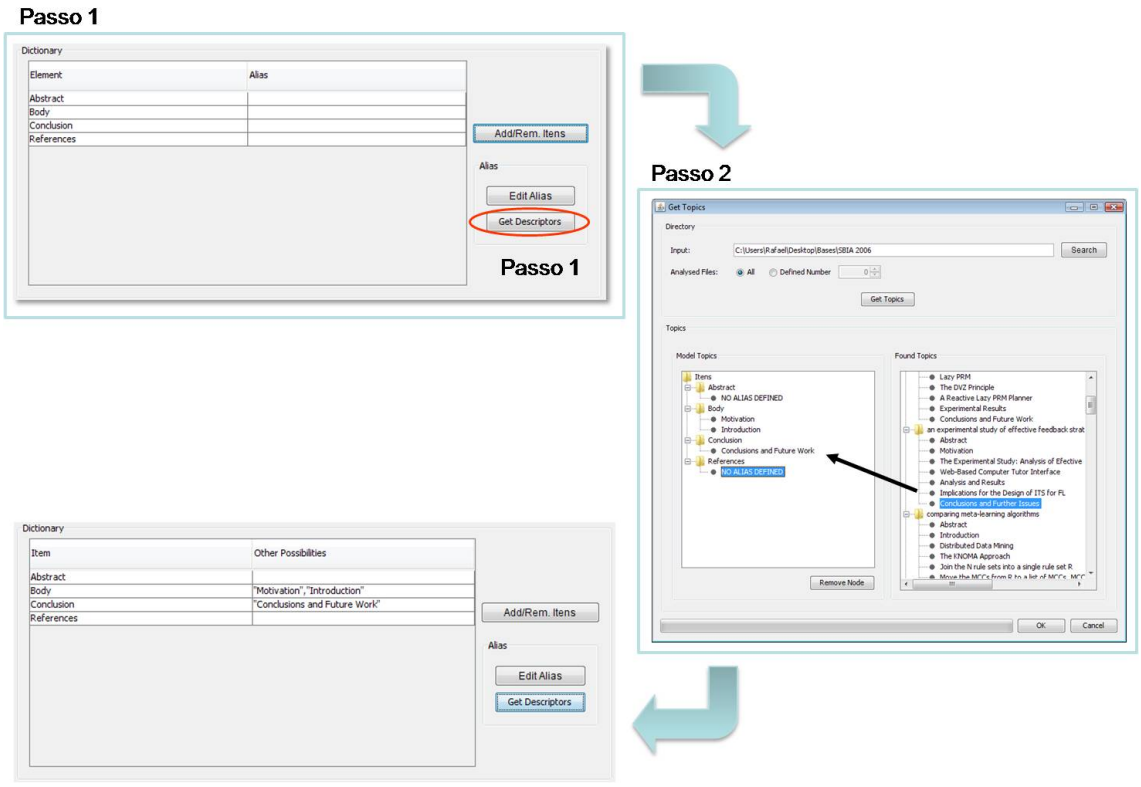

Figura 19. Passos para inserir significados aos elementos do modelo manualmente por meio do detector dos descritores de seções dos artigos científicos.

Para este exemplo, foi construído um modelo analisando apenas 3 documentos da coleção de artigos científicos, e utilizando a opção para detectar os descritores de seções para editar o significado dos elementos. Após esse passo, o modelo resultante ficou como apresentado na Figura 20.

Nome: Exemplo Descrição: Exemplo de Uso 1
Elementos: - Title; - Authory; - Abstract; - Body: *Introduction*; - Conclusion: *Discussion and Outlook, Applications and Future Work*; - References

Figura 20. fig:Exemplo do modelos criado para a desmonstração da ferramenta.

As opções referentes ao formato dos arquivos resultantes do processo de extração de metadados foram `XML`, `Multiple Files`, e `Unparsed Reference`. A tela inicial do sistema apresentada para este exemplo é apresentada na Figura 21.

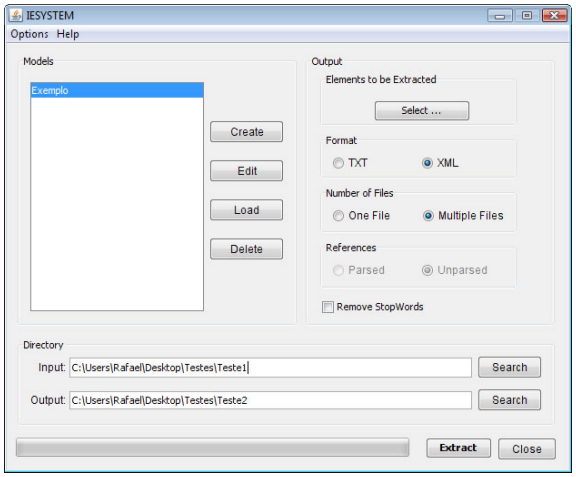

Figura 21. Tela inicial para o exemplo de uso da ferramenta `IESystem`.

Antes de iniciar o processo de extração de metadados, é necessário definir quais metadados devem ser gravados nos arquivos resultantes do processo de extração. Par isso, basta acionar o botão *Select* na tela principal e selecionar os elementos desejados, como ilustrado na Figura 22.

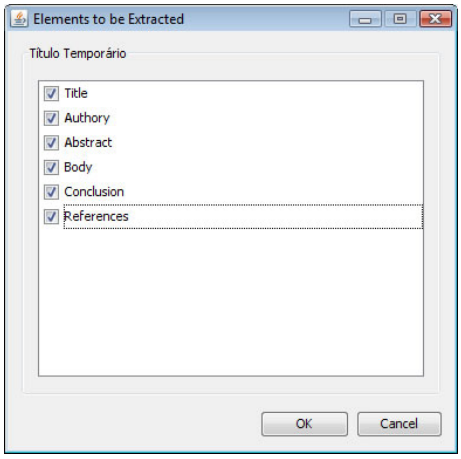

Figura 22. Seleção dos metadados que serão gravados resultantes do processo de extração.

Para exemplificar as situações que possam ocorrer durante a execução do IESystem, foram excluídos propositalmente da base de nomes de autores os nomes `Henry` e `Andre`, pois esses nomes são autores de dois artigos científicos da base. Após acionar o botão `Extract`, o processo de extração inicia-se. A primeira metade da barra de progresso é destinada ao processo de conversão e aplicação de filtros, e a segunda metade da barra de progresso é destinada ao processo de extração de elementos. Se for detectada a ocorrência de um nome de autor que não esteja na base, a tela apresentada na Figura 23 é exibida ao usuário.

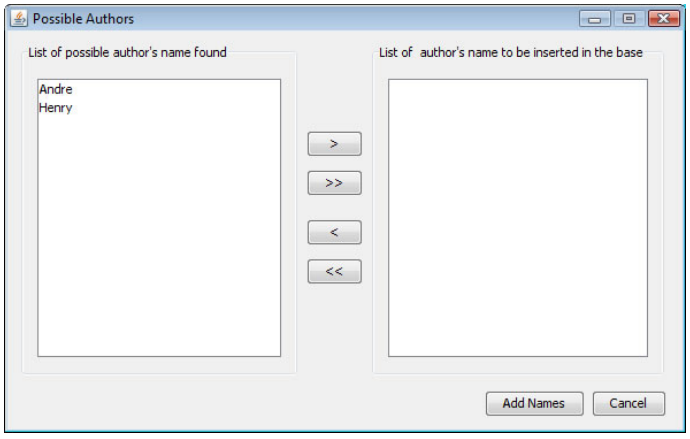

Figura 23. Possíveis autores encontrados no exemplo de uso.

Nota-se que foram detectados como possíveis nomes de autores os nomes excluídos propositalmente da base. Para adicioná-los basta selecioná-los individualmente e acionar o botão >, ou acionar o botão >> para adicionar todos os nomes detectados.

Após o processo de extração de metadados, é exibida ao usuário a tela contendo os resultados do processo, apresentada na Figura 24. Neste exemplo, os itens que apareceram na tela de resultados foram:

- *Exemplo*: nome do modelo utilizado. Os arquivos que foram extraídos baseados nos elementos deste modelo formam os nós folhas do nó `Exemplo`.
- *Errors*: arquivos que não pertencem a nenhum dos modelos fornecidos pelo usuário. As folhas desses nós são os nomes dos arquivos. Ao acessar uma folha deste nó é exibido um botão nomeado `Show the erros with that file` que caso seja acionado exibe um tela informando ao usuário quais elementos do texto o sistema não conseguiu extrair para os modelos usados no processo.
- *Files to take a look*: são indicados os arquivos que que tiveram seus metadados extraídos mas que possivelmente podem conter erros na extração. Automaticamente os arquivos que possuem possíveis autores indicados no passo anterior são inseridos neste nó.

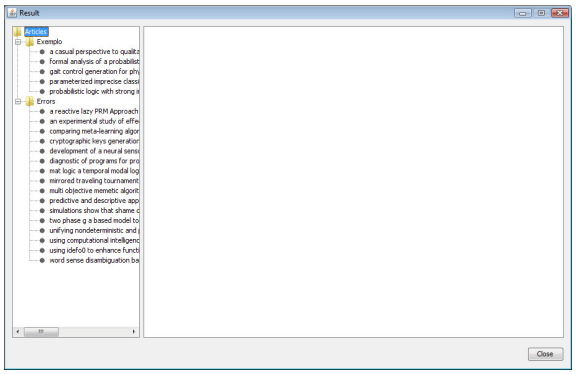

Figura 24. Resultado obtido na primeira iteração do exemplo de uso.

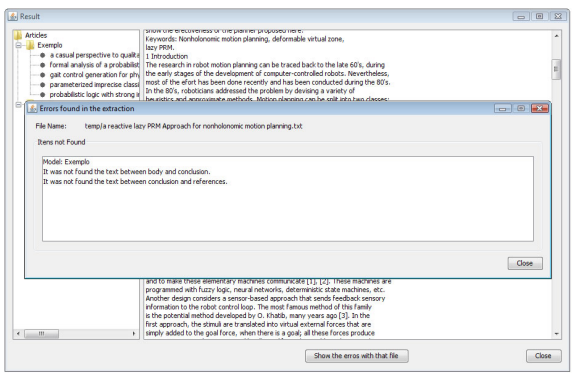

Figura 25. Exemplo de erros que ocorreram durando o processamento do exemplo de uso.

Após inserir os nomes nas bases, verificar os erros, e os arquivos indicados pelo sistema para serem verificados manualmente, o modelo criado para este teste foi atualizado, ficando ilustrado na Figura 26.

Nome: Exemplo
Descrição: Exemplo de Uso 1

Elementos:
- Title;
- Authory;
- Abstract;
- Body: *Introduction*, *Automatic Initialization for Instrumentation Design Studies*, *Motivation*;
- Conclusion: *Discussion and Outlook*, *Conclusions*, *Conclusions and Future Work*, *Applications and Future Works*, *Concluding Remarks*, *Conclusion and Further Issues*, *Conclusions and Remarks*;
- References

Figura 26. fig:Exemplo do modelos criado para a desmonstração da ferramenta.

Após modificar o modelo, o processo de extração foi refeito e o resultado é apresentado na Figura 27.

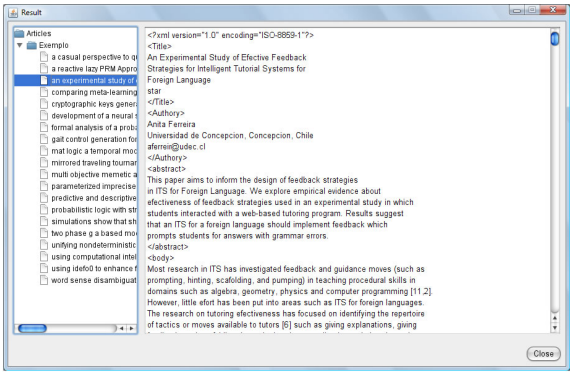

Figura 27. Resultado obtido para o exemplo de uso após a segunda iteração.

Pode-se notar que o sistema não indicou nenhum erro e nenhum arquivo para verificar manualmente.

6. Considerações Finais

As tarefas de organização, gerenciamento e extração de conhecimento de artigos científicos pode ser melhorada com o uso de metadados. Visando

isso, a ferramenta IESystem provê uma série de funcionalidades para que o usuário possa realizar a extração de metadados de artigos científicos. A base para a extração de metadados são os *modelos*, os quais permitem ao usuário definir a(s) estrutura(s) dos artigos científicos presentes em uma coleção, possibilitando assim a extração de metadados de artigos oriundos de diferentes fontes, que possuem diferentes estruturas, e até escritos em diferentes línguas.

Os metadados extraídos utilizando a ferramenta IESYSTEM podem ter diversas utilidades. Em processos de Mineração de Textos, podem ser utilizados termos pertencentes à alguns dos metadados extraídos, ou podem ser atribuídos pesos diferentes a termos que apareçam em determinados metadados. Na Recuperação de Informação, os documentos recuperados podem ser ranqueados de acordo com o metadado em que a *string* de busca fornecida pelo usuário apareça.

Referências

Lopes, A. A., Pinho, R., Paulovich, F., e Minghim, R. (2007). Visual text mining using association rules. *Computers and Graphics*, 31(3):316–326. Citado na página 4.

Álvarez, A. C. (2007). Extração de informação de artigos científicos: uma abordagem baseada em indução de regras de etiquetagem. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação. Citado na página 24.

Moura, M. F., Marcacini, R. M., Nogueira, B. M., da Silva Conrado, M., e Rezende, S. O. (2009). Uma abordagem completa para a construção de taxonomias de tópicos em um domínio. Relatório técnico, Instituto de Ciências Matemáticas e Computação - USP and Embrapa Informática Agropecuária. Citado na página 24.

Nahm, U. Y. e Mooney, R. J. (2000). A mutually beneficial integration of data mining and information extraction. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence(AAAI-2000)*, páginas 627–623. Citado na página 4.

A. Exemplos de Expessões Regulares Utilizadas na Ferramenta IESystem

A seguir são apresentadas algumas das expressões regulares utilizadas na ferramenta IESystem juntamente com suas explicações.

- Expressão regular que casa uma sequência de digitos e pontos que seguem de uma quebra de linha (sequência de 5 espaços), ou uma sequência de algarismos romanos precedidos por um ponto, uma string referente a um elemento de um documento, uma sequência de caracteres qualquer, uma sequência de digitos e pontos, ou uma sequência de algarismos romanos precedidos por um ponto que seguem uma quebra de linha, seguido por uma string referente à um elemento do documento que apareça em sequencia no modelo criado, e uma quebra de linha.

  /(?i) {5,}([[:digit:][:punct:]]+|[IVXLCDM]+?)(*?)$var1(.)( )5,(.*?)( ){5,}([[:digit:][:punct:]]+|[IVXLCDM]+?)(*?)$var2( ){5,}/

- Expressão regular casa uma string referente a um elemento do documento que segue uma quebra de linha, uma sequência de caracteres qualquer, e novamente uma string referente a um elemento do documento que segue uma quebra de linha.

  /(?i){5,$var1( ){5,(.*?)( ){5,$var2( ){5,}/

- Esta expressão regular casa uma string referente a um elemento do documento que segue uma quebra de linha, uma sequência de caracteres qualquer, uma sequência de digitos e pontos, ou uma sequência de algarismos romanos precedidos por um ponto que seguem uma quebra de linha, seguido por uma string referente à um elemento do documento que apareça em sequencia no modelo criado, e uma quebra de linha.

  /(?i){5,}$parsig1(.)(.*?)( ){5,}([[:digit:][:punct:]]+|[IVXLCDM]+?)(*?)$parsig2( ){5,}/

- Expressão regular que casa uma sequência de digitos e pontos, ou uma sequência de algarismos romanos precedidos por um ponto, uma string referente a um elemento de um documento seguido por uma quebra de linha, uma sequência de caracteres qualquer, uma sequência de digitos e pontos, ou uma sequência de algarismos romanos precedidos por um ponto que seguem uma quebra de linha, seguido por uma string referente à um elemento do documento que apareça em sequencia no modelo criado seguido

por uma quebra de linha.

/(?i)([[:digit:][:punct:]]+|[IVXLCDM]+?)(*?)$parsig1(.)( ){5,}(.*?){5,}$parsig2( ){5,/