

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Informática

Dissertação de Mestrado

Extração de Passagens de Texto
Usando um Método Independente de Domínio

Welmisson Jammesson da Silva

Campina Grande - PB
Julho de 2009

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Informática

Extração de Passagens de Texto
Usando um Método Independente de Domínio

Welmisson Jammesson da Silva

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Informática da Universidade Federal de Campina Grande – Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Sistemas de Informação e Banco de Dados

Prof. Marcus Sampaio
(Orientador)

Campina Grande, Paraíba, Brasil.

© Welmisson Jammesson da Silva, Julho de 2009.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

S586e

2009 Silva, Welmisson Jammesson da.

Extração de passagens de texto usando um método independente de domínio / Welmisson Jammesson da Silva. — Campina Grande, 2009. 91 f.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática. Referências.

Orientador: Prof. Dr. Marcus Costa Sampaio.

1. Extração de Informação. 2. Dados não-estruturados. 3. Método de Extração Supervisionado. 4. Similaridade Estrutural. 5. Similaridade Textual. I. Título.

CDU – 004.775(043)

**"EXTRAÇÃO DE INFORMAÇÃO NÃO ESTRUTURADA, USANDO UM MÉTODO
SUPERVISIONADO E INDEPENDENTE DE DOMÍNIO"**

WELMISSON JAMMESSON DA SILVA

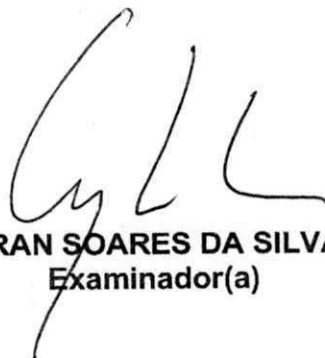
DISSERTAÇÃO APROVADA EM 11.08.2009



MARCUS COSTA SAMPAIO, DR.
Orientador(a)



ULRICH SCHIEL, DR.
Examinador(a)



ALTIGRAN SOARES DA SILVA, DR.
Examinador(a)

CAMPINA GRANDE - PB

Resumo

Extração de Informação (EI) é uma coleção de métodos e técnicas que têm como objetivo extrair, de fontes semi-estruturadas ou não-estruturadas, informação relevante. Um sistema de EI é capaz de extrair, de fontes de informação textuais, apenas informação que seja do interesse dos usuários do sistema, as partes que não são interessantes aos usuários não são extraídas.

Nesta dissertação, é proposto um novo método supervisionado de EI em que a informação extraída, partes de um texto, não é estruturada; isto representa um avanço em relação à EI ‘tradicional’, em que a informação extraída é estruturada segundo um *template* definido por usuário. Sendo supervisionada, a extração de informação de novos documentos é induzida de uma coleção prévia de documentos com suas partes relevantes assinaladas — conjunto de treinamento —; porém, o método inova no sentido de que o conjunto de treinamento pode ser muito pequeno em termos absolutos, resultando em um baixo custo de preparação do mesmo. Outra novidade do método está em sua técnica de extração, que é uma adequada combinação de técnicas existentes. Independência de domínio e de formato de documentos são outras duas importantes características do método.

Para a validação do método, o sistema TIES — *Textual Information Extraction System* — foi desenvolvido e testado com dois domínios díspares, um sobre sistemas elétricos de potência e o outro sobre legislação para administração pública: os resultados dos testes, para os dois domínios, revelaram-se promissores.

Abstract

Information Extraction (IE) is a collection of methods and techniques that have as objective to extract, from semi-structured or non-structured data sources, relevant information. An EI system is able to extract, from textual information sources, only information that is of interest to system users, the parts that are not interesting to users are not extracted.

In this work, a new supervised IE method is proposed where the extracted information, text parts, is non-structured; this represents a progress in relation to 'traditional' IE, where the extracted information is structured according to a user-defined template. Being supervised, information extraction from new documents is induced from a previous collection of documents with their marked relevant parts — training set —; however, the method innovates in the sense that the training set can be very small in absolute terms, this way propitiating low cost of its preparation. Another innovation of the method is its extraction technique, that is an appropriate combination of existent techniques. Domain independence and independence of format of documents are other two important characteristics of the method.

For the validation of the method, the system TIES — Textual Information Extraction System — was developed and tested with two disparate domains, one on electric power systems and the another on legislation for public administration: the results of the tests, for the two domains, were promising.

Agradecimentos

Primeiramente, agradeço a Deus, e a seu Filho Jesus, por terem me dado oportunidade e forças para esta conquista. A Maria Santíssima, por sua valorosa intercessão em meu favor a Seu Filho.

A minha mãe, que é a principal responsável na terra pelo que eu sou, por ter me apoiado em todas as etapas nesta conquista e por suas orações em meu favor. A Seu João Henrique, pela união abençoada, por Deus, com minha mãe. Aos meus irmãos Wellington e Wilton, pela força e parceria de muito tempo. As minhas tias: Beta, Bia, Fátima, e Lenice, pelo apoio ainda na graduação.

A minha parceira, linda e maravilhosa namorada, Amanda, pelo carinho, força, paciência e compreensão nos momentos bons e ruins. Agradeço também a seu pai, mãe e irmãos, por me fazerem sentir que era parte da família.

A minha prima Avany e a sua família, minha também, pelo apoio e conselhos, não apenas nesta etapa.

Ao professor Marcus Sampaio, pela ótima orientação, ensinamentos e experiências, adquiridos com as reuniões e atividades no mestrado.

Ao professor Cláudio Baptista, por ter sido meu orientador durante a graduação e primeiro a me estimular ao mestrado.

A professora Joseana Fechine, pela orientação enquanto foi participante do PET-Computação. Ao grupo PET-Computação da UFCG, especialmente a todos os participantes que tenham trabalho comigo, pelas experiências adquiridas e contribuições em minha formação acadêmica.

A Aninha e Vera, secretárias da COPIN, por toda ajuda na parte burocrática.

Aos meus amigos de graduação: Ana Esther, Cláudio, Daniel, Danilo, Hugo, Ianna, Neto, Vinício e Yuri, pelo companheirismo, apoio e ajuda na “etapa” anterior.

Aos meus novos amigos: Felipe, Fernando, Gilson, Marcos Fábio e Rute (além de alguns anteriores: Danilo, Neto, Hugo e Yuri), pelo companheirismo e ajuda durante o mestrado.

Finalmente, a todos aqueles que porventura eu tenha esquecido, mas que de alguma forma ajudaram na produção deste trabalho, minhas sinceras desculpas e gratidão.

Conteúdo

Capítulo 1	1
Introdução	1
1.1 Estrutura de Fontes de Informação	1
1.2 Extração de Informação	1
1.2.1 Estrutura dos Dados Fontes	2
1.2.2 Visualização da Informação Extraída	2
1.2.3 Classes de Métodos de Extração de Informação	4
1.2.4 Métricas de Avaliação de Extração	6
1.3 Objetivos da Dissertação	7
1.3.1 Objetivos Gerais	7
1.3.2 Objetivos Específicos	8
1.4 Estrutura da dissertação	8
Capítulo 2	9
Técnicas de Extração de Informação	9
2.1 Extração de Informação por Análise de Similaridade	9
2.1.1 Extração Automática de Notícias na Web	12
2.1.2 Análise de Similaridade de Documentos Normativos	12
2.1.3 Extração Visual de Informação	13
2.2 Extração de Passagens Relevantes	14
2.2.1 Extração de Passagens Coerentes Usando Modelos de Markov	15
2.2.2 Recuperação de Passagem usando Similaridade entre Vértices de um Grafo	16
2.3 Segmentação Topicamente Coerente de Texto	17
2.3.1 TextTiling	17
2.3.2 Minimum Cut	18
2.4 Conclusões	19
Capítulo 3	20
Proposta de um Novo Método de Supervisão e de uma Nova Técnica de Extração de Informação	20
3.1 O Método de Extração de Informação	20
3.2 A Técnica de Extração de Informação	21
3.2.1 Marcação de Estrutura	26
3.2.2 Marcação de Passagem	35
3.3 O Protótipo TIES – Textual Information Extraction System	49
3.3.1 Arquitetura do TIES	50
3.4 Conclusões	54
Capítulo 4	55
Avaliação Experimental do TIES	55
4.1 Plano de Testes	55
4.1.1 Hipóteses a Verificar	55
4.1.2 <i>Corpora</i> de Testes	56
4.1.3 Métricas de Avaliação	60
4.1.4 Calibragem do Sistema TIES	61
4.2 Resultados dos Testes	63

4.2.1 Classificação dos Documentos-Consulta	63
4.2.2 Resultados para o <i>Corpus</i> Chesf.....	64
4.2.3 Resultados para o <i>Corpus</i> Legislativo.....	69
4.3 Conclusões.....	74
Capítulo 5	76
Conclusões.....	76
5.1 Contribuições.....	76
5.2 Conclusões.....	77
5.3 Trabalhos Futuros	77
Referências	79

Lista de Figuras

Figura 1.1: (a) Página Web de classificados de veículos; (b) <i>template</i> vazio; (c) <i>template</i> preenchido.	3
Figura 1.2: Texto com relacionamento implícito: ehCapitalDe(Paris, França).....	3
Figura 1.3: Texto com uma passagem destacada.	4
Figura 1.4: Texto com passagem relevante e passagem extraída.	6
Figura 2.1: Transformação de árvore em outra.	10
Figura 3.1: Documento com seções e subseções.	23
Figura 3.2: Documento com marcações de estrutura.	24
Figura 3.3: Árvore de estrutura do documento da Figura 3.2.	24
Figura 3.4: Documento com delimitações de seções e subseções.	25
Figura 3.5: Exemplo de documento-consulta.	27
Figura 3.6: Documento-consulta com primeiras marcações de estrutura induzidas.	29
Figura 3.7: Documento-consulta com todas as marcações de estrutura induzidas.....	30
Figura 3.8: (a) Documento com apenas uma marcação de título induzida; (b) documento sem marcações de títulos induzida.	32
Figura 3.9: Árvores não equivalentes.	33
Figura 3.10: Árvores equivalentes.	34
Figura 3.11: Documento com passagem relevante marcada.	36
Figura 3.12: Documentos com várias passagens marcadas.	38
Figura 3.13: Documentos com marcação na seção raiz.	39
Figura 3.14: (a) Documento-treinamento com passagem marcada; (b) documento-consulta com seção indução destacada.	40
Figura 3.15: Seção indução segmentada.	41
Figura 3.16: Expansão do segmento inicial.	44
Figura 3.17: (a) Divisão do segmento inicial em quatro subsegmento; (b) expansão do subsegmento mais similar.	45
Figura 3.18: Documento-consulta com passagem induzida marcada.	47
Figura 3.19: Arquitetura do TIES.	50
Figura 4.1: Documento da Chesf com passagem relevante a seus operadores.....	58
Figura 4.2: Documento com indicação de passagem relevante (<i>gold standard</i>) e passagem induzida.	60
Figura 4.3: Gráficos para os resultados do <i>corpus</i> Chesf.	69
Figura 4.4: Gráficos para os resultados do <i>corpus</i> Legislativo.	73

Lista de Tabelas

Tabela 3.1: Correspondência entre títulos: documento-treinamento e documento-consulta.....	29
Tabela 3.2: Correspondência final entre título: documento-treinamento e documento-consulta.....	30
Tabela 4.1: Quantidades de documentos-treinamento e documentos-consulta para cada subclasse do <i>corpus</i> Chesf.....	57
Tabela 4.2: Quantidades de documentos-treinamento e documentos-consulta para cada classe do <i>corpus</i> Legislativo.....	59
Tabela 4.3: Possíveis configurações do protótipo TIES.....	62
Tabela 4.4: Resultados para um documento-consulta do grupo 1.....	64
Tabela 4.5: Resultados para um documento-consulta do grupo 2.....	65
Tabela 4.6: Resultados para um documento-consulta do grupo 3.....	66
Tabela 4.7: Resultados gerais para os três grupos - Chesf.....	68
Tabela 4.8: Resultados para um documento-consulta do grupo 1.....	70
Tabela 4.9: Resultados para um documento-consulta do grupo 2.....	71
Tabela 4.10: Resultados gerais para os dois grupos – Legislativo.....	72

Lista de Códigos

Algoritmo 3.1: Pseudocódigo para o algoritmo TSI.	31
Algoritmo 3.2: Pseudocódigo para o algoritmo SSA	34
Algoritmo 3.3: Pseudocódigo para a equivalência entre árvores.	35
Algoritmo 3.4: Pseudocódigo para as atividades do Marcador de Estrutura.....	35
Algoritmo 3.5: Pseudocódigo para a expansão de segmento.	44
Algoritmo 3.6: Pseudocódigo para a redução-expansão de segmento	46
Algoritmo 3.7: Pseudocódigo para as atividades do Marcador de Passagem.	48
Algoritmo 3.8: Pseudocódigo para o algoritmo RPI.	48

Capítulo 1

Introdução

É sobejamente constatado que, graças à Internet e às *intranets*, vive-se sob um verdadeiro *tsumani* de dados, isto é, um crescimento exponencial da produção de informação digital. A consequência disso é que, encontrar manualmente informações específicas torna-se uma tarefa cada vez mais difícil, ou até mesmo impraticável. Surge então a necessidade de tecnologias de *extração de informação*, capazes de automaticamente recuperar apenas informação relevante, segundo critérios definidos pelo usuário.

Para discorrer sobre extração de informação, o primeiro passo é relembrar a taxionomia comumente adotada de estruturas de fontes de informação.

1.1 Estrutura de Fontes de Informação

As fontes de informação caracterizam-se pela estrutura que seus dados apresentam, podendo ser: estruturada, semi-estruturada e não-estruturada. Em fontes de dados estruturadas as informações estão normalmente dispostas em tabelas, controladas por software de gerenciamento de banco de dados (SGBD). As informações em fontes semi-estruturadas se apresentam entre marcadores (*tag*), que podem ser reconhecidos ou processados por máquinas; exemplos: páginas HTML e documentos XML. Em fontes não-estruturadas as informações se apresentam em linguagem natural, isto é, sem estrutura tabular e sem marcação; exemplos: relatórios e memorandos produzidos por uma organização em suas atividades.

1.2 Extração de Informação

Extração de Informação (EI) é um conjunto de técnicas que têm como objetivo extrair, de fontes semi-estruturadas ou não-estruturadas, informação selecionada [Etzioni 2008].

Dado uma fonte de dados, como um documento textual, o usuário estaria interessado em somente partes dela. Identificar estas partes específicas, que se caracterizam como informações relevantes aos usuários, é tarefa para EI.

Sistemas de software que implementam técnicas de extração de informação são conhecidos como **Sistemas de Extração de Informação (SEI)**. SEIs podem ser caracterizados por pelo menos três critérios ortogonais: estrutura dos dados fonte, visualização da informação extraída e classe do método de extração.

1.2.1 Estrutura dos Dados Fontes

Sistemas de extração utilizam elementos em um texto que possam ajudar na identificação de informação relevante. Por sua vez, a forma de como identificar a informação depende da estrutura da fonte de informação.

- **Extração de fontes semi-estruturadas:** No caso de extração de fontes semi-estruturadas, notadamente páginas HTML, a informação relevante se encontra entre alguns tipos de marcação (*tag*). No fundo, a procura não é pela informação desejada, mas sim por um padrão de seqüência de marcações em uma coleção temática que delimitam esta informação [Hammer 1997].
- **Extração de fontes não-estruturadas:** Quanto à extração em fontes não-estruturadas, estas não possuem marcadores para que, orientando-se por eles, uma informação relevante possa ser encontrada por um SEI. Isto torna ainda mais complexa a extração de informação. Para complicar ainda mais a tarefa de extração, informação relevante aparece geralmente fragmentada. Em conseqüência, SEIs para extração de informação de fontes de dados não-estruturadas necessitam de sofisticados meios de identificação da informação desejada, tais como processamento natural de linguagem, análise de similaridade textual e contextos de formatação de texto (estilo e tamanho de fonte, alinhamento de parágrafos, etc.) [Aumman 2006].

1.2.2 Visualização da Informação Extraída

A visualização da informação extraída pode ser estruturada, ou não. No primeiro caso, trata-se de representação de instâncias de entidades previamente definidas, e de

relacionamentos entre instâncias de entidades. No segundo caso, trata-se de identificar as passagens relevantes de texto.

- **Extração de Instâncias de Entidade:** É fornecido um formulário estruturado (*template*), que deve ser preenchido por um SEI por uma coleção de instâncias de uma ou mais entidades [Etzioni 2005]. Como exemplo, em páginas de classificados de veículos na Web a informação de interesse dos usuários (compradores) são instâncias das entidades: descrição, cor, combustível, ano, modelo, preço vendedor e localidade do veículo. Na Figura 1.1a há um exemplo de um trecho de página de classificado de veículos (retirado do site: <http://www.rodao.com.br>). A Figura 1.1b mostra um *template* que deve ser preenchido com valores na página da Figura 1.1a, e a Figura 1.1c exhibe o *template* preenchido com os dados do primeiro veículo na Figura 1.1a. Os valores de entidades, geralmente, são textos curtos, ou algumas palavras.



Figura 1.1: (a) Página Web de classificados de veículos; (b) *template* vazio; (c) *template* preenchido.

- **Extração de Relacionamentos entre Instâncias de Entidade:** Além de instâncias de entidades, muitas vezes é de interesse extrair as instâncias dos relacionamentos entre as entidades [Etzioni 2005]. Para ilustrar, pode-se imaginar um texto que contenha nomes de países e suas capitais; o interesse é preencher o relacionamento $ehCapitalDe(cidade, país)$. A Figura 1.2 mostra um exemplo de texto de onde poderia ser extraída a instância $ehCapitalDe(Paris, França)$ do relacionamento.

"Paris" capital da França, situada em ambas as margens do rio Sena, cerca de 190 km a SE do Havre e 170 km do mar. Um dos maiores centros culturais e intelectuais do mundo, ...

Figura 1.2: Texto com relacionamento implícito: $ehCapitalDe(Paris, França)$.

- **Extração de Passagem de Texto:** Uma passagem de texto é basicamente um texto não-estruturado, de qualquer tamanho e formato (mas não é o texto total de um documento) [Jiang 2006]. A Figura 1.3 mostra parte do texto de um documento e uma passagem delimitada por um retângulo.

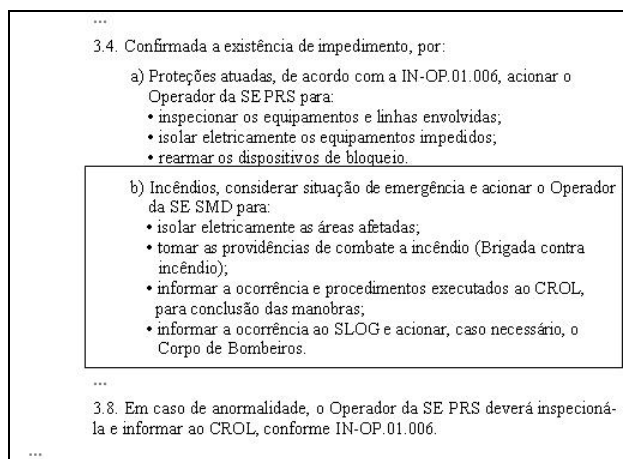


Figura 1.3: Texto com uma passagem destacada.

1.2.3 Classes de Métodos de Extração de Informação

Podem ser identificadas na literatura quatro classes de métodos de extração de informação, em ordem cronológica de seus surgimentos: engenharia do conhecimento, método supervisionado, método auto-supervisionado e extração aberta de informação.

- **Engenharia do Conhecimento.** Esta abordagem requer um especialista para escrever regras de extração (expressões regulares) em uma linguagem específica, para ajudar um SEI a extrair informação de uma coleção de documentos [Hammer, 1997]. Um problema com este método é que a aprendizagem da linguagem específica e a escrita das regras de extração podem representar tarefas muito complexas, exigindo muito tempo despendido pelo especialista.
- **Extração Supervisionada.** Nesta classe, os SEIs aprendem automaticamente a extrair informação, com o auxílio de exemplos de documentos de um domínio marcados por especialistas do domínio. Os exemplos são chamados de *conjunto de treinamento*. Percebe-se o objetivo de minimizar o esforço do especialista, porque um conjunto de treinamento, embora ainda possa ser grande em termos absolutos, é sempre comparativamente pequeno frente à coleção inteira de documentos do domínio. Além do mais, marcar documentos é uma tarefa muito

mais fácil que escrever regras genéricas de extração, muitas vezes complexas. A preparação do conjunto de treinamento, porém, requer esforço e aprendizagem especiais, quanto maior for o conjunto maior será o esforço. Avanços recentes, visando a reduzir o custo do esforço manual, incluem indução automática de marcações apoiada em máquinas de estado finito condicionalmente treinadas [McCallum 2003], ou em redes lógicas de Markov [Poon 2007]. No fundo, nestas abordagens, o esforço de marcação de *corpora* grandes é substituído pelo esforço de construção de complexos modelos de marcação, além de terem uma eficácia de extração longe de satisfatória [Poon 2007].

- **Extração Auto-Supervisionada.** Com o fim de reduzir os custos dos métodos supervisionados, um novo método de extração de informação funciona essencialmente assim: dada uma entidade, um (pequeno) conjunto de padrões genéricos é usado para automaticamente marcar um conjunto de treinamento com anotações específicas da entidade ([Etzioni 2005], [Feldman 2006]). Embora seja auto-supervisionado e “escalável”, o método é dependente de entidade: pode exigir, da parte de especialistas, um laborioso trabalho de identificação das entidades relevantes a um domínio, e a especificação de padrões genéricos para cada uma das entidades escolhidas.

Extração Aberta de Informação. Em [Etzioni 2008], O. Etzioni e outros autores discutem um novo paradigma de extração de informação, fundamentado na idéia de que, pelo menos para algumas línguas, como a inglesa, é viável caracterizar um conjunto compacto de padrões independentes de entidade. O interesse disso reside no fato de que, certos *corpora* — enciclopédias e a própria Web — não podem ser restringidos a um pequeno número de entidades e a domínios específicos pré-selecionados. Em [Schubert 2002], L. Schubert aparece como a primeira tentativa de construção de um SEI aberto, mas suas precisão, revocação e escalabilidade não foram medidas.

Das quatro classes de métodos apresentadas, a supervisionada é ainda a que tem a melhor relação qualidade da extração / custo. Muito esforço ainda terá que ser despendido para que os métodos auto-supervisionados e / ou abertos venham finalmente a se impor.

1.2.4 Métricas de Avaliação de Extração

Quanto à mensuração da qualidade da extração, existem duas métricas muito utilizadas na avaliação da eficácia de um sistema de extração de informação, são as bem conhecidas: precisão e revocação. Precisão é a razão entre a quantidade de informação relevante extraída e a quantidade total de informação extraída. Revocação é a razão entre a quantidade de informação relevante extraída e quantidade total de informação relevante. Em outras palavras: precisão é a porção da informação extraída que de fato é relevante, e revocação é a porção da informação relevante que é extraída [Moens 2006]. Formalmente:

$$\text{precisão} = \frac{\text{quant_relevante_extraída}}{\text{quant_total_extraída}} \quad \text{revocação} = \frac{\text{quant_relevante_extraída}}{\text{quant_total_relevante}}$$

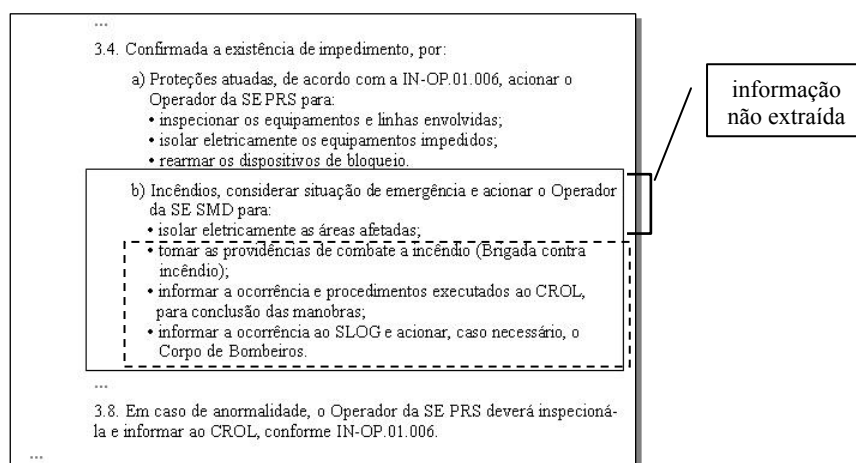


Figura 1.4: Texto com passagem relevante e passagem extraída.

O texto Figura 1.4 possui duas passagens destacadas. Considerando que a passagem dentro do retângulo de linha contínua é a informação relevante, e que a passagem dentro do retângulo de linha tracejada é a informação extraída. O resultado da extração obteve precisão de 100% (a informação extraída é toda relevante), porém, a revocação foi inferior a 100% (parte da informação relevante não foi extraída).

Outra métrica muito utilizada na avaliação da eficácia de um sistema de extração e a medida F. Esta é calculada pela média harmônica entre os valores de precisão e revocação, e retorna um balanceamento entre as duas. Para atingir um alto valor de medida F é necessário que ambos os valores das outras métricas (precisão e revocação) sejam altos. A medida F entre precisão e revocação é formalizada como segue:

$$medida F = \frac{2}{\frac{1}{precisão} + \frac{1}{cobertura}}$$

Definido o contexto da dissertação, Extração de Informação, e os desafios postos à tarefa de extrair informação, passa-se aos objetivos da dissertação.

1.3 Objetivos da Dissertação

Nesta seção são apresentados os objetivos desejados para este trabalho, divididos em gerais e específicos.

1.3.1 Objetivos Gerais

O objetivo geral da dissertação é a proposta de uma nova instância – modelo – de método de extração de informação. As características do modelo são as seguintes:

- *Extração de passagens relevantes de documentos não-estruturados.* O modelo extrai passagem de texto, ou seja, trata-se de informação não estruturada.
- *A solução de extração de informação independente de domínio e formato.* O modelo é aplicável a qualquer domínio, onde existam passagens a serem extraídas de documentos não-estruturados. Quanto à independência de formato, o modelo pode ser aplicado a qualquer formato de arquivo textual.
- *O modelo é uma instância da classe de método supervisionado.* Ou seja, o modelo é um método e a extração das passagens é baseada em um conjunto de treinamento, fornecido pelo usuário. A classe de método supervisionado foi escolhida por ser a de melhor relação qualidade de extração / custo, entre as classes de métodos de extração (Seção 1.2.3).
- *Eficiência vs. Custo.* Como discutido anteriormente, a preparação do conjunto de treinamento pode exigir um esforço de alto para o usuário, e a tentativa de automatizar esta preparação pode reduzir a eficácia de extração. Assim, as inovações do modelo proposto consistem em oferecer uma forma eficiente de extração supervisionada e com baixo custo para o usuário.

1.3.2 Objetivos Específicos

O principal objetivo específico é o desenvolvimento de um protótipo que implementa o modelo proposto: **TIES** (*Textual Information Extraction System*), este deve atender às especificações do modelo.

Os demais objetivos são relacionados ao TIES e estão listados abaixo:

- *Minimização do trabalho do usuário.* O usuário deve ter pouco esforço na preparação do conjunto de treinamento. Seu trabalho será marcar as informações relevantes em uma quantidade pequena de documentos.
- *Validação do protótipo do TIES.* A validação do protótipo deve ser realizada em pelo menos dois domínios de documento. Este requisito é imprescindível para atestar a independência de domínio do TIES. Foram utilizados dois *corpora* de documentos.
 - *Corpus Chesf:* Conjunto de documentos sobre instruções operacionais e normativas de sistemas elétricos. O modelo foi desenvolvido inicialmente para realizar extração de passagens neste *corpus*.
 - *Corpus Legislativo:* Conjunto de documentos sobre leis e decretos da administração de estados brasileiros.
- *Eficácia na Extração.* Quanto à eficácia do TIES, ela deve atingir valores satisfatórios de revocação e precisão nas extrações. Estes valores devem indicar que o sistema poderá ser utilizado em alguma empresa ou organização que tenha passagens relevantes a serem extraídas de documentos não-estruturados.

1.4 Estrutura da dissertação

Os próximos capítulos deste trabalho estão organizados da seguinte forma. O capítulo 2 discorre sobre algumas técnicas para a extração de informação e alguns trabalhos que implementam tais técnicas. O capítulo 3 apresenta o novo modelo para extração de informação: seu método e técnicas empregadas, finalizando com um protótipo de sistema que implementa o modelo. O capítulo 4 exhibe os resultados de testes que experimentalmente avaliaram o protótipo. Finalmente, o capítulo 5 apresenta as conclusões de todo o trabalho e discute sobre trabalhos futuros.

Capítulo 2

Técnicas de Extração de Informação

Neste capítulo, são apresentadas algumas das técnicas de extração de informação mais referenciadas na literatura concernente. É importante ressaltar, desde já, que técnicas e métodos de EI são conceitos ortogonais. Por exemplo, qualquer técnica de EI pode em princípio ser usada no contexto de um método supervisionado, discutido no Capítulo 1; o problema então fica restrito à escolha da melhor técnica, dada a especificidade de uma aplicação de EI. Por razões didáticas, as técnicas são divididas em três categorias: análise de similaridade, passagens relevantes de texto, e segmentação topicamente coerente de texto.

2.1 Extração de Informação por Análise de Similaridade

Identificam-se três classes principais de análise de similaridade: similaridade estrutural, similaridade textual e similaridade de formatação.

Similaridade Estrutural

Documentos são compostos por objetos, exemplo, documentos XML e HTML são compostos por *tags* (marcações) específicas. Medir a similaridade estrutural entre dois documentos é verificar a semelhança, na organização dos objetos, entre os documentos. As estruturas de documentos podem ser representadas por modelos estruturais que reflitam a organização de seus objetos. À guisa de ilustração, a estrutura de um documento HTML pode ser representada por uma árvore, cujos nodos representam suas *tags*.

Existem várias funções que medem a similaridade entre duas árvores: uma das mais conhecidas é *distância de edição entre árvores* [Valiente 2002], que verifica a quantidade de operações, sobre os nodos, necessária para transformar uma árvore em outra — quando menor a quantidade de operações, maior será a similaridade. As

operações são: inclusão, exclusão e substituição de nodo. A Figura 2.1 mostra três árvores: A, B e C. A árvore A pode ser transformada na árvore B pela exclusão do nodo “b” e pela inclusão do nodo “g” — 2 operações no total; já para transformar A em C é necessário excluir os nodos “d”, “e” e b — 3 operações no total. Sendo assim, entre as árvores B e C, a árvore mais similar a A é a B, pela distância de edição.

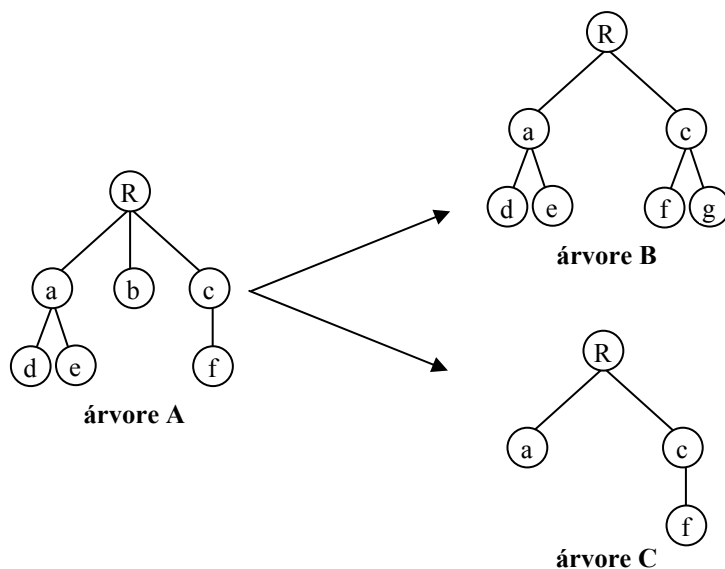


Figura 2.1: Transformação de árvore em outra.

Em [Flesca 2005], é apresentada uma forma de medir a similaridade estrutural entre documentos XML. Neste trabalho, a estrutura de um documento XML é representada como uma série temporal, em que cada ocorrência de uma *tag* corresponde a um pulso na série. Depois, com base na série, um sinal é gerado. Dois documentos XML podem ser comparados por análise da Transformada de Fourier de seus sinais, esta transforma cada sinal em uma frequência. A similaridade entre os documentos é dada pela diferença na magnitude das frequências de seus sinais: quanto menor a diferença, mais similares serão as estruturas dos documentos.

Similaridade Textual

Medir a similaridade textual entre dois documentos é verificar o quando seus textos são semelhantes lexicamente. Mais precisamente, quanto maior for a quantidade de termos em comum, maior será a similaridade, e quanto maior for a quantidade de termos exclusivos, menor será a similaridade.

Uma técnica simples para similaridade textual é a Jaccard [Manning 1999]. Ela representa um texto como um conjunto de termos (*tokens*). A similaridade entre dois textos, com respectivos conjuntos de termos, é medida pela razão entre a quantidade de termos da intercessão dos conjuntos, pela quantidade de termos da união dos mesmos conjuntos.

Outra função, bem conhecida, que mede a similaridade textual é a função *co-seno* [Salton 1968]. Nesta, textos são representados por vetores no espaço vetorial, e o co-seno do ângulo formado por dois vetores é a medida de similaridade entre seus textos.

Assim como para estrutura existe a distância de edição entre árvores, para texto existe a distância de edição entre *strings* [Levenshtein, 1966]. Esta técnica não é baseada em termos, como as duas anteriores. Dois textos (*strings*) são vistos como duas seqüências de caracteres. A distância de edição entre eles é medida pela quantidade de operações, sobre os caracteres, necessária para transformar um texto no outro — quando menor a quantidade de operações, maior será a similaridade. As operações podem ser: inclusão, exclusão e substituição de caracteres.

A fim de realizar uma análise de similaridade mais voltada para a semântica dos textos, estes podem passar por um ou mais pré-processamentos, exemplo: retirada das *stop words* (termos comuns e não relevantes no texto), *stemming* (redução dos termos ao seu radical) e *thesaurus* (substituição de termos sinônimos por um mais genérico). Estes pré-processamentos podem ser empregados antes de medir a similaridade entre textos, utilizando qualquer função de similaridade textual.

Similaridade de Formatação

Uma técnica de similaridade de formatação de dois textos considera somente as características visuais dos mesmos. As características visuais mais exploradas são: tipo, estilo e tamanho de fonte; alinhamento do texto em um documento, quantidade de colunas do texto, etc. Uma função de similaridade com base em características visuais foi utilizando no sistema *Visual Information Extraction* [Aumann 2006], que será apresentado pela Seção 2.1.3.

Na seqüência, são discutidos três trabalhos de EI que fazem análise de similaridade segundo as três classes de técnicas apresentadas.

2.1.1 Extração Automática de Notícias na Web

Reis [Reis et al. 2004] apresenta uma abordagem, dependente de domínio, para extração não-supervisionada de informação não estruturada em páginas de notícias da web. A solução é baseada em análise de similaridade estrutural. Para realizar esta análise foi desenvolvido o algoritmo RTDM – *Restricted Top-Down Mapping*, que é uma variação do algoritmo de distância de edição entre árvores.

No processo, são selecionadas várias páginas HTML e geradas árvores que representam as estruturas de cada uma delas. Para a geração da árvore que representa uma página, informação entre *tags* é um nodo. As páginas são agrupadas formando *clusters*, de acordo com a similaridade, entre suas árvores, medida pelo RTDM. Este algoritmo também faz um mapeamento entre nodos nas árvores. Depois, para cada *cluster* é gerada uma expressão regular – *ne-pattern (node extraction pattern)* – que também é uma árvore e casa com todas as árvores do *cluster*. Na extração das informações de uma nova página de notícias, primeiro é selecionado o *ne-pattern* mais apropriado para a árvore da página, utilizando a similaridade do RTDM. Pelo casamento entre o *ne-pattern* e a árvore da página, algumas informações, ricas em texto, são extraídas da nova página.

O objetivo específico do trabalho foi a extração dos títulos e dos corpos das notícias nas páginas. Estas informações são ricas em texto, assim, são extraídas pelo processo anterior. Dos textos extraídos, da nova página de notícia, um é o título e outro é o corpo da notícia. Heurísticas são utilizadas para identificar quais são os textos que representam estas informações: (a) aquele que possuir mais de 100 palavras será o corpo; e (b) aquele que possuir entre 1 e 20 palavras, além da máxima interseção com o corpo, será o título.

A abordagem foi avaliada em 4088 páginas de notícias na web, todas de sites de notícias brasileiros. Apesar da heurística muito simples na identificação de títulos e corpos, a solução obteve uma boa acurácia de identificação.

2.1.2 Análise de Similaridade de Documentos Normativos

Lau [Lau et al 2003] descreve um sistema de mineração de documentos de regulamentações governamentais. O objetivo do trabalho é encontrar, em um documento, uma seção de texto que seja mais similar textualmente a uma seção

específica de um segundo documento. A abordagem é útil quando o interesse é identificar seções que tratam do mesmo assunto em documentos de regulamentações distintos, e que foram produzidos por diferentes fontes (ex.: federal, estadual, municipal).

Para medir a similaridade entre as seções dos documentos, o sistema primeiro transforma-os em documentos semi-estruturados com a adição de *tags* XML, documentos de regulamentações geralmente são não-estruturados. Assim, tem-se a produção de novos documentos XML. Cada seção dos documentos originais estará em um elemento no XML. Depois, cada documento XML é representado como uma árvore que descreve sua estrutura, e assim, cada nodo na árvore representa uma seção no documento. A identificação de uma seção específica é transformada em identificação de um nodo específico em uma árvore. Supondo a seguinte tarefa de identificação: seja um nodo A pertencente à árvore X, deseja-se encontrar em uma árvore Y o nodo B mais similar a A. O sistema calcula a similaridade entre A e todos os nodos em Y, identificando o que for mais similar a A. No cálculo da similaridade entre dois nodos, são consideradas a similaridade entre eles e a similaridade entre seus nodos vizinhos (pai, irmãos e filhos). A função de similaridade utilizada é a co-seno. A similaridade entre dois nodos (seções) não é feita diretamente utilizando os textos neles contidos, mas sim frases substantivas extraída dos textos. Para a tarefa de extração das frases substantivas, foram utilizadas ferramentas de processamento natural de linguagem.

O sistema foi avaliado utilizando documentos de regulamentações sobre acessibilidade, oriundos de órgãos americanos e britânicos. No entanto, não foi apresentado, no artigo, nenhum valor quantitativo que comprovasse a eficácia da solução, apenas foi demonstrado um único exemplo onde a solução obteve sucesso.

2.1.3 Extração Visual de Informação

O sistema *Visual Information Extraction* (VIE) [Aumman 2006] extrai informações relevantes de documentos não-estruturados. A extração é independente de domínio e supervisionada – é necessário que o usuário marque as informações relevantes em um conjunto de documentos de treinamento. Com base nas marcações o VIE identifica automaticamente as informações relevantes em novos documentos. A solução utiliza uma abordagem com análise de similaridade de formatação entre os documentos.

No processo de extração de informação em um novo documento, o VIE modela as estruturas dos documentos (novo e treinamento) em árvores, onde as folhas são objetos primitivos de texto (termos), e os nodos internos são objetos compostos de primitivos (linhas, parágrafos). Esta árvore é a *O-Tree* (*Object-Tree*), e ela obedece a uma hierarquia de tipos de objetos: os objetos de mais alto nível são grupos de objetos de níveis inferiores. Depois, é selecionada a *O-Tree* do conjunto treinamento que é mais similar à *O-Tree* do novo documento. Na comparação, é considerada a similaridade entre objetos de mesmo nível nas árvores. A função de similaridade usada é baseada nas características de formatação dos objetos (tamanho e tipo de fonte, localização, etc.). Em seguida, para cada objeto marcado como relevante na *O-Tree* selecionada, é procurado, na *O-Tree* do novo documento, o objeto mais similar à ele, utilizando a mesma função de similaridade. Se o valor de similaridade for maior que um limiar k , o objeto encontrado será extraído como informação relevante.

O VIE foi testado em um conjunto de documentos de relatórios analíticos sobre investimentos bancários, de quatro fontes distintas. Foram extraídas informações como nome do autor, título, subtítulo, entre outras. O VIE obteve uma alta acurácia nos testes. No entanto, a solução apresenta duas grandes limitações: (a) O VIE só pode identificar informação com características visuais distintas (formatação) dos textos que as cercam; e (b) a solução é para documentos que sejam bastante similares em estrutura e formatação, não sendo indicada quando estas características são muito variáveis entre os documentos.

Concluída a seção sobre extração de informação por análise de similaridade, a próxima apresenta técnicas de extração de passagens relevantes.

2.2 Extração de Passagens Relevantes

Passagens de texto são informações de maior granularidade em comparação a instâncias de entidades. Mais precisamente, uma passagem é um segmento, de qualquer tamanho, de um texto; Por exemplo, uma passagem poderia compreender um ou mais parágrafos. As passagens relevantes de um documento são aquelas que possuem informação de interesse do usuário, ou seja, são relevantes ao usuário. Os trabalhos discutidos propõem técnicas de extração de passagens relevantes de documentos.

2.2.1 Extração de Passagens Coerentes Usando Modelos de Markov

Em [Jiang 2006], é proposta uma técnica de extração de passagens coerentes, relevantes e de tamanho variado, em documentos não-estruturados. Coerência de uma passagem diz respeito ao(s) tema(s) que ela trata, o texto de uma passagem coerente trata de um único tema. A solução é baseada em modelos de Markov (*Hidden Markov Models* – HMMs) [Rabiner 1990]. Em resumo, HMM é uma máquina probabilística de estados finitos, que tem entradas e saídas.

Dado um consulta do usuário, a solução identifica em um documento qual é a passagem de texto que é mais relevante à consulta. A idéia central é que, uma passagem relevante é uma seqüência de palavras que pode ser reconhecida por um HMM. Seja uma consulta q e um documento d . Este documento é representado como uma seqüência de palavras $d = (w_1, w_2, \dots, w_n)$ e a passagem relevante é uma subseqüência de d que é mais relevante a q . A passagem relevante é reconhecida por um estado relevante, que reconhece as palavras pertencentes à ela, de acordo com uma distribuição de probabilidade. Em oposição, as partes não relevantes do documento são reconhecidas por estados não relevantes, de acordo com outra distribuição de probabilidade. A distribuição de probabilidade do estado relevante é altamente relacionada à consulta do usuário. Na tarefa de extração de uma passagem relevante, o modelo verifica, para cada palavra do documento, se ela é reconhecida pelo estado relevante ou por um estado não relevante. A seqüência de palavras no texto do documento reconhecidas pelo estado relevante é a passagem relevante procurada.

Para os testes, foram implementadas três versões da solução, que diferem na forma de como foram estimados os parâmetros para as probabilidades do estado relevante. Estas três versões foram confrontadas com outras soluções para a extração de passagens relevantes, chamadas no artigo de métodos *baseline*. Foram utilizados dois corpora para os testes: um sintético e o TREC HARD [Allan 2003], corpus muito utilizado para avaliar sistema de extração de passagens relevantes. Os testes demonstram um melhor desempenho do método proposto, em relação aos *baseline*. Entretanto, a eficácia da solução ainda é baixa.

2.2.2 Recuperação de Passagem usando Similaridade entre Vértices de um Grafo

Dkaki [Dkaki 2007] apresenta um modelo de recuperação de informação capaz de identificar as k passagens de um documento mais similares a uma consulta de usuário. O modelo representa os textos das passagens e da consulta como vértices de um grafo, e através de uma função para medir a similaridade entre os vértices, são identificadas as passagens mais similares à consulta.

No processo de identificar as passagens de um documento mais similares a uma consulta, o modelo constrói um grafo bi-partido. Todas as passagens do documento, a consulta e os termos (palavras presentes nas passagens e consulta) são representados por um vértice no grafo. Vértices são de dois tipos: os vértices que representam passagens e consulta são “vértices texto” e os que representam termos são “vértices termo”. As arestas incluídas no grafo ligam um vértice termo a um vértice texto, indicando que o termo, do vértice termo, é encontrado no texto do vértice texto. No grafo não existem arestas que liguem vértices de mesmo tipo. Depois, o modelo constrói uma matriz de similaridade que, inicialmente, registra os valores de similaridade de co-seno entre todos os vértices texto, cada célula registra a similaridade entre dois vértices. Em seguida, o valor de cada célula é alterado dando lugar a um novo valor, este é calculado por uma nova e complexa função de similaridade. Finalmente, as linhas na matriz que correspondem aos vértices do texto da consulta são excluídas, ficando somente os vértices do texto do documento, então a matriz é ordenada de modo que os vértices texto com maior similaridade ocorram primeiro, os k primeiros vértices texto na matriz representam as k passagens do documento que são mais similares à consulta.

O método foi testado utilizando o corpus TREC Novelty 2004 [Soboroff 2004], que é utilizado para avaliar sistemas de extração. Foi definido que para os testes, uma passagem de um documento é na verdade uma sentença do mesmo. Foram utilizadas técnicas de pré-processamento das sentenças: remoção das *stop words* e *stemming*. Os resultados dos testes foram confrontados com resultados obtidos por outro modelo – considerado *baseline* – no mesmo corpus. Os valores comprovam que o novo método é bem superior ao *baseline* quando k é pequeno, reduzindo a eficácia à medida que k aumenta.

A seção seguinte apresenta soluções para a segmentação topicamente coerente de textos.

2.3 Segmentação Topicamente Coerente de Texto

Algumas técnicas de EI têm o objetivo de dividir um texto, geralmente longo, em segmentos topicamente coerentes. Tópico é relativo ao tema de uma parte do texto, deste modo, estas técnicas delimitam um texto nos pontos em que o tema muda. Dentro de cada segmento há uma coesão léxica, termos (palavras) são semelhantes, isto é o que determina sua delimitação. A seguir são apresentados dois trabalhos que realizam a segmentação topicamente coerente de texto.

2.3.1 TextTiling

TextTiling [Hearst 1997] é um dos mais usados e conhecidos algoritmos de segmentação topicamente coerente de texto. Segmentos são delimitados de acordo com padrões de co-ocorrência léxica e sua distribuição em cada texto.

Para realizar a segmentação de um texto, inicialmente este é pré-processado: remoção das *stop words* e *stemming*. Em seguida, o algoritmo divide o texto em blocos de tamanho fixo (tamanho é parâmetro). Utilizando uma janela deslizante, um escore de similaridade léxica entre cada par de blocos adjacentes é medido. Nos cálculos de similaridade, o *TextTiling* pode utilizar três diferentes meios para analisar a co-ocorrência léxica, que são: similaridade de co-seno, adição de novos vocábulos e cadeias léxicas.

Similaridade de co-seno já foi discutida na Seção 2.1. Na adição de novos vocábulos, é verificada a soma dos novos termos – que são observados pela primeira vez no texto – nos blocos adjacentes, este valor é normalizado pelo total de termos nos dois blocos. Quanto às cadeias léxicas, o *TextTiling* procura pela ocorrência de seqüências de termos similares que ocorram entre os blocos de texto, estas são as cadeias léxicas; a similaridade entre dois blocos é medida pela quantidade de cadeias que cruzam estes blocos.

Utilizando qualquer um dos três métodos de análise de co-ocorrência léxica, o *TextTiling* gera um gráfico de curvas onde os valores de similaridade são “plotados”. Os pontos no eixo das ordenadas representam os valores de similaridade obtidos. Caso seja

constatada a presença de vales na curva (baixos valores de similaridade), isto indica a mudança de tópicos e onde delimitadores de segmentos devem ser inseridos.

Para avaliar a segmentação do *TextTiling*, foram utilizados textos em artigos de revistas. Para analisar a eficácia do algoritmo, as posições dos delimitadores de segmentos inseridos nos textos pelo *TextTiling* foram confrontadas com as posições de delimitadores inseridos por humanos. Em outros testes, vários artigos inteiros foram concatenados em um único documento, então foi verificada a capacidade que o algoritmo tinha de delimitar o fim de um artigo e o início do subsequente. Os resultados da avaliação relevam que o *TextTiling* é uma boa solução à segmentação de texto em partes topicamente coerentes.

2.3.2 Minimum Cut

Malioutov [Malioutov et al 2006] transforma o problema de dividir um texto, por segmentação topicamente coerente, em uma tarefa de partição de grafos. Para tal tarefa, foi desenvolvido o novo algoritmo de segmentação textual chamado de *Minimum Cut*.

Na tarefa de segmentação de um texto, é criado um grafo valorado e totalmente conectado (grafo completo) que representa o texto a ser segmentado. Neste grafo, as sentenças do texto são seus vértices, e o valor de cada aresta é o escore da similaridade entre as sentenças dos vértices ligados por ela. Para efeito de menor esforço computacional, todos os vértices cujos valores não alcançam um determinado linear x são excluídos do grafo. A função de similaridade é a co-seno e pré-processamentos são realizados nas sentenças antes das avaliações de similaridade: remoção das *stop words* e *stemming*.

A solução funciona da seguinte forma: seja um grafo $G = \{V, E\}$, de V vértices e E arestas, *Minimum Cut* começa dividindo G em duas partições, A e B . A divisão é orientada de forma que a similaridade entre os vértices de A seja maior que a similaridade dos vértices de A com os vértices de B e vice-versa. O objetivo é minimizar a soma dos valores das arestas que cruzam as partições, e maximizar esta soma para as arestas dentro de cada partição. Para encontrar as partições que atendam a este requisito, o algoritmo faz uso de programação dinâmica. O processo de divisão é repetido até dividir o grafo em k partições (k é parâmetro do algoritmo). Cada partição representa um conjunto de sentenças seqüências no texto, ou seja, um segmento.

O *Minimum Cut* foi avaliado utilizando três corpora distintos de texto: textos sobre física, textos sobre inteligência artificial e um corpus sintético produzido por Choi em [Choi 2000] e muito utilizado na avaliação de algoritmos de segmentação. Segundo a avaliação, a análise global de similaridade entre todas as sentenças do texto melhora a eficácia da segmentação. Os resultados dos testes mostraram que a solução é uma boa alternativa à divisão de textos em segmentos topicamente coerentes.

2.4 Conclusões

Este capítulo apresentou algumas técnicas que são utilizadas em métodos de extração de informação. Um método pode contemplar uma única técnica ou uma combinação delas, sejam de uma ou mais categorias.

A combinação de mais de uma técnica pode ser empregada de acordo com os requisitos do método. Como exemplo, se o requisito for alta eficácia de extração, a combinação de análises de similaridade (estrutural e textual) pode resultar em uma eficácia maior que a resultante de um único método. Outro exemplo, se o requisito for extração de passagens topicamente coerentes, a técnica apresentada na Seção 2.2.1 (Extração de Coerentes Passagens Usando Modelos de Markov) poderá ser utilizada, no entanto, sua eficácia ainda é baixa. Uma outra alternativa é a combinação formada pela técnica de segmentação topicamente coerente de texto com a técnica de extração de passagens relevantes, desta é possível extrair passagens de texto que sejam topicamente coerentes.

O capítulo seguinte apresenta um novo método de extração que emprega uma combinação de várias técnicas de extração.

Capítulo 3

Proposta de um Novo Método de Supervisão e de uma Nova Técnica de Extração de Informação

O objetivo deste capítulo é apresentar um novo método de extração de informação, da classe de métodos supervisionados, de documentos não-estruturados. Este novo método tem como principais características: eficácia na extração e baixo custo na preparação do conjunto de treinamento, além de ser independente de domínio e formato de arquivo. Além disso, ele emprega uma técnica de extração que é uma combinação de técnicas existentes. Ao final do capítulo, é apresentado um sistema de extração que implementa o método proposto e foi utilizado na validação do método e de sua técnica.

3.1 O Método de Extração de Informação

Como mencionado ao final do Capítulo 1, os métodos de extração supervisionados são a classe de melhor relação eficácia / custo. Entretanto, a preparação do conjunto de treinamento pode ainda requerer um esforço de alto custo para o usuário, e a tentativa de diminuir este esforço pode resultar em redução na eficácia de extração.

Nesta seção, uma solução para este problema é proposta. Trata-se de um novo método de extração que é uma instância da classe supervisionada, e suas principais características são: extração de passagens relevantes de documentos não-estruturados, eficácia na extração, pequeno custo na preparação do conjunto de treinamento, independência de domínio e independência de formato.

- Em relação ao método supervisionado, este é do tipo *lazy learner* [Feldman and Sanger 2007], que posterga a indução de padrão de extração para o momento de extrair a passagem. A tarefa de extração pode ser resumida da seguinte forma: dado um conjunto de documentos onde as passagens relevantes estão marcadas

(destacadas), o método identifica e extrai estas passagens de novos documentos. Os documentos onde as passagens já estão marcadas formam o conjunto de treinamento, e os novos documentos são chamados documentos de consulta.

- Para reduzir o custo do esforço do usuário, o método não exige muitos exemplos de treinamento. Para a extração de passagens relevantes de documentos de um domínio, o usuário necessita preparar um conjunto de treinamento pequeno, ou muito pequeno, podendo até ter um único documento.
- Com relação a sua eficácia, o novo método alcança altos valores de revocação e precisão nas extrações. As definições destas métricas, no contexto de extração de informação, foram apresentadas na Seção 1.2.4.
- Sobre a independência de domínio, o método é aplicável a qualquer domínio, onde existam passagens a serem extraídas e um conjunto de treinamento possa ser preparado.
- O método pode ser aplicado para processar qualquer formato de arquivo de documento, desde que seja possível recuperar (extrair) seus textos.

Todas as características apresentadas serão comprovadas no capítulo de avaliação, nesta dissertação. A seção seguinte apresenta a técnica de extração utilizada no método proposto.

3.2 A Técnica de Extração de Informação

A identificação das passagens relevantes, nos documentos de consulta, é feita por análises de similaridade entre o texto destes documentos com os textos dos documentos no conjunto de treinamento. Extração por análise de similaridade foi discutida na Seção 2.1.

O novo método de extração emprega uma nova técnica de extração, que é uma combinação de técnicas existentes: similaridade estrutural, similaridade textual e segmentação topicamente coerente de texto. Estas técnicas são executadas em dois diferentes processos, descritos resumidamente a seguir.

- A análise de similaridade estrutural ocorre no processo chamado de Marcação de Estrutura. Dado um conjunto de treinamento e um documento de consulta, este

processo identifica, no conjunto de treinamento, qual é o documento mais similar estruturalmente ao documento de consulta. Uma vez identificado este documento, ele e o documento de consulta passam ao processo que analisa a similaridade textual.

- O processo chamado de Marcação de Passagem realiza a análise de similaridade textual entre o documento de consulta e o de treinamento, passados pelo processo anterior. Por análises de similaridade textual, as passagens relevantes e marcadas no documento treinamento são induzidas, e posteriormente marcadas como relevantes, no documento de consulta. Apesar da técnica ser de “extração” de passagem, o método propõe que as passagens induzidas sejam marcadas (destacadas) no documento de consulta, assim como são marcadas no documento de treinamento. A análise de similaridade textual é apoiada por técnica de segmentação topicamente coerente de texto. Segmentação topicamente coerente foi discutida na Seção 2.3.

Neste ponto, é importante deixar bem claro que os documentos de treinamento (conjunto de treinamento) já devem possuir suas passagens relevantes marcadas, eles formam a base de conhecimento do método. Quando um documento de consulta passa pelos dois processos, do método de extração, suas passagens são automaticamente identificadas, com base nas passagens marcadas nos documentos de treinamento. Portanto, identificar passagens específicas em documentos de consulta é o objetivo do método. Cada execução do método é sobre um documento de consulta, a marcação de passagens neste documento é o resultado da execução.

Antes de apresentar em detalhes os processos que compõem o método, será apresentada uma discussão sobre estrutura implícita e seções em documentos não-estruturados, conceitos necessários ao entendimento dos processos.

Estrutura implícita de documentos não-estruturados

Apesar de documentos não-estruturados não possuírem uma estrutura bem definida, alguns podem apresentar o texto organizado de tal forma que seja possível identificar uma estrutura implícita. Esta estrutura implícita pode ser descrita por uma hierarquia de níveis que é formada pelo aninhamento dos títulos das seções e subseções

do documento, se ele as possuir. Como exemplo, para a Figura 3.1 a estrutura implícita deste pode ser:

1. PROCEDIMENTOS INICIAIS
2. REENERGIZAÇÃO DA INSTALAÇÃO
 - 2.1. Reenergização sem Impedimento
 - 2.2. Reenergização com Impedimento
3. PROCEDIMENTOS GERAIS

1.	PROCEDIMENTOS INICIAIS
1.1	Identificar se o desligamento foi "geral", caracterizado pela falta total de tensão na instalação, (barras e linhas/links) exceto serviços auxiliares, e a inexistência de carregamento nas linhas/trafos/links. Caso seja "geral", proceder a partir do item 1.2. Caso contrário, prosseguir conforme os procedimentos descritos na IN-OP.01.006.
...	...
1.13	Caso haja impedimento para reenergização da instalação, prosseguir a partir do item 2.2.
2.	REENERGIZAÇÃO DA INSTALAÇÃO
2.1	Reenergização sem Impedimento
	Rearmar as chaves de bloqueio atuadas, independente de autorização do Operador de Sistema.
	Proceder a preparação e reenergização na seqüência descrita nos Anexos, efetuando contato com o CROL no item que exigir sua autorização.
2.2	Reenergização com Impedimento
	Informar ao CROL:
	<ul style="list-style-type: none">• equipamentos impedidos ou outros impedimentos, e motivo dos mesmos;• principais sinalizações indicadas, chaves de bloqueio e proteções atuadas;• resultados da inspeção no pátio e as ações de isolamento adotadas;• providências tomadas e configuração atual da Instalação.
	Preparar e restabelecer a Instalação sob orientação do CROL.
3.	PROCEDIMENTOS GERAIS
3.1	Efetuar inspeção geral, anotar relés e bandeirolas atuadas, registrando os resultados nos formulários "Lay-out Chassi Proteção" (Anexo III da IN-OP.01.006).
...	...
3.7	Caso haja uma tentativa de reenergização sem sucesso, devido a ocorrências na própria Instalação, torna-se necessário contatar o CROL.

Figura 3.1: Documento com seções e subseções.

Os títulos numerados por 1, 2, e 3 pertencem ao 1º nível na hierarquia, os numerados por 2.1 e 2.2 pertencem ao 2º nível. A estrutura do documento da Figura 3.1 pode ser indicada por marcações com *tags* XML. A Figura 3.2 mostra o mesmo documento, agora com as marcações de estrutura pelas *tags*: <NIVEL1> e <NIVEL2>. A *tag* <DOCUMENTO> marca o texto de todo o documento, ela representa a raiz da estrutura.

```

<DOCUMENTO>

<NIVEL1> 1.  PROCEDIMENTOS INICIAIS </NIVEL1>
1.1  Identificar se o desligamento foi "geral", caracterizado pela falta total de tensão na
instalação, (barras e linhas/links) exceto serviços auxiliares, e a inexistência de
carregamento nas linhas/trafos/links. Caso seja "geral", proceder a partir do item 1.2.
Caso contrário, prosseguir conforme os procedimentos descritos na
IN-OP.01.006.
...
1.13. Caso haja impedimento para reenergização da instalação, prosseguir a partir do item 2.2.
<NIVEL1> 2.  REENERGIZAÇÃO DA INSTALAÇÃO </NIVEL1>
<NIVEL2> 2.1  Reenergização sem Impedimento </NIVEL2>
Rearmar as chaves de bloqueio atuadas, independente de autorização do Operador de
Sistema.
Proceder a preparação e reenergização na sequência descrita nos Anexos, efetuando
contato com o CROL no item que exigir sua autorização.
<NIVEL2> 2.2  Reenergização com Impedimento </NIVEL2>
Informar ao CROL:
• equipamentos impedidos ou outros impedimentos, e motivo dos mesmos;
• principais sinalizações indicadas, chaves de bloqueio e proteções atuadas;
• resultados da inspeção no pátio e as ações de isolamento adotadas;
• providências tomadas e configuração atual da instalação.
Preparar e restabelecer a instalação sob orientação do CROL.
<NIVEL1> 3.  PROCEDIMENTOS GERAIS </NIVEL1>
3.1.  Efetuar inspeção geral, anotar relés e bandeirolas atuadas, registrando os resultados nos
formulários "Lay-out Chassi Proteção" (Anexo III da IN-OP.01.006).
...
3.7.  Caso haja uma tentativa de reenergização sem sucesso, devido a ocorrências na própria
instalação, torna-se necessário contatar o CROL.

</DOCUMENTO>

```

Figura 3.2: Documento com marcações de estrutura.

A estrutura de um documento pode ser representada de maneira simples como uma **árvore de estruturas**. A Figura 3.3 mostra a árvore de estrutura para o documento da Figura 3.2. Cada título marcado na estrutura do documento é representado por um nodo na árvore. Toda árvore possui um nodo raiz, no caso é o nodo <DOCUMENTO>. Neste trabalho, os termos “estrutura de documento” e “árvore de estrutura” serão tratados como sinônimos.

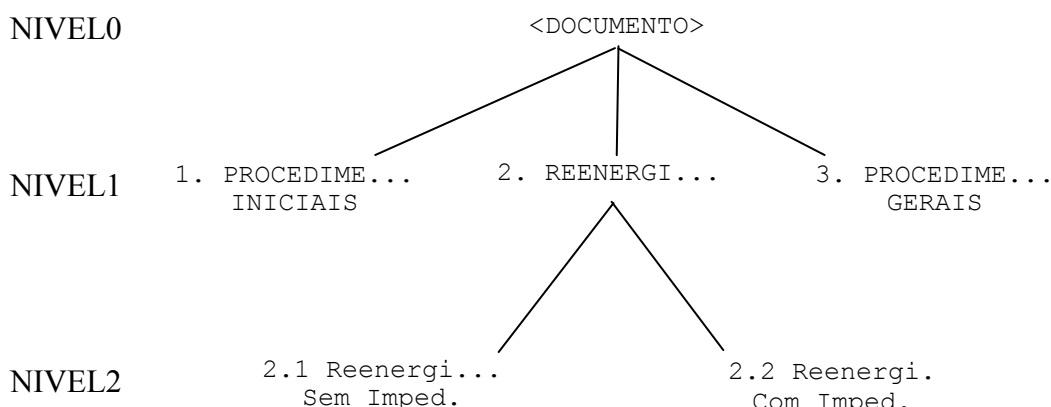


Figura 3.3: Árvore de estrutura do documento da Figura 3.2.

Definição de Seção e Subseção de Texto

Seção é um fragmento de texto composto de título e seguido de corpo: o título pertence a um nível na hierarquia da estrutura do documento; o corpo é a porção de texto entre o título da seção e o título imediatamente subsequente e de mesmo nível. Seções que se apresentam dentro de outra são chamadas de subseções.

A Figura 3.4 destaca as delimitações das seções e subseções do documento da Figura 3.1. As seções estão delimitadas por retângulos tracejados e as subseções por retângulos pontilhados.

1.	PROCEDIMENTOS INICIAIS
1.1	Identificar se o desligamento foi "geral", caracterizado pela falta total de tensão na instalação, (barras e linhas/links) exceto serviços auxiliares, e a inexistência de carregamento nas linhas/trafos/links. Caso seja "geral", proceder a partir do item 1.2. Caso contrário, prosseguir conforme os procedimentos descritos na IN-OP.01.006.
...	...
1.13	Caso haja impedimento para reenergização da instalação, prosseguir a partir do item 2.2.
2.	REENERGIZAÇÃO DA INSTALAÇÃO
2.1	Reenergização sem impedimento
	Rearmar as chaves de bloqueio atuadas, independente de autorização do Operador de Sistema.
	Proceder a preparação e reenergização na seqüência descrita nos Anexos, efetuando contato com o CROL no item que exigir sua autorização.
2.2	Reenergização com impedimento
	Informar ao CROL:
	<ul style="list-style-type: none">• equipamentos impedidos ou outros impedimentos, e motivo dos mesmos;• principais sinalizações indicadas, chaves de bloqueio e proteções atuadas;• resultados da inspeção no pátio e as ações de isolamento adotadas;• providências tomadas e configuração atual da instalação.
	Preparar e restabelecer a instalação sob orientação do CROL.
3.	PROCEDIMENTOS GERAIS
3.1	Efetuar inspeção geral, anotar relés e bandeirolas atuadas, registrando os resultados nos formulários "Lay-out Chassi Proteção" (Anexo III da IN-OP.01.006).
...	...
3.7	Caso haja uma tentativa de reenergização sem sucesso, devido a ocorrências na própria instalação, torna-se necessário contatar o CROL.

Figura 3.4: Documento com delimitações de seções e subseções.

Definidas estrutura e seções de documento, os processos Marcação de Estrutura e Marcação de Passagem podem ser explicados em detalhes.

Por motivos de simplificação de escrita, será usado o termo documento-consulta de em vez de "documento de consulta", documento-treinamento em vez de "documento de treinamento" e documentos-treinamento em vez de "conjunto de treinamento". Além disso, o termo "seção" será genérico e indicará seções ou subseções.

3.2.1 Marcação de Estrutura

Este processo recebe um documento-consulta e documentos-treinamento, que podem ser fornecidos pelo usuário. Os documentos-treinamento, quando recebidos, devem ter suas estruturas marcadas. A marcação pode ser feita por *tags* XML, como no exemplo da Figura 3.2. O documento-consulta não precisa ter marcação de estrutura.

Os principais objetivos do processo são: induzir marcação de estruturas para o documento-consulta e identificar o documento-treinamento que é mais similar estruturalmente ao documento-consulta. A execução deste processo é dividida em três fases: A) Indução de Estrutura, B) Análise de Similaridade Estrutural e C) Equivalência entre Árvores.

A) Indução de Estrutura

Quando um documento-consulta é recebido para marcação de estrutura, são induzidas várias estruturas para ele, cada uma é construída com base na estrutura de um documento-treinamento. A indução de cada estrutura é realizada pelo algoritmo **TSI** (*Tree-Structure Inducer*), e sua execução é explicada a seguir.

Considerando que o documento da Figura 3.2, com árvore de estrutura representada pela Figura 3.3, pertence aos documentos-treinamento e considerando o documento da Figura 3.5 um documento-consulta: quando o documento-consulta passa pelo processo de Marcação de Estrutura, o algoritmo TSI irá induzir uma árvore de estrutura baseada na árvore do documento-treinamento. Na verdade, o TSI irá induzir marcação de estrutura para o documento-consulta.

1.	PROCEDIMENTOS INICIAIS
1.1	Comunicar ao Operador de Sistema, em até três minutos, as informações preliminares (flash), baseado nas informações disponibilizadas pelos meios de supervisão existentes: <ul style="list-style-type: none"> • Horário da ocorrência; • Equipamentos e/ou linhas que desligaram.
...	...
1.10.	Caso haja impedimento para reenergização da Instalação, prosseguir a partir do item 2.2.
2.	REENERGIZAÇÃO
2.1.	Sem Impedimento
2.1.1.	As chaves de bloqueios devem ser rearmadas
2.1.2.	Executar seqüência descrita nos Anexos, efetuando contato com o CROL no item que exigir.
2.2.	Com Impedimento
2.2.1.	O operador deve informar ao CROL os seguintes itens: <ul style="list-style-type: none"> • Equipamentos e linhas de transmissão impedidas, assim como seus motivos; • Sinalizações indicativas, chaves de bloqueio e proteções; • Ações de isolamento tomadas
2.2.2.	Comunicar ao CROL em até dez minutos, as informações complementares descritas acima.
2.2.3.	Preparar e restabelecer a Instalação sob orientação do CROL.
3.	PROCEDIMENTOS GERAIS
3.1.	Caso ocorra impedimento de equipamento durante a reenergização cujo item não possa ser pulado ou substituído, contatar com o CROL.
...	...
3.9.	Caso haja uma tentativa de reenergização sem sucesso, devido a ocorrências externas à Instalação, e não havendo impedimento, prepará-la novamente e reenergizá-la conforme Anexo.

Figura 3.5: Exemplo de documento-consulta.

Para cada título de seção marcado na estrutura do documento-treinamento, o TSI irá procurar, no documento-consulta, um fragmento de texto que seja semelhante textualmente a ele. TSI leva em conta somente o texto do título e fragmento, ignorando caracteres que não sejam literais. Para medir a similaridade entre os textos, qualquer função poderá ser utilizada, algumas funções foram comentadas na Seção 2.1 (desta dissertação). No protótipo construído para validar o método (Seção 3.3), foi utilizado a função de similaridade de co-seno, que retorna um escore entre 0 e 1. Uma vez a estrutura do documento-consulta induzida, tem-se uma correspondência entre os elementos nesta estrutura com elementos na estrutura do documento-treinamento.

Na versão corrente, TSI é um algoritmo *greedy* (guloso) [Cormen 2001] e executa da seguinte forma:

Passo 1: TSI monta uma lista com os títulos marcados na estrutura do documento-treinamento, na ordem em que eles aparecem no documento. No caso do exemplo da Figura 3.3, a lista é formada pelos títulos:

1. PROCEDIMENTOS INICIAIS
2. REENERGIZAÇÃO DA INSTALAÇÃO

2.1. Reenergização sem Impedimento

2.2. Reenergização com Impedimento

3. PROCEDIMENTOS GERAIS

Passo 2: Para cada título da lista do passo 1, pertencente ao primeiro nível na hierarquia (no exemplo são os numerados por 1, 2 e 3), TSI procura no documento-consulta um fragmento de texto que seja X% similar a ele e que ainda não tenha marcação de estrutura.

- A procura considera o documento inteiro, ou seja, começa no início do documento-consulta e pára quando encontrar o primeiro fragmento que atenda ao requisito de similaridade ou quando atingir o final do documento. No último caso de parada, é considerado que não foi encontrado um fragmento similar ao título.
- O fragmento encontrado é marcado pela mesma *tag* que marca o título do documento-treinamento, ex.: <NIVEL1> para <NIVEL1>. Após a marcação, o fragmento passa a ser um título de seção na estrutura do documento-consulta e um nodo na sua árvore. O título do documento-treinamento e o marcado no documento-consulta são chamados **títulos correspondentes**, assim como suas respectivas seções são **seções correspondentes** e seus respectivos nodos são **nodos correspondentes**. Isto porque, todo “nodo” representa um “título” de uma “seção” de um documento.
- Ao final do passo 2, o documento-consulta estará com marcações indicando seções no nível 1.
- Para o exemplo, os títulos correspondentes gerados pelo passo 1 são mostrados na Tabela 3.1, a Figura 3.6 mostra o documento-consulta com as primeiras marcações de estrutura.

Tabela 3.1: Correspondência entre títulos: documento-treinamento e documento-consulta

Correspondência entre títulos	
Título no documento-treinamento	Título no documento-consulta
1. PROCEDIMENTOS INICIAIS	1. PROCEDIMENTOS INICIAIS
2. REENERGIZAÇÃO DA INSTALAÇÃO	2. REENERGIZAÇÃO
3. PROCEDIMENTOS GERAIS	3. PROCEDIMENTOS GERAIS

<NIVEL1> 1. PROCEDIMENTOS INICIAIS </NIVEL1>	
1.1	Comunicar ao Operador de Sistema, em até três minutos, as informações preliminares (flash), baseado nas informações disponibilizadas pelos meios de supervisão existentes: <ul style="list-style-type: none"> • Horário da ocorrência; • Equipamentos e/ou linhas que desligaram.
...	...
1.10.	Caso haja impedimento para reenergização da Instalação, prosseguir a partir do item 2.2.
<NIVEL1> 2. REENERGIZAÇÃO </NIVEL1>	
2.1.	Sem Impedimento
2.1.1.	As chaves de bloqueios devem ser rearmadas
2.1.2.	Executar seqüência descrita nos Anexos, efetuando contato com o CROL no item que exigir.
2.2.	Com Impedimento
2.2.1.	O operador deve informar ao CROL os seguintes itens: <ul style="list-style-type: none"> • Equipamentos e linhas de transmissão impedidas, assim como seus motivos; • Sinalizações indicativas, chaves de bloqueio e proteções; • Ações de isolamento tomadas
2.2.2.	Comunicar ao CROL em até dez minutos, as informações complementares descritas acima.
2.2.3	Preparar e restabelecer a Instalação sob orientação do CROL
<NIVEL1> 3. PROCEDIMENTOS GERAIS </NIVEL1>	
3.1.	Caso ocorra impedimento de equipamento durante a reenergização cujo item não possa ser pulado ou substituído, contatar com o CROL
...	...
3.9.	Caso haja uma tentativa de reenergização sem sucesso, devido a ocorrências externas à Instalação, e não havendo impedimento, prepará-la novamente e reenergizá-la conforme Anexo.

Figura 3.6: Documento-consulta com primeiras marcações de estrutura induzidas.

Passo 3: O passo 2 é repetido para os títulos nos níveis 2, 3, e assim por diante, estes são os títulos de subseções. Entretanto, para cada título de subseção, somente será procurado um título correspondente a ele no documento-consulta, se seu nodo pai (na árvore) já possuir um correspondente; e a procura considerará apenas o texto na seção do correspondente a seu pai, e não o documento inteiro.

- No exemplo, correspondentes (no documento-consulta) aos títulos do documento-treinamento numerados por 2.1 e 2.2, devem ser procurados apenas na seção, do documento-consulta, cujo título corresponde ao título numerado por 2 no documento-treinamento.

- Após este último passo, a correspondência entre os títulos é mostrada pela Tabela 3.2; e a Figura 3.7 mostra o documento-consulta com todas das marcações de estrutura induzidas, além da tag <DOCUMENTO> que envolve todo o texto e define a seção raiz.

Tabela 3.2: Correspondência final entre título: documento-treinamento e documento-consulta.

Correspondência entre títulos	
Título no documento-treinamento	Título no documento-consulta
1. PROCEDIMENTOS INICIAIS	1. PROCEDIMENTOS INICIAIS
2. REENERGIZAÇÃO DA INSTALAÇÃO	2. REENERGIZAÇÃO
2.2 Reenergização sem Impedimento	2.2 Sem Impedimento
2.3 Reenergização com Impedimento	2.3 Com Impedimento
3. PROCEDIMENTOS GERAIS	3. PROCEDIMENTOS GERAIS

<DOCUMENTO>	
<NIVEL1>	1. PROCEDIMENTOS INICIAIS </NIVEL1>
1.1	Comunicar ao Operador de Sistema, em até três minutos, as informações preliminares (flash), baseado nas informações disponibilizadas pelos meios de supervisão existentes: <ul style="list-style-type: none"> • Horário da ocorrência; • Equipamentos e/ou linhas que desligaram;
...	...
1.10.	Caso haja impedimento para reenergização da Instalação, prosseguir a partir do item 2.2.
<NIVEL1>	2. REENERGIZAÇÃO </NIVEL1>
<NIVEL2>	2.1 Sem Impedimento </NIVEL2>
2.1.1.	As chaves de bloqueios devem ser rearmadas
2.1.2.	Executar seqüência descrita nos Anexos, efetuando contato com o CROL no item que exigir.
<NIVEL2>	2.2 Com Impedimento </NIVEL2>
2.2.1.	O operador deve informar ao CROL os seguintes itens: <ul style="list-style-type: none"> • Equipamentos e linhas de transmissão impedidas, assim como seus motivos; • Sinalizações indicativas, chaves de bloqueio e proteções; • Ações de isolamento tomadas
2.2.2.	Comunicar ao CROL em até dez minutos, as informações complementares descritas acima.
2.2.3	Preparar e restabelecer a Instalação sob orientação do CROL.
<NIVEL1>	3. PROCEDIMENTOS GERAIS </NIVEL1>
3.1.	Caso ocorra impedimento de equipamento durante a reenergização cujo item não possa ser pulado ou substituído, contatar com o CROL.
...	...
3.9.	Caso haja uma tentativa de reenergização sem sucesso, devido a ocorrências externas à Instalação, e não havendo impedimento, prepará-la novamente e reenergizá-la conforme Anexo.
</DOCUMENTO>	

Figura 3.7: Documento-consulta com todas as marcações de estrutura induzidas.

Para resumir os passos anteriores, o Algoritmo 3.1 apresenta um pseudocódigo para o TSI.

Algoritmo 3.1: Pseudocódigo para o algoritmo TSI.

TSI	
INPUT	docTreinamento, docConsulta
OUTPUT	docConsultaEstruturaMarcada <i>(títulos marcados)</i>
1	- FOR EACH título IN docTreinamento DO <i>(cada marcação de estrutura é um título)</i>
2	- procurar fragmento em docConsulta que seja pelo menos X% similar ao texto de título
3	- marcar fragmento com a mesma <i>tag</i> que marca título <i>(fragmento passa a ser um título)</i>
4	- estabelecer correspondência entre os títulos dos dois documentos
5	- RETURN docConsultaEstruturaMarcada

Considerações

- A calibragem do valor de X (passo 2) depende do domínio dos documentos processados. Estes valores podem ser obtidos por análises experimentais do método no domínio.
- Na Tabela 3.2, as correspondências entre os títulos obedecem à mesma ordem que eles aparecem nos seus respectivos documentos: primeiro título do documento-treinamento corresponde ao primeiro do documento-consulta, segundo do documento-treinamento corresponde ao segundo do documento-consulta, e assim por diante. No entanto, esta ordem não é obrigatória, o *n*-ésimo título de um documento pode corresponder a um título que esteja em qualquer posição no outro documento.
- O problema de decidir onde começa e onde termina um fragmento candidato a título, no documento-consulta, é resolvido com a análise de seu *contexto*, como o faz Kruschwitz em [Kruschwitz 2005]: exploração de numeração, de caracteres (itálico, negrito, etc.), espaçamento, e outras características não semânticas.
- Caso nenhum título tenha sido induzido no documento-consulta, apenas as *tags* raízes <DOCUMENTO> e </DOCUMENTO> serão incluídas no documento, englobando todo o seu texto, assim será uma árvore de um único nodo, o nodo raiz. Os nodos raízes das duas árvores são correspondentes.

A Figura 3.8 mostra dois novos exemplos de documentos com estruturas induzidas, tendo como base o documento-treinamento da Figura 3.3. Na Figura 3.8a,

apenas uma marcação de título foi induzida; e na Figura 3.8b, nenhuma marcação de título foi induzida.

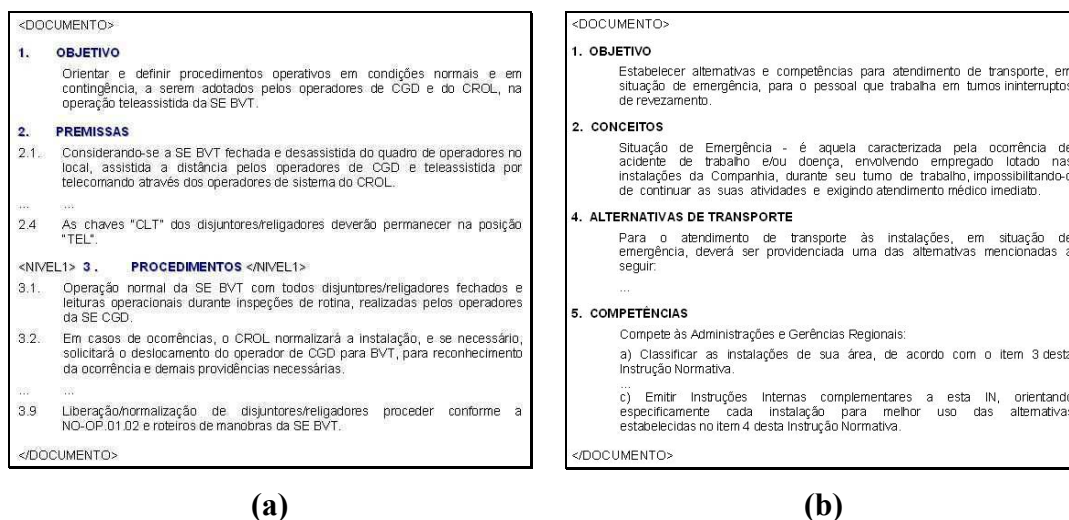


Figura 3.8: (a) Documento com apenas uma marcação de título induzida; (b) documento sem marcações de títulos induzida.

O algoritmo TSI é executado para todos os documentos-treinamento, ou seja, para cada documento-treinamento uma estrutura é induzida para o documento-consulta. Assim, para cada árvore de documento-treinamento existe uma que foi induzida para o documento-consulta, estas são chamadas árvores correspondentes. Ao final desta fase, é selecionado o par de árvores correspondentes que possuem maior similaridade estrutural, entre elas. A fase de medir a similaridade estrutural entre as árvores é mostrada a seguir.

B) Análise de Similaridade Estrutural

A análise de similaridade estrutural tem o objetivo de medir a similaridade entre duas árvores que representam as estruturadas de dois documentos. Quanto mais similares são as árvores, mais similares estruturalmente são seus documentos. A análise é executada pelo algoritmo SSA (*Structural Similarity Analyzer*) da seguinte forma:

Sejam duas árvores, A e B, com nodos $a \in A$ e $b \in B$, e o conjunto das triplas $\langle a, b, \text{escore_de_similaridade} \rangle$, em que a e b são nodos (não raízes) de títulos correspondentes, e *escore_de_similaridade* é a similaridade textual entre os dois títulos, já medido na fase anterior. O *escore de similaridade* entre as duas árvores, *SiS (Similarity Score)*, é a média dos *escores das triplas*, isto é:

$$SiS = \frac{\sum S_i}{n}, \text{ onde } S_i \text{ é o escore da tripla } i, \text{ e } n \text{ é a quantidade de triplas.}$$

Entretanto, SiS não leva em conta a equivalência entre as duas árvores. Duas árvores são *equivalentes* se suas estruturas são isomorfas (mesmo número de níveis, mesmo número de nodos, e mesma disposição dos nodos).

Como ilustrado na Figura 3.9, no cálculo de SiS nodos da árvore marcada (treinamento) podem não ter correspondentes na árvore induzida (consulta). Conseqüentemente, o valor de SiS precisa ser ajustado, com um fator de ajuste de similaridade.

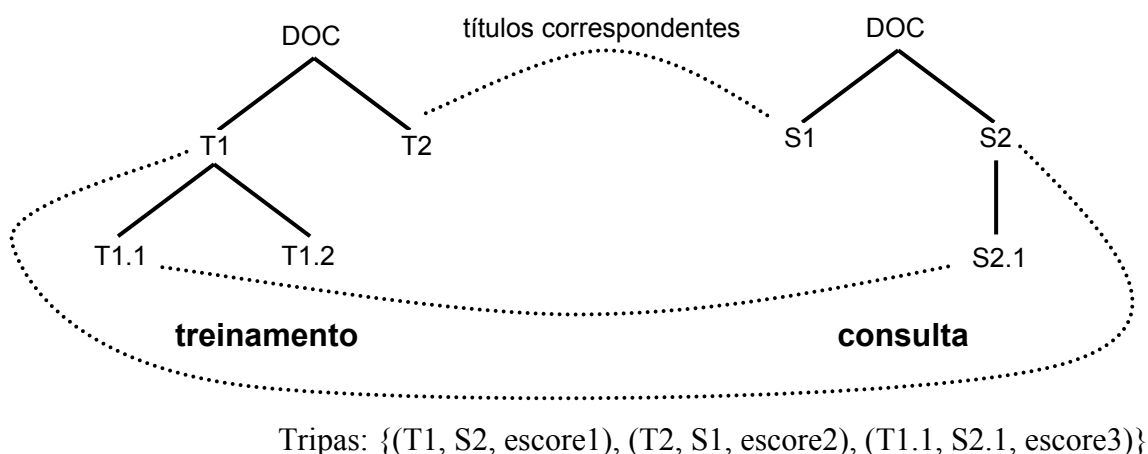


Figura 3.9: Árvores não equivalentes.

Considerando duas árvores, o Fator de Ajuste de Similaridade, SF (*Similarity Adjustment Factor*), é medido por:

$$SF = \frac{m}{n}, \text{ onde } m \text{ é a quantidade total de nodos correspondentes nas duas}$$

árvores e n é a quantidade total de nodos das duas árvores. Por exemplo, para as árvores da Figura 3.9 tem-se $SF = 6/7$.

A similaridade estrutural de duas árvores, StS (*Structural Similarity*), é o produto do escore de similaridade pelo fator de ajuste de similaridade, ou seja:

$$StS = SiS \times SF .$$

Este produto retorna um valor real entre 0 e 1: se o valor for 1, significa que as árvores são equivalentes, e todos os pares de títulos similares têm similaridade igual a 1

(100%); por outro lado, se o valor for zero, a interpretação é que não existe similaridade entre pares de nodos. De qualquer forma, quanto mais próximo de 1 for o valor, maior será a similaridade.

Um pseudocódigo para o algoritmo SSA, que resume sua execução, é mostrado pelo Algoritmo 3.2.

Algoritmo 3.2: Pseudocódigo para o algoritmo SSA

SSA	
INPUT	árvoreA, árvoreB
OUTPUT	escoreSimEstrutural
1	- FOR EACH tripla <a, b, escoreSimTextual> DO <i>(a e b são nodos</i>
2	- somaEscore = somaEscore + escoreSimTextual <i>correspondentes, a ∈ árvoreA e</i>
3	- SiS = somaEscore / quant. de triplas <i>b ∈ árvoreB; escoreSimTextual</i>
4	- SF = (quant. de triplas * 2) / (quant. de nodos em A + <i>é a sim. entre os títulos dos</i> quant. nodos em B) <i>nodos)</i>
5	- RETURN SiS * SF

O par formado pelo documento-consulta e pelo documento-treinamento, estruturalmente mais similar vão para a última fase do processo de Marcação de Estrutura, a Equivalência entre Árvores.

C) Equivalência entre Árvores

A idéia do processo de equivalência é que cada nodo de uma das árvores deve ter um único nodo correspondente na outra árvore, e vice-versa. Assim, todos os nodos da árvore do documento-treinamento que não tiverem correspondentes são descartados, incluindo todos os seus descendentes. Voltando às árvores da Figura 3.9, depois do processo de equivalência, a árvore do documento-treinamento é transformada naquela mostrada na Figura 3.10. Na transformação, o nodo T1.2 foi descartados.

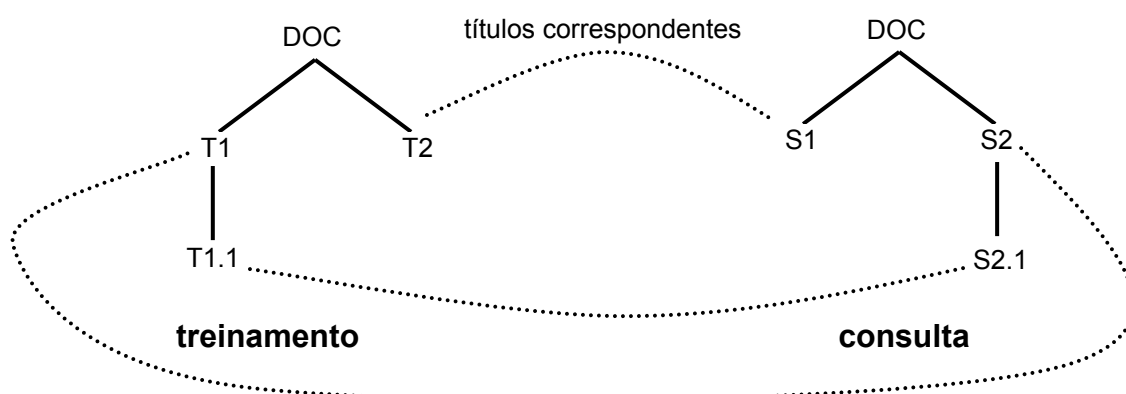


Figura 3.10: Árvores equivalentes.

Como será visto, árvores equivalentes assim construídas facilitam a tarefa do processo de Marcação de Passagem.

O Algoritmo 3.3 mostra um pseudocódigo para a fase de equivalência entre árvores.

Algoritmo 3.3: Pseudocódigo para a equivalência entre árvores.

EQUIV (EQUIVALÊNCIA)	
INPUT	árvoreDocConsulta, árvoreDocTreinamento
OUTPUT	árvoreDocTreinamento equivalente à árvoreDocConsulta
1	- FOR EACH nodo IN árvoreDocTreinamento DO
2	- IF nodo não possui correspondente em árvoreDocConsulta THEN
3	- remove subárvore de árvoreDocTreinamento com raiz em nodo
4	- RETURN árvoreDocTreinamento

Finalização do Processo

Uma vez definidos e apresentados todas as fases e algoritmos empregado no processo de Marcação de Estrutura, pode-se resumir as atividades gerais deste em um pseudocódigo mostrado pelo Algoritmo 3.4. Em sua execução, este algoritmo executa os algoritmos anteriores (TIS, SSA e EQUIV).

Algoritmo 3.4: Pseudocódigo para as atividades do Marcador de Estrutura

Processo de Marcação de Estrutura	
INPUT	conjTreinamento, docConsulta
OUTPUT	docConsEstruturaMarcada, docTreinEstrutMaisSimilar
1	- FOR EACH docTreinamento _i IN conjTreinamento DO
2	- docConsEstruturaMarcada _i = TIS (docTreinamento _i , docConsulta)
3	- simEstrutural _i = SSA (arvore (docTreinamento _i), arvore (docConsEstruturaMarcada _i))
4	- SELECT o par (docTreinamento _i , docConsEstruturaMarcada _i) com maior valor de simEstrutural _i DO
5	- docTreinEstrutMaisSimilar = docTreinamento _i
6	- árvore (docTreinEstrutMaisSim) = EQUIV (arvore (docTreinEstrutMaisSimilar), arvore (docConsulta))
7	- RETURN docConsEstruturaMarcada _i , docTreinEstrutMaisSimilar
Função: árvore (documento): retorna a árvore de estrutura do documento	

Ao final de todas as fases deste processo, tem-se um documento-consulta, um documento-treinamento, ambos com estruturas marcadas e árvores equivalentes. Em seguida, estes são passados ao processo Marcação de Passagem.

3.2.2 Marcação de Passagem

Este processo recebe um documento-consulta e um documento-treinamento, ambos com estruturas marcadas. Sua função é induzir e marcar passagens relevantes no documento-consulta, com base nas passagens relevantes marcadas no documento-treinamento.

As passagens relevantes do documento-treinamento devem estar marcadas (destacadas). Uma forma de marcá-las é utilizar *tags* XML, como na Figura 3.11: passagem marcada pela *tag* <imped-incendio>.

```
<DOCUMENTO>
...
<NIVEL1> 3. PROCEDIMENTOS GERAIS </NIVEL1>
  3.1. Em caso de desligamento, o CROL normalizará a Instalação e, se necessário, solicitará o deslocamento do Operador da SE PRS até a SE STD para reconhecimento da ocorrência e demais providências necessárias.
  ...
  <NIVEL2> 3.4. Confirmada a existência de impedimento, por:</NIVEL2>
    a) Proteções atuadas, de acordo com a IN-OP.01.006, acionar o Operador da SE PRS para:
      • inspecionar os equipamentos e linhas envolvidas;
      • isolar eletricamente os equipamentos impedidos;
      • rearmar os dispositivos de bloqueio.
    <imped-incendio>
    b) Incêndios, considerar situação de emergência e acionar o Operador da SE SMD para:
      • isolar eletricamente as áreas afetadas;
      • tomar as providências de combate a incêndio (Brigada contra incêndio);
      • informar a ocorrência e procedimentos executados ao CROL, para conclusão das manobras;
      • informar a ocorrência ao SLOG e acionar, caso necessário, o Corpo de Bombeiros.
    </imped-incendio>
    ...
    3.8. Em caso de anormalidade, o Operador da SE PRS deverá inspecioná-la e informar ao CROL, conforme IN-OP.01.006.
  ...
</DOCUMENTO>
```

Figura 3.11: Documento com passagem relevante marcada.

A indução das passagens é realizada pelo algoritmo **RPI** (*Relevant Passage Inducer*). Este se apóia nas árvores equivalentes dos documentos para otimizar a indução de passagens similares textualmente àquelas marcadas no documento-treinamento. Uma vez induzida, uma passagem é marcada.

Qualquer função de similaridade textual pode ser utilizada pelo RPI. Algumas funções de similaridade foram comentadas na Seção 2.1. No protótipo construído para validar o método (Seção 3.3), foi utilizado a função de similaridade de co-seno.

A execução do RPI ocorre da seguinte forma:

Para cada passagem marcada no documento-treinamento uma única passagem é induzida no documento-consulta, e as etapas do RPI são: A) Identificação da Seção indução, B) Segmentação da Seção Indução, C) Identificação do Segmento Mais

Similar, D) Otimização do Seguimento Mais Similar e E) Marcação da Passagem Induzida. Cada etapa será explicada, em detalhes, a seguir.

A) Identificação da Seção Indução

Nesta etapa, o RPI identifica a seção no documento-consulta, onde é mais provável de se encontrar a passagem que é mais similar à passagem marcada no documento-treinamento. A esta seção e dado o nome de **seção indução**.

O primeiro passo é identificar em qual seção no documento-treinamento está a passagem relevante marcada. Antes de identificar qual é a seção indução, é necessário identificar em qual seção, no documento-treinamento, está a passagem relevante marcada.

Uma passagem relevante pode ser o texto de uma parte de uma seção, um seção inteira, várias seções, ou até mesmo o documento inteiro. Identificar em qual seção uma passagem relevante está, depende de sua marcação. As possíveis variações de marcação de passagens são:

1. Quando a marcação envolve todo o texto de uma seção e nada além dele, a passagem relevante está na própria seção.

Ex.: Na Figura 3.12, a marcação <geral_1> marca uma seção inteira (1. PROCEDIMENTOS INICIAIS) e nada além dela.

```

<DOCUMENTO>
  <geral_1>
  <NIVEL1> 1.      PROCEDIMENTOS INICIAIS </NIVEL1>
  1.1      Identificar se o desligamento foi "geral", caracterizado pela falta total de tensão na
            instalação, (barras e linhas/links) exceto serviços auxiliares, e a inexistência de
            carregamento nas linhas/trafos/links. Caso seja "geral", proceder a partir do item 1.2.
            Caso contrário, prosseguir conforme os procedimentos descritos na
            IN-OP.01.006.
  ...
  1.13     Caso haja impedimento para reenergização da Instalação, prosseguir a partir do item 2.2.
  </geral_1>
  <NIVEL1> 2.      REENERGIZAÇÃO DA INSTALAÇÃO </NIVEL1>
  <geral_2>
  <NIVEL2> 2.1     Reenergização sem Impedimento </NIVEL2>
            Reamar as chaves de bloqueio atuadas, independente de autorização do Operador de
            Sistema.
            Proceder a preparação e reenergização na sequência descrita nos Anexos, efetuando
            contato com o CROL no item que exigir sua autorização.
  <NIVEL2> 2.2     Reenergização com Impedimento </NIVEL2>
            Informar ao CROL:
            • equipamentos impedidos ou outros impedimentos, e motivo dos mesmos;
            • principais sinalizações indicadas, chaves de bloqueio e proteções atuadas;
  </geral_2>
            • resultados da inspeção no pátio e as ações de isolamento adotadas;
            • providências tomadas e configuração atual da Instalação.
            Preparar e restabelecer a Instalação sob orientação do CROL.
  <NIVEL1> 3.      PROCEDIMENTOS GERAIS </NIVEL1>
  <geral_3>
  3.1      Efetuar inspeção geral, anotar relés e bandeirolas atuadas, registrando os resultados nos
            formulários "Lay-out Chassi Proteção" (Anexo III da IN-OP.01.006).
  </geral_3>
  ...
  3.7      Caso haja uma tentativa de reenergização sem sucesso, devido a ocorrências na própria
            Instalação, toma-se necessário contatar o CROL.
  </DOCUMENTO>

```

Figura 3.12: Documentos com várias passagens marcadas.

- Quando a marcação envolve parte do texto de uma única seção, a passagem relevante está na própria seção.

Ex.: Na Figura 3.12, a marcação **<geral_3>** marca uma parte do texto de uma seção (3. PROCEDIMENTOS GERAIS).

- Quando a marcação envolve mais de uma seção, seja totalmente ou em parte, a passagem relevante está na seção cujo nodo na árvore de estrutura é o ancestral mais próximo de todos os nodos das seções interceptadas pela marcação.

Ex.: Na Figura 3.12, a marcação **<geral_2>** marca toda uma seção (2.1 Reenergização sem Impedimento) e parte de outra (2.2 Reenergização com Impedimento). Neste caso, a passagem relevante esta na seção numerada por 2 (2. Reenergização da Instalação).

4. Quando a marcação envolve um texto que está fora de qualquer seção no primeiro nível do documento, a passagem está na seção raiz do documento, *tag* <DOCUMENTO>.

Ex.: Na Figura 3.13, a marcação <ficha> marca uma passagem que está fora de qualquer seção do primeiro nível do documento.

```
<DOCUMENTO>

<ficha>

    CHESF          INSTRUÇÃO NORMATIVA          IN-OP.01.006
    SISTEMA   :   OPERAÇÃO
    SUBSISTEMA:   OPERAÇÃO DO SISTEMA E INSTALAÇÕES
    ASSUNTO   :   REENERGIZAÇÃO DE EQUIPAMENTOS E LINHAS DE TRANSMISSÃO

</ficha>

<NIVEL1> 1. OBJETIVO </NIVEL1>
    Estabelecer conceitos e procedimentos a serem adotados pelos Operadores de
    Sistema e de Instalação, visando a reenergização ou caracterização de impedimento
    de equipamentos e linhas de transmissão.

<NIVEL1> 2. ABRANGÊNCIA </NIVEL1>
    Esta Instrução Normativa se aplica para todos os casos de desligamentos em
    Instalações.
    ...

<NIVEL1> 5. PROCEDIMENTOS GERAIS APÓS DESLIGAMENTO NA INSTALAÇÃO </NIVEL1>
    A seguir, são apresentadas situações de caráter geral e os respectivos
    procedimentos a serem adotados pelos Operadores de Instalação e Sistema, após
    ocorrência de desligamento na instalação:

<NIVEL2> 5.1. Ocorrência de Desligamento na Instalação </NIVEL2>

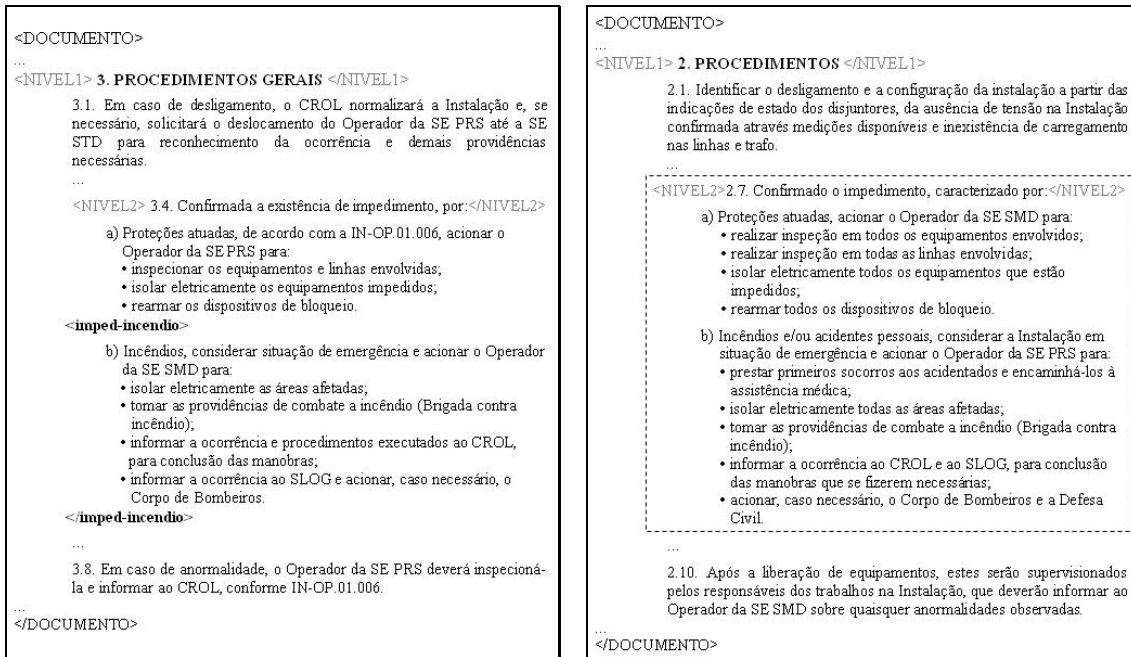
<NIVEL3> 5.1.1. O Operador de Instalação deve: </NIVEL3>
    a) silenciar os alarmes sonoros e registrar o horário da ocorrência;
    b) identificar se o desligamento foi "geral", caracterizado pela falta total de
    tensão na Instalação, (barras, linhas e links) exceto serviços auxiliares, e
    a inexistência de carregamento nas linhas, trafos e links.
    Caso seja "geral" proceder conforme Instruções de Operação de Reenergização
    específicas;
    ...

</DOCUMENTO>
```

Figura 3.13: Documentos com marcação na seção raiz.

Uma vez identificada a seção no documento-treinamento onde está a passagem relevante, a seção correspondente a ela será a **seção indução**. A idéia é que: a passagem, no documento-consulta, mais similar à marcada (no documento-treinamento) provavelmente está na seção correspondente à seção da passagem marcada. Ou seja, se a passagem marcada está na seção X do documento-treinamento, a seção Y do documento-consulta e que corresponde a X, possivelmente tem a passagem que é a mais similar à marcada. Assim, a indução da passagem mais similar pode ser feita considerando apenas o texto da seção indução, desprezando todo o resto do documento.

A Figura 3.14 mostra um exemplo de um documento-treinamento com passagem marcada (a) e um exemplo de um documento-consulta (b) com uma seção indução indicada por um retângulo tracejado.



(a)

(b)

Figura 3.14: (a) Documento-treinamento com passagem marcada; (b) documento-consulta com seção indução destacada.

Consideração

- Caso a passagem marcada esteja em uma seção que não possua correspondente no documento-consulta, esta será ignorada, ou seja, não induzirá marcação no documento-consulta. O nodo desta seção já foi descartado pela fase de equivalência entre árvores, no processo de Marcação de Estrutura.

Uma vez identifica a seção indução, a próxima etapa do algoritmo RPI é a segmentação desta seção.

B) Segmentação da Seção Indução

Nesta etapa, o RPI divide a seção indução em segmentos de textos, e cada segmento abrange parte seqüencial do texto da seção. A Figura 3.15 mostra a seção indução do documento da Figura 3.14b dividida em três segmentos (1, 2 e 3), as linhas tracejadas delimitam os segmentos formados.

Segmentos são formados por conjuntos de unidades de texto; pode-se citar como unidades de texto: sentenças e parágrafos. Técnicas de segmentação topicamente

coerente de texto (discutidas na Seção 2.3) dividem um texto nos pontos em que o tema muda.

RPI utiliza técnicas para dividir o texto da seção indução em segmentos topicamente coerentes, no entanto, ele não foi projetado para realizar esta tarefa. São utilizadas ferramentas existentes, integradas ao RPI, para a tarefa de segmentação. Esta característica é um ponto de extensão do método de extração proposto, que pode utilizar diferentes soluções para a segmentação da seção indução. No protótipo construído para validar o método (Seção 3.3), foram utilizados dois segmentadores: *TextTiling* [Hearst 1997] e *Minimum Cut* [Malioutov 2006]. No exemplo da Figura 3.15, a segmentação foi realizada pelo algoritmo *TextTiling*.

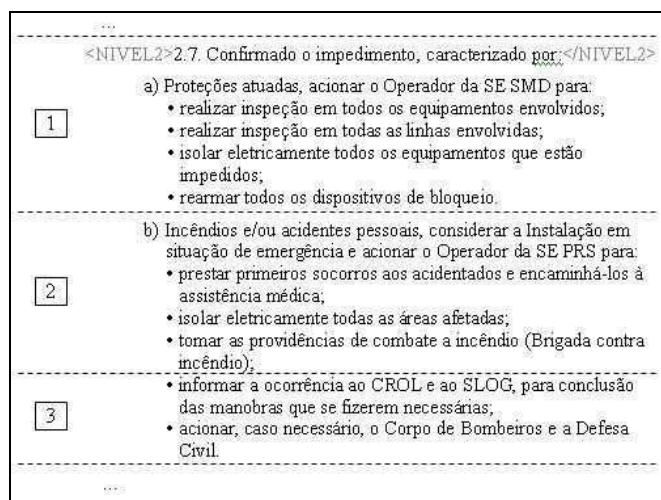


Figura 3.15: Seção indução segmentada.

C) Identificação do Segmento Mais Similar

Nesta etapa, é identificado o segmento de texto da seção indução que é mais similar textualmente à passagem marcada no documento-treinamento. O RPI mede a similaridade textual entre a passagem marcada e o texto de cada segmento formado na etapa anterior. O segmento que tiver maior similaridade com a passagem será selecionado e chamado de **segmento mais similar**.

No exemplo, o segmento numerado por 2, na Figura 3.15, é o mais similar à passagem marcada na Figura 3.14a.

Após a identificação do segmento mais similar, a próxima etapa é encontrar um novo segmento, derivado do anterior, que seja ainda mais similar à passagem marcada.

D) Otimização do Segmentando mais Similar

Nesta etapa, a passagem procurada no documento-consulta (que é a mais similar a marcada) está totalmente ou em parte dentro do segmento mais similar. Quatro casos podem ocorrer: (1) o segmento coincide com a passagem procurada, (2) o segmento contém apenas parte da passagem procurada, (3) o segmento contém toda a passagem e parte que não é dela, e (4) o segmento contém apenas parte da passagem e parte que não é dela. Para os três últimos casos, o tamanho deste segmento será alterado, gerando um novo segmento que contenha apenas a passagem procurada.

O RPI não sabe qual dos quatro casos ocorre, por isso ele escolhe três segmentos: o segmento mais similar e outros dois derivados do primeiro; a intenção é que um dos três tenha toda e apenas a passagem mais similar a marcada.

O primeiro segmento escolhido é o mais similar identificado pela etapa anterior, agora ele será chamado de **segmento inicial**, o valor da similaridade dele com a passagem marcada será referenciado por **Si** (similaridade inicial). Este segmento contempla o caso em que a passagem procurada já está completamente nele, e ele contém apenas ela.

No exemplo da Figura 3.15, o segmento 2 é o segmento inicial.

Os segmentos derivados são gerados pela expansão e redução-expansão do segmento inicial. A expansão tenta resolver o caso em que o segmento inicial contém apenas parte da passagem procurada (1). Já a redução-expansão tenta resolver os casos em que o segmento inicial contém toda a passagem e parte que não é dela (3) ou apenas parte da passagem e parte que não é dela (4). Procedimentos de expansão e redução-expansão são descritos a seguir.

Expansão do Segmento Inicial

Neste procedimento, um novo segmento é derivado da expansão para cima e / ou para baixo do segmento inicial. Unidades de texto adjacentes ao segmento são a ele incorporadas, e cada ação de expansão incorpora uma unidade por vez.

Uma unidade de texto pode ser uma sentença ou um parágrafo, apenas um dos dois tipos pode ser escolhido. A definição de qual tipo tratar vai depender do tipo que é tratado pela solução de segmentação utilizada, se os segmentos são formados por

conjunto de sentenças, então a cada expansão uma sentença será incorporada; caso os segmentos sejam formados por conjunto de parágrafos, a cada expansão um parágrafo será incorporado. Os passos realizados pelo RPI, na expansão do seguimento inicial, são os seguintes:

Passo 1: O RPI começa gerando um segmento que inicialmente é igual ao segmento inicial, este novo segmento é chamado **segmento expandido**. O valor da similaridade textual entre o segmento expandido e a passagem marcada, no documento-treinamento, é referenciado por **Se** (similaridade de expansão).

Passo 2: Depois, o RPI vai incorporando, ao segmento expandido, unidades de texto adjacentes e acima dele. Após cada incorporação, o valor de **Se** é re-calculado.

As incorporações obedecem a seguinte restrição: uma unidade só será incorporada se com ela o valor de **Se** for no mínimo **T%** do maior valor de **Se** registrado durante todo o processo de expansão. Uma vez a restrição anterior não sendo atendida, o processo de expansão pára.

- O valor escolhido de **T%** vai depender do domínio dos documentos tratados, podendo ser definido segundo análises experimentais. Um valor de 100% indica que durante o processo de expansão **Se** nunca poderá reduzir. Um valor menor que 100% indica uma tolerância (**T%** – percentual mínimo de tolerância), permitindo que **Se** seja reduzido durante o processo de expansão, porém, a redução está condicionada a um valor anterior de **Se**, o maior registrado.

Passo 3: Após a expansão para cima, o RPI expande o segmento expandido com unidades adjacentes e abaixo dele. O procedimento é o mesmo empregado na expansão para cima, além de obedecer à mesma restrição para incorporação.

Ao final, tem-se um segmento expandido e um valor para **Se**, resultantes da expansão do segmento inicial.

No exemplo, o segmento 2' da Figura 3.16 é derivado de todo o procedimento de expansão do segmento inicial.

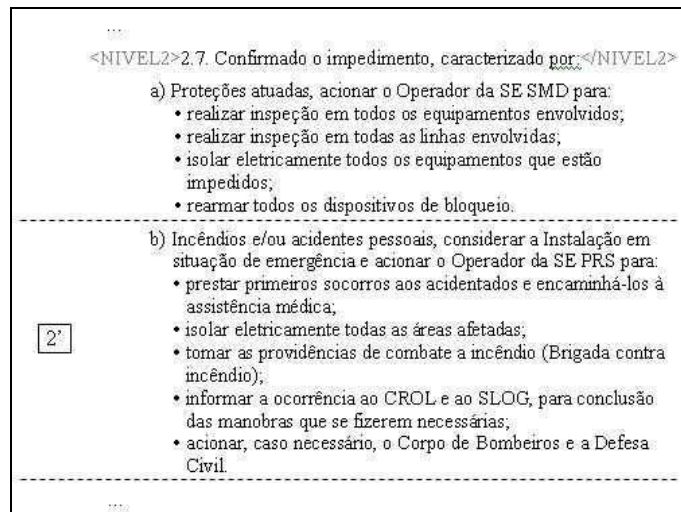


Figura 3.16: Expansão do segmento inicial.

O Algoritmo 3.5 mostra um pseudocódigo para a expansão de segmento. Em seguida é mostrado o procedimento de redução-expansão.

Algoritmo 3.5: Pseudocódigo para a expansão de segmento.

ExpansãoSeg		
INPUT	segmento, direção, passagem	<i>(direção: acima ou abaixo; passagem: passagemMarcada)</i>
OUTPUT	segExpandido	
1	- simExpansão = simTextual (segmento, passagem)	
2	- segExpandido = segmento	
3	- WHILE simTextual (expandeSeg (segmento, direção), passagem) >= T% do maior valor registrado para simExpansão DO	<i>(só continuará expandido, se com a expansão a similaridade for pelo menos T% da maior registrado durante a execução do algoritmo)</i>
4	- segmento = expandeSeg (segmento, direção)	
5	- RETURN segmento	

Funções: **simTextual**(texto1, texto2): retorna o escore de similaridade textual entre dois textos
expandeSeg(segmento, direção): incorpora a unidade de texto adjacente ao segmento à direção indicada; retorna o segmento com a unidade incorporada

Redução-expansão do Segmento Inicial

Neste procedimento, um novo segmento é derivado da redução e conseqüente expansão do segmento inicial. Os passos para a redução-expansão são semelhantes às etapas anteriores do RPI, e são mostradas a seguir:

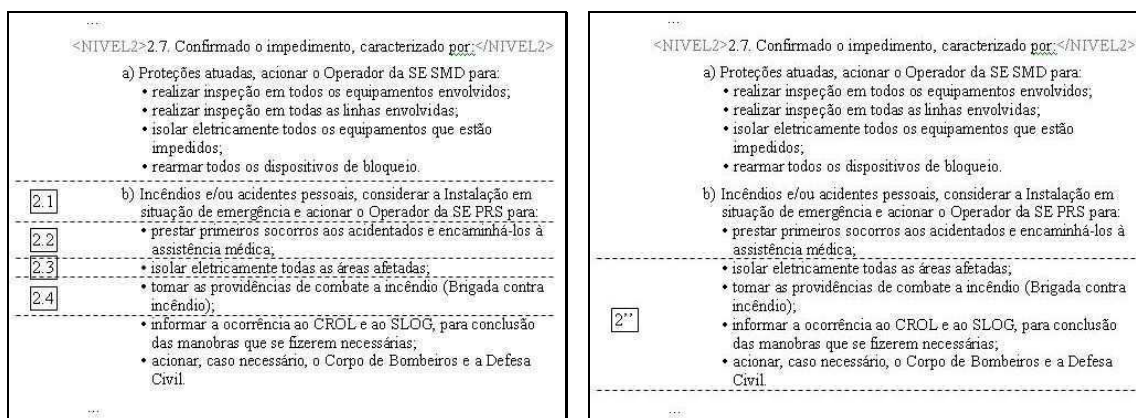
Passo 1: O segmento inicial é dividido em novos subsegmentos, cada um é formado por uma única unidade de texto. Neste caso, não é utilizada uma ferramenta para segmentação, apenas é definido que cada unidade será um segmento. A unidade de texto (sentença ou parágrafo) considerada para formar os segmentos é a mesma tratada no processo de expansão.

Passo 2: Após a segmentação, é escolhido o subsegmento mais similar textualmente à passagem marcada no documento-treinamento. O valor da similaridade entre o subsegmento e a passagem é referenciado por **Sr** (similaridade de redução-expansão).

Passo 3: Por fim, o subsegmento mais similar passa pelo mesmo processo de expansão, descrito anteriormente. Em cada incorporação de unidade, o valor de Sr é recalculado e o requisito é o mesmo, ou seja, uma unidade só será incorporada ao subsegmento se com ela o valor de Sr for no mínimo T% do maior valor registrado para Sr durante todo o processo de redução-expansão.

Ao final, tem-se um segmento reduzido-expandido e um valor de Sr, resultantes de todo o procedimento de redução-expansão.

Para o exemplo, a Figura 3.17a mostra a divisão do segmento inicial em quatro subsegmentos. O mais similar à passagem marcada é o 2.4, da expansão deste é derivado o segmento reduzido-expandido: segmento 2'' da Figura 3.17b.



(a)

(b)

Figura 3.17: (a) Divisão do segmento inicial em quatro subsegmento; (b) expansão do subsegmento mais similar.

Todo o procedimento de redução-expansão é resumido em um pseudocódigo mostrado pelo Algoritmo 3.6. Percebe-se que este executa o algoritmo de expansão de segmento (Algoritmo 3.5).

Algoritmo 3.6: Pseudocódigo para a redução-expansão de segmento

ReduçãoExpansãoSeg		
INPUT	segmento, direção, passagem	<i>(direção: acima ou abaixo;</i>
OUTPUT	segReduzidoExpandido	<i>passagem: passagemMarcada)</i>
1	- subSegmentos = subSegmentar (segmento)	
2	- FOR EACH subSeg _i IN subSegmentos DO	
3	-simTxt _i = simTextual (subSeg _i , passagem)	
4	- SELECT o par (subSeg _i , simTxt _i) com maior valor de simTxt _i DO	
5	- subSeg _i = ExpansãoSeg (subSeg _i , acima, passagem)	
6	- subSeg _i = ExpansãoSeg (subSeg _i , abaixo, passagem)	
7	- RETURN subSeg _i	
Função: subSegmentar (segmento): divide um segmento em subsegmentos, que são unidades de texto		
simTextual (texto1, texto2): retorna o escore de similaridade textual entre dois textos		

E) Marcação da Passagem Induzida

Concluídas todas as etapas anteriores, executadas pela RPI, existem três segmentos (segmento inicial, segmento expandido e segmento reduzido-expandido) e seus respectivos valores de similaridade (S_i , S_e e S_r). O segmento com maior valor de similaridade será a passagem induzida relevante no documento-consulta, pois é a passagem identificada mais similar à marcada no documento-treinamento.

Para o exemplo, os segmentos são: segmento inicial (segmento 2 – Figura 3.15) com similaridade S_i , segmento expandido (segmento 2' – Figura 3.16) e segmento reduzido-expandido (segmento 2'' – Figura 3.17b). No caso, $S_e > S_r > S_i$, assim sendo, o segmento expandido é a passagem induzida relevante e será marcado com a mesma *tag* que marca a passagem relevante do documento-treinamento (<imped-incendio>). A Figura 3.18 mostra o documento consulta com esta marcação.

```

<DOCUMENTO>
...
<NIVEL1> 2. PROCEDIMENTOS </NIVEL1>
    2.1. Identificar o desligamento e a configuração da instalação a partir das
    indicações de estado dos disjuntores, da ausência de tensão na Instalação
    confirmada através medições disponíveis e inexistência de carregamento
    nas linhas e trafo.
    ...
    <NIVEL2>2.7. Confirmado o impedimento, caracterizado por:</NIVEL2>
        a) Proteções atuadas, acionar o Operador da SE SMD para:
            • realizar inspeção em todos os equipamentos envolvidos;
            • realizar inspeção em todas as linhas envolvidas;
            • isolar eletricamente todos os equipamentos que estão
              impedidos;
            • rearmar todos os dispositivos de bloqueio.
        <imped-incendio>
        b) Incêndios e/ou acidentes pessoais, considerar a Instalação em
        situação de emergência e acionar o Operador da SE PRS para:
            • prestar primeiros socorros aos acidentados e encaminhá-los à
              assistência médica;
            • isolar eletricamente todas as áreas afetadas;
            • tomar as providências de combate a incêndio (Brigada contra
              incêndio);
            • informar a ocorrência ao CROL e ao SLOG, para conclusão
              das manobras que se fizerem necessárias;
            • acionar, caso necessário, o Corpo de Bombeiros e a Defesa
              Civil.
        </imped-incendio>
        ...
        2.10. Após a liberação de equipamentos, estes serão supervisionados
        pelos responsáveis dos trabalhos na Instalação, que deverão informar ao
        Operador da SE SMD sobre quaisquer anormalidades observadas.
    ...
</DOCUMENTO>

```

Figura 3.18: Documento-consulta com passagem induzida marcada.

Finalização do Processo

O algoritmo RPI é executado uma vez para cada uma das passagens marcadas no documento-treinamento. Após a marcação de todas as passagens no documento-consulta, as marcações de estrutura são retiradas deste documento, o processo Marcação de Passagem finaliza e seu resultado é o documento-consulta com suas passagens relevantes marcadas.

Todas as etapas do RPI e do módulo Marcador de Passagem estão resumidas em pseudocódigos mostrados pelo Algoritmo 3.7 (Marcador de Passagem) e pelo Algoritmo 3.8 (RPI). Nestes pseudocódigos, percebe-se que na execução do Marcador de Passagem o algoritmo RPI é executado. Da mesma forma, na execução do RPI, os algoritmos anteriores: ExpansãoSeg (Algoritmo 3.5) e ReduçãoExpansãoSeg (Algoritmo 3.6) são executados.

Algoritmo 3.7: Pseudocódigo para as atividades do Marcador de Passagem.

Marcador de Passagem	
INPUT	docTreinamento, docConsultaEstrutMarcada
OUTPUT	docConsultaPassagensMarcadas <i>(sem marcações de estrutura)</i>
1	- FOR EACH passagemMarcada IN docTreinamento DO
2	- docConsPassagensMarcadas = RPI (docTreinamento, passagemMarcada, docConsultaEstrutMarcada)
3	- docConsPassagensMarcadas = retiraMarcasEstrutura (docConsPassagensMarcadas)
4	- RETURN docConsPassagensMarcadas

Funções: **retiraMarcasEstrutura**(documento): retira as marcações de estrutura do documento

Algoritmo 3.8: Pseudocódigo para o algoritmo RPI.

RPI	
INPUT	docTreinamento, passagemMarcada, docConsultaEstrutMarcada
OUTPUT	docConsultaPassagensMarcadas
1	- identificar seçãoIndução em docConsultaEstrutMarcada
2	- segmentos = segmentar (seçãoIndução)
3	- FOR EACH seg _i IN segmentos DO
4	- simTxt _i = simTextual (seg _i , passagemMarcada)
5	- SELECT o par (seg _i , simTxt _i) com maior valor de simTxt _i DO
6	- segInicial = seg _i
7	- Si = simTxt _i <i>(similaridade inicial)</i>
8	- segExpandido = ExpansãoSeg (segInicial, acima, passagemMarcada)
9	- segExpandido = ExpansãoSeg (segExpandido, abaixo, passagemMarcada)
10	- Se = simTextual (segExpandido, passagemMarcada) <i>(similaridade de expansão)</i>
11	- segReduzidoExpand = ReduçãoExpansãoSeg (segInicial, passagemMarcada)
12	- Sr = simTextual (segReduzidoExpand, passagemMarcada) <i>(sim. de redução-expansão)</i>
13	- docConsultaEstrutMarcada = docConsultaPassagensMarcadas
14	- S _{MAX} = máximo (Si, Se, Sr)
15	- IF Se = S _{MAX} THEN
16	- marcar segExpandido em docConsultaPassagensMarcadas
17	- ELSE IF Si = S _{MAX} THEN
18	- marcar segInicial em docConsultaPassagensMarcadas
19	- ELSE IF Sr = S _{MAX} THEN
20	- marcar segReduzidoExpand em docConsultaPassagensMarcadas
21	- RETURN docConsultaPassagensMarcadas

Funções: **segmentar**(texto): divide um texto em segmentos (ex.: TextTiling e Minimum Cut)

máximo(conjunto de valores): retorna o valor máximo no conjunto

Concluída toda a apresentação do novo método proposto, a próxima seção apresenta um sistema de extração de informação que é um protótipo para o método.

3.3 O Protótipo TIES – Textual Information Extraction System

TIES (Textual Information Extraction System) é um sistema de extração automática de informação relevante de documentos não-estruturados. Ele é um protótipo que implementa o método de extração de informação proposto na seção anterior, e foi construído com o propósito de validar este método.

Como o método do TIES é da classe supervisionada, é necessário que a ele seja fornecido um conjunto de documentos de treinamento, com passagens relevantes marcadas. Com base nestas passagens, o TIES induz passagens relevantes em novos documentos.

O sistema utiliza uma combinação de técnicas de extração: análises de similaridade estrutural e textual, segmentação topicamente coerente de texto, além de induzir passagens de texto.

Por utilizar um método supervisionado com a combinação de técnicas de extração, o TIES consegue ser eficaz na extração e reduzir o custo na preparação do conjunto de treinamento. Estas afirmações serão devidamente verificadas no capítulo seguinte, sobre a avaliação o protótipo.

Outras importantes características do TIES são: independência de domínio, independência de formato, aplicabilidade e extensibilidade, a saber:

- Quanto à independência de domínio, o sistema é capaz processar documentos de qualquer domínio, onde haja passagens relevantes a serem extraídas e um conjunto de treinamento possa ser preparado.
- Pela independência de formato, o sistema não impõe restrição sobre o formato de arquivo dos documentos fornecidos a ele. Qualquer formato pode ser utilizado, desde que seja possível implementar uma forma de recuperar (extrair) o texto dos documentos.
- A aplicabilidade do TIES é bastante ampla. Empresas e organizações possuem um grande número de documentos digitais e não-estruturados, frutos de suas atividades operacionais. O TIES pode ser utilizado no ambiente operacional de

qualquer empresa ou organização, onde usuários desejam extrair passagens específicas (relevantes) em seus documentos.

- Algumas funcionalidades do sistema são configuráveis. As mudanças nas configurações ocorrem pela adição de novos elementos ao sistema, com novas funcionalidades, ou a substituição de elementos de mesma funcionalidade. Nestas configurações, observa-se a característica de extensibilidade do TIES.

A seção seguinte traz a arquitetura do sistema e a descrição de seus módulos

3.3.1 Arquitetura do TIES

A arquitetura do TIES é mostrada na Figura 3.19. Ela possui documentos de texto fornecidos pelos usuários, que são as entradas do sistema, e é composta de dois grandes módulos: Marcador de Estrutura e Marcador de Passagem. A seguir, os módulos e suas respectivas entradas e saídas serão explicados.

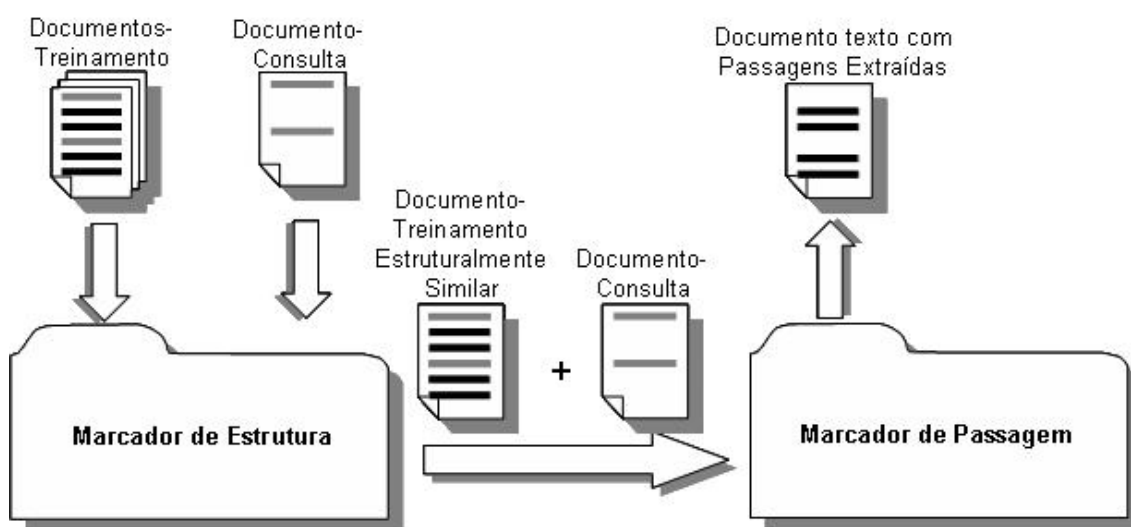


Figura 3.19: Arquitetura do TIES.

Marcador de Estrutura

O módulo Marcador de Estrutura implementa o processo de Marcação de Estrutura proposto na Seção 3.2.1.

A este módulo, o usuário do sistema fornece um conjunto de treinamento (Documentos-Treinamento na Figura 3.19). Os documentos-treinamento devem ter marcadas suas estruturas e passagens relevantes, como no exemplo da Figura 3.14a. Também são fornecidos, pelo usuário, os documentos de consulta (Documento-Consulta na Figura 3.19), onde as passagens serão induzidas.

Como para o processo de Marcação de Estrutura, as principais tarefas realizadas pelo módulo são: induzir estruturas para o documento-consulta e identificar o documento-treinamento que é mais similar estruturalmente ao documento-consulta.

No módulo, foram implementadas as três fases do processo de Marcação de Estrutura, assim como seus algoritmos:

A) Indução de Estrutura

- Algoritmo TSI (*Tree-Structure Inducer*)

B) Análise de Similaridade Estrutural

- Algoritmo SSA (*Structural Similarity Analyzer*)

C) Equalibidade de Árvores

- Algoritmo de Equalibidade de Árvores

As saídas do Marcador de Estrutura é o documento-consulta, com estrutura marcada (Documento-Consulta Com Estrutura Marcada na Figura 3.19), e o documento-treinamento mais similar ao documento-consulta (Documento-Treinamento Estruturalmente Mais Similar na Figura 3.19). Estas saídas são entradas para o próximo módulo: Marcador de Passagem.

Marcador de Passagem

O módulo Marcador de Passagem implementa o processo de Marcação de Passagem proposto na Seção 3.2.2.

Este recebe do módulo anterior um documento-treinamento, com passagens relevantes marcadas, e um documento-consulta. A tarefa do Marcador de Passagem é induzir, no documento-consulta, as passagens mais similares às marcadas no documento-treinamento. As passagens induzidas são as passagens relevantes do documento-consulta.

Foram implementadas as fases do processo de Marcação de Passagem, e seus algoritmos:

A) Identificação da Seção Indução

B) Segmentação da Seção Indução

C) Identificação do Segmento Mais Similar

D) Otimização do Segmentando mais Similar

- Algoritmo de expansão de segmento
- Algoritmo de redução-expansão de segmento

E) Marcação da Passagem Induzida

- No protótipo, as passagens induzidas não são marcadas no documento-consulta, elas são apenas extraídas do documento e gravadas em um arquivo texto, que é a saída deste módulo e saída do sistema. Os usuários consultam este arquivo para ver as passagens. Esta restrição não afeta a eficácia do sistema.

A saída deste módulo é um documento texto com todas as passagens relevantes que foram induzidas do documento-consulta.

Para a extração de passagens relevantes de documentos que pertencem a um domínio, o TIES recebe um conjunto documentos de treinamento para o domínio, esta será a base de conhecimento do sistema. A execução do sistema é feita sobre um documento de consulta, por vez, onde há passagens relevantes e o usuário deseja que sejam identificadas.

Considerações

- O protótipo foi todo implementado utilizando a linguagem de programação orientada a objetos Java¹. Na versão corrente, são processados documentos de texto no formato do Microsoft Word 2003 e em PDF (*Portable Document Format*). Os textos no formato do Microsoft Word são recuperados dos documentos por meio da ferramenta Apache POI². Na recuperação dos textos dos documentos em PDF é utilizada a ferramenta Apache PDFBox³. Este é um ponto de extensão do TIES – para que o protótipo processe outros formatos, é necessária apenas a inclusão (implementação), no sistema, de uma ferramenta que recupere os textos dos documentos, seja qual for o formato.

¹ <http://java.sun.com/javase/>. Acessado em julho 2009.

² <http://jakarta.apache.org/poi/>. Acessado em julho 2009.

³ <http://incubator.apache.org/pdfbox/>. Acessado em julho 2009.

- Como foi dito anteriormente, o TIES realiza a extração das passagens relevantes por análises de similaridade estrutural e textual, entre documento-consulta e documentos-treinamento. A arquitetura do sistema não exige que os documentos fornecidos sejam similares em estrutura ou texto. No entanto, para extrair as passagens relevantes de um documento-consulta, é necessário que exista pelo menos um documento-treinamento, que seja do mesmo domínio do documento-consulta.
- Para a segmentação da seção indução, foram utilizadas duas técnicas existentes de segmentação topicamente coerente de textos. O TIES pode ser configurado para que uma das duas seja usada por vez. A primeira é o algoritmo de segmentação mais conhecido na literatura, o *TextTiling* [Hearst 1997], a outra é o algoritmo *Minimum Cut* [Malioutov 2006], ambos discutidos na Seção 2.3. Estes algoritmos foram escolhidos para esta tarefa por terem uma implementação em código Java, disponível e livre (*TextTiling*⁴ - [Choi 1999], *Minimum Cut*⁵), e por terem atingido bons resultados em suas avaliações. Ambos os algoritmos dividem um texto em segmentos formados por conjunto de parágrafos.
- Nas medições de similaridade textual, em ambos os módulos, foi utilizada a similaridade de co-seno. Como já dito, na função co-seno os textos são transformados em vetores, e o co-seno do ângulo entre dois vetores é o escore de similaridade entre seus textos. Para cada texto um vetor é formado, e cada termo (palavra) do texto possui um peso (valor) no vetor. Existem várias formas de pesar os termos nos vetores, no protótipo foram implementados duas destas formas: *Binary* e *TF (Term-Frequency)* [Salton 1988]; uma opção de configuração define qual das duas será utilizada no protótipo.
- Antes de medir a similaridade entre dois textos, estes passam por pré-processamentos conhecidos:
 - Retirada de todos os caracteres não literais.
 - Retirada de todos dos algarismos romanos.

⁴ <http://myweb.tiscali.co.uk/freddyychoi/>. Acessado em julho de 2009.

⁵ <http://people.csail.mit.edu/igorm/>. Acessado em julho de 2009.

- Retirada das *stop words* (termos comuns) – uma lista de termos (palavras) comuns é fornecida ao sistema. Nas análises de similaridade, estes termos são desconsiderados.
- *Stemming* (redução dos termos a seus radicais) – uma lista com termos e respectivos radicais é fornecida ao sistema. Nos textos, cada termo é substituído por seu respectivo radical.

3.4 Conclusões

Neste capítulo foi proposto um novo método de extração de informação, que é capaz de extrair passagens relevantes em documentos não-estruturados de texto. O método tem duas características importantes: é da classe de métodos supervisionados e apresenta uma nova técnica de extração de informação.

A nova técnica de extração proposta utiliza uma combinação de técnicas de extração existentes. Esta combinação objetiva a eficácia de extração do método. A construção do sistema TIES mostra que o método é uma solução de extração de informação implementável por alguma linguagem de programação.

O próximo capítulo traz a avaliação do TIES, que por consequência avalia o novo método.

Capítulo 4

Avaliação Experimental do TIES

Este capítulo descreve a avaliação experimental do protótipo TIES, descrito no Capítulo 3. A avaliação tem o propósito de verificar a eficácia do sistema, no que diz respeito a tarefas de extração de passagens relevantes de documentos. São apresentados, pela ordem: plano de testes, resultados obtidos e discussão sobre os resultados.

4.1 Plano de Testes

O plano de testes consiste de: (1) hipóteses a verificar; (2) escolha dos *corpora* de documentos; definição das métricas de avaliação; e (4) calibragem do sistema TIES.

4.1.1 Hipóteses a Verificar

As três hipóteses levantadas contemplam, respectivamente: domínio, formato de arquivo, eficácia de extração e custo de treinamento do TIES.

Hipótese 1: Independência de Domínio

O TIES deve extrair passagens relevantes de documentos de qualquer domínio. Para verificar esta hipótese, o TIES precisa ser testado com ao menos dois domínios díspares de documentos.

Hipótese 2: Independência de Formato de Arquivo

O TIES deve ser capaz de processar mais de um formato de arquivo de documento. Para provar esta hipótese, o sistema necessita ser testado processando arquivos de pelo menos dois formatos distintos.

Hipótese 3: Uso em Aplicações Exigindo Alta Eficácia de Extração

O TIES deve ser útil mesmo em aplicações exigindo um alto grau de eficácia de extração de informação. Para verificar esta hipótese, o TIES precisa ser confrontado com pelo menos uma aplicação de missão crítica.

Hipótese 4: Custo de Treinamento Baixo

É importante que o esforço de treinamento do TIES possa ter um custo baixo. A verificação desta hipótese pode se dar com conjuntos de treinamento de pequeno volume de dados, em termos absolutos: quanto menor um conjunto de treinamento, menor o esforço de especialistas para treinar o TIES.

4.1.2 Corpora de Testes

Tendo em vista a hipótese 1, dois *corpora* foram utilizados nos testes do TIES, o primeiro sobre sistemas elétricos de potência, e o segundo, da área legislativa..

Corpus Chesf

Foram utilizados documentos da Companhia Hidroelétrica do São Francisco (Chesf), do Sistema Eletrobrás, a qual tem a missão de gerar e distribuir energia elétrica para a região Nordeste do Brasil. Os documentos do *corpus* consistem em instruções normativas e operacionais para os operadores dos sistemas elétricos da Chesf.

O *corpus* é composto de 70 documentos, Cada um com tamanho variando entre 5 e 14 páginas⁶. Os documentos são de uma das classes: Instruções Operacionais (IO) – 50 documentos – e Instruções Normativas (IN) – 20 documentos.

Documentos de uma mesma classe foram divididos em subclasses. Subclasses são constituídas de documentos que tratam de um mesmo assunto; exemplo: documentos de instruções operacionais sobre operações em *linhas de transmissão elétrica* e sobre operações em *disjuntores elétricos* formam duas subclasses, distintas, de instruções operacionais.

A classe de Instruções Operacionais foi dividida em três subclasses, que possuíam 16, 23 e 11 documentos.

⁶ Tipicamente com páginas de tamanho A4 e fonte Times 12.

Do total de documentos do *corpus*, dez foram escolhidos para formar conjuntos de treinamento, seis IOs e quatro INs. Os demais documentos (60 no total) foram utilizados como documentos-consulta, para que suas passagens relevantes fossem induzidas pelo TIES. Para cada subclasse, um conjunto de treinamento foi formado, este possui entre 1 e 3 documentos. A escolha dos documentos-treinamento, em cada subclasse, foi aleatória.

A quantidade total de documentos, a quantidade de documento no conjunto de treinamento e a quantidade de documentos-consulta (documentos que tiveram suas passagens induzidas utilizando o conjunto de treinamento), em cada subclasse, são mostradas pela Tabela 4.1.

Tabela 4.1: Quantidades de documentos-treinamento e documentos-consulta para cada subclasse do *corpus* Chesf.

Subclasse	Total Documentos	Quantidade Doc-Trein.	Quantidade Doc-Cons.
Instruções Operacionais			
Subclasse 1	16	1	15
Subclasse 2	23	3	20
Subclasse 3	11	2	9
Instruções Normativas			
Subclasse 4	8	1	7
Subclasse 5	11	3	9

Em cada documento-treinamento foram marcadas manualmente as passagens relevantes, em seções distintas. As passagens são instruções que os operadores do sistema elétrico da Chesf devem seguir para a manutenção deste sistema. A Figura 4.1 mostra uma IO onde uma passagem relevante aos operadores está destacada por um retângulo, esta passagem descreve os procedimentos que devem ser tomados no caso de impedimento de uma subestação, causado por incêndio.


 Companhia Hidro Elétrica de São Francisco	MANUAL DA OPERAÇÃO - MO DOCUMENTO	VIGÊNCIA:
	INSTRUÇÃO DE OPERAÇÃO	19/03/2008
ORIGEM DIVISÃO DE METODIZAÇÃO E SUPORTE DA OPERAÇÃO		
OPERAÇÃO TELEASSISTIDA DA SE SANTA CRUZ II		
1. OBJETIVO Orientar e definir procedimentos operativos em condições normais e em contingência, a serem adotados pelos Operadores da SE Paraíso (PRS) e do CROL, na operação teleassistida da SE Santa Cruz II (STD). ...		
3. PROCEDIMENTOS GERAIS		
3.1. Em caso de desligamento, o CROL normalizará a Instalação e, se necessário, solicitará o deslocamento do Operador da SE PRS até a SE STD para reconhecimento da ocorrência e demais providências necessárias. ...		
3.4. Confirmada a existência de impedimento, por:		
a) Proteções atuadas, de acordo com a IN-OP.01.006, acionar o Operador da SE PRS para:		
<ul style="list-style-type: none"> • inspecionar os equipamentos e linhas envolvidas; • isolar eletricamente os equipamentos impedidos; • rearmar os dispositivos de bloqueio. 		
b) Incêndios, considerar situação de emergência e acionar o Operador da SE SMD para:		
<ul style="list-style-type: none"> • isolar eletricamente as áreas afetadas; • tomar as providências de combate a incêndio (Brigada contra incêndio); • informar a ocorrência e procedimentos executados ao CROL, para conclusão das manobras; • informar a ocorrência ao SLOG e acionar, caso necessário, o Corpo de Bombeiros. 		
...		
3.8. Em caso de anormalidade, o Operador da SE PRS deverá inspecioná-la e informar ao CROL, conforme IN-OP.01.006. ...		

Figura 4.1: Documento da Chesf com passagem relevante a seus operadores.

Os documentos-treinamento continham entre 3 e 6 passagens relevantes, com tamanho variando entre 30 e 200 palavras, aproximadamente. Estas passagens foram selecionadas por engenheiros e operadores de sistemas elétricos que trabalhavam na Chesf.

Manualmente também foram marcados os títulos das seções e subseções dos documentos do conjunto de treinamento. A marcação é manual porque não foi implementada uma forma, no TIES, de identificar estes títulos nos documentos de treinamento.

Corpus Legislativo

Documentos legislativos, sobre leis e decretos da administração de estados brasileiros, formaram o segundo *corpus*.

Foram escolhidas duas leis estaduais: Estatuto dos Servidores Públicos Estaduais e Processo Administrativo no Âmbito da Administração Pública Estadual, caracterizando duas classes de documentos (Estatuto dos Servidores e Processo

Administrativo). De cada classe, foram utilizados 20 documentos que tratavam da lei em 20 estados brasileiros, um documento para cada estado. Assim, tem-se um total de 40 documentos utilizados nos testes. A classe de Processos Administrativos foi formada por documentos variando entre 8 e 15 páginas⁷. Já a classe de estatuto dos servidores possuía documentos com aproximadamente 35 a 80 páginas, mas para compor o *corpus* de teste algumas páginas foram excluídas, ficando as 15 primeiras em cada documento.

Os documentos deste *corpus* não formaram subclasses, pois documentos de uma mesma classe tratam sobre uma mesma lei (assunto).

Para cada classe, dois documentos foram escolhidos, aleatoriamente, para formar o conjunto de treinamento da classe. Os demais documentos foram utilizados como documento-consulta.

A quantidade total de documentos, a quantidade documentos de treinamento e a quantidade de documentos-consulta, em cada classe, são mostradas pela Tabela 4.2.

Tabela 4.2: Quantidades de documentos-treinamento e documentos-consulta para cada classe do *corpus* Legislativo.

Total Documentos	Quantidade Doc-Trein.	Quantidade Doc-Cons.
Estatuto dos Servidores		
20	2	18
Processo Administrativo		
20	2	18

Em cada documento-treinamento, cinco passagens relevantes e em seções distintas foram selecionadas e marcadas manualmente. A escolha das passagens foi aleatória, com a restrição que suas seções sempre possuíssem correspondentes nos demais documentos. As passagens tinham entre 120 e 350 palavras. Manualmente também foram marcados os títulos das seções e subseções de cada documento-treinamento.

Formato dos Arquivos

Em ambos os *corpora*, existiam documentos no formato de arquivo do Microsoft Word e no formato de arquivo PDF (*Portable Document Format*).

⁷ Tipicamente com páginas de tamanho A4 e fonte Times 12.

Revocação

É a razão entre a quantidade de unidades de texto comuns às duas passagens (relevante e induzida), pela quantidade de unidades de texto na passagem relevante. Em outras palavras, revocação é a porção da informação relevante que foi induzida, sendo formalizada como segue:

$$revocação = \frac{|unidades_passagem_induzida \cap unidades_passagem_relevante|}{|unidades_passagem_relevante|}$$

A unidade de texto considerada para os cálculos de revocação foi “parágrafo”, uma vez que os processos de expansão / redução-expansão, nos testes do protótipo, consideram parágrafo como unidade de texto.

Com esta definição de revocação e precisão, pode-se observar que a passagem induzida na Figura 4.2 obteve precisão de 100%, mas a revocação foi inferior a este valor.

Medida F

A medida F entre precisão e revocação é formalizada como segue:

$$medida\ F = \frac{2}{\frac{1}{precisão} + \frac{1}{cobertura}}$$

4.1.4 Calibragem do Sistema TIES

O sistema pode ser configurado de várias maneiras, dependendo das escolhas do segmentador e da forma de pesar os vetores. Também, são definidos os percentuais de similaridade mínimos para as análises de similaridade, executadas pelos algoritmos TSI e RPI. As configurações finalizam com as opções de pré-processamento: *stop words* e *stemming*.

Segmentador e Peso-vetor

Conforme informado na seção que descreve o protótipo do sistema, existem duas opções de configuração para a segmentação textual, são os segmentadores: *Minimum Cut* e *TextTiling*. Como o TIES utiliza similaridade de co-seno, o sistema oferece duas

opções para pesar os termos dos textos, nos vetores, para os cálculos de similaridade textual: *Binary* e *TF*.

Combinando estas opções, quatro configurações para o sistema são possíveis, como mostra a Tabela 4.3:

Tabela 4.3: Possíveis configurações do protótipo TIES

Configuração	Segmentador	Peso-vetor
Configuração 1	<i>Minimum Cut</i>	<i>Binary</i>
Configuração 2	<i>Minimum Cut</i>	<i>TF</i>
Configuração 3	<i>TextTiling</i>	<i>Binary</i>
Configuração 4	<i>TextTiling</i>	<i>TF</i>

Pré-processamento do TIES

Foi fornecida ao sistema uma lista básica de *stop words*, para serem desconsideradas nos cálculos de similaridade textual, além de uma lista de palavras e seus respectivos radicais para os processos de *stemming*. Ambas as listas são derivadas de outras disponíveis na Web (SnowBall⁹). Para a lista de palavras e radicais, foram adicionados alguns poucos termos frequentemente presentes nos domínios dos *corpora* de testes, exemplos: concessionária, disjuntor, religação, desenergizar, reenergizar, etc. (Chesf); e concurso, investidura, nomeação, etc. (Legislativo). A produção das listas, com termos característicos de um domínio, pode fazer com que o TIES seja mais eficaz na indução de passagens relevantes para o domínio.

Percentuais Mínimos de Similaridade

Para o valor de percentual de similaridade mínimo, X%, nos cálculo de similaridade textual (entre os títulos dos documentos) executados pelo algoritmo TSI (Seção 3.2.1) implementado no protótipo, X foi configurado em 50 (0,5) de similaridade de co-seno. Este baixo valor foi utilizado para ambos os *corpora*, e foi baixo porque os títulos das seções dos documentos dos dois *corpora* são pequenos (geralmente uma sentença pequena). Uma pequena diferença entre dois pequenos textos poderá resultar em um valor pequeno de similaridade de co-seno, entre eles.

⁹ <http://snowball.tartarus.org/algorithms/portuguese/stemmer.html>. Acessado em julho de 2009.

Quanto ao percentual mínimo de tolerância (T%) empregado pelo algoritmo RPI (Seção 3.2.2), para a redução dos valores de similaridade de expansão (Se) e redução-expansão (Sr), este foi 90%. Este valor foi definido por análises experimentais e preliminares em ambos os *corpora*. Nestas análises, outros valores para T foram testados, no entanto, constatou-se que com T igual a 90 resultava em extração com melhores valores de revocação sem afetar muito a precisão.

Apresentado o plano de testes, os *corpora* de testes e configurações do sistema, a próxima seção mostra os resultados das execuções destes testes.

4.2 Resultados dos Testes

Esta seção discorre sobre os resultados e análises dos testes para o TIES, nos dois *corpora* anteriormente descritos. Entre seus objetivos, está a verificação das hipóteses levantadas na Seção 4.1.1.

Antes de começar a apresentar os resultados, será vista uma forma de classificar os documentos-consulta, que ajudará nas análises dos resultados.

4.2.1 Classificação dos Documentos-Consulta

Para mostrar melhor os resultados obtidos nos testes, os documentos-consulta foram divididos em três grupos, de acordo com a similaridade com seus respectivos documentos-treinamento. Para cada documento-consulta, existe um documento-treinamento que foi utilizado para induzir as passagens suas passagens.

As similaridades definidas são as seguintes:

- **Similaridade em Conteúdo:** Para um documento-consulta, é a média aritmética dos escores de similaridade textual (co-seno com peso-vetor *Binary*) entre cada seção indução (do documento-consulta) e sua seção correspondente, no respectivo documento-treinamento. No cálculo da similaridade, são considerados título e corpo de cada seção.
- **Similaridade em Estrutura:** É o escore de similaridade estrutural entre as estruturas de um documento-consulta com seu respectivo documento-treinamento, sendo calculado pelo algoritmo de similaridade estrutural SSA.

A pertinência de um documento-consulta a um grupo é definida da seguinte:

- **Grupo 1:** Documento-consulta **muito similar** em conteúdo (SC) e estrutura (SE) ao seu documento-treinamento. Definições: $SC \geq 0,8$ e $SE \geq 0,8$
- **Grupo 2:** Documento-consulta **similar** em conteúdo e estrutura ao seu documento-treinamento. Definições: $0,8 > SC \geq 0,6$ e $SE < 0,8$.
- **Grupo 3:** Documento-consulta **pouco similar** em conteúdo, mas **similar** em estrutura ao seu documento-treinamento. Definições $SC < 0,6$ e $SE < 0,8$

É interessante notar que os três grupos coincidem em um ponto: a similaridade de estrutura; isso não é mera coincidência, pois um documento-consulta e seu documento-treinamento têm uniformidade temática. Documentos que versam sobre o mesmo tema, geralmente, são similares em conteúdo e estrutura. As diferenças no conteúdo podem ser até significativas, por questões principalmente de estilo, de disposição dos assuntos, de termos sinônimos, de polissemia, etc. Quanto às diferenças na estrutura, são conseqüências de diferenças nos títulos das seções.

4.2.2 Resultados para o *Corpus Chesf*

Os documentos-consulta do *corpus* Chesf foram divididos entre os três grupos,. As quantidades de documentos nos grupos 1, 2 e 3 foram 24, 21 e 15, respectivamente.

A Tabela 4.4 mostra os resultados para um documento-consulta do grupo 1: IONTD01.doc, com respectivo documento-treinamento IOAGD01.doc. São exibidos os valores das métricas de **precisão**, **revocação** e **medida F**, obtidos pelas **passagens induzidas**. Os resultados são agrupados pelas possíveis configurações do protótipo em segmentador e peso-vetor. Uma média **global**, por configuração, é calculada para os valores das métricas, seguida de desvio-padrão (média / desvio).

Tabela 4.4: Resultados para um documento-consulta do grupo 1.

Doc-consulta	Doc-treinamento	Passagem Induzida	Precisão	Revocação	medida F
IONTD01.doc	IOAGD01.doc				
Configuração - Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>Binary</i>		barra-subestacao	1,00	0,50	0,67
		linha-de-transmissao	1,00	1,00	1,00
		reator	1,00	1,00	1,00
		global	1,00 / 0,00	0,83 / 0,29	0,89 / 0,19
Configuração - Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>TF</i>		barra-subestacao	1,00	0,50	0,67
		linha-de-transmissao	1,00	1,00	1,00
		reator-	1,00	1,00	1,00
		global	1,00 / 0,00	0,83 / 0,29	0,89 / 0,19

Doc-consulta	Doc-treinamento	Passagem Induzida	Precisão	Revocação	medida F
IONTD01.doc	IOAGD01.doc				
Configuração - Segmentador: <i>TextTiling</i> - Peso-vetor: <i>Binary</i>		barra-subestacao	1,00	0,75	0,85
		linha-de-transmissao	1,00	1,00	1,00
		reator	1,00	1,00	1,00
		global	1,00 / 0,00	0,92 / 0,14	0,95 / 0,09
Configuração - Segmentador: <i>TextTiling</i> - Peso-vetor: <i>TF</i>		barra-subestacao	1,00	1,00	1,00
		linha-de-transmissao	1,00	1,00	1,00
		reator	1,00	1,00	1,00
		global	1,00 / 0,00	1,00 / 0,00	1,00 / 0,00

Os excelentes resultados, para o exemplo da Tabela 4.4, é conseqüência do alto grau de similaridade entre as estruturas e conteúdo destes documentos. Já nestes resultados, pode ser percebida a influência das variações de configuração do sistema. Utilizando o segmentador *Minimum Cut*, os valores foram os mesmos em ambas as formas de peso-vetor. Com o *TextTiling* os resultados foram diferentes em relação aos com o *Minimum Cut*, e variariam com a mudança de peso-vetor na primeira passagem induzida (barra-subestacao).

A Tabela 4.5 mostra os resultados para um documento-consulta do grupo 2: documento-consulta IO-OC.NE.5LE.01.doc, com respectivo documento-treinamento IO-OC.NE.5LE.02.doc.

Tabela 4.5: Resultados para um documento-consulta do grupo 2.

Doc-consulta	Doc-treinamento	Passagem Induzida	Precisão	Revocação	medida F
IO-OC.NE.5LE.01.doc	IO-OC.NE.5LE.02.doc				
Configuração - Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>Binary</i>		trafo1	0,97	0,50	0,66
		trafo2	0,97	1,00	0,98
		trafo3	1,00	0,41	0,58
		global	0,98 / 0,02	0,64 / 0,32	0,74 / 0,21
Configuração - Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>TF</i>		trafo1	0,97	0,50	0,66
		trafo2	0,97	1,00	0,98
		trafo3	1,00	0,44	0,62
		global	0,98 / 0,02	0,65 / 0,31	0,75 / 0,20
Configuração - Segmentador: <i>TextTiling</i> - Peso-vetor: <i>Binary</i>		trafo1	1,00	1,00	1,00
		trafo2	0,97	1,00	0,98
		trafo3	1,00	0,41	0,58
		global	0,99 / 0,02	0,80 / 0,34	0,85 / 0,24
Configuração - Segmentador: <i>TextTiling</i> - Peso-vetor: <i>TF</i>		trafo1	1,00	1,00	1,00
		trafo2	0,97	1,00	0,98
		trafo3	1,00	0,70	0,83
		global	0,99 / 0,02	0,90 / 0,17	0,94 / 0,09

Os documentos do grupo 2, apesar de serem muito similares em conteúdo, são apenas similares em estrutura. Isto reflete em similaridade de conteúdo menor entre os documentos deste grupo, quando comparada à similaridade de conteúdo entre os

documentos do grupo 1. Esta afirmação explica o motivo dos resultados para o exemplo do grupo 2 não terem sido tão bons quando os resultados para o exemplo do grupo 1 (Tabela 4.4). Ainda assim, os resultados na Tabela 4.5 são muito bons, apesar de baixos valores registrados na revocação das passagens “trafo1” e “trafo3”.

A Tabela 4.6 mostra os resultados para um documento-consulta do grupo 3: documento-consulta IO-L-12.pdf, com respectivo documento-treinamento IO-L-16.pdf.

Tabela 4.6: Resultados para um documento-consulta do grupo 3.

Doc-consulta	Doc-treinamento	Passagem Induzida	Precisão	Revocação	medida F
IO-L-12.pdf	IO-L-16.pdf				
Configuração - Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>Binary</i>		protecao	0,93	1,00	0,96
		configuracao	0,99	0,33	0,50
		normalizacao1	0,95	0,21	0,34
		liberacao1	1,00	0,11	0,20
		normalizacao2	0,99	0,55	0,70
		liberacao2	1,00	0,12	0,21
		global	0,98 / 0,03	0,39 / 0,35	0,49 / 0,30
Configuração - Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>TF</i>		protecao	0,89	1,00	0,94
		configuracao	1,00	1,00	1,00
		normalizacao1	0,95	0,21	0,34
		liberacao1	1,00	0,11	0,19
		normalizacao2	0,98	0,41	0,58
		liberacao2	1,00	0,12	0,21
		global	0,97 / 0,04	0,48 / 0,42	0,54 / 0,36
Configuração - Segmentador: <i>TextTiling</i> - Peso-vetor: <i>Binary</i>		protecao	0,93	1,00	0,96
		configuracao	0,93	0,05	0,10
		normalizacao1	0,98	0,60	0,75
		liberacao1	1,00	0,11	0,19
		normalizacao2	0,99	0,55	0,70
		liberacao2	1,00	0,16	0,28
		global	0,97 / 0,03	0,41 / 0,37	0,50 / 0,35
Configuração - Segmentador: <i>TextTiling</i> - Peso-vetor: <i>TF</i>		protecao	0,99	0,50	0,66
		configuracao	0,93	0,05	0,10
		normalizacao1	0,96	0,46	0,62
		liberacao1	1,00	0,32	0,49
		normalizacao2	0,99	0,50	0,66
		liberacao2	1,00	0,16	0,28
		global	0,98 / 0,03	0,33 / 0,19	0,47 / 0,23

A pouca similaridade no conteúdo entre os documentos do grupo 3 explica os baixos valores de revocação apresentados pela Tabela 4.6. Em todas as passagens induzidas, houve variação nos valores de revocação entre as configurações. Uma variação maior ocorreu na passagem “configuracao”. Para esta, a segunda configuração

do sistema *Minimum Cut* e *TF* foi a melhor, atingindo 1,00 (100%) em precisão e revocação. Entretanto, se a configuração for *TextTiling* e *TF*, o resultado na revocação é quase mínimo, atingindo 0,05.

Na análise dos valores de precisão, mostrados nas três tabelas, nota-se que em todos os casos (de documento-consulta, de configuração e de passagem) estes foram excelentes. As passagens induzidas apresentaram um percentual grande de intercessão com seus respectivos *gold standard*, traduzindo-se em elevados valores de precisão.

Apesar das passagens induzidas serem identificadas como as mais similares textualmente às passagens marcadas, nos documentos-treinamento, a indução de uma passagem não quer dizer que ela seja o *gold standard*. Quando maior for o percentual do texto do *gold standard* que se intercepta com a passagem induzida, maior será a revocação da indução. Este percentual será maior quanto maior for a similaridade textual entre a passagem marcada e o *gold standard*. Por este motivo, quando a similaridade entre os documentos (consulta e treinamento) vai diminuindo, a revocação segue. Isto explica a redução nos valores de revocação, nos resultados das tabelas, de um grupo para outro.

Os resultados dos testes, sumarizados, para todos os documentos e em cada grupo, são mostrados pela Tabela 4.7. Os resultados são agrupados por configuração. Observando-os, pode-se concluir que a configuração que obteve os melhores resultados para o primeiro grupo foi *TextTiling* e *TF*, já para o grupo 2 foi *Minimum Cut* e *TF* e para o terceiro grupo, *Minimum Cut* e *Binary*.

Tabela 4.7: Resultados gerais para os três grupos - Chesf

Grupo 1	Sumarizados		
Configuração	Precisão	Revocação	medida F
- Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>Binary</i>	1,00 / 0,00	0,99 / 0,04	0,99 / 0,03
- Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>TF</i>	1,00 / 0,00	0,99 / 0,04	0,99 / 0,03
- Segmentador: <i>TextTiling</i> - Peso-vetor: <i>Binary</i>	1,00 / 0,00	1,00 / 0,02	1,00 / 0,01
- Segmentador: <i>TextTiling</i> - Peso-vetor: <i>TF</i>	1,00 / 0,00	1,00 / 0,00	1,00 / 0,00
Grupo 2	Sumarizados		
Configuração	Precisão	Revocação	medida F
- Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>Binary</i>	0,99 / 0,01	0,90 / 0,14	0,93 / 0,10
- Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>TF</i>	0,99 / 0,02	0,91 / 0,13	0,94 / 0,09
- Segmentador: <i>TextTiling</i> - Peso-vetor: <i>Binary</i>	0,99 / 0,01	0,88 / 0,19	0,91 / 0,12
- Segmentador: <i>TextTiling</i> - Peso-vetor: <i>TF</i>	0,99 / 0,03	0,87 / 0,14	0,90 / 0,12
Grupo 3	Sumarizados		
Configuração	Precisão	Revocação	medida F
- Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>Binary</i>	0,99 / 0,01	0,69 / 0,18	0,77 / 0,15
- Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>TF</i>	0,99 / 0,01	0,69 / 0,24	0,76 / 0,23
- Segmentador: <i>TextTiling</i> - Peso-vetor: <i>Binary</i>	0,99 / 0,01	0,64 / 0,19	0,71 / 0,17
- Segmentador: <i>TextTiling</i> - Peso-vetor: <i>TF</i>	0,99 / 0,02	0,61 / 0,27	0,69 / 0,22

A Figura 4.3 mostra gráficos de barras com os resultados de medida F e revocação, por configuração. No gráfico, nota-se a queda nos valores de revocação, e por conseqüência na medida F, à medida que a similaridade entre os documentos nos grupos vai diminuindo (similaridade: grupo 1 > grupo 2 > grupo 3).

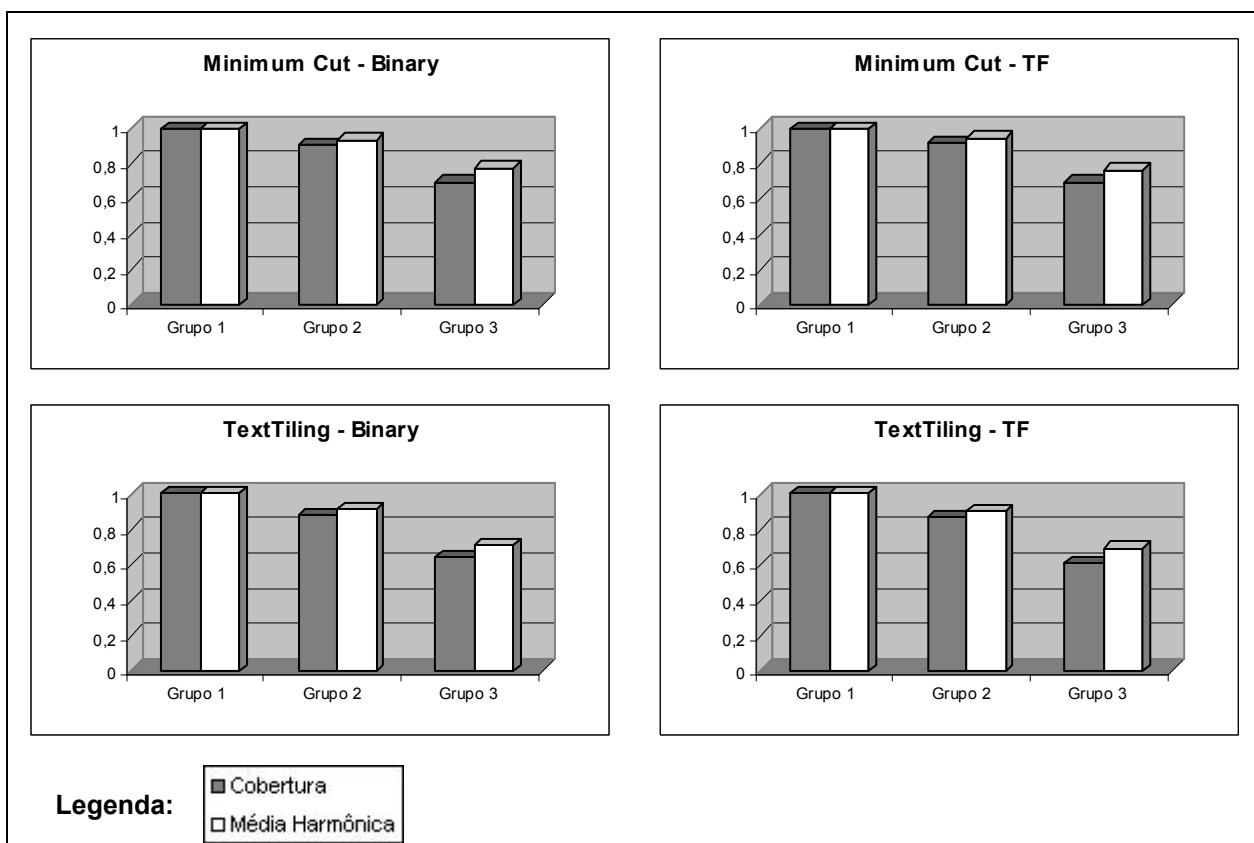


Figura 4.3: Gráficos para as resultados do corpus Chesf

A escolha de qual configuração definir no TIES, para indução de passagens de um domínio específico, pode depender da quantidade de documentos que o domínio tenha em cada grupo. No caso do *corpus* do domínio Chesf, a maior parte dos documentos concentra-se no grupo 1, o que justifica a escolha pela configuração *TextTiling* e *TF*.

4.2.3 Resultados para o *Corpus* Legislativo

Assim como para o *corpus* Chesf, os documentos-consulta do *corpus* Legislativo foram divididos em grupos de similaridade. No entanto, foram divididos apenas entre os grupos 1 e 2, pois neste *corpus* os documentos-consulta são sempre similares, ou muito similares em conteúdo a seus respectivos documentos-treinamento. As quantidades de documentos nos grupos 1 e 2 foram 18 e 18, respectivamente.

A Tabela 4.8 mostra os resultados para um documentos consulta do grupo 1: *procAdm-GO.doc* (Processo Administrativo, do Estado de Goiás) com documento-treinamento *procAdm-AL.doc* (Processo Administrativo, do Estado de Alagoas).

Tabela 4.8: Resultados para um documento-consulta do grupo 1.

Doc-consulta	Doc-treinamento	Passagem Induzida	Precisão	Revocação	medida F
procAdm-MG.doc	procAdm-AL.doc				
Configuração - Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>Binary</i>		instrucao	1,00	1,00	1,00
		recurso-revisao	1,00	1,00	1,00
		disp-gerais	1,00	1,00	1,00
		competencia	1,00	1,00	1,00
		forma-tempo-lugar	1,00	0,75	0,86
		global	1,00 / 0,00	0,95 / 0,11	0,97 / 0,06
Configuração - Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>TF</i>		instrucao	1,00	1,00	1,00
		recurso-revisao	1,00	1,00	1,00
		disp-gerais	1,00	1,00	1,00
		competencia	1,00	0,55	0,71
		forma-tempo-lugar	1,00	0,30	0,46
		global	1,00 / 0,00	0,77 / 0,33	0,83 / 0,24
Configuração - Segmentador: <i>TextTiling</i> - Peso-vetor: <i>Binary</i>		instrucao	1,00	1,00	1,00
		recurso-revisao	1,00	1,00	1,00
		disp-gerais	1,00	1,00	1,00
		competencia	1,00	1,00	1,00
		forma-tempo-lugar	1,00	0,30	0,46
		global	1,00 / 0,00	0,86 / 0,31	0,89 / 0,24
Configuração - Segmentador: <i>TextTiling</i> - Peso-vetor: <i>TF</i>		instrucao	1,00	1,00	1,00
		recurso-revisao	1,00	1,00	1,00
		disp-gerais	1,00	1,00	1,00
		competencia	1,00	0,55	0,71
		forma-tempo-lugar	1,00	0,30	0,46
		global	1,00 / 0,00	0,77 / 0,33	0,83 / 0,24

Assim como nos resultados para o *corpus* anterior, o fato dos dois documentos serem muito similares tanto em conteúdo quando em estrutura explica os excelentes resultados em precisão e revocação. Exceção para as duas últimas passagens induzidas, os textos são bem diferentes de um documento para o outro.

A Tabela 4.9 mostra os resultados para um documento-consulta do grupo 2: estatServ-RR.pdf (Estatuto dos Servidores, do Estado de Roraima) com documento-treinamento estatServ-MT.doc (Estatuto dos Servidores, do Estado de Mato Grosso).

Tabela 4.9: Resultados para um documento-consulta do grupo 2.

Doc-consulta	Doc-treinamento	Passagem Induzida	Precisão	Revocação	medida F
estatServ-RR.pdf	estatServ-MT.doc				
Configuração - Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>Binary</i>		disp-preliminares	1,00	0,67	0,80
		nomeacao	1,00	1,00	1,00
		posse	1,00	1,00	1,00
		reintegracao	0,97	0,25	0,40
		vacancia	1,00	1,00	1,00
		global	0,99 / 0,01	0,78 / 0,33	0,84 / 0,26
Configuração - Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>TF</i>		disp-preliminares	1,00	0,67	0,80
		nomeacao	1,00	1,00	1,00
		posse	1,00	1,00	1,00
		reintegracao	0,97	0,25	0,40
		vacancia	1,00	0,56	0,72
		global	0,99 / 0,01	0,70 / 0,32	0,78 / 0,25
Configuração - Segmentador: <i>TextTiling</i> - Peso-vetor: <i>Binary</i>		disp-preliminares	1,00	0,67	0,80
		nomeacao	1,00	0,29	0,44
		posse	1,00	1,00	1,00
		reintegracao	0,97	0,50	0,66
		vacancia	1,00	1,00	1,00
		global	0,99 / 0,01	0,69 / 0,31	0,78 / 0,24
Configuração - Segmentador: <i>TextTiling</i> - Peso-vetor: <i>TF</i>		disp-preliminares	1,00	0,67	0,80
		nomeacao	1,00	0,29	0,44
		posse	1,00	1,00	1,00
		reintegracao	1,00	1,00	1,00
		vacancia	1,00	0,56	0,72
		global	1,00 / 0,00	0,70 / 0,30	0,79 / 0,23

Nesta tabela, percebe-se que com a redução na similaridade dos documentos do grupo 2, em relação aos do grupo 1, os valores de revocação são reduzidos.

A Tabela 4.10 exibe os resultados dos testes, sumarizados, para todos os documentos, em cada grupo. Os resultados são agrupados por configuração.

Tabela 4.10: Resultados gerais para os dois grupos – Legislativo

Grupo 1	Sumarizados		
Configuração	Precisão	Revocação	medida F
- Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>Binary</i>	1,00 / 0,00	0,99 / 0,02	1,00 / 0,01
- Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>TF</i>	1,00 / 0,00	0,97 / 0,09	0,98 / 0,06
- Segmentador: <i>TextTiling</i> - Peso-vetor: <i>Binary</i>	1,00 / 0,00	0,97 / 0,05	0,98 / 0,04
- Segmentador: <i>TextTiling</i> - Peso-vetor: <i>TF</i>	1,00 / 0,00	0,96 / 0,09	0,97 / 0,06
Grupo 2	Sumarizados		
Configuração	Precisão	Revocação	medida F
- Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>Binary</i>	1,00 / 0,01	0,85 / 0,10	0,89 / 0,07
- Segmentador: <i>Minimum Cut</i> - Peso-vetor: <i>TF</i>	0,99 / 0,01	0,78 / 0,10	0,83 / 0,08
- Segmentador: <i>TextTiling</i> - Peso-vetor: <i>Binary</i>	0,99 / 0,01	0,81 / 0,12	0,86 / 0,09
- Segmentador: <i>TextTiling</i> - Peso-vetor: <i>TF</i>	0,99 / 0,01	0,77 / 0,11	0,84 / 0,08

A Figura 4.4 mostra um gráfico de barras com os resultados de medida F e revocação “plotados”, por configuração.

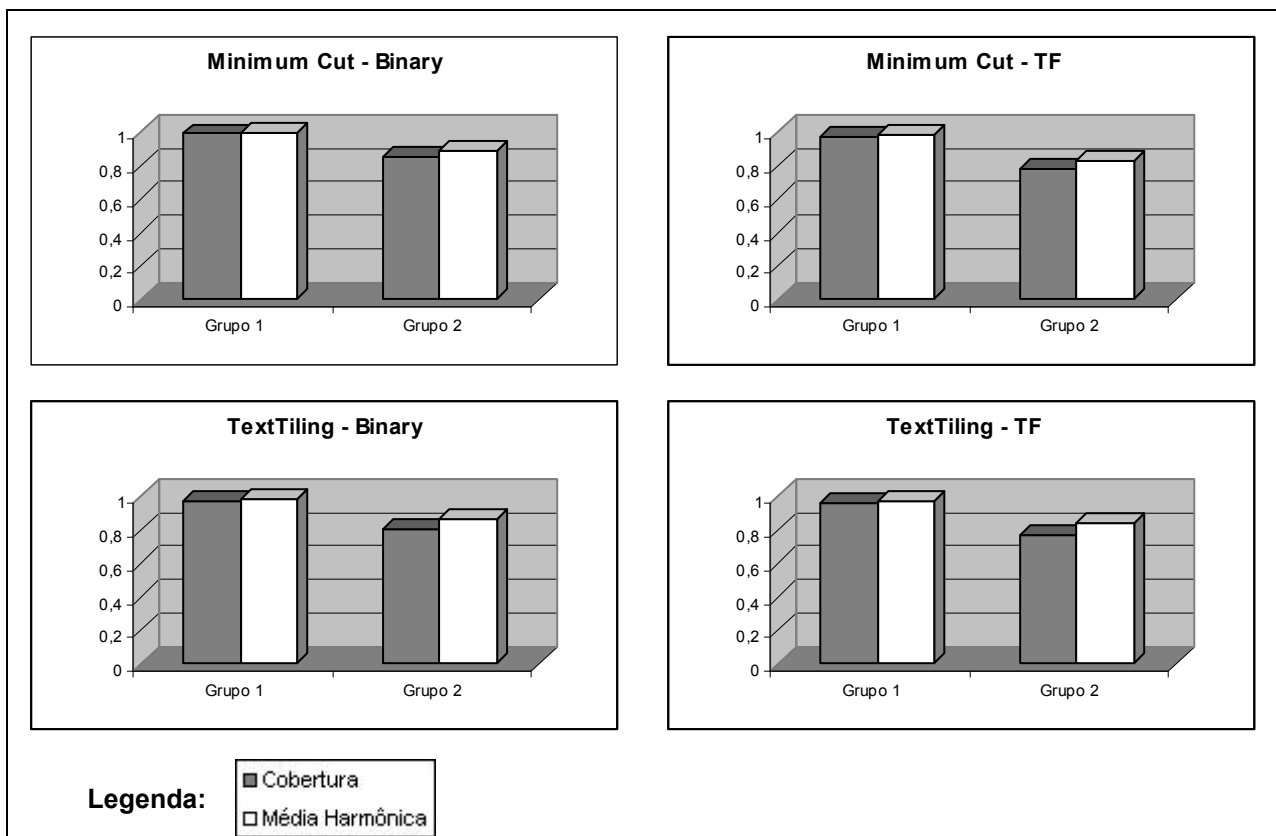


Figura 4.4: Gráficos para os resultados do *corpus* Legislativo

Apesar de em ambos os grupos os documentos-consulta serem muito similares em conteúdo com seus respectivos documentos-treinamento, os documentos do grupo 1 são ainda mais similares que os do grupo 2. Isto porque, como para o *corpus* Chesf, documentos “muito similares em estrutura” tendem a ser mais similares em conteúdo, que documentos que são apenas “similares em estrutura”. Esta afirmação explica a queda nos valores de revocação, do grupo 1 para o grupo 2.

Como é percebido nos gráficos acima, para o *corpus* Legislativo os resultados foram excelentes para os documentos do grupo 1 e muito bons para o grupo 2, considerando qualquer configuração. Em ambos os grupos, a primeira configuração (*Minimum Cut* e *Binary*) foi a que obteve os melhores valores nas métricas.

A seção seguinte apresenta algumas discussões gerais sobre os resultados nesta seção apresentados.

4.3 Conclusões

Esta seção apresenta algumas discussões sobre os resultados da avaliação do TIES, comprovação das hipóteses e apresentação de uma ambiente operacional real onde o TIES pode atuar.

Dos resultados apresentados, destacam-se as altas precisões obtidas em todos os testes. Quanto à revocação, esta foi alta para testes com os documentos dos grupos 1 e 2, em ambos os *corpus*, e obteve valores significativos para o grupo 3 do *corpus* Chesf. A conclusão de todos estes valores é que, como o TIES é um sistema de extração por análises de similaridade, quando maior for a similaridade entre os documentos melhor será a eficácia da extração.

Um problema que afetou os resultados das extrações foi a falsa dissimilaridade entre termos nos textos (sinonímia). Termos sinônimos em dois textos, possuem similaridade semântica, entretanto, contribuem para a redução dos valores de similaridade textual.

No geral, os valores de precisão e revocação, com conseqüente medida F, foram muito satisfatórios.

Com relação às hipóteses levantadas pela Seção 4.1.1, estas foram devidamente comprovadas:

Este capítulo mostrou que o protótipo do TIES foi testado em dois domínios distintos: Chesf e Legislativo. Para ambos os domínios, foram obtidos altos valores de precisão, revocação e medida F. Estes testes e conseqüentes resultados comprovam a hipótese do TIES ser independente de domínio.

Nos testes foram utilizados documentos no formato do Microsoft Word e no formato PDF. Comprovando a independência de formato do TIES.

Os documentos e passagens marcadas no *corpus* Chesf são relevantes e críticos à manutenção do sistema elétrico da empresa. Os altos valores de precisão e revocação, pelo menos para os documentos dos grupos 1 e 2, neste *corpus* comprovam que o TIES é útil para extrair informação em aplicação que exijam eficácia na extração.

O custo de treinamento se resume na quantidade de documentos-treinamento necessária para atingir uma boa acurácia nas extrações. Para ambos os domínios (Chesf

e Legislativo) esta quantidade foi pequena em todos os conjuntos de treinamento formatos (variando entre 1 e 3 documentos por conjunto). Assim, fica comprovada a hipótese de que o TIES tem custo de treinamento baixo.

Capítulo 5

Conclusões

De uma maneira geral, extrair manualmente informação específica de documentos é uma tarefa demasiadamente trabalhosa, ou até mesmo impraticável. Visando a contribuir para a solução do problema de extração de informação, esta dissertação investe em novos métodos e técnicas de extração de informação relevante de documentos textuais, ou não-estruturados.

5.1 Contribuições

É proposto um novo método de extração de informação de documentos não-estruturados, da classe de métodos supervisionados. Como todo método supervisionado, o usuário necessita marcar as passagens relevantes de cada documento de um conjunto de documentos de um certo domínio, conhecido como conjunto-treinamento. Com base no conjunto-treinamento, o método permite extrair passagens relevantes de novos documentos. O problema com métodos supervisionados é que o custo de marcação dos documentos de um conjunto-treinamento pode ser muito alto, mesmo que, como acontece normalmente, o tamanho do conjunto-treinamento seja pequeno, comparativamente à coleção inteira de documentos do domínio.

A principal novidade do método reside na sua técnica de extração de informação, que é uma adequada combinação de técnicas existentes e bem conhecidas: análise de similaridade estrutural, análise de similaridade textual e segmentação topicamente coerente de texto. O método também permite que os conjuntos de treinamento possam ser muito pequenos em termos absolutos, desta forma contribuindo para reduzir o esforço de marcação dos documentos.

Otras importantes características do método são: independente de domínio e independência de formato de documentos.

Para a validação do método, o mesmo foi implementado sob a forma de um novo sistema de extração de informação alcunhado de TIES (*Textual Information Extraction System*).

5.2 Conclusões

Com base nos resultados experimentais com o TIES, pode-se concluir que:

- O método de extração de informação proposto provou ser eficaz, pelo menos para os dois domínios díspares testados: os altos valores das métricas de precisão e revocação comprovam esta afirmação;
- A eficácia de extração de informação depende da boa escolha de um conjunto-treinamento: quanto mais similares forem os documentos do conjunto-treinamento com os documentos a serem marcados, maior será a eficácia de extração;
- O método foi testado com documentos em dois formatos, Microsoft Word e PDF, obtendo a mesma eficácia para cada um dos formatos, o que atesta a independência de formato do método de extração de informação.

A eficácia do TIES para um dos domínios testados, sobre sistemas elétricos de potência, enseja a sua utilização no contexto de uma aplicação real: trata-se de um sistema de apoio à decisão dos operadores da CHESF - SAD¹⁰; toda a parte concernente à extração de informação de documentos normativos e operacionais da CHESF ficará parcialmente a cargo do TIES, restando aos especialistas da CHESF a tarefa (pequena) de ajustar as marcações do TIES.

5.3 Trabalhos Futuros

Como trabalhos futuros podem ser citados:

- Investigação de novas técnicas de marcação automática dos documentos de conjuntos-treinamento, visando a praticamente anular o custo de marcação.

¹⁰ <http://sad.dsc.ufcg.edu.br/pub/Main/ProjetoSAD/SAD-resumo.pdf>. Acessado em julho de 2009.

Já existem trabalhos sobre marcação automática dos documentos-treinamento: em [McCallum 2003] a indução de marcação é apoiada por máquinas de estado finito condicionalmente treinadas; outra abordagem é a utilização de redes lógicas de Markov [Poon 2007]. Entretanto, seus resultados ainda são insuficientes, em termos de eficácia.

- Investigação da utilização de um *thesaurus* [Jones 1993] para as análises de similaridade textual.

A eficácia do TIES poderá ser melhorada, se for utilizado um *thesaurus* para resolver problemas de sinonímia entre termos nos textos.

- Implementação de uma *interface* gráfica com o usuário.

Na versão atual, a *interface* do TIES é via linha de comandos. Por meio de uma interface gráfica com o usuário, este poderia interagir com o sistema: fornecer as entradas, visualizar as saída e ajustar as passagens induzidas quando não tiverem eficácia de 100%.

Referências

- Allan, J. (2003). Hard track overview in trec 2003: High accuracy retrieval from documents. In *Proceedings of the 12th Text REtrieval Conference (TREC)*. pages 24-37, Gaithersburg, Maryland.
- Aumann, Y., Feldman, R., Liberzon, Y., Rosenfeld, B., Schler, J. (2006). Visual Information Extraction. *Knowledge and Information Systems*, 10(1):1-15.
- Choi, Freddy. (1999). JTextTile: A free platform independent text segmentation algorithm.
- Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 26-33, Seattle, Washington. Morgan Kaufmann Publishers Inc.
- Cormen, T H., Leiserson, C. E., Rivest, R. L., Stein, C. (2001). Introduction to Algorithms, 2nd edition. MIT Press.
- Dkaki, T., Mothe, J., Truong, Q. D. (2007). Passage Retrieval Using Graph Vertices Comparison. In *Proceedings of the 2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based (SITIS)*, pages 71-76. IEEE.
- Etzioni, O., Cafarella, M., Downey, D., Popescu, A., Shaked, T., Soderland, S., Weld, D. S., Yates, A. (2005). Unsupervised Named-entity Extraction from the Web: An Experimental Study. *Artificial Intelligence*, 165(1):91-134.
- Etzioni, O., Banko, M., Soderland, S. Weld D. S. (2008). Open Information Extraction form the Web, *Communications of the ACM*, 51(12):68-74.
- Feldman, R., Rosenfeld, B., Soderland, S., Weld, D., Etzioni, O. (2006). Self-Supervised Relation Extraction from the Web, In *Proceedings of the International Symposium on Methodologies for Intelligent Systems*. pages. 755-764.
- Feldman, R., Sanger, J. (2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.

- Flesca, S., Manco, G., Masciari, E., Pontieri, L., Pugliese, A. (2005). Fast detection of XML structural similarity. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):160–175.
- Hammer, J., Garcia-Molina, H., Cho, J., Aranha, R., Crespo, A. (1997). Extracting Semistructured Information from the Web. In *Workshop on Management of Semistructured Data (PODS/SIGMOD'97)*, pages 18-25.
- Hearst, M. (1997). TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33-64.
- Jiang, J., Zhai, C. (2006). Extraction of Coherent Relevant Passages Using Hidden Markov Models. *ACM Transactions on Information Systems (TOIS)*, 24(3):295-319.
- Jones, S. (1993). A Thesaurus Data Model for an Intelligent Retrieval System. *Journal of Information Science*, 19(3)167-178.
- Kruschwitz, U. (2005). Intelligent Document Retrieval – Exploiting Markup Structures. Springer.
- Lau, G. Law, K., Wiederhold G. (2003). Similarity analysis on government regulations. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, Washington, D.C. pages 711-716. ACM.
- Levenshtein, I. V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Cybernetics and Control Theory*, 10(8):707-710.
- Malioutov, I., Barzilay, R. (2006). Minimum Cut Model for Spoken Lecture Segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 25-32, Sydney, Australia. Association for Computational Linguistics.
- Manning, C. D. and Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.
- McCallum, A. (2003). Efficiently Inducing Features of Conditional Random Features. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, pages 403-410.

- Moens, Marie-Francine. (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context*. Springer.
- Poon, H., Domingos, P. (2007). Joint Inference in Information Extraction. In *Proceedings of the 22nd National Conference on Artificial Intelligence*, pages 913-918.
- Reis, D. C., Golgher P. B., Silva, A. S., Laender, A. F. (2004). Automatic web news extraction using tree edit distance. In *Proceedings of the 13th international conference on World Wide Web*, New York, NY, USA, pages 502-511. ACM.
- Rabiner, L. R. (1990). A tutorial on hidden Markov models and selected applications in speech recognition. *Readings in speech recognition*, pages 267-296. Morgan Kaufmann Publishers Inc.
- Salton, G. (1968). *Automatic Information Organization and Retrieval*. New York: McGraw Hill, page 421.
- Salton, G., and C. Buckley. (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5):513-23.
- Schubert, L. (2002). Can We Derive General World Knowledge from Texts?. In *Proceedings of the second international conference on Human Language Technology Research*, pages 94-97, San Diego, California. Morgan Kaufmann Publishers Inc.
- Soboroff, I. (2004). Overview of the TREC 2004 Novelty Track, In *Proceedings of Text Retrieval Conference (TREC)*.
- Valiente, G. (2002). Tree edit distance and common subtrees. Research Report LSI-02-20-R, Universitat Politècnica de Catalunya, Barcelona, Spain.
- Moens, Marie-Francine. (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context*. Springer.