3
Proposta para Análise de Requisitos utilizando PLN

A proposta aqui apresentada está fundamentada no uso de técnicas de processamento da linguagem natural para apoio às atividades associadas à verificação e validação de requisitos expressos em linguagem natural. Na medida do possível, abstraimos aspectos da implementação de cada estratégia proposta.

Os principais processos desenvolvidos são apresentados brevemente a seguir, e detalhados no decorrer deste capítulo.

Geração de visões dos requisitos: desenvolvemos uma estratégia para extração semi-automática de uma taxonomia a partir do próprio documento de requisitos. A taxonomia obtida é utilizada para agrupar requisitos associados a determinadas características do sistema. Os agrupamentos também podem ser definidos a partir de solicitações dos interessados que, na representação de um certo grupo de usuários do futuro sistema, têm interesse em avaliar o conjunto de requisitos associados a uma propriedade ou característica específica do sistema.

Uma vez obtidos os agrupamentos, visões textuais ou gráficas são oferecidas aos participantes dos processos de verificação e validação (V&V). A visão gráfica possibilita a identificação rápida das associações e dependências entre requisitos; a explicitação de tais dependências é fundamental para atividades de evolução do sistema, principalmente na análise de impacto que uma determinada alteração deverá provocar.

Construção ou atualização do léxico da aplicação: nossa proposta envolve a identificação de termos relevantes do domínio da aplicação, através da análise de documentos relacionados ao sistema em desenvolvimento. A estratégia proposta busca: a) termos próprios do domínio da aplicação, através da identificação de termos não incluídos em dicionários, e b) a identificação de atores e de objetos manipulados no sistema, através de uma estratégia baseada em sintagmas nominais.

Esta estratégia também é útil na reengenharia de sistemas, possibilitando a obtenção de informações a partir de manuais de usuário, de documentos originais

de requisitos ou de solicitações de alteração do sistema.

Detecção de discrepâncias, erros e omissões em requisitos: utilizamos duas abordagens, uma delas visando à identificação de duplicidades em requisitos, e a outra visando à indicação de requisitos não funcionais ausentes do documento de requisitos. A primeira abordagem está baseada em medidas de similaridade entre documentos, e a segunda utiliza a metodologia de análise de conteúdo.

A seguir cada um desses processos é apresentado em detalhe; a arquitetura de suporte a esta proposta incorpora os processos detalhados neste capítulo, e vai além, utilizando agentes pessoais e apoiando o trabalho dos participantes nas atividades de V&V. A arquitetura está detalhada no próximo capítulo.

3.1. Geração de visões dos requisitos

Documentos de requisitos muitas vezes são escritos em linguagem natural. Ali estão registrados características e propriedades importantes do sistema a ser desenvolvido, bem como necessidades e expectativas dos usuários.

O processo de análise de requisitos [Leite94] inclui atividades de verificação e validação. Muitas técnicas foram propostas e estão em uso corrente na indústria para atividades de V&V em documentos de requisitos [Fagan86] [Leite91] [Davis93] [Porter95] [Wilson97] [Shull00]. No entanto, aspectos tais como competitividade e novos paradigmas para o processo de desenvolvimento de software têm gerado maior pressão pela diminuição dos prazos para estas atividades. Pesquisadores tanto na academia quanto na indústria têm buscado técnicas e ferramentas que possibilitem agilizar o processo de desenvolvimento, sem perda da qualidade do produto gerado.

Para sistemas com grande número de requisitos, as atividades de V&V são difíceis de realizar devido ao volume de informações a verificar ou validar. Verificação da incompletude, por exemplo, é uma tarefa árdua se o conjunto de requisitos a avaliar é da ordem de centenas ou mesmo milhares de requisitos. Se conseguirmos trabalhar com grupos menores, a complexidade destas tarefas tende a diminuir. Requisitos não podem ser agrupados ao acaso: precisamos de uma estratégia que nos auxilie a encontrar conjuntos de requisitos que realmente contenham características comuns.

Agrupamentos de requisitos com características similares, ou visões que possam ser associadas a características ou propriedades do sistema tornam-se necessários [Palmer92] [Gruenbacher01]. Diferentes visões do documento de requisitos, geradas a partir de critérios previamente definidos, possibilitam que o participante escolha um conjunto de requisitos para analisar. No contexto deste trabalho, denominamos visão dos requisitos a um conjunto de requisitos com características em comum, apresentado de forma textual ou gráfica. A visão é obtida pela aplicação de uma estratégia que combina técnicas de processamento da linguagem natural e uso de taxonomias.

A necessidade de visões de requisitos é maior em ambientes distribuídos de desenvolvimento, onde os participantes estão geograficamente distantes e sujeitos às dificuldades de comunicação, dentre outras. Diferentes visões de requisitos podem ser atribuídas a diferentes participantes, para os processos de verificação e validação. Acreditamos que a identificação de características relevantes da aplicação, associadas a visões textuais ou gráficas dos requisitos relacionados, possa auxiliar também o trabalho do gerenciamento de requisitos, principalmente em ambientes distribuídos de desenvolvimento. Dificuldades inerentes a este ambiente podem tornar mais trabalhosas as atividades relacionadas ao processo de verificação e validação de requisitos.

A identificação de grupos de requisitos relacionados pode também ser utilizada nas fases posteriores do processo de desenvolvimento, como por exemplo para apoio às tarefas de alocação de requisitos a componentes, ou na definição de casos de testes de aceitação do sistema pelos usuários. Agrupamentos de requisitos relacionados também podem ser utilizados para apoiar a alocação de requisitos a componentes ou mesmo a incrementos, no caso de desenvolvimento incremental.

A construção de visões é fundamentada na identificação de grupos de requisitos relacionados entre si ou a uma determinada característica do sistema. O agrupamento está fundamentado no princípio de que agrupamentos de requisitos relacionados a uma determinada característica do sistema facilitam a identificação de erros relacionados a conflitos, inconsistências, incompletude e ambigüidade em documentos de requisitos, conforme proposto por Palmer e Liang [Palmer92]. Desta forma, a determinação automática de agrupamentos de requisitos associados a características comuns apóia a verificação e a validação, pois permite focalizar

em conjuntos menores e possivelmente menos complexos de requisitos.

3.1.1. Categorização de requisitos e geração de visões

A indústria costuma utilizar documentos de requisitos escritos em linguagem natural, e um dos modelos mais comuns é o que utiliza sentenças numeradas e algumas vezes classificadas segundo uma taxonomia geral (por exemplo: requisitos funcionais, requisitos inversos e requisitos não-funcionais [Leite01]). Outras representações, como casos de uso, estórias do usuário e cenários também utilizam linguagem natural. Nosso desafio inicial está relacionado tanto à definição do número de agrupamentos a identificar, quanto à escolha do processo a adotar para reunir requisitos relacionados. O processo definido deve ainda ser independente dos modelos utilizados para o registro dos requisitos.

Para o número de agrupamentos a utilizar, seguimos um preceito da área de classificação de documentos. É comum que documentos sejam referidos através de palavras, expressões ou frases que caracterizam o tema ou o conteúdo principal de um documento [Baeza-Yates99]. De maneira geral, espera-se que poucos temas sejam suficientes para referir assuntos relevantes num documento, independente do fato da escolha de temas ou conceitos significativos ter sido feita de forma manual ou de forma semi-automática através de ferramentas de manipulação da linguagem natural. Para o nosso trabalho, definimos inicialmente que o número de grupos seria um valor próximo de cinco, conforme discutido na seção 3.1.2.

Um primeiro experimento para gerar esses agrupamentos considerou a identificação de termos significativos no documento de requisitos, através de medidas estatísticas para identificar termos relevantes. Foram utilizadas medidas como *tf* (*term frequency*), *tfidf* (*term frequency inverse document frequency*) e *relevância* [Gonzales05], com resultados similares. Para gerar os agrupamentos, identificamos requisitos relacionados utilizando um procedimento determinístico: para cada um dos termos obtidos pela aplicação das medidas referidas obtinha-se seu *stem* ou radical e varria-se o documento de requisitos procurando pela ocorrência desse radical. Requisitos contendo o radical eram agrupados e esse

grupo identificado pelo termo utilizado.

Os agrupamentos obtidos desta forma mostraram-se consistentes com relação à característica identificada pelo termo, porém um aspecto não atendido por esta abordagem envolveu expressões ou termos relacionados a requisitos não funcionais, como por exemplo termos relativos à segurança e à confiabilidade. Expressões ou termos relacionados a estes requisitos tendem a ter pouca presença no texto, sendo portanto preteridos no momento da escolha dos termos a utilizar. De maneira geral podemos dizer que requisitos não funcionais não foram adequadamente atendidos por esta abordagem, que se mostrou inadequada para os objetivos que buscávamos atingir.

Uma outra tentativa realizada envolveu a utilização de algoritmos de clusterização para a identificação dos grupos de requisitos. Neste experimento foi utilizado o pacote *Weka* disponibilizado pela Universidade de Waikato [Witten00]. Trabalhamos com o *k-means*, um dos mais utilizados algoritmos de particionamento iterativo. Este algoritmo utiliza, como medida de similaridade entre documentos, a distância Euclidiana, que mede a similaridade entre documentos considerando os atributos em comum [Wives04].

Utilizamos o *k-means* aplicando-o a um documento de requisitos de uma aplicação da área do turismo. Este documento passou por um processo de preparação que incluiu a separação dos requisitos em arquivos do tipo texto puro e a geração de vetores termo-documento. Estes vetores registram, para cada documento, a freqüência dos *t* termos definidos para avaliação. Os termos costumam ser aqueles mais relevantes, ou aqueles de maior freqüência.

Para a execução do algoritmo *k-means*, o número de agrupamentos é um dos argumentos fornecidos pelo usuário. Apesar de já termos uma definição inicial para este número, resolvemos pesquisar uma alternativa que pudesse confirmar ou não nossa escolha. Utilizamos então o algoritmo EM (Expectation-Maximization) [Manning99].

Este algoritmo, EM, inicialmente agrupa os documentos em dois grupos e calcula um conjunto de valores que inclui média e desvio padrão dos termos utilizados para representação desses documentos. A implementação utilizada também calcula o coeficiente *LogLikelihood*, que mede a similaridade global. Na segunda iteração o número de grupos é incrementado e os demais procedimentos são repetidos até obter um coeficiente ótimo para os indicadores utilizados. Nesse

momento a execução é encerrada e o número ideal de grupos é fornecido.

No nosso caso, a execução do algoritmo EM para o documento já referido indicou cinco grupos, o que corroborou nossa definição inicial para o número de grupos. Utilizamos como atributos todos os substantivos com freqüência por documento maior ou igual a quatro [Daile96], e uma etapa de pré-processamento gerou um conjunto de vetores que, para cada requisito individual, trazia a freqüência dos termos selecionados. Trabalhamos com substantivos pois eles podem expressar funções gramaticais como sujeito e objeto, funções semânticas como agente e instrumento ou funções retóricas como tópico ou tema [Thrane80].

Definido o número ideal de grupos, continuamos o experimento executando o *k-means* para então obter os agrupamentos de requisitos. O documento utilizado, já referido, era oriundo de uma aplicação na área do turismo e possuía cento e sessenta e nove requisitos. Dos cinco grupos resultantes do processo de clusterização, um apresentava apenas dois requisitos, enquanto outro apresentava cento e cinco. A este último grupo, particularmente, não se conseguiu identificar claramente um tema que perpassasse todos os requisitos do grupo.

Realizamos duas outras tentativas com o *k-means*, modificando, a cada vez, os atributos considerados: ora utilizamos como atributos aqueles termos próximos da média de freqüência no conjunto de documentos [Manning99] [Luhn58], ora limitando o conjunto de atributos aos mais freqüentes. Não obtivemos resultados significativamente melhores que o anterior.

Uma das premissas que nos guiou nesta investigação considera que requisitos podem estar associados a mais de um grupo, ou seja, os agrupamentos a serem obtidos não serão necessariamente disjuntos. Um exemplo claro desta característica resultou da análise dos agrupamentos do documento de requisitos da área do turismo; este documento continha requisitos que deveriam ser manipulados por usuários de diferentes áreas da organização, no processo de V&V. Alguns requisitos deveriam ser avaliados por pessoal da área de atendimento ao cliente, por tratarem de aspectos da venda de um pacote turístico, mas também interessavam ao pessoal da área financeira, por tratarem de questões relacionadas à forma de pagamento utilizada pelo cliente.

Optamos então por investigar alternativas com o uso de taxonomias. Taxonomias têm sido utilizadas para classificação de documentos. Bibliotecas constituem um excelente exemplo de uso de taxonomias para classificação de

diferentes tipos de documentos. Na área de requisitos, encontramos um exemplo do uso de taxonomias na classificação de requisitos organizacionais que impactam no projeto de produtos [Gershenson99].

Nossa estratégia está baseada no uso de uma taxonomia estruturada em dois níveis. Um conjunto mínimo de temas relevantes da aplicação compõe o nível mais alto da taxonomia; esta taxonomia é a base para identificação automática, no documento de requisitos, de termos associados que serão utilizados para a categorização dos requisitos. Estes termos irão compor o segundo nível da taxonomia, e sua seleção combina a descoberta de associações (através da análise do contexto no qual constam os temas do primeiro nível da taxonomia) e a sua avaliação através de medidas estatísticas como o escore T e a informação mútua. Associações são descobertas através de colocações, que são grupos de palavras próximas mas não necessariamente contíguas, que co-ocorrem no texto numa freqüência maior que o esperado pelo acaso [Manning99] [Sardinha04].

A estratégia proposta é esquematizada utilizando uma perspectiva de processo (um actigrama SADT [Ross77]) conforme mostra a Figura 12. Esse processo é composto por 3 atividades: a) identificação de termos relevantes para enriquecer a taxonomia; b) categorização dos requisitos com base na taxonomia enriquecida pelos termos e c) geração das visões dos requisitos.

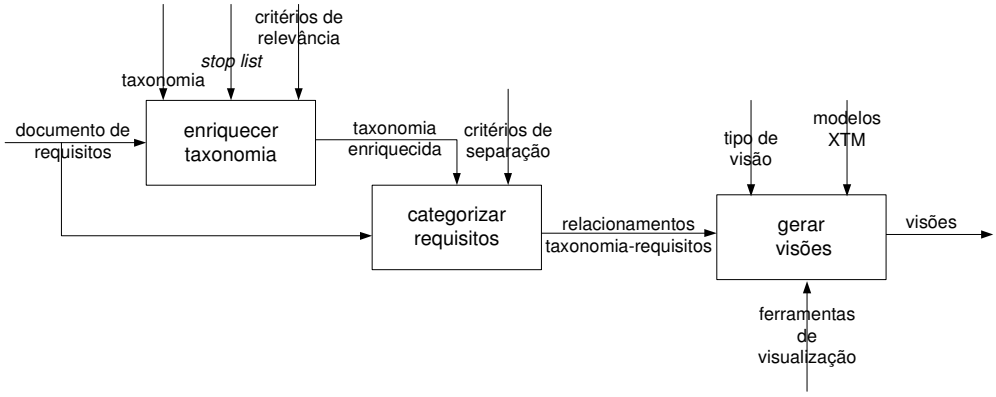


Figura 12 - Visão geral do processo para categorização dos requisitos e geração de visões

A taxonomia enriquecida é utilizada para categorização dos requisitos, gerando agrupamentos de requisitos relacionados. Estes conjuntos de requisitos associados à taxonomia possibilitam diferentes visões dos requisitos aos participantes do processo. Essas visões serão utilizadas nos processos de análise

de requisitos que se quer apoiar. A seguir descreveremos cada uma das etapas do processo proposto.

3.1.2. Identificação de termos para enriquecimento da taxonomia

Como apresentado na Figura 12, o processo pode ser dividido em 3 grandes atividades, ou sub-processos. Primeiro, usando tecnologias de tratamento de linguagem natural, procuramos identificar termos para compor nossa taxonomia. Em seguida procede-se a categorização dos requisitos de acordo com a taxonomia enriquecida e finalmente utilizamos os agrupamentos para compor as visões e apresentá-las tanto textualmente como graficamente.

Os temas que irão compor o primeiro nível da taxonomia definem a quantidade de agrupamentos para os requisitos. Uma referência forte para esta definição refere-se ao número inicial de características para guiar o processo de elicitação proposto pelo método PreView [Sommerville98]. Nesse método, características representam necessidades de alto nível para a aplicação e influenciam os requisitos derivados de cada ponto de vista. Fundamentado no uso de pontos de vista e de características para elicitação de requisitos, o PreView estipula que o número de características deve ser baixo, da ordem de cinco.

Outra referência utilizada consta no trabalho de Hoskinson [Hoskinson05], que trata da extração de conceitos de textos para posterior geração de taxonomias. Hoskinson estabelece que o número de conceitos para os nodos de mais alto nível da taxonomia deve estar no intervalo de três a oito. Adotamos este último critério, pois ele é mais geral que a nossa definição inicial e abarca o proposto pelo método PreView. Acreditamos, porém, que o engenheiro de requisitos deva usar o valor aqui apresentado apenas como guia, avaliando a cobertura dos requisitos em relação aos temas selecionados e trabalhando o número de temas de mais alto nível da taxonomia de forma a atingir cobertura completa do conjunto de requisitos, ou de sua maior parte.

Para a identificação dos termos do segundo nível, consideramos que a frase (ou contexto) que apresenta um tema da taxonomia inicial deve conter outros termos que são relacionados a ele e, portanto, também são relevantes. O processo de busca de contexto procura pelo tema da taxonomia inicial, e retira para análise

uma sub-sentença que é composta pelo tema considerado e pelas palavras imediatamente à sua esquerda e à sua direita. Para cada um dos temas da taxonomia inicial vamos então obter um conjunto de sub-sentenças que será analisado para a extração de colocações, ou de termos para o segundo nível da taxonomia. Este processo pode ser visualizado na Figura 13.

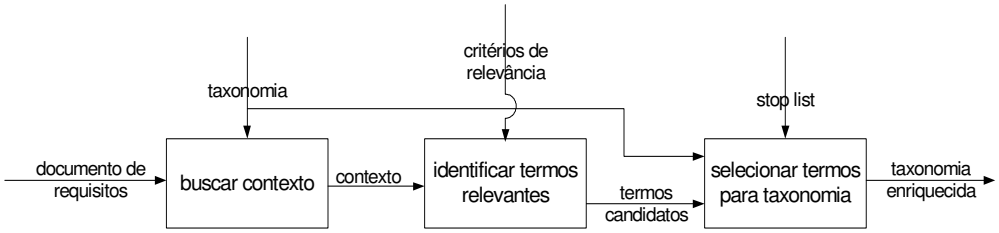


Figura 13- Processo para enriquecer a taxonomia

A busca de contexto trabalha com cada um dos termos da taxonomia inicial, varrendo o texto do documento de requisitos e buscando identificar a presença do termo. Quando isto acontece, é extraída uma sub-sentença composta do termo, das cinco palavras à sua esquerda e das cinco palavras à sua direita. Esta quantidade pode ser menor se, por exemplo, for encontrado o final do parágrafo antes que esse número seja atingido. Procedimento similar é executado em relação ao início do parágrafo. A Figura 14 mostra um exemplo de sub-sentença contendo o termo **reserva**.

a	proposta	e	confirma	a	**reserva**	porém	ainda	não	efetua	nenhum

Figura 14 - Sub-sentença apresentando contexto para o termo "reserva"

Na identificação das sub-sentenças é utilizado o radical da palavra de busca, então para o termo reserva será utilizado o radical **reserv**. As sub-sentenças identificadas compõem um novo documento. Esse documento será analisado buscando a identificação de colocações, que serão avaliadas de acordo com medidas de associação para identificar se a colocação é relevante ou se sua presença no texto é devida ao acaso. Utilizamos como medidas de associação o escore T e a Informação Mútua. A forma de cálculo para o escore T é apresentada na equação (1), e a fórmula para a informação mútua é apresentada na equação (2).

$$T(x,y) = \frac{freq(x,y) - \frac{freq(x) * freq(y)}{N}}{\sqrt{freq(x,y)}} \quad (1)$$

$$im(x,y) = Log_2\left(\frac{P(x,y)}{P(x) * P(y)}\right) \quad (2)$$

Nestas medidas, *x* e *y* representam os termos em análise. Em (1), *freq(x,y)* corresponde à freqüência conjunta de *x* e *y*, *freq(x)* indica a freqüência do termo *x*, *freq(y)* a freqüência do termo *y* e *N* indica quantidade de termos. Em (2), *P(x,y)* indica a probabilidade da ocorrência conjunta de *x* e *y*, *P(x)* e *P(y)* indicam respectivamente a probabilidade de ocorrência do termo *x* ou do termo *y* no documento, e N corresponde ao total de termos do documento.

A informação mútua compara a probabilidade da ocorrência conjunta de x e y com a probabilidade de observar x e y independentemente - ela é dada pela razão entre o valor observado e o esperado. Se houver uma associação genuína entre x e y, a probabilidade conjunta P(x,y) será muito maior do que P(x)*P(y), e conseqüentemente *im(x,y)*≈0. Se não houver nenhum relacionamento interessante entre x e y, então P(x, y)≈P(x)*P(y), e assim, *im(x;y)*≈0. A probabilidade conjunta, P(x,y), é estimada normalizando-se o valor de *freq(x,y)*, ou seja, dividindo *freq(x,y)* por N; as probabilidades P(x) e P(y) são calculadas também normalizando-se as freqüências.

O escore T considera freqüências não normalizadas, e portanto é um indicador absoluto das colocações - indica pares que são por si só muito freqüentes. Valores de informação mútua tendem a ser maiores quando as palavras analisadas aparecem sempre próximas no texto, ou seja, a presença de uma delas é forte indicador da presença da outra (as palavras tem pouca presença independente). Valores maiores do escore T são indicativos de maior número de co-ocorrência das palavras analisadas.

Para que as colocações sejam consideradas significativas é necessário observar o patamar mínimo para cada uma delas. Os valores de corte aceitos e normalmente utilizados para estas medidas de associação são respectivamente três para a informação mútua e dois para o escore T [Sardinha04]. Assim, todas as colocações cujas medidas de associação sejam iguais ou maiores que os patamares definidos são consideradas relevantes para enriquecimento da taxonomia. Na identificação das colocações também trabalhamos com o radical das palavras.

Os termos obtidos neste processo podem ser compostos de até três palavras (trigramas). Esses termos são denominados *termos candidatos*, e são analisados num processo de seleção que desconsidera termos que iniciem ou terminem por palavra relacionada na lista de *stopwords*. Finalmente, as *stopwords* eventualmente presentes nos termos selecionados são também retiradas, gerando

um padrão que permitirá que expressões como *reserva de vôo*, *reserva do vôo*, *reservar o vôo* ou *reservar os vôos* sejam consideradas similares para fins do processo de categorização e correspondam a um mesmo padrão.

A participação do engenheiro de requisitos no processo de definição dos termos a serem incluídos na *stoplist* é muito importante, dado que é provável que termos técnicos da área da TI e mesmo do domínio da aplicação também devam ser desconsiderados no processo de seleção de termos candidatos.

3.1.3. Categorizar requisitos

A identificação dos requisitos relacionados aos termos da taxonomia é centrada no processo de categorização, que utiliza a taxonomia enriquecida. O processo utilizado para categorização é determinístico e trabalha com o reconhecimento de padrões. Consultas por padrões possibilitam a identificação de documentos que atendem a propriedades pré-especificadas; expressões regulares, padrões estendidos e radicais estão entre os tipos de padrão mais utilizados na recuperação de informações [Baeza-Yates99]. A Figura 15 esquematiza o processo de categorização.

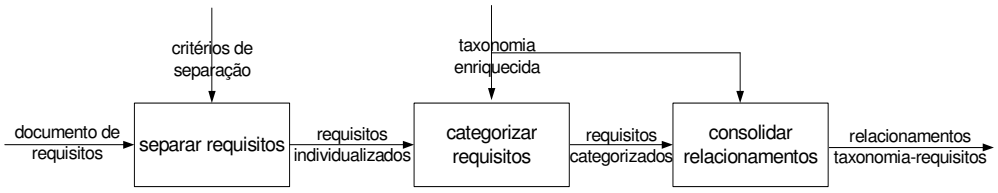


Figura 15 - Processo para categorização de requisitos

Separar requisitos consiste em individualizar requisitos que estão reunidos num único documento. Se não existir uma delimitação clara do conjunto de frases que descrevem um requisito, a própria identificação do requisito pode ser utilizada para esta finalidade. Esta etapa na verdade é uma preparação ao processo de categorização, e pode ser dispensada se os requisitos já estiverem individualizados.

Para categorizar requisitos utilizamos a taxonomia enriquecida, e a base para este processo utiliza os termos associados identificados pelas colocações relevantes (veja seção 2.5.5). Para cada um destes termos associados é aplicada a técnica de reconhecimento de padrões ao texto dos requisitos; é considerado como padrão o radical das palavras que compõem o termo. Se o requisito analisado

atende o padrão, então sua identificação é incluída na lista de relacionamentos para o termo sendo avaliado.

O uso do radical das palavras para compor o padrão propicia a recuperação de termos no singular ou no plural, no masculino ou no feminino, de palavras de diferentes categorias gramaticais e também de expressões similares. Observamos que como o *stemmer* utilizado é dirigido para a língua portuguesa, termos em outras línguas serão mantidos como apresentados, ou seja, na íntegra. Isto deve ser observado pois é comum, em documentos de requisitos, encontramos termos da área de tecnologia da informação que são utilizados na língua inglesa, mesmo que exista uma tradução para a língua portuguesa. Este é o caso de termos como *password*, *logon*, *logoff* e *home page.*

Ao final desta atividade cada termo estará associado a uma lista de requisitos relacionados, e a última etapa consolida as listas correspondentes a cada um dos temas de alto nível da taxonomia, pois um mesmo requisito poderá atender a diversos padrões. A consolidação irá gerar um conjunto de requisitos associados a cada tema de alto nível da taxonomia.

Pelo processo utilizado, esse conjunto reflete aqueles requisitos cuja associação foi determinada pelo fato do texto do requisito conter ao menos um dos padrões considerados relevantes para aquele tema da taxonomia. Como um mesmo requisito poderá estar associado a mais de um ramo da taxonomia, o número de associações será maior que o número de requisitos do sistema. Estas associações, se representadas graficamente, resultarão num grafo onde as interdependências entre requisitos estarão explicitadas. Essa representação é descrita a seguir.

3.1.4. Gerar visões

Diferentes apresentações das visões dos requisitos devem ser propiciadas, para que cada um dos interessados possa escolher aquela mais apropriada às suas necessidades. As visões podem ser apresentadas textualmente ou estruturadas num grafo, possibilitando uma visão mais abrangente de conjuntos de requisitos relacionados. A Figura 16 apresenta o processo para geração das visões.

A atividade de expandir as listas de relacionamentos gera conjuntos <termo-

da-taxonomia> <requisitos relacionados e atributos>. Os atributos a serem utilizados são selecionados dentre aqueles que estão sendo registrados no projeto em desenvolvimento, por exemplo origem do requisito, data de inserção e *rationale*. A descrição dos requisitos é essencial para as visões a serem apresentadas, e portanto ao menos este atributo é utilizado.

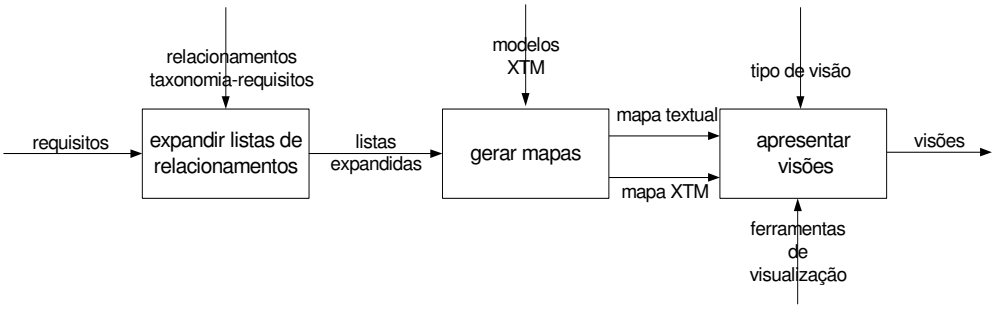


Figura 16 - Processo para geração das visões dos requisitos

As listas expandidas serão utilizadas na geração de mapas que são a base para as visões textuais e gráficas dos requisitos. O mapa textual é estruturado na forma de uma tabela, onde a cada tema da taxonomia são registrados os requisitos associados e seus atributos. O mapa XTM (XML *Topic Maps*) é necessário para a apresentação das visões gráficas e é gerado com utilização dos modelos apresentados logo a seguir.

Visões textuais são apresentadas através de um visualizador para textos, e serão utilizadas em atividades tais como identificação de duplicidades entre requisitos e verificação de completude. Visões gráficas foram construídas porque possibilitam percepção visual das interdependências entre requisitos. Se esta visão permitir ainda a navegação entre os nodos, será possível explorar as associações entre requisitos e entre requisitos e taxonomia.

O modelo que escolhemos para apresentação das visões gráficas está baseado na utilização de *topic maps*, ou mapas de tópicos. Mapas de tópicos podem ser caracterizados como um padrão para descrever estruturas de conhecimento e associá-las a recursos de informação [Pepper00]. As entidades básicas presentes em tais mapas são tópicos de informação, associações e ocorrências. Um tópico pode representar uma entidade ou um objeto; uma associação relaciona dois tópicos, de acordo com um critério, e uma ocorrência pode representar um atributo de tópicos ou de associações.

No contexto deste trabalho, itens da taxonomia e requisitos foram caracterizados como tópicos, e associações foram utilizadas para registrar as ligações de cada termo da taxonomia aos requisitos relacionados. Denominamos relacionado_a à associação entre termos da taxonomia e requisitos associados. No contexto de requisitos, associações também podem ser utilizadas para registro da rastreabilidade e consultadas para análise de impacto em eventos de mudanças: se o item representa, por exemplo, um requisito não funcional, os requisitos funcionais a ele associados também devem ser avaliados quando ocorre uma alteração na política ou procedimentos correlatos ao requisito não-funcional.

Mapas de tópicos podem ser representados com uso do XML, e a linguagem XTM (XML *Topic Maps*) foi criada para facilitar a troca de *Topic Maps* entre aplicações [Pepper01]. Estruturas previamente definidas podem ser utilizadas para definição dos tópicos e suas associações, agilizando a criação dos mapas e a geração das visões. O modelo (*template*) utilizado para os itens da taxonomia pode ser visualizado na Figura 17, o modelo para requisitos é apresentado na Figura 18 e o modelo para associações na Figura 19 (os modelos aqui apresentados estão simplificados).

```
<topic>
  <instanceOf> <topicRef xlink:href="#termos"/> </instanceOf>
  <baseName>
         <baseNameString> termo da taxonomia </baseNameString>
  </baseName>
</topic>
```

Figura 17 - Modelo XTM para termos da taxonomia

```
<topic>
  <instanceOf> <topicRef xlink:href="#req_func"/> </instanceOf>
  <baseName>
         <baseNameString> id_do_requisito </baseNameString>
  </baseName>
</topic>
```

Figura 18 - Modelo XTM para requisitos funcionais

```
<association>
  <instanceOf> <topicRef xlink:href="#relaciona"/> </instanceOf>
  <member>
         <topicRef xlink:href="#id_requisito"/> </member>
  <member>
         <topicRef xlink:href="#id_termo"/>  </member>
</association>
```

Figura 19 - Modelo XTM para associações entre temas e requisitos

A escolha de uso de uma ferramenta de mapas de tópicos para a visualização das associações entre requisitos e a taxonomia deveu-se principalmente às características dinâmicas oferecidas por estas ferramentas. A visualização dos agrupamentos, por si só, já é uma característica importante para a

compreensão das interdependências entre requisitos [Baniassad04] [Gruenbacher01], mas o uso destas ferramentas possibilita também a navegação entre requisitos relacionados. Esta última característica foi apontada como necessária [Dag01], pois possibilita uma forma visual e ágil para a percepção das interdependências entre requisitos.

Exemplo de um mapa de tópicos identificando um conjunto de requisitos associado ao tema pagamento pode ser visualizado na Figura 20.

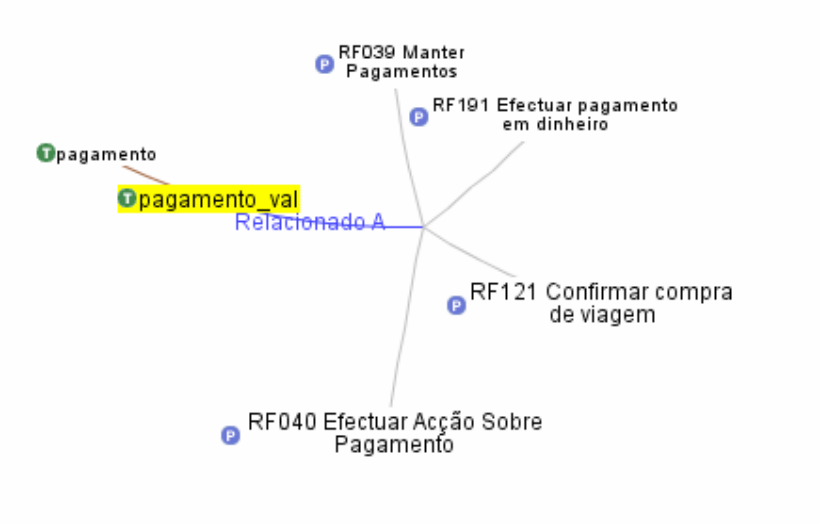


Figura 20 - Mapa visual mostrando requisitos relacionados ao tema **pagamento**

3.2. Construção do léxico da aplicação

Atividades de apoio à construção ou atualização do Universo de Informações são baseadas na identificação de palavras ou expressões próprias do domínio da aplicação. Nosso desafio está relacionado à identificação de símbolos que irão compor o léxico da aplicação e contribuir para a construção do vocabulário do domínio da organização.

Processos de desenvolvimento de software, como por exemplo o RUP (Rational Unified Process) definem glossários como um dos artefatos a serem gerados no processo de requisitos. Um glossário pode ser entendido como uma forma simplificada de léxico, estruturado de forma linear e contendo termos e suas definições. Os termos a serem inseridos num glossário são aqueles utilizados pelos participantes do processo para fazer referências à características da aplicação, visando facilitar o entendimento entre eles. O glossário deve ser

construído durante a elaboração do modelo de negócios ou modelo de domínio. Já o Léxico Ampliado da Linguagem, ou LAL [Leite93], é uma forma mais elaborada de registro de termos próprios do domínio da aplicação, fornecendo mais informações que simplesmente a definição de um termo. Detalharemos o LAL, segundo proposta apresentada em [Leite93].

Termos registrados no LAL são tipificados, e esta é a primeira diferença em relação a um glossário de termos do domínio da aplicação. Termos inseridos no LAL representam símbolos característicos do domínio da aplicação, e correspondem a um de quatro tipos: sujeito, objeto, verbo ou estado. Símbolos do LAL possuem noção e denotação. A noção de um símbolo é o que o define, e a d2enotação registra os impactos que o símbolo provoca ou recebe, no domínio considerado. A Tabela 6 registra os quatro tipos de símbolos e impactos correspondentes.

Tabela 6 - Símbolos do LAL, noção e impactos [Leite90]

Tipo do símbolo	Noção	Impacto
Sujeito	quem é o sujeito	ações que executa
Verbo	quem realiza, quando acontece e quais os procedimentos	quais os reflexos das ações no ambiente e novos estados decorrentes
Objeto	definir o objeto e identificar outros objetos com os quais ele se relaciona	ações que podem ser aplicadas ao objeto
Estado	o que indica e ações que levaram a esse estado	identificar outros estados que podem ocorrer a partir do estado que se descreve

Sujeitos correspondem a entidades ativas, atores com papel relevante para a aplicação; um sujeito pode ser um ator, um componente ou um outro sistema com o qual deverá ocorrer interação. Verbos registram ações ou funcionalidades a serem desempenhadas pelos sujeitos ou pelo sistema em desenvolvimento, com algum impacto ou reflexo no ambiente operacional. Objetos são entidades passivas utilizadas ou necessárias a uma ação ou conjunto de ações, e estados são caracterizados por atributos significativos que registram valores em diferentes momentos da execução do sistema. A Tabela 7 apresenta exemplo de entrada para o léxico de um sistema para bibliotecas de uma universidade.

Tabela 7 – Exemplo de símbolo de um léxico do tipo LAL

Léxico Ampliado da Linguagem - Sistema de Bibliotecas
Usuário tipo do símbolo: sujeito noção: pessoa que pode utilizar a biblioteca; pode ser um aluno, professor ou funcionário da universidade impactos: usuário é cadastrado no sistema usuário é retirado do cadastro de usuários usuário retira obras da biblioteca usuário devolve obras anteriormente retiradas usuário renova datas para devolução de obras anteriormente retiradas

Optamos pela utilização do léxico Ampliado da Linguagem [Leite90] para representação do léxico das aplicações, e definimos uma estratégia para a avaliação de documentos da organização e extração de termos para compor o léxico. Nossa proposta utiliza duas diferentes abordagens para a construção do léxico da organização: a primeira está baseada na identificação de termos ou expressões não-dicionarizados, e a segunda utiliza extração de sintagmas nominais.

3.2.1. Termos ou expressões não incluídos em dicionários

O processo para extração de termos não-dicionarizados baseia-se na premissa que termos próprios do domínio da aplicação não deverão estar presentes em dicionários usuais da língua. Exemplos de termos incluem acrônimos para outros sistemas, palavras técnicas de tecnologia de informação em língua inglesa e até mesmo diferentes denominações para representar um mesmo objeto. Este último caso é especialmente interessante quando os participantes do processo de requisitos compartilham uma mesma língua, mas diferenças culturais devidas às diferentes localizações geográficas provocam o uso de vários termos para um mesmo objeto ou sujeito (por exemplo, celular e telemóvel).

Termos que atendem a uma dessas características têm forte indicação para serem registrados no léxico da aplicação, evitando dificuldades na leitura de documentos originadas por desconhecimento de termos ou por ambigüidades. O processo para extração de termos não dicionarizados está representado na Figura 21.

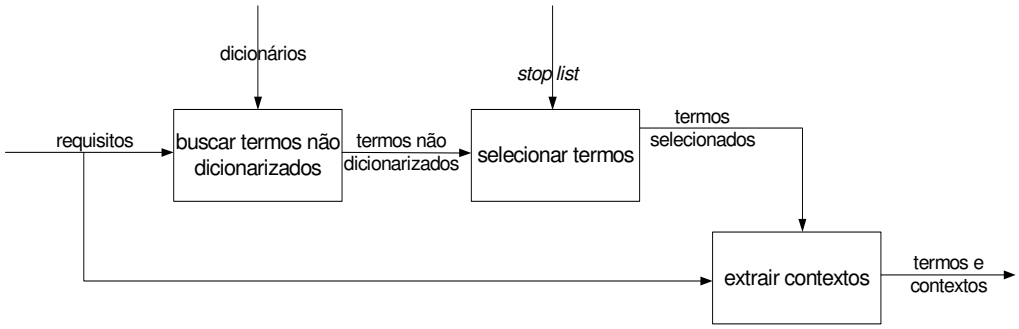


Figura 21 - Processo para extração de termos não-dicionarizados

3.2.1.1. Buscar termos não incluídos em dicionários

Nesta etapa o texto é *tokenizado* e os *tokens* são utilizados como expressão de busca num conjunto de dicionários para a língua portuguesa. Estes dicionários não são dicionários no sentido estrito do termo, pois não contêm definições ou sinônimos para os termos. Eles podem ser pensados como uma lista de termos, ordenados em ordem alfabética; cada dicionário é restrito a uma única classe gramatical de termos. O conjunto inclui dicionários específicos para categorias como substantivos, adjetivos, verbos, abreviaturas, pronomes, artigos, conjunções e interjeições.

Os dicionários utilizados foram criados e cedidos por Akeo Tanabe, pesquisador associado ao LES/DI-PUC-Rio, e estão descritos em [Tanabe06]. Esses dicionários abrangem aproximadamente dezoito mil adjetivos, mais de cinco mil substantivos e trezentas abreviaturas. Termos não relacionados nos dicionários são extraídos dos requisitos, pois são candidatos a comporem o léxico da aplicação.

3.2.1.2.Selecionar termos

Os termos não dicionarizados são agora avaliados em relação a uma *stop list*, que deverá conter os termos do léxico, se este já existir, e quaisquer outros termos considerados irrelevantes pelos participantes do processo.

O processo de seleção considera todos os termos, independente da freqüência com que eles possam estar presentes no texto. Isto significa que qualquer termo explicitado no documento e que não conste dos dicionários ou

stoplist utilizados não é um termo corrente da língua. Para evitar que esses termos possam dar margem a diferentes interpretações nos processos de V&V, eles deverão ser relacionados no léxico da aplicação.

Apenas termos não dicionarizados e que não constem da *stoplist* específica desta etapa deverão ser processados pela próxima etapa, que trata da extração de contextos.

3.2.1.3. Buscar noção e impactos

A busca de contexto trabalha com cada um dos símbolos selecionados, varrendo os requisitos e buscando identificar a presença do termo. Este processo utiliza um concordanceador ligeiramente modificado para extrair não apenas uma sub-sentença que contenha o termo, mas extrai o parágrafo completo.

O parágrafo extraído pode corresponder à definição ou a impactos desse termo no contexto da aplicação. Os parágrafos extraídos serão agrupados junto ao símbolo candidato, e utilizados posteriormente pelo engenheiro de requisitos para inserção no léxico da aplicação.

3.2.2. Identificação de sujeitos e objetos

Sintagmas nominais são definidos como (a) classe gramatical com comportamento sintático de sujeito, de objeto direto e, se precedido de preposição, de adjunto adnominal ou de objeto indireto [Perini98] [Vieira01]; (b) enunciado que representa um conceito ou uma entidade (abstrata ou concreta) identificada por nomes próprios ou sintagmas nominais descritivos; pode ainda representar um papel [Liberato97].

No contexto da engenharia de software, Booch et al [Booch00] afirmam que um ator representa um papel que um ser humano, um dispositivo de hardware ou mesmo um outro sistema desempenha [Booch00]: ele é o sujeito da ação. Em documentos técnicos, atores são registrados através de identificadores que incluem nomes próprios, profissões e papéis. Em termos lingüísticos podemos associar atores e sintagmas nominais. Recursos correspondem a objetos que ocupam um espaço no mundo real ou virtual e são utilizados ou gerados por

ações. Recursos e objetos podem ser identificados por substantivos e, portanto também podem ser lingüisticamente identificados por sintagmas nominais.

O processo geral de extração de símbolos proposto é baseado na extração de sintagmas nominais e é apresentado na Figura 22.

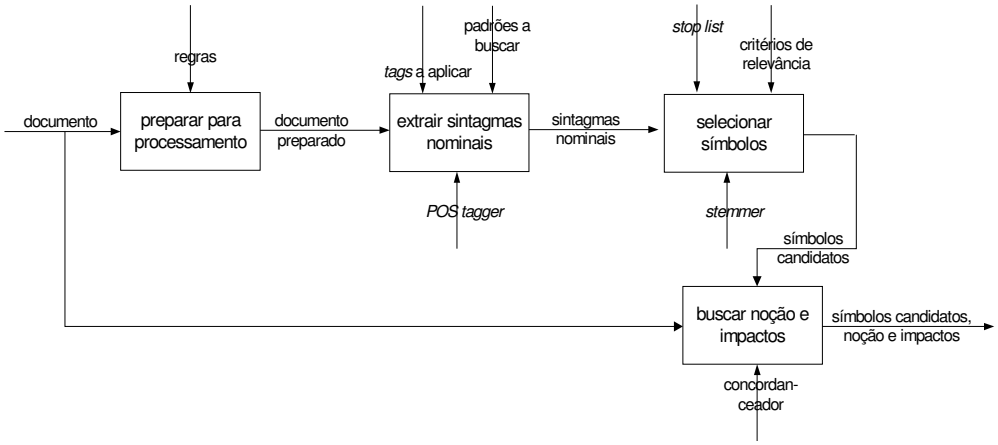


Figura 22 - Processo geral para extração de símbolos

O processo de extração de símbolos está estruturado em quatro atividades ou sub-processos. Inicialmente o documento é preparado para posteriormente ser manipulado por um *Part of Speech tagger* (*POS tagger*). Após a inserção das *tags* (etiquetas) são extraídos sintagmas nominais que correspondem a padrões pré-definidos; estes sintagmas nominais passarão por um processo de seleção, que irá considerar apenas aqueles que atendem a critérios de relevância estabelecidos. A etapa final extrai, para cada um dos sintagmas, contextos que possam vir a ser utilizados como noção e impactos. A seguir detalharemos cada um desses sub-processos.

3.2.2.1. Preparar para processamento

O trabalho de preparação dos documentos é necessário, pois as ferramentas utilizadas trabalham com arquivos do tipo texto puro. Os documentos da organização podem estar em diferentes formatos, como por exemplo, *pdf* (Portable Document Format), *doc* (Microsoft Word document) ou *html* (HyperText Markup Language). Documentos podem ainda incluir figuras e tabelas, que não serão processadas pelas ferramentas de tratamento de texto. O processo de preparação inicialmente retira figuras, transforma tabelas em texto,

retira possíveis *tags* de formatação e gera texto puro, ou seja, arquivos do tipo *txt*.

Após a obtenção do arquivo em formato *txt*, faz-se necessário a *tokenização* do documento, ou seja, a colocação de um *token* por linha (inclusive dos caracteres de pontuação). A separação dos *tokens* é uma exigência do etiquetador utilizado e poderia ser dispensada caso a ferramenta utilizada não exigisse este formato. O documento de requisitos, agora já no formato texto puro, é *tokenizado* e pode então ser trabalhado para a extração de sintagmas nominais.

3.2.2.2. Extrair sintagmas nominais

O processo de extração de sintagmas nominais é baseado na utilização do etiquetador morfossintático QTAG, que analisa o texto pré-processado e associa uma etiqueta a cada um dos *tokens*. Após a etiquetagem são extraídos e contabilizados os sintagmas nominais que atendem a padrões pré-estabelecidos, e que corresponderão a atores e objetos registrados no documento avaliado.

Sintagmas nominais possuem uma estrutura bem definida. Perini [Perini98] coloca que sintagmas nominais possuem duas estruturas básicas: a estrutura à esquerda do núcleo do sintagma é composta por posições que podem ser ocupadas por determinantes, possessivos, quantificadores e outras classes de palavras. A estrutura à direita do núcleo é composta por modificadores, que por sua vez podem ser classes abertas ou mesmo outros sintagmas nominais. Neste trabalho não utilizamos a estrutura à esquerda, pois a linguagem de documentos de requisitos é objetiva, não deve utilizar figuras de linguagem e neste contexto o que nos interessa é o núcleo do sintagma nominal.

Para a extração dos sintagmas utilizamos alguns padrões pré-definidos, pois nem todo sintagma nominal presente no texto é interessante, no contexto deste trabalho.

Extração de Atores ou sujeitos: padrões utilizados

Na língua portuguesa funções ou papéis desempenhados por pessoas ou entidades são identificados por substantivos com terminações específicas. Alguns exemplos com terminação *ente*: gerente, presidente; terminação *or*: gestor, diretor, trabalhador; terminação *ário*: usuário, funcionário. Na extração de sintagmas

nominais correspondendo a atores/sujeitos, utilizamos um conjunto de 82 padrões, pois para cada terminação consideramos o singular, o plural, feminino e masculino, na prática multiplicando cada terminação por quatro. Exemplos de padrões utilizados são apresentados na Tabela 8; os anexos B e C trazem a relação completa.

Tabela 8 - Padrões para extração de sujeitos e atores

N*ente PRP* N* / N*entes PRP* N* / N*enta PRP* N* / N*entas PRP* N*
N*or PRP* N* / N*ores PRP* N* /N*ora PRP* N* /N*oras PRP* N*
N*eiro PRP* N*/ N*eiros PRP* N*/ N*eira PRP* N*/ N*eiras PRP* N*
N*ente N*/N*entes N* /N*enta N*/N*entas N*
N*or N*/N*ores N*/N*ora N*/N*oras N*

Nos padrões relacionados acima o símbolo * indica qualquer quantidade de caracteres. Como exemplo, o padrão *N*ente PRP N** deverá extrair sintagmas compostos por um nome seguido por preposição e outro nome, sendo que o primeiro nome deverá ter a terminação *ente*. Os padrões foram definidos tendo por base um dicionário de terminações da língua portuguêsa e revisados após avaliação de textos etiquetados nos nossos experimentos iniciais.

Extração de Recursos ou Objetos: padrões utilizados

A extração de recursos/objetos utilizou padrões mais genéricos que os utilizados na etapa de extração de atores/sujeitos. Foram utilizados nove padrões, relacionados na Tabela 9.

Tabela 9 - Padrões para extração de objetos e recursos

N* PRP* N* PRP* N*	N* PRP* N*	N* N*
N* PRP* N* N*	N* PRN* N*	N* PART*
N* CPR* N* N*	N* CPR* N*	N* ADJ*

Os padrões utilizados foram definidos empiricamente após avaliação de diferentes tipos de documentos gerados ou manipulados no processo de requisitos. As principais fontes para este trabalho foram documentos de requisitos de diferentes domínios de aplicação e manuais de usuário. Buscamos identificar padrões gerais, mas certamente o conjunto utilizado deve ser revisto e modificado para atender especificidades de uma determinada organização de domínio de aplicação.

3.2.2.3. Selecionar símbolos

O processo de seleção dos símbolos a partir dos sintagmas candidatos desconsidera sintagmas que contenham termos relacionados na *stoplis*t, que é composta por termos da língua geral ou mesmo do domínio, mas que não são relevantes para o léxico. É importante observar que a *stoplist*, neste contexto, não é formada por artigos, pronomes, preposições, e outros normalmente incluídos em listas desse tipo.

A etapa seguinte realiza a *stemização* dos sintagmas extraídos. Isto é necessário pois, nos nossos experimentos, identificamos a necessidade de agrupar singular e plural, feminino e masculino. Por exemplo, em nossos estudos de caso, foram separadamente extraídos e contabilizados os sintagmas nominais *gestor de escala* e *gestor de escalas*, o mesmo ocorrendo com os sintagmas *cessionária* e *cessionário*.

Os sintagmas candidatos são então agrupados e contabilizados após o processo de *stemização*, e adotamos o masculino singular para a representação, quando fosse o caso. A última atividade do processo de seleção descarta os sintagmas cuja freqüência seja menor que quatro, conforme procedimento usual na área de extração de terminologia [Daile96].

3.2.2.4. Buscar noção e impactos

A busca de contexto trabalha com cada um dos símbolos selecionados, varrendo o documento buscando identificar a presença do símbolo, e para esta tarefa foi utilizado um concordanceador modificado. Quando a ferramenta identifica a presença de um símbolo, é extraído o parágrafo completo, pois este pode corresponder à definição ou a impactos do símbolo.

Na identificação das sentenças é utilizado o radical (*stem*) do símbolo de busca; os parágrafos extraídos serão agrupados junto ao símbolo candidato, e utilizados posteriormente pelo engenheiro de requisitos para inserção no léxico da aplicação. Nos estudos de caso observamos que a identificação da noção (ou definição) dos símbolos é sensível ao contexto da aplicação ou mesmo a padrões da organização.

3.3. Detecção de discrepâncias, erros e omissões em requisitos

O apoio à detecção de discrepâncias, erros e omissões em requisitos é apoiado em parte pelo agrupamento de requisitos, que possibilita aos participantes do processo de verificação ou validação trabalhar em conjuntos menores de requisitos. Isto agiliza o processo: por exemplo, para identificar incompletude em requisitos relacionados a um determinado tema manipula-se apenas o conjunto de requisitos associado a esse tema, e não todo o conjunto de requisitos do sistema.

Outra formas de apoio que estamos propondo estão relacionadas à automação da identificação de duplicidade em requisitos e à identificação de omissões em RNF´s.

Em ambientes distribuídos de desenvolvimento, usuários, clientes e engenheiros de requisitos estão dispersos geograficamente. Se o processo de elicitação ocorre em diferentes locais, usuários distantes geograficamente podem passar a diferentes engenheiros de requisitos informações que irão gerar requisitos similares. Isto só será percebido quando os requisitos forem colocados num único documento, mas esta detecção poderá ser dificultada se houver um grande número de requisitos. Se a verificação de duplicidade é automatizada, a descoberta de requisitos candidatos à duplicidade pode ser feita rapidamente.

Alguns modelos de registro de requisitos, como casos de uso, são mais voltados à especificação de funcionalidades do sistema. Organizações que usem um modelo de especificação de requisitos baseadas em casos de uso costumam registrar requisitos não funcionais numa especificação de requisitos suplementares. Mesmo com diversos catálogos de requisitos não funcionais disponíveis para uso, a prática mostra que o registro dos requisitos não funcionais não é tarefa adequadamente atendida em muitas organizações de desenvolvimento de software.

Nossa proposta para identificação de duplicidade em requisitos e de identificação de omissões em requisitos não funcionais está descrita a seguir.

3.3.1.
Duplicidade em requisitos

A identificação de duplicidade nos requisitos é aplicada nos agrupamentos de requisitos, e está baseada em medidas de similaridade. Entre medidas de similaridade aplicáveis a documentos, optamos por utilizar os coeficientes do coseno, Jaccard e Dice para identificar requisitos candidatos à verificação de duplicidade. A abordagem utilizada neste processo para representação dos documentos é a conhecida por *bag-of-words*, e a estrutura do processo para identificação de requisitos similares é apresentada na Figura 23.

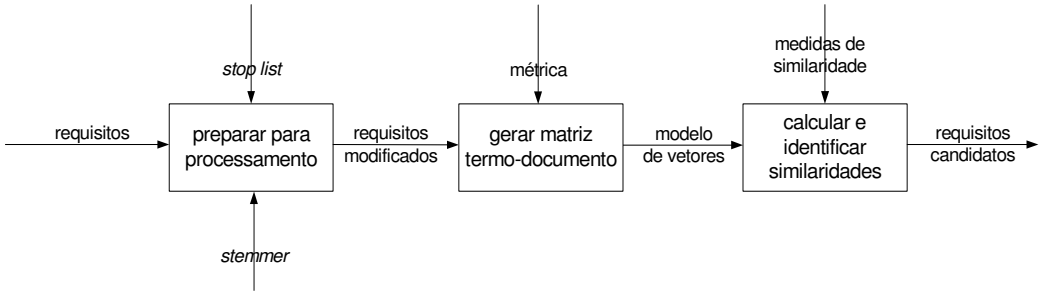


Figura 23 - Processo para identificação de requisitos similares

Para o cálculo das similaridades é utilizada uma matriz termo-documento: as linhas representam os requisitos da aplicação e as colunas os termos utilizados na definição desses requisitos. As células de tais matrizes contém uma métrica relacionada à freqüência do termo no documento, e para o nosso caso utilizamos a freqüência pura. Tais matrizes podem apresentar problemas de alta dimensionalidade (quantidade de termos) e conter dados esparsos (células vazias ou com valor zero). Para tratar esse problema, utilizamos uma etapa de pré-processamento descrita a seguir.

3.3.1.1.
Preparar para processamento

O pré-processamento dos documentos está baseado na utilização de uma *stoplist* e de um *stemmer* para obtenção dos radicais (*stems*) dos termos. A *stop list* utilizada inclui um conjunto de composto por artigos, preposições, conjunções, advérbios. Termos que façam parte da *stoplist* serão retirados do

texto dos requisitos. O uso do stemmer propicia o agrupamento de termos com proximidade conceitual.

A aplicação deste processo diminui o número de termos únicos nos documentos de requisitos, contribuindo para redução da dimensão da matriz termo-documento e melhor desempenho dos algoritmos para cálculo da similaridade. Os termos que compõem a *stop list*, conforme já visto no capítulo 2, costumam estar presentes em grande parte dos documentos e possuem baixo poder discriminatório em relação aos demais.

3.3.1.2. Gerar matriz termo-documento

Para a geração da matriz cada um dos requisitos é analisado, os *stems* são contabilizados e também colocados num vetor comum a todos os requisitos. Este processo é repetido para todos os requisitos, e ao final obteremos uma matriz onde cada requisito estará representado por um vetor de termos, e o peso associado a cada termo será obtido pela freqüência do termo no requisito correspondente. Cada um dos *n* requisitos estará representado por um vetor de tamanho *t* (t correspondendo ao número total de termos únicos).

Como o objetivo nesta etapa é identificar similaridades entre requisitos, são desconsiderados apenas os termos relacionados na *stoplist*, não importando a freqüência com que eles possam estar presentes no conjunto de requisitos.

3.3.1.3. Calcular e identificar similaridades

O cálculo dos coeficientes de similaridade considera pares de requisitos: cada requisito é avaliado contra cada um dos demais requisitos do grupo. Utilizamos os coeficientes de Dice, do coseno e de Jaccard, conforme discutido no capítulo 2. Os coeficientes calculados retornam sempre valores entre 0 e 1; quanto mais próximo de 1 o valor, mais similar é o par de requisitos. Quanto mais próximo de 0, menos similar é o par de requisitos.

No nosso trabalho, ao estruturarmos os requisitos na forma de matrizes termo-documento, utilizamos uma *stoplist* composta apenas por artigos, pronomes, advérbios, conjunções e outras palavras sem valor relevante. Não

foram incluídos termos técnicos ou do domínio da aplicação.

Nosso cálculo correlaciona os valores obtidos para a obtenção de um índice de similaridade único, considerando-se a média aritmética desses coeficientes. Os requisitos candidatos à análise de duplicidade são então apontados, considerando o limiar de 0,90, ou seja, todos os pares cujo índice de similaridade seja igual ou maior que 0,90 serão avaliados para efeitos de identificação de duplicidade. Este limiar pode ser ajustado para atender a particularidades de uma dada aplicação.

Para a definição do limiar foram executados diversos experimentos, considerando documentos de requisitos de diferentes domínios de aplicação. O limiar definido foi aquele que apresentou, considerando o conjunto de experimentos, maiores taxas de acerto na identificação de requisitos em duplicidade e menor taxa de falsos positivos.

3.3.2. Omissões em requisitos

Uma das tarefas mais difíceis em processos de verificação de requisitos é a identificação de omissões. Como identificar aquilo que não está expresso nos requisitos? É muito difícil viabilizar esta verificação de forma automática para requisitos funcionais, expressos em linguagem natural, dado que cada aplicação possui características e necessidades que a diferenciam de qualquer outra. Mesmo sem termos realizado uma avaliação sistemática e qualitativa, podemos inferir que os conjuntos de requisitos funcionais em documentos de requisitos de diferentes domínios diferem de forma significativa. Não há um conjunto de referência que possa ser utilizado para a verificação de omissões em requisitos funcionais.

Por outro lado, sistemas de informação costumam utilizar um conjunto reduzido de requisitos não funcionais. Acreditamos que a identificação da presença ou ausência de requisitos não funcionais seja viável, desde que o processo utilizado consiga resolver ou contornar dificuldades inerentes ao uso da linguagem natural no documento de requisitos.

Um dos problemas a ser resolvido está diretamente relacionado ao grau de detalhamento aplicado no registro dos requisitos. Para sistemas com muitas funcionalidades a serem atendidas, ou cujos clientes e usuários estão geograficamente separados, é possível supor o cenário onde um grupo de

engenheiros de requisitos atua, buscando identificar junto a clientes e usuários as características e funcionalidades a serem atendidas pelo sistema a ser desenvolvido. As diferentes capacidades lingüísticas dos engenheiros de requisitos podem levar ao registro de requisitos em diferentes estilos, e então teremos que lidar com a avaliação de requisitos expressos em diferentes graus de abstração ou mesmo utilizando diferentes termos para um mesmo conceito.

Outro problema a ser abordado está na própria classificação geral dos requisitos: como requisitos não funcionais serão implementados através de funcionalidades no sistema, alguns engenheiros de requisitos os classificam (incorretamente) como funcionais. Nos documentos de requisitos que analisamos, foi freqüente o caso de requisitos não funcionais de controle de acesso serem classificados como requisitos funcionais. Avaliar então a ausência de requisitos não funcionais implica na avaliação de todo o conjunto de requisitos, dado que eles podem estar classificados incorretamente.

Como avaliar de forma objetiva o conjunto de requisitos, resultando em informações claras sobre a presença ou ausência de requisitos não funcionais no documento avaliado? Não existem variáveis concretas a serem medidas, nem mesmo um vocabulário padrão para uso no registro dos requisitos. A solução encontrada utiliza a metodologia da análise de conteúdo- para avaliar um documento de requisitos e identificar possíveis omissões, considerando requisitos não funcionais.

Como foi visto no capítulo 2, métodos de análise de conteúdo são baseados na utilização de um dicionário de categorias; estas categorias representam idéias ou conceitos que se deseja investigar se estão ou não presentes no texto.

No contexto de requisitos, dicionários manipulados para a análise de conteúdo podem ser construídos a partir dos catálogos públicos de requisitos não funcionais. Cada entrada no dicionário é composta por um requisito não funcional; a cada entrada são associados os termos e expressões usados no contexto da organização para o registro daquele requisito não funcional.

O processo completo para a identificação de omissões em RNF´s no documento de requisitos é apresentado na Figura 24 e inclui a etapa de criação do dicionário, a instanciação para uma dada organização e a execução da análise de conteúdo do documento de requisitos, gerando um relatório sobre omissões relacionadas a RNF´s.

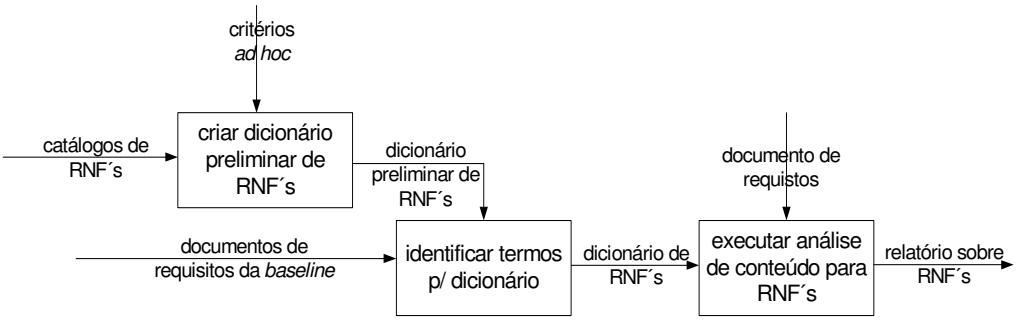


Figura 24 - Processo para análise de conteúdo para RNF´s

3.3.2.1. Criar dicionário preliminar de RNF´s

O dicionário de RNF´s a ser utilizado para análise do documento de requisitos gerado nesta etapa não pretende ser um dicionário completo; o objetivo na sua construção é registrar, para uma dada organização, os RNF´s necessários à gama de aplicações visualizada para a organização. Tendo portanto como entradas catálogos públicos de RNF´s, como por exemplo aqueles extraídos de [Sommerville04] [IEEE98] [SWEBOK04], uma avaliação conduzida por especialistas da organização deverá selecionar os RNF´s que deverão constar no dicionário. Neste processo deverão ser considerados aspectos relacionados a padrões de qualidade da organização e aos ambientes de execução dos sistemas, entre outros.

É importante ter em mente a utilização deste dicionário: ele deverá ser utilizado em atividades de verificação e validação, e não em atividades de elicitação. Essa diferença é enfatizada, pois catálogos de RNF´s para uso na etapa de elicitação ou modelagem de requisitos devem ser abrangentes, possibilitando a quem os utiliza um mapeamento amplo dos RNF´s possíveis para uma vasta gama de aplicações. A Figura 25 apresenta o dicionário inicial, fortemente baseado em [Sommerville04] [IEEE98] [SWEBOK04].

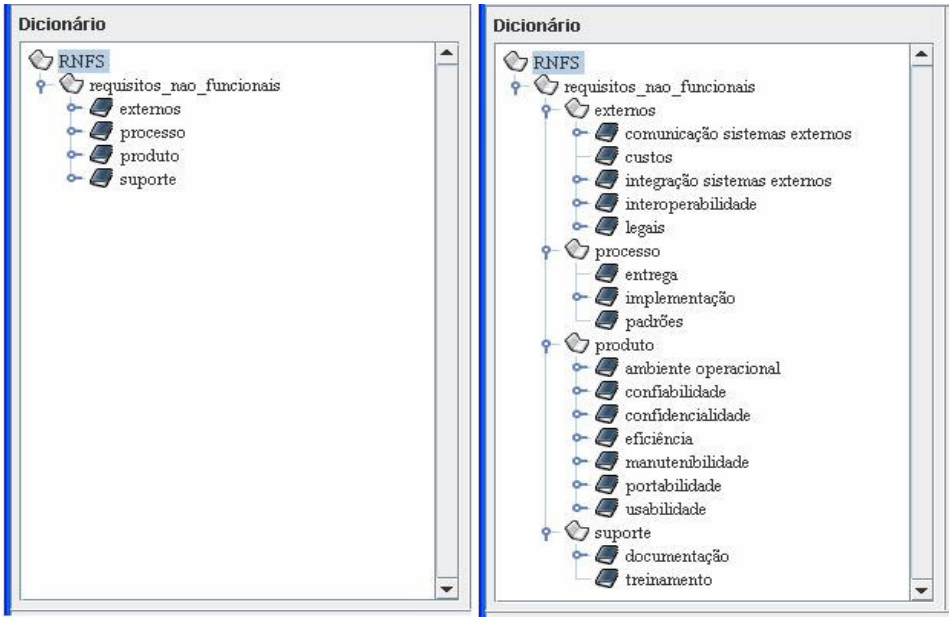


Figura 25 - Estrutura do dicionário de RNF´s para identificação de omissões

Na Figura 25, à esquerda, é exibido o primeiro nível de categorias do dicionário criado, e à direita o detalhamento de cada uma dessas categorias iniciais. Os requisitos não funcionais que serão avaliados estão agrupados nas categorias externos, processo, produto e suporte. Cada uma dessas categorias está sub-dividida em várias outras, para permitir uma boa estruturação do dicionário. A categoria externos agrega o conjunto de termos que está relacionada a requisitos externos ao sistema, como por exemplo requisitos legais, de comunicação com sistemas já existentes, de custos. A categoria processo agrega o conjunto de termos relacionados ao processo de desenvolvimento do software, como restrições de implementação, padrões a serem utilizados nas atividades e artefatos do processo de desenvolvimento, requisitos de entrega do software. A categoria produto agrega termos que estão associados a propriedades ou características que o sistema deve apresentar, como portabilidade, usabilidade, confiabilidade, confidencialidade e outros. E a categoria de suporte abriga termos relacionados à documentação que deverá acompanhar o software e ao treinamento necessário para o grupo de usuários.

A saída desta atividade é um dicionário preliminar de RNF´s; as entradas do dicionário conterão apenas o nome e uma breve descrição de cada RNF.

3.3.2.2. Identificar termos para o dicionário de RNF´s

Uma vez obtido o dicionário preliminar de RNF´s, deve-se agora instanciá-lo para uma dada organização. Isso significa inserir, para cada um dos RNF´s do dicionário, entradas que registrem termos ou expressões já utilizados em documentos de requisitos da *baseline* da organização. Este parte do processo é feita de forma semi-automatizada: os participantes analisam documentos da baseline da organização e identificam termos que são usuais no registro de requisitos não funcionais. Este processo deve ser suportado pelo léxico da organização. O ideal é que os participantes desta atividade sejam profissionais já com boa experiência em processos de requisitos, com uma visão ampla da área de tecnologia da informação da empresa, e com bom desempenho em atividades do processo de requisitos.

Os termos e expressões selecionados serão inseridos nas categorias definidas na etapa anterior. O suporte de software para esta atividade deve ter flexibilidade para permitir o registro de palavras simples e de expressões. Para possibilitar que uma única entrada no dicionário possa representar um conjunto de termos, é desejável que se possa também utilizar o caracter * para indicar uma ou múltiplas ocorrências de caracteres alfabéticos. Uma única inserção deve representar singular, plural, masculino, feminino, entre outras variações.

Exemplificando, na ferramenta implementada o termo conec* permitirá a representação de palavras como conecta, conectar, conectando, conectou. O caracter * também poderá estar presente substituindo uma palavra interna ao termo, como por exemplo em *password * acesso*, expressão que irá representar password de acesso, password para acesso, password do acesso. Os termos podem ser registrados tanto em maiúsculas como em minúsculas.

Ao final desta atividade, teremos um dicionário onde cada entrada corresponde a um RNF, sua descrição e um conjunto de termos, que será utilizado com a finalidade de identificar a presença ou ausência desse RNF em documentos de requisitos.

3.3.2.3. Executar análise de conteúdo para RNF´s

A etapa de análise de conteúdo para identificar presença ou ausência de RNF´s em documentos de requisitos deve ser executada a cada modificação no documento de requisitos. O documento de requisitos é analisado, buscando-se identificar a presença de cada uma das expressões registradas no dicionário de RNF´s. Os resultados deste processo são expressos em forma de tabelas, a mais simples uma tabela de freqüência, conforme pode ser visualizado na Figura 26.

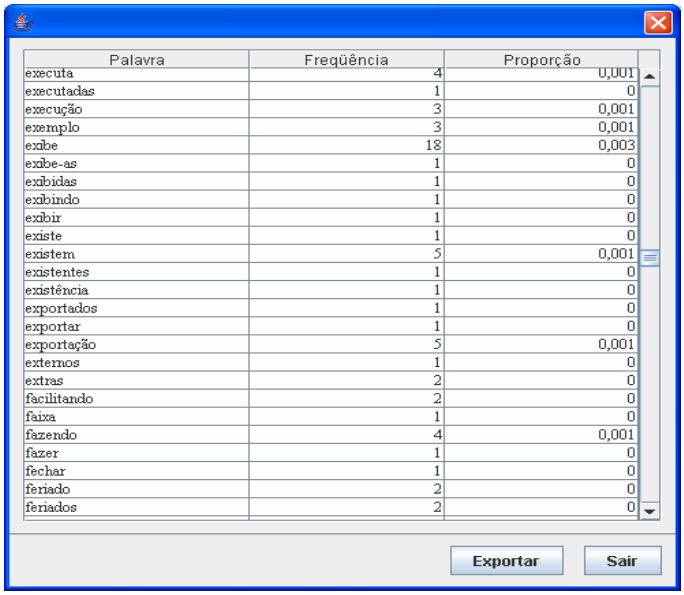
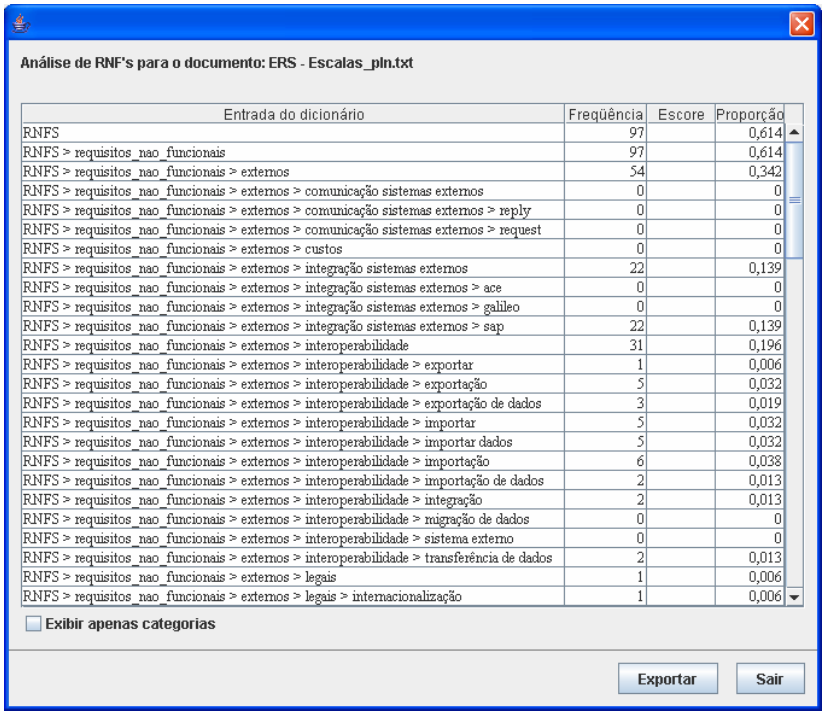


Figura 26 - Freqüência das palavras no documento de requisitos

É gerado um relatório apresentando, para cada uma das categorias, o detalhamento de freqüência de cada expressão registrada e o escore correspondente. O relatório consolida as informações de acordo com a hierarquia associada a cada um dos RNF´s, conforme apresentado na Figura 27.



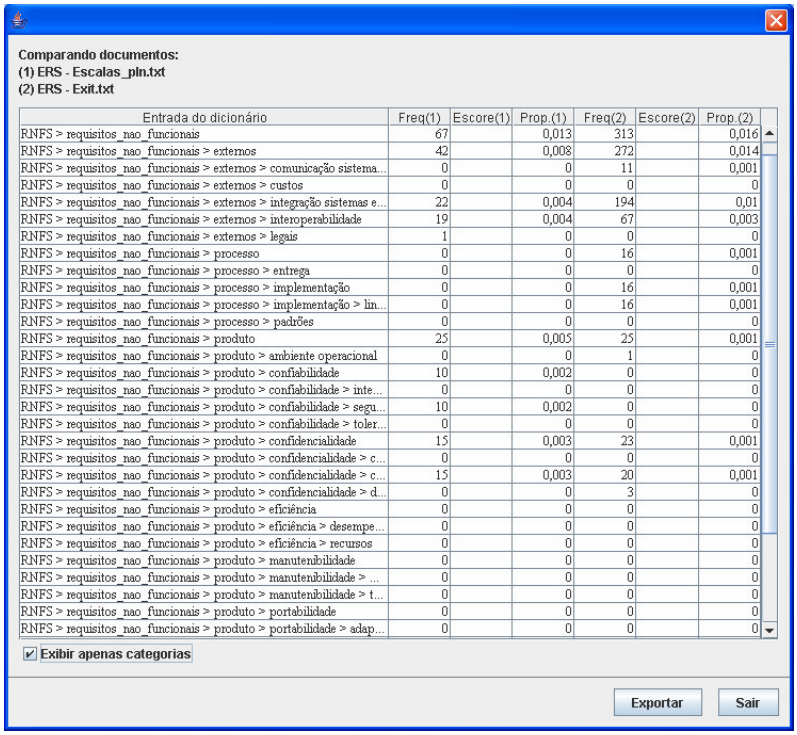
Entrada do dicionário	Freqüência	Escore	Proporção
RNFS	97		0,614
RNFS > requisitos_nao_funcionais	97		0,614
RNFS > requisitos_nao_funcionais > externos	54		0,342
RNFS > requisitos_nao_funcionais > externos > comunicação sistemas externos	0		0
RNFS > requisitos_nao_funcionais > externos > comunicação sistemas externos > reply	0		0
RNFS > requisitos_nao_funcionais > externos > comunicação sistemas externos > request	0		0
RNFS > requisitos_nao_funcionais > externos > custos	0		0
RNFS > requisitos_nao_funcionais > externos > integração sistemas externos	22		0,139
RNFS > requisitos_nao_funcionais > externos > integração sistemas externos > ace	0		0
RNFS > requisitos_nao_funcionais > externos > integração sistemas externos > galileo	0		0
RNFS > requisitos_nao_funcionais > externos > integração sistemas externos > sap	22		0,139
RNFS > requisitos_nao_funcionais > externos > interoperabilidade	31		0,196
RNFS > requisitos_nao_funcionais > externos > interoperabilidade > exportar	1		0,006
RNFS > requisitos_nao_funcionais > externos > interoperabilidade > exportação	5		0,032
RNFS > requisitos_nao_funcionais > externos > interoperabilidade > exportação de dados	3		0,019
RNFS > requisitos_nao_funcionais > externos > interoperabilidade > importar	5		0,032
RNFS > requisitos_nao_funcionais > externos > interoperabilidade > importar dados	5		0,032
RNFS > requisitos_nao_funcionais > externos > interoperabilidade > importação	6		0,038
RNFS > requisitos_nao_funcionais > externos > interoperabilidade > importação de dados	2		0,013
RNFS > requisitos_nao_funcionais > externos > interoperabilidade > integração	2		0,013
RNFS > requisitos_nao_funcionais > externos > interoperabilidade > migração de dados	0		0
RNFS > requisitos_nao_funcionais > externos > interoperabilidade > sistema externo	0		0
RNFS > requisitos_nao_funcionais > externos > interoperabilidade > transferência de dados	2		0,013
RNFS > requisitos_nao_funcionais > externos > legais	1		0,006
RNFS > requisitos_nao_funcionais > externos > legais > internacionalização	1		0,006

Figura 27 - Freqüência das categorias e expressões no documento de requisitos

Caso a organização disponha de um documento de requisitos considerado que seja considerado como modelo em termos de RNF's, esse documento poderá ser base para comparações, conforme visualizado na Figura 28.

Esses relatórios subsidiam o engenheiro de requisitos na avaliação das omissões no documento de requisito e elaboração de outras análises, entre as quais destacamos:

- coerência entre interoperabilidade com sistemas externos x desempenho
- coerência nos RNF´s de cadastrar usuários, conectar e desconectar
- ausência de RNF´s de documentação
- ausência de RNF´s de processo



Entrada do dicionário	Freq(1)	Escore(1)	Prop.(1)	Freq(2)	Escore(2)	Prop.(2)
RNFS > requisitos_nao_funcionais	67		0,013	313		0,016
RNFS > requisitos_nao_funcionais > externos	42		0,008	272		0,014
RNFS > requisitos_nao_funcionais > externos > comunicação sistema...	0		0	11		0,001
RNFS > requisitos_nao_funcionais > externos > custos	0		0	0		0
RNFS > requisitos_nao_funcionais > externos > integração sistemas e...	22		0,004	194		0,01
RNFS > requisitos_nao_funcionais > externos > interoperabilidade	19		0,004	67		0,003
RNFS > requisitos_nao_funcionais > externos > legais	1		0	0		0
RNFS > requisitos_nao_funcionais > processo	0		0	16		0,001
RNFS > requisitos_nao_funcionais > processo > entrega	0		0	0		0
RNFS > requisitos_nao_funcionais > processo > implementação	0		0	16		0,001
RNFS > requisitos_nao_funcionais > processo > implementação > lin...	0		0	16		0,001
RNFS > requisitos_nao_funcionais > processo > padrões	0		0	0		0
RNFS > requisitos_nao_funcionais > produto	25		0,005	25		0,001
RNFS > requisitos_nao_funcionais > produto > ambiente operacional	0		0	1		0
RNFS > requisitos_nao_funcionais > produto > confiabilidade	10		0,002	0		0
RNFS > requisitos_nao_funcionais > produto > confiabilidade > inte...	0		0	0		0
RNFS > requisitos_nao_funcionais > produto > confiabilidade > segu...	10		0,002	0		0
RNFS > requisitos_nao_funcionais > produto > confiabilidade > toler...	0		0	0		0
RNFS > requisitos_nao_funcionais > produto > confidencialidade	15		0,003	23		0,001
RNFS > requisitos_nao_funcionais > produto > confidencialidade > c...	0		0	0		0
RNFS > requisitos_nao_funcionais > produto > confidencialidade > c...	15		0,003	20		0,001
RNFS > requisitos_nao_funcionais > produto > confidencialidade > d...	0		0	3		0
RNFS > requisitos_nao_funcionais > produto > eficiência	0		0	0		0
RNFS > requisitos_nao_funcionais > produto > eficiência > desempe...	0		0	0		0
RNFS > requisitos_nao_funcionais > produto > eficiência > recursos	0		0	0		0
RNFS > requisitos_nao_funcionais > produto > manutenibilidade	0		0	0		0
RNFS > requisitos_nao_funcionais > produto > manutenibilidade > ...	0		0	0		0
RNFS > requisitos_nao_funcionais > produto > manutenibilidade > t...	0		0	0		0
RNFS > requisitos_nao_funcionais > produto > portabilidade	0		0	0		0
RNFS > requisitos_nao_funcionais > produto > portabilidade > adap...	0		0	0		0

Figura 28 - Comparação de documentos tendo por base o dicionário de RNF's

O protótipo utilizado nesta etapa do trabalho pode ser utilizado para explorar outras características do documento de requisitos. Por exemplo, os termos ou expressões relacionados no dicionário podem ser destacados através de uma opção da ferramenta, e o texto do documento de requisitos pode ser percorrido, com destaque para o termo escolhido.