PROGRAMA DE PÓS-GRADUAÇÃO
MESTRADO ACADÊMICO EM SISTEMAS E COMPUTAÇÃO

JOSÉ ALBERTO SOUSA TORRES

SUMARIZAÇÃO AUTOMÁTICA DE ARTIGOS CIENTÍFICOS DE ENGENHARIA DE SOFTWARE COMO SUPORTE AO PROCESSO DE REVISÃO SISTEMÁTICA

Salvador
2011

JOSÉ ALBERTO SOUSA TORRES

SUMARIZAÇÃO AUTOMÁTICA DE ARTIGOS CIENTÍFICOS DE ENGENHARIA DE SOFTWARE COMO SUPORTE AO PROCESSO DE REVISÃO SISTEMÁTICA

Dissertação apresentada à Universidade Salvador como parte das exigências do Curso de Mestrado Acadêmico em Sistemas e Computação, área de concentração Engenharia de Software e Gerência do Conhecimento, para obtenção do título de "Mestre em Sistemas e Computação".

Orientador: Profa. Dra. Laís do Nascimento Salvador
Co-Orientadora:Profª. Drª. Daniela Soares Cruzes

Salvador
2011

FICHA CATALOGRÁFICA

(Elaborada pelo Sistema de Bibliotecas da Universidade Salvador - UNIFACS)

Torres, José Alberto Sousa

Sumarização automática de artigos científicos de engenharia de software como suporte ao processo de revisão sistemática/ José Alberto Sousa Torres. – 2011.

105 f. : il.

Dissertação (Mestrado) - Universidade Salvador – UNIFACS.
Mestrado em Sistemas e Computação, 2011.
Orientadora: Profª. Drª. Laís do Nascimento Salvador
Co- Orientadora: Profª. Drª. Daniela Soares Cruzes

1. Mineração de dados. 2. Teorema de Bayes. 3. Ontologias. I. Salvador, Laís do Nascimento, orient. II. Cruzes, Daniela Soares Cruzes, co-orient. III. Título.

CDD: 004. 6

JOSÉ ALBERTO SOUSA TORRES

SUMARIZAÇÃO AUTOMÁTICA DE ARTIGOS CIENTÍFICOS DE ENGENHARIA DE SOFTWARE COMO SUPORTE AO PROCESSO DE REVISÃO SISTEMÁTICA

Dissertação apresentada à Universidade Salvador como parte das exigências do Curso de Mestrado Acadêmico em Sistemas e Computação, área de concentração em Engenharia de Software e Gerência do Conhecimento, para obtenção do título de "Mestre em Sistemas e Computação.

Laís do Nascimento Salvador – Orientadora _____
Doutora em Engenharia da Computação pela Universidade São Paulo (USP)
Universidade Federal da Bahia – UFBA

Cláudio Nogueira Sant'Anna _____
Pós-Doutor pela Universidade pela Universidade Federal da Bahia – UFBA
Universidade Federal da Bahia – UFBA

Glauco de Figueiredo Carneiro _____
Doutor em Ciência da Computação - Ufba
UNIFACS Universidade Salvador

Salvador de junho de 2011

Dedico este trabalho à minha família: meus pais, irmãs e minha esposa, figuras fundamentais na construção de minha personalidade e eternos incentivadores do meu crescimento intelectual.

AGRADECIMENTOS

Foram muitas as dificuldades durante a construção deste trabalho, assim como foram muitos os momentos em que pensei em desistir. A desgastante carga de trabalho, as inúmeras viagens a serviço e o estresse cotidiano por pouco não fizeram com que este estudo não saísse do campo das idéias.

Os problemas enfrentados fazem com que esta seção fuja do caráter meramente figurativo e se torne uma verdadeira homenagem a um grupo de incentivadores que me apoiaram diariamente e me mostraram que eu era capaz de realizar este sonho.

Agradeço, primeiramente, a minha amada esposa Iza que, de uma só vez, assumiu os papéis de revisora oficial deste trabalho e líder do clube de incentivadores. Sem o seu apoio incondicional, certamente não teria conseguido.

Agradeço aos meus pais Lurdinha e Betinho, membros vitalícios do clube e meus conselheiros espirituais nos momentos de verdadeira dificuldade. Às minhas irmãs Tati e Dani, pelo apoio e palavras de incentivo.

À professora Daniela, que me acompanhou desde o momento que este projeto ainda era um embrião, e a professora Laís, que me adotou em um momento em que quase já cruzava a tênue linha entre prosseguir e desistir, meus sinceros agradecimentos. Espero ter retribuído a expectativa que depositaram neste trabalho.

Aos meus amigos Ricássio – e seu eterno jeito engraçado que consegue fazer sorrir até o anão zangado; Brenno, com seu coração grande que não cabe no peito; e Fábio, enrolado mas gente boa, que sempre acreditaram que eu conseguiria chegar até aqui. Não posso esquecer do chato do James e suas eternas mensagens cobrando o fim da dissertação, que se tornaram um grande incentivo para que eu conseguisse terminar este trabalho.

Agradeço, ainda, aos professores Roberto Pinho, Ibekwe-SanJuan e Agarwal, sempre atenciosos na resposta aos e-mails, que forneceram acesso a métodos e

ferramentas desenvolvidas em seus estudos e que foram fundamentais para a construção desta dissertação.

Não poderia esquecer, ainda, do meu eterno conselheiro e uma das mentes mais incríveis que conheço, ao Prof. Cláudio Amorim, mais uma vez, obrigado.

Quando tudo está perdido, sempre existe uma caminho, quando tudo está perdido, sempre existe uma luz...

(Renato Russo – Via láctea)

RESUMO

O elevado esforço e a grande quantidade de tempo necessários à construção de uma revisão sistemática tem sido os principais obstáculos a um incremento ainda maior na quantidade publicada deste tipo de estudo. Diversos autores tem proposto métodos e desenvolvido ferramentas com o objetivo de reduzir o tempo para construção destes estudos, através da automatização de partes do processo. Nesta linha, o nosso estudo busca automatizar uma das etapas mais custosas do processo de revisão sistemática, a de "Identificação de Resultados do Estudo", atividade na qual os pesquisadores lêem o texto completo de um conjunto de artigos científicos com o objetivo de identificar e isolar as sentenças que indicam "resultados". O principal objetivo é o de reduzir a quantidade de texto a ser lido pelos pesquisadores para conseguir identificar os "resultados", o que levaria, consequentemente, a uma diminuição do tempo necessário para construção do estudo e a redução da carga de trabalho dos pesquisadores envolvidos. O método que propomos analisa o texto completo dos artigos de entrada e classifica automaticamente as sentenças em uma de duas categorias pré-definidas: "resultados" e "não resultados", selecionando, em seguida, apenas os parágrafos que contenham sentenças definidas como "resultados". O texto final resumido, a ser analisado pelos pesquisadores, é montado a partir da listagem sequencial destes parágrafos selecionados. Nos testes realizados, o algoritmo produziu sumários com 20% do tamanho dos documentos originais, em média. Observou-se, ainda, um percentual de acerto na classificação dos parágrafos da ordem de 74%, resultado que consideramos positivo, levando-se em consideração o percentual de redução do texto original e a dificuldade inerente à análise de documentos em linguagem natural.

Palavras-chave: Mineração de Textos. Extração da Informação. Revisão Sistemática. Sumarização de textos.

ABSTRACT

The high effort and large amount of time required to build a systematic review has been the main obstacles to increase the amount published of this type of study. Several authors have proposed methods and developed tools aiming to reduce the construction time for these studies through automating parts of the process. In this vein, our study seeks automate one of the most costly stages of the systematic review process - "Identification of the Study Results", activity in which researchers read the full text of a set of articles in order to identify and isolate the sentences that show "results". The main objective is to reduce the amount of text to be read by researchers aiming to identify results, reducing, consequently, the time necessary for build the study and reducing the workload of researchers involved. The method that we propose analyzes the complete text of Articles and automatically classifies incoming sentences in one of two pre-defined categories: "results" and "no results", selecting only paragraphs containing sentences defined as "results". The final summary, to be analyzed by the researchers, is assembled from a list of sequential selected paragraphs. In the test performed, the algorithm produced summaries with 20% of the size of the original documents, on average. It got also a percentage of correct classification of paragraphs of the order of 74%, a result which is good, taking into account the percentage of reduction of the original text and difficulty inherent in the analysis of documents in natural language.

Keywords: Text Mining. Information Extraction. Systematic Reviews. Text Summarization.

LISTA DE FIGURAS

Figura 2.1 - Etapas do pré-processamento do texto 24
Figura 2.2 - Taxonomia para Algoritmos de Stemming 25
Figura 2.3 - Método de classificação vizinho mais próximo 28
Figura 2.4 - Casos típicos encontrados com a aplicação do SVM 29
Figura 2.5 - Grupo de Objetos divididos em duas classes 30
Figura 2.6 - Classificação de novos dados utilizando Naive Bayes 31
Figura 2.7 - Exemplo de Árvore de Decisão 34
Figura 2.8 - Cálculo da distância das palavras de uma sentença 37
Figura 2.9 - Estruturas para indicadores do grupo B (simplificado) 42
Figura 2.10 - Patamares de tamanho das sentenças 43
Figura 5.1 - Esquema do método de sumarização do texto 63
Figura 5.2 - Esquema do método classificação 66
Figura 7.1 - Diagrama de Fluxo do Experimento 88

LISTA DE GRÁFICOS

Gráfico 5.1 - Distribuição dos valores da frequência das palavras-chave nas sentenças resultado e não resultados em relação ao total de sentenças do texto..... 68
Gráfico 5.2 - Distribuição dos valores das palavras sinalizadoras entre as sentenças resultado e não resultados, relativo a cada categoria, isoladamente..... 69
Gráfico 5.3 - Distribuição dos valores das palavras sinalizadoras, das sentenças resultado e não resultados, em relação ao total de sentenças 1
Gráfico 5.4 - Distribuição dos valores do atributo "Setor no Parágrafo" entre as sentenças do corpus 70
Gráfico 5.5 - Distribuição dos valores do atributo "TF-IDF" entre as sentenças do corpus 70
Gráfico 5.6 - Distribuição dos valores do atributo "TF-ISF" entre as sentenças do corpus 1
Gráfico 5.7 - Distribuição dos valores do atributo "Tamanho da Sentença" entre as sentenças do corpus 71
Gráfico 5.8 - Distribuição dos valores do atributo "Gist Sentence" entre as sentenças do corpus 72
Gráfico 5.9 - Distribuição dos valores do atributo "Média da conectividade léxica" entre as sentenças do corpus 1
Gráfico 5.10 - Distribuição dos valores do atributo "Média da conectividade léxica" entre as sentenças..... 73
Gráfico 5.11 - Distribuição dos valores do atributo "Tempo verbal da sentença" entre as sentenças..... 73
Gráfico 5.12 - Distribuição das citações entre resultados e não resultados 1
Gráfico 5.13 - Distribuição dos resultados e não resultados entre as diferentes seções do texto..... 74
Gráfico 5.14. - Divisão das sentenças pertencentes a cada uma das seções entre as duas categorias em análise 75
Gráfico 5.15 - Distribuição dos resultados e dos não resultados entre os possíveis valores das Frase auto indicativas..... 75
Gráfico 5.16 - Presença de comparação nas sentenças do texto..... 76
Gráfico 5.17 - Distribuição dos resultados e não resultados em relação à presenção ou não de comparação..... 76
Gráfico 5.18 - Análise da presença de número nas sentenças..... 77
Gráfico 5.19 - Análise da presença de indicador de porcentagem nas sentenças..... 1

LISTA DE TABELAS

Tabela 2.1 - Condições e Ações da tabela de decisão modelo (S = Atende o critério e N = Não atende) 33
Tabela 2.2 - Condições e Ações da tabela de decisão modelo 34
Tabela 2.3 - Dicionário de Palavras 39
Tabela 2.4 - Lista de paradigmas e grupos de palavra-chave 42
Tabela 3.1 - Análise de modelos para publicação de artigos 47
Tabela 3.2 - Análise estatística da estrutura de artigos científicos 48
Tabela 3.3 - Análise da estrutura de artigos em Ciência da Computação 48
Tabela 3.4 - Descritivo de Seções do artigo cientifico 48
Tabela 3.5 - Número de segmentos (linhas) em cada seção (colunas) 49
Tabela 3.6 - Distribuição dos resultados encontrados em artigos por sessão 49
Tabela 4.1 - Categorias para classificação de Sentenças do texto científico 53
Tabela 4.2 - Características das Sentenças do texto científico 54
Tabela 4.3 - Precisão e Abrangência do algoritmo de Teufel 55
Tabela 4.4 - Exemplo de Padrões léxico-sintáticos 56
Tabela 4.5 - Resultado do experimento de classificação das sentenças 57
Nota: (Legenda: A – Abrangência e P - Precisão) 57
Tabela 4.6 - Análise de termos retirados do Corpus de Agarwal 58
Tabela 4.7 - Classificação de Sentenças utilizando o método de Agarwal 58
Tabela 4.8 - Importância dos termos na definição da categoria das sentenças 59
Tabela 4.9 - Frequência de termos em artigos retirados do Corpus de Cruzes 61
Tabela 5.1 - Resultado da classificação das sentenças 78
Tabela 7.1 - Classificação das sentenças após aplicação do conjunto de regras 1, 2 e 3. 90

LISTA DE QUADROS

Quadro 2.1 - Representação do texto em função dos termos e de seus pesos..... 23
Quadro 2.2 - Distância entre as instâncias xi e xj..... 27
Quadro 2.3 - Restrições do SVM..... 29
Quadro 2.4 - Fórmula de Bayes aplicada à classificação de documentos..... 32
Quadro 2.5 - Quadrantes da tabela de decisão 32
Quadro 2.6 - Fórmula para cálculo do abrangência..... 35
Quadro 2.7 - Fórmula para cálculo da precisão..... 36
Quadro 2.8 - Definição de peso pelo método da localização..... 39
Quadro 2.9 - Fórmula para Cálculo do TF-IDF 41
Quadro 2.10 - Fórmula do Cálculo do TF-ISF..... 41
Quadro 5.1 - Heurística do algoritmo de detecção de Sentenças 64
Quadro 5.2 - Fórmula de Cálculo do TF-ISF..... 65
Quadro 5.3 - Conjunto de Regras 1 67
Quadro 5.4 - Conjunto de Regras 2 67
Quadro 5.5 - Conjunto de Regras 3 67

SUMÁRIO

1 INTRODUÇÃO 17
1.1 REVISÃO SISTEMÁTICA EM ENGENHARIA DE SOFTWARE 18
1.2 ESTUDOS SOBRE AUTOMATIZAÇÃO DO PROCESSO DE REVISÃO SISTEMÁTICA 20
1.3 OBJETIVO 20
1.3.1 Objetivos Secundários 20
1.4 LIMITAÇÕES 21
1.5 ORGANIZAÇÃO DO TRABALHO 21
2 MINERAÇÃO DE TEXTO 23
2.1 ETAPAS DA MINERAÇÃO DE TEXTO 23
2.1.1 Pré-Processamento 23
2.1.2 Classificação de Textos 27
2.1.3 Desempenho da Classificação 35
2.2 CLASSIFICAÇÃO DE SENTENÇAS DO TEXTO CIENTÍFICO 36
2.2.1 Freqüência das Palavras-chave (*frequency keywords*) 36
2.2.2 Método das Palavras Sinalizadoras (*cue method*) 38
2.2.3 Método da Localização (*location method*) 39
2.2.4 Método das Palavras-chave no Título (title-keyword method) 40
2.2.5 Método da Palavra Temática (ThematicWord Feature) 40
2.2.6 Frase Auto-indicativa 41
2.2.7 Tamanho da Sentença (Sentence Lenght) 43
2.2.8 Presença de Palavras da "Gist Sentence" na Sentença 43
2.2.9 Método da Inicial Maiúscula (*UppercaseWord Feature*) 44
2.2.10 Métodos da Linha Base (*baseline*) e da Média da Conectividade Léxica (*average lexical connectivity*) 44
2.3 CONSIDERAÇÕES FINAIS 44
3 Análise do Artigo Científico 45
3.1 ANÁLISE DAS NORMAS ESTILÍSTICAS DE ESCRITA 45
3.2 ORGANIZAÇÃO DO CONTEÚDO EM ARTIGOS CIENTÍFICOS 47
3.3 POR QUE PADRONIZAR AS SEÇÕES DO CORPO DO ARTIGO 48
3.4 ANOTAÇÃO DE ARTIGOS CIENTÍFICOS 49
3.5 DISCUSSÃO 50
3.6 CONCLUSÃO 51
4 CLASSIFICAÇÃO DE SENTENÇAS 52
4.1 ARGUMENTAÇÃO BASEADA EM DISCURSO EM ARTIGOS CIENTÍFICOS 53
4.2 IDENTIFICANDO INFORMAÇÕES ESTRATÉGICAS EM ARTIGOS CIENTÍFICOS 55
4.3 CLASSIFICAÇÃO AUTOMÁTICA DE SENTENÇAS EM TEXTOS COMPLETOS DE ARTIGOS 56
4.4 CONCLUSÃO 61
5 MÉTODO PROPOSTO 63
5.1 IMPORTAÇÃO DOS ARTIGOS E SEGMENTAÇÃO DAS SENTENÇAS 63
5.2 CÁLCULO DOS ATRIBUTOS 64
5.3 CLASSIFICAÇÃO 66
5.3.1 Classificação Baseada em Regras 67
***j) Tempo verbal da sentença* 1**
5.3.2 Classificação Baseada Em aprendizado de Máquina 78

5.4 SELEÇÃO DOS PARÁGRAFOS 79
5.5 CONCLUSÃO 79
6 MÉTODO PROPOSTO 81
6.1 DESCRIÇÃO DAS FUNCIONALIDADES 81
6.2 CONCLUSÃO 84
7 EXPERIMENTO 86
7.1 DEFINIÇÃO DO EXPERIMENTO 86
7.1.1 Questões 87
7.1.2 Métricas 87
7.1.3 Seleção de Variáveis 87
7.1.4 Recursos 88
7.2 PLANEJAMENTO DO EXPERIMENTO 88
7.2 EXECUÇÃO DO EXPERIMENTO 89
7.3 RESULTADOS DO EXPERIMENTO 90
7.4 CONCLUSÃO 91
8 RESULTADOS E DISCUSSÃO 92
8.1 ANÁLISE DA ESTRUTURA DO ARTIGO EM ENGENHARIA DE SOFTWARE 92
8.2 ANÁLISE DOS MÉTODOS DE CLASSIFICAÇÃO DE SENTENÇAS 93
8.3 DEFINIÇÃO DO MÉTODO DE SUMARIZAÇÃO DE TEXTO 94
8.3.1 Ferramenta para pré-processamento textual 95
9 CONCLUSÃO 97
REFERÊNCIAS 99
Anexo A - Lista dos Artigos Utilizados nesta Dissertação oriundos do Corpus de Cruzes (2007)

Capítulo 1

INTRODUÇÃO

Este capítulo tem por objetivo contextualizar a pesquisa, fornecer uma visão geral do ambiente e apontar a relevância deste estudo, justificando a sua execução. Ainda neste capítulo são definidas as limitações do trabalho e é descrita, de forma resumida, a sequência de atividades necessárias para o seu desenvolvimento.

A história da revisão sistemática começa no início do século XX, embora sua popularidade tenha crescido somente no final da década de 90. A primeira revisão deste tipo que se tem registro foi publicada por Pearson (1904) no *British Medical Journal* e sintetizava resultados de apenas dois estudos.

A revisão sistemática é uma revisão de estudos primários que, através de uma abordagem sistemática, objetiva reduzir viés, ou seja, evitar a ocorrência de um erro sistemático ou tendenciosidade. Este estudo sintetiza informações sobre efeitos de tecnologias – tamanho do efeito e para quem o efeito ocorreu – produzidas pelos estudos disponíveis de boa qualidade, fornecendo uma base científica para a tomada de decisão a respeito do tema abordado (SILVA, 2003). O processo de condução deste tipo de revisão guia-se por uma estrita e bem definida sequência de passos metodológicos, que seguem um rígido protocolo definido antes do início das atividades. A estratégia de busca de evidências, a questão, o foco e todos os detalhes metodológicos são formalmente explicitados, de forma que outros profissionais possam reproduzir o mesmo protocolo (BIOLCHINI, 2005).

As revisões sistemáticas utilizam protocolos cada vez mais elaborados que procuram reduzir os numerosos vieses possíveis no processo de estimar efeitos. Nesse sentido, utilizam estratégias de busca de estudos exaustivas, examinam a qualidade do desenho, da execução do estudo e fazem referência sobre a força da evidência que cada estudo proporciona, além da validade de sua inclusão na revisão (SILVA, 2003).

1.1 REVISÃO SISTEMÁTICA EM ENGENHARIA DE SOFTWARE

Uma considerável quantidade de estudos primários com preocupação metodológica vem sendo desenvolvidos na área de Engenharia de Software nos últimos anos. Entretanto, estes trabalhos, isoladamente, não possuem evidências suficientes para comprovar a aplicabilidade ou qualidade de uma determinada técnica, em função da grande possibilidade de variações nos ambientes onde se desenvolveram cada um dos experimentos.

Uma das formas de solucionar este problema seria através da análise de um conjunto de artigos que tratem de uma mesma questão de pesquisa e a posterior síntese dos seus resultados. Nesta linha, existe uma série de iniciativas questionando como a Engenharia de Software poderia ser beneficiada com a adoção de uma metodologia baseada em evidências, que seriam obtidas através da construção de revisões sistemáticas.

Diferentemente da área médica, onde a construção destas revisões são amplamente difundidas, a Engenharia de Software tem algumas especificidades que podem tornar mais difícil a síntese das pesquisas para obter evidências. Uma das maiores diferenças é que a maioria dos métodos e técnicas de Engenharia de Software devem ser aplicados por profissionais de Software qualificados que conheçam os métodos e técnicas que estão sendo aplicados. Já na área da saúde, apesar dos profissionais médicos serem qualificados, o tratamento prescrito (por exemplo, uso de um determinado medicamento) não requer a sua presença para ser corretamente administrado por outro profissional ou recebido pelo paciente. Tomemos como exemplo o experimento padrão ouro na avaliação de questões terapêuticas em saúde, que é o ensaio clínico controlado randomizado duplo-cego. Neste tipo de estudo, nem o médico nem o paciente sabem qual tratamento está sendo administrado, de forma a prevenir que as suas expectativas distorçam os resultados. Este protocolo é impossível de ser aplicado na Engenharia de Software, pois para que o profissional desta área possa executar uma tarefa, ele precisa conhecer o método ou ferramenta envolvida (BIOLCHINI, 2005).

Uma outra diferença entre estas duas áreas é o fato de existir, na área de saúde, uma grande quantidade de instituições que fomentam a construção e divulgam os resultados das revisões sistemáticas, a exemplo da Colaboração Cochrane[1] ou do Centro de Revisões e Disseminação da Universidade de York[2]. Ainda não existem organizações similares na área de Engenharia de Software.

---

1 http://www.cochrane.org/
2 http://www.york.ac.uk/inst/crd/

Um dos primeiros estudos a definir as diretrizes para construção de revisões sistemáticas em Engenharia de Software foi proposto por Kitchenham (2004) e baseou-se em diretrizes estabelecidas para a área médica. Um outro trabalho de definição de método para realização deste tipo de estudo secundário na área de Engenharia de Software foi proposto por Cruzes (2007) – o método SecESE, e será utilizado como base desta Dissertação.

1.1.1 SecESE - Análise Secundária em ESE

A SecESE, proposta por Cruzes (2007), utiliza métodos explícitos e reproduzíveis que são executados sistematicamente, de forma a assegurar que o procedimento da análise seja visível e reprodutível por outros pesquisadores.

Como pode ser visualizado na Figura 2.1, a abordagem SecESE é dividida em quatro atividades principais, são elas:

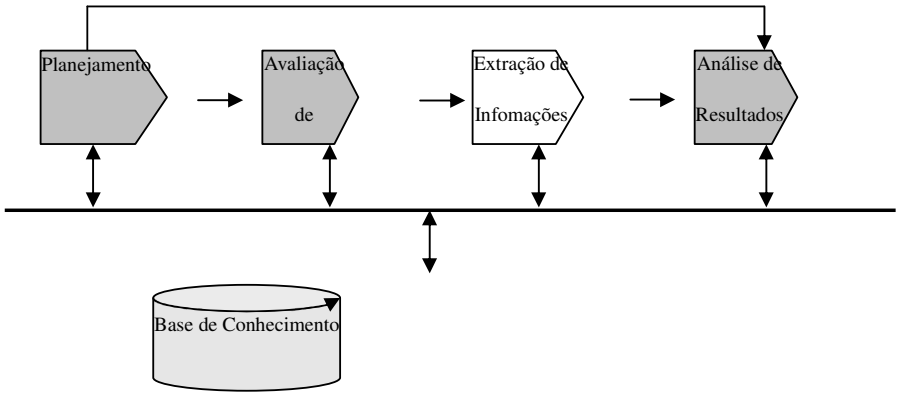
(1) Planejamento - *definição e planejamento da análise e da explicitação dos critérios de busca e seleção dos objetos de análise;*

(2) Avaliação de qualidade - a*pós a seleção dos objetos de análise, deve-se avaliar a qualidade dos mesmos;*

(3) Extração – *é realizada a extração das informações codificadas dos artigos;*

(4) Análise de resultados - *é realizada a análise das informações coletadas durante todo o processo.*

Figura 1.1 - Atividades da SecESE



Fonte: Cruzes (2007)

Destacamos, na Figura 1.1, a fase de Extração de Informações, por ser a etapa base para o desenvolvimento do objeto desta dissertação. Segundo a metodologia de Cruzes, esta atividade se divide em duas sub-atividades: a extração de informações de contexto e a extração de resultados. A primeira trata da recuperação de determinados itens do artigo selecionados pela autora do trabalho, como o título, o tópico de estudo, as variáveis dependentes e independentes e uma série de outras informações que tratam de aspectos relevantes do estudo. A segunda etapa trata da identificação das sentenças que indicam resultados no texto científico.

1.2 ESTUDOS SOBRE AUTOMATIZAÇÃO DO PROCESSO DE REVISÃO SISTEMÁTICA

Estudos sobre ferramentas de apoio às etapas de "Avaliação de Qualidade" e "Extração das Informações" foram produzidos por Malheiros (2007) – com o PexExplorer, e Silva Rocha (2009) - com o ContextExtractor.

Ambas as ferramentas foram construídas com base em técnicas de mineração visual de textos. O PexExplorer com o intuito de auxiliar os pesquisadores na seleção dos artigos a serem utilizados na construção da revisão sistemática e o ContextExtractor com o objetivo de recuperar automaticamente as informações de contexto.

Estes dois estudos demonstram que é viável a utilização de suporte computacional ao processo de Revisão Sistemática, com redução do tempo e do custo necessários para realização deste tipo de estudo.

1.3 OBJETIVO

Esta dissertação tem por objetivo principal propor um método para sumarização de artigos científicos em Engenharia de Software, com ênfase na localização e identificação de sentenças que representem os resultados dos artigos científicos em análise, de forma a reduzir o tempo gasto pelos pesquisadores ao realizar as revisões sistemáticas. O método proposto atuará na automatização de parte da etapa de extração de informações da SecESE, exibida na figura 1.1.

1.3.1 Objetivos Secundários

a) Identificar padrões estrututrais em artigos científicos, com base na análise da organização do seu conteúdo e formatação visual do texto – Capítulo 3;

b) Avaliar e comparar as técnicas de classificação de sentenças em artigos científicos

– Capítulo 4;

c) Especificar e construir um protótipo de ferramenta para dar suporte à mineração visual dos artigos científicos e à classificação de suas sentenças – Capítulo 5;

d) Modelar e executar um experimento controlado de forma a avaliar a precisão e abrangência do método proposto – Capítulo 6.

1.4 LIMITAÇÕES

Não faz parte do escopo deste trabalho propor melhorias a metodologia para construção de estudos secundários utilizada – a SecESE.

O escopo é limitado, ainda, à análise e criação de sumários de artigos da área de Engenharia de Software. Neste trabalho não será realizado qualquer experimento da aplicação desta técnica em artigos de outras áreas, principalmente por não possuirmos um corpus de artigos anotados (com a marcação explícita da categoria das sentenças) que nos possibilite executar estes testes. Este fato, entretanto, não exclui a possibilidade de que a técnica desenvolvida possa ter resultados satisfatórios se utilizada em tais documentos, apenas não garantimos que os resultados sejam semelhantes aos apresentados para os artigos de Engenharia de Software.

1.5 ORGANIZAÇÃO DO TRABALHO

Este estudo encontra-se dividido nos seguintes capítulos:

a) Capítulo 2 – Mineração de Textos – descrição dos conceitos e técnicas de mineração de texto utilizados neste trabalho;

b) Capítulo 3 - Análise do Artigo Científico – análise da estrutura de formatação visual e composição do conteúdo de Artigos Cientificos de Engenharia de Software;

c) Capítulo 4 – Classificação das Sentenças – análise dos trabalhos de Teufel (1999), Ibekwe-SanJuan (2008) e Agarwal (2009) sobre técnicas de classificação de sentenças em artigos científicos;

d) Capítulo 5 – Método Proposto – descrição do método utilizado neste trabalho;

e) Capítulo 6 – Ferramenta de Avaliação – descrição das funcionalidades da ferramenta utilizada no experimento deste estudo;

f) Capítulo 7 – Experimento – descrição do experimento realizado para validação do método proposto;

g) Capítulo 8 – Resultados e Discussão – análise dos resultados encontrados neste estudo;

h) Capítulo 9 – Conclusão – aponta as conclusões obtidas a partir da análise dos resultados deste estudo.

Capítulo 2

MINERAÇÃO DE TEXTO

Este capítulo contém uma breve revisão de literatura sobre a mineração de textos, em especial as etapas de pré-processamento – análise léxica, remoção de "stop words", redução morfológica, normalização de sinônimos, atribuição de pesos e redução de dimensionalidade; e classificação – análise dos métodos "naive bayes", vizinho mais próximo, SVM e tabela e árvores de decisão. Serão analisadas, ainda, técnicas para cálculo de pesos e classificação de sentenças do texto..

A Mineração de textos, também conhecida como mineração de dados do texto, refere-se, geralmente, ao processo de extração de informações de documentos de texto não estruturados, podendo ser visto como uma extensão da Mineração de Dados (do inglês Data Mining) (TAN,1999).

O processo de mineração de textos foi definido por Martins Junior (2003) em função de atividades como a Obtenção dos Documentos, Pré-Processamento dos Documentos, Extração do Conhecimento ou Classificação e Avaliação do Conhecimento.

2.1 ETAPAS DA MINERAÇÃO DE TEXTO

Neste tópico, detalharemos as etapas de Pré-processamento e Classificação dos documentos textuais, por estarem relacionadas ao conteúdo deste trabalho durante toda a etapa de processamento dos documentos textuais que representam os artigos científicos até a fase de criação do resumo do texto.

2.1.1 Pré-Processamento

Em processos de mineração, um texto é usualmente representado como um vetor de palavras (também chamado vetor de termos) com pesos associados. A representação deste vetor pode ser observada no Quadro 2.1, onde *d* representa o documento, *w* o conjunto de palavras extraídas do texto e *p* os pesos associados.

Quadro 2.1 - Representação do texto em função dos termos e de seus pesos

$$d_1 = (w_1p_1, w_2p_2, \ldots w_np_n)$$

Este vetor sofre uma série de transformações durante o processo, de forma que, ao final, passa a conter uma série de informações que possibilitam a utilização das mais variadas técnicas de extração da informação. Torres (2005) definiu um processo para realização do pré-processamento do documento textual. As sete atividades sequenciais são exibidas na Figura 2.1 e descritas nas seções subsequentes.

Figura 2.1 - Etapas do pré-processamento do texto

Fonte: Torres (2005).

2.1.1.1 Análise Léxica

É a primeira etapa do processo e tem por objetivo definir o conceito de "termo" a ser adotado. Tipicamente, os termos são definidos como as palavras presentes no texto, mas há autores que utilizam as sentenças como indicadores de termos, acreditando que esta representação apresenta melhor semântica.

2.1.1.2 Remoção de Stop Words

É a etapa em que são eliminados os termos indesejados, presentes no vetor inicial. "Stop Words" são os termos que aparecem com alta frequência no texto, mas não contribuem para a diferenciação entre categorias, mostrando-se irrelevantes para o processo de classificação e gerando uma carga adicional de processamento.

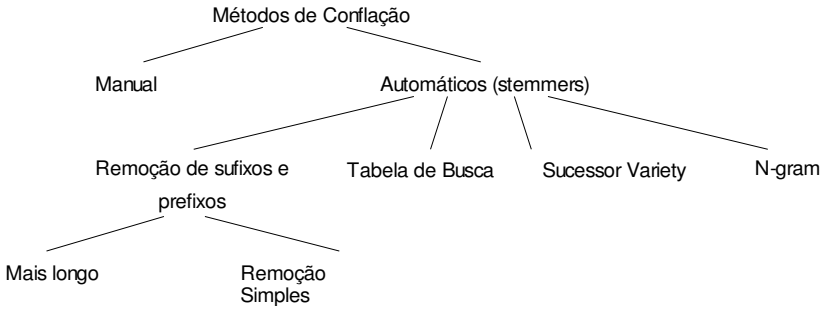
2.1.1.3 Redução Morfológica (Stemming)

O “Stemming” é um método de padronização de palavras amplamente utilizado, desenhado para permitir a redução de termos à menor unidade semântica possível (PAICE, 1990). A definição é que uma palavra típica possui um radical (“stem”) que se refere a alguma ideia central, e que certos prefixos e sufixos são utilizados para modificar o seu significado ou

adequar a palavra a uma determinada regra sintática. O "stemming" promove a remoção destes prefixos e sufixos de modo a reduzir a palavra a sua "essência".

Os algoritmos de Stemming possuem características próprias que os levam a uma separação por classes. Frakes e Baeza-Yates (1992) propuseram uma taxonomia para classificação dos algoritmos de Stemming, exibida na Figura 2.2. Dentro das categorias propostas, os algoritmos de remoção de sufixos e prefixos são os mais encontrados na literatura. Dentre estes, o snowball[3], implementação do Stemming de Porter (1980), é o mais conhecido e utilizado.

Figura 2.2 - Taxonomia para Algoritmos de Stemming



Fonte: Frakes e Baeza-Yates (1992).

O Stemming de Porter é baseado na remoção de sufixos das palavras. Alguns dos destaques é o tamanho reduzido, a rapidez no processamento e o raciocínio simplório. Neste algoritmo, os sufixos são removidos simplesmente para prover uma melhoria na performance dos sistemas de recuperação da informação e mineração de dados, e não como um exercício linguístico. Não são sempre óbvias as circunstâncias em que um sufixo deve ser removido e a avaliação do valor da remoção de um sufixo não é algo trivial.

O método funciona a partir de duas listas, uma com os sufixos e outra com as condições de remoção. Quando a condição é satisfeita, a operação sobre a palavra é realizada. Quando a palavra possui poucas letras, não é realizada a remoção do sufixo, mesmo que exista uma condição que indique a operação. O tamanho da palavra é dado pela contagem das

3 http://www.tartarus.org/snowball

letras, não havendo uma base linguística para esta observação. Sufixos complexos são removidos em diferentes etapas.

2.1.1.4 Normalização de Sinônimos

Por vezes, palavras são resumidas em siglas, sendo que ambas possuem o mesmo significado. Banco Central do Brasil e BACEN, por exemplo, devem ser tratados como uma mesma palavra para efeito de classificação. O tratamento destes casos é realizado com uma lista de palavras de mesmo significado.

2.1.1.5 Atribuição de Pesos

Frequentemente, podemos definir que alguns termos são mais importantes para a definição de um texto do que outros. Dentro deste contexto, a possibilidade de atribuir pesos permitiria que essa percepção pudesse ser aplicada também na atividade processamento automático.

Diversos são os atributos que podem ser definidos para cálculo do peso do termo. Detalharemos alguns destes na Seção 2.2.

2.1.1.6 Redução de Dimensionalidade

O alto tamanho do vetor de termos é um dos problemas mais significantes na classificação de textos. Por conta disso, antes que se inicie o processo de classificação, deve-se tentar reduzir ao máximo o tamanho do vetor sem, no entanto, influenciar negativamente na categorização. Sebastianini (1999) distingue o processo de redução de dimensionalidade em *por seleção de termos* e *por extração de termos*.

A redução de dimensionalidade por seleção de termos é uma técnica que seleciona, a partir do conjunto original T, um sub-conjunto T', de termos que, quando utilizados para classificação do documento, apresente o melhor resultado possível. Uma técnica simples é a da frequência do termo, onde são mantidos apenas os termos que mais aparecem em um determinado documento, sendo removidos os demais.

Já a técnica de redução de dimensionalidade por extração de termos atenta para a geração de um conjunto de termos "artificiais" - T', do original T, sendo que o subconjunto T' está contido em T (T' □ T). O objetivo é reduzir o tempo de processamento sem comprometer o resultado da classificação. Uma das técnicas utilizadas é o agrupamento de termos (Term Clustering), que tenta agrupar as palavras que possuem alguma correlação. Estes grupos de termos, ou um dos representantes deles, poderiam, então, ser utilizados como dimensão dos vetores, ao invés dos termos originais.

2.1.2 Classificação de Textos

O processo de classificação de textos envolve a tentativa de, através de técnicas de aprendizagem de máquina, permitir aos computadores classificar novos documentos em categorias pré-definidas, a partir da inferência sobre um conjunto de documentos previamente classificados por humanos, chamado de grupo de treinamento.

Para Lacerda e Braga (2004) apud Cherkasskg e Mulier (1998), podemos definir um método de aprendizagem como:

> Um algoritmo (usualmente implementado em software) que estima um mapeamento desconhecido (dependência) entre entradas dos sistemas e saídas baseado nos dados disponíveis, isto é, nas amostras (entradas, saídas) conhecidas. Uma vez que a dependência tem sido estimada, ela pode ser usada na predição das saídas futuras do sistema baseado nos valores de entrada conhecidos.

Yang e Liu (1999) e Joachims (1998) realizaram um estudo comparativo de diversas técnicas de Aprendizado de Máquina no contexto da classificação de textos e, dentre as diversas técnicas apresentadas, duas se destacaram positivamente: a técnica do Vizinho mais próximo (k-nearest neighbord) e a de Máquinas de Vetores de Suporte (Support Vector Machines). Oliveira (2004) apresentou estudo sobre classificação com o algoritmo Naive Bayes com índices também positivos. Detalharemos estes três métodos, além dos métodos de Tabelas de Regras de Decisão e do algoritmo de árvore de decisão.

2.1.2.1 Vizinho mais próximo

O método do vizinho mais próximo vem sendo estudado no reconhecimento de padrões há algumas décadas. O raciocínio indutivo deste algoritmo corresponde a admitirmos que a classificação de uma instância "*x*" será muito similar à classificação de uma outra instância que se encontre próxima a esta, levando em consideração a distância euclidiana como conceito de proximidade.

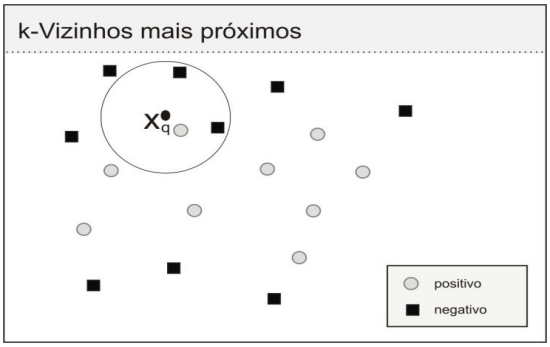
Tomemos como base a instância "*x*", supracitada, descrita pelo vetor de características $V = \{a_1(x), a_2(x), ..., a_n(x)\}$, onde $a_n(x)$ é o valor do enésimo atributo da instância "*x*". Desta forma, a distância entre duas instâncias x_i e x_j é definida no quadro 2.2.

Quadro 2.2 - Distância entre as instâncias xi e xj

$$d(xi, xj) \equiv \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2}$$

Observemos, por exemplo, o espaço bidimensional representado na figura 2.3, que possui um ponto x_q a ser classificado como positivo ou negativo. Tomando como base os três vizinhos mais próximos, podemos claramente classificá-lo como negativo, já que a maioria dos vizinhos selecionados possui tal característica.

Figura 2.3 - Método de classificação vizinho mais próximo



Fonte: Torres (2005)

O método do vizinho mais próximo tem se mostrado um método de inferência indutiva muito efetivo para diversos problemas práticos, sendo bastante robusto quando lhe é provido um conjunto de treinamento suficientemente grande.

Um dos problemas encontrados com este método, no entanto, é o fato de se utilizar todos os atributos da instância no momento da classificação. Em uma instância com 15 atributos onde apenas 3 são realmente relevantes, a distância entre os vizinhos levará em consideração todos os atributos, fazendo com que a classificação possa ser definida através dos 12 atributos que são irrelevantes, o que ficou conhecido como "maldição da dimensionalidade" *(curse of dimensionality)*.

Soluções têm sido propostas como forma de resolver o problema da dimensionalidade. Dentre as encontradas, a utilização de técnicas para eliminação de atributos insignificantes e a utilização de pesos para diferenciar a importância dos termos tem trazido bons resultados.

2.1.2.2 Máquinas de Vetores Suporte (Support Vector Machines - SVM)

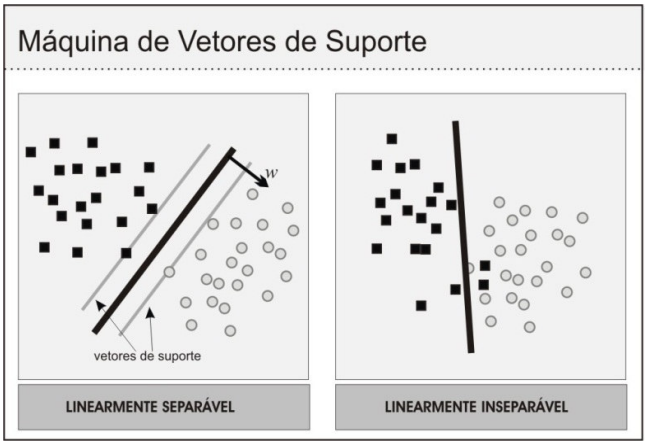
O método em questão foi desenvolvido com base na teoria do aprendizado estatístico. O método das máquinas de vetores de suporte, doravante chamado SVM, é utilizado para o reconhecimento de padrões definidos sobre o espaço vetorial, onde o problema consiste em

encontrar uma superfície de decisão que melhor separe os dados em duas categorias, a classe dos positivos e a dos negativos (MARTINS JÚNIOR, 2003).

Dado um conjunto de treinamento ***S,*** que contém pontos de duas classes, a meta é estabelecer a equação do hiperplano que divide ***S,*** deixando todos os pontos da mesma classe do mesmo lado, enquanto maximiza a distância mínima entre estas classes e o hiperplano.

Nem sempre essa separação dar-se-á de forma direta. Apenas nos casos linearmente separáveis será possível, a partir do conjunto de entrada, traçar um hiperplano que consiga dividir corretamente todos os integrantes de ambas as classes em seus respectivos subconjuntos. Nos casos linearmente inseparáveis, tanto o hiperplano quanto os vetores de suporte são obtidos através da resolução de um problema de otimização. Exemplos de ambos podem ser visualizados na figura 2.4.

Figura 2.4 - Casos típicos encontrados com a aplicação do SVM



Fonte: Torres (2005)

Considerando o caso linearmente separável, podemos definir a superfície de decisão do SVM como: $\vec{w} \cdot \vec{x} - b = 0$, onde $\vec{x}$ é um ponto de dados arbitrário, que representa o conteúdo do documento a ser classificado, e o vetor $\vec{w}$ e a constante *b* devem ser aprendidos do conjunto de treinamento $S_n = (\vec{x}_1, y_1), (\vec{x}_2, y_2), ..., (\vec{x}_n, y_n)$, de tamanho n e com $y_i \in \{-1, +1\}$. Como explanado, o problema do SVM é encontrar w e *b* que satisfaçam as restrições encontradas no quadro 2.3.

Quadro 2.3 - Restrições do SVM

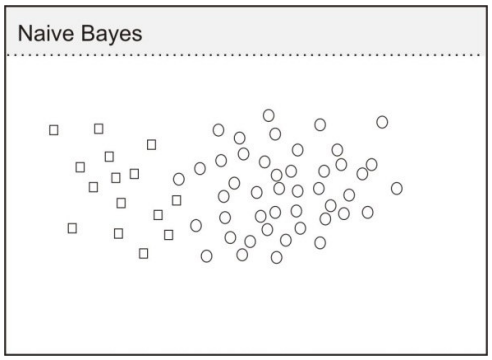
$$\vec{w} \cdot \vec{x}_i - b >= +1 \text{ para } y \in \text{classe } 1$$

$$\vec{w} \cdot \vec{x}_i - b <= -1 \text{ para } y \in \text{classe } -1$$

2.1.2.3 Naive Bayes

O teorema de Bayes, fundamentado na teoria da probabilidade, teve origem com o reverendo Thomas Bayes no século XVIII. De forma a facilitar o entendimento do teorema, discutiremos o exemplo descrito em Statsoft (2009). A figura 2.5. mostra um conjunto de objetos classificados em classes distintas, quadrado e círculo. O objetivo é descobrir a probabilidade que um novo objeto, aleatório, teria de pertencer a uma das classes dadas, baseado no conjunto de objetos existentes.

Figura 2.5 Grupo de Objetos divididos em duas classes

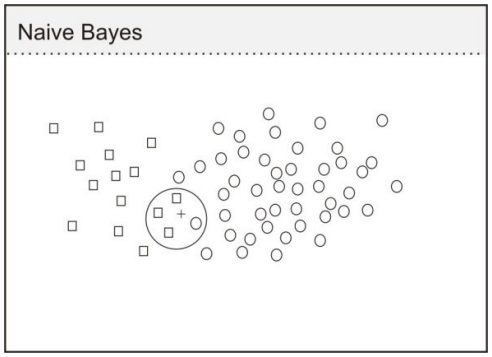


Fonte: Torres (2005).

Uma vez que existem mais círculos do que quadrados, é correto acreditar que um novo caso possui mais chances de se encaixar entre os círculos. Na análise bayesiana, esta certeza é conhecida como probabilidade *a priori,* sendo baseada em experiências prévias, já que o conjunto de treinamento é conhecido. Pode-se, portanto, definir as probabilidades a priori como: *P(Círculo) = Número de círculos / número total de objetos* e *P(Quadrado) = Número de quadrados / número total de objetos*. Sabendo que existem 60 objetos, sendo 45 círculos e 15 quadrados: *P(Círculo) = 45/60* e *P(Quadrado) = 15/60*

Imagine, no entanto, que surja um novo ponto, uma cruz, entre os círculos e os quadrados, e que seja necessário determinar a probabilidade condicional dele pertencer a um dos grupos existentes, conforme a figura 2.6.

Figura 2.6 - Classificação de novos dados utilizando Naive Bayes



Fonte: Torres (2005).

Supondo que os objetos estão agrupados, pode-se assumir que quanto mais objetos de determinada classe estejam próximos ao novo ponto, maiores as chances deste ponto estar associado a esta classe. Para medir esta probabilidade (do inglês *likelihood*), deve-se traçar uma circunferência ao redor da cruz, isolando internamente aqueles pontos que estejam mais próximos. A partir da contagem destes pontos vizinhos, por classe a que pertencem, pode-se obter a probabilidade que a cruz tem de ser círculo ou quadrado: *P(cruz, círculo) = Número de círculos vizinhos a cruz / total de círculos* e *P(cruz, quadrado) = número de quadrados vizinhos a cruz / total de quadrados*. Observando que existem dentro desta circunferência um círculo e três quadrados, tem-se: *P(cruz, círculo)= 1/45* e *P(cruz, quadrado)=3/15*.

O teorema de Bayes trabalha, justamente, com a associação entre essas duas probabilidades. Na análise bayesiana, a probabilidade final vem da combinação de ambas as fontes de informação, formando a probabilidade *a posteriori*: *P(quadrado,cruz) = P(cruz, quadrado) * P(quadrado),* ***P(quadrado,cruz)*** *= 3/15 * 1/4 =* ***1/20****, e P(círculo, cruz) = P(cruz, círculo) * P(círculo),* ***P(círculo, cruz)*** *= 1/45 * 3/4=* ***1/60****.*

Desta forma, pode-se, finalmente, definir que a cruz deveria ser definida como quadrado, já que a probabilidade a posteriori do quadrado foi superior à do círculo.

O teorema de Bayes vem sendo utilizado como suporte ao processo de categorização de textos. No quadro 2.4, encontra-se a fórmula utilizada para estimar a probabilidade de um documento pertencer a uma classe (OLIVEIRA, 2004).

Quadro 2.4 - Fórmula de Bayes aplicada à classificação de documentos

$$P(C = c_i \mid \vec{x}) = \frac{P(\vec{x} \mid C = c_i) \times P(C = c_i)}{P(\vec{x})}$$, **onde**:

$\vec{x}$ representa um vetor de termos e c_i representa uma classe..

2.1.2.4 Tabela de Regras de Decisão

Tanto as tabelas quanto as árvores de decisão não tem sua origem na área de sistemas de computação. Eles tem uma aplicação muito mais abrangente, englobando áreas como a Biologia e Teoria da Informação, além das Ciências da Computação.

Uma tabela de decisão é um modelo alternativo para um problema composto por condições, ações e regras. Condições são variáveis cujos valores são avaliados para a tomada de decisão. Ações representam o conjunto de operações a serem executadas condicionadas pelas respostas das condições. Regras representam o conjunto de situações, que são verificadas em resposta às condições (TCHEMRA, 2009).

Uma tabela de decisão é composta de linhas e colunas separadas em quatro quadrantes distintos, exemplificados no quadro 2.5. No quadrante superior esquerdo, cada linha corresponde à variável do problema que será analisada no processo de decisão – as condições; o superior direito corresponde aos valores das variáveis, das condições do primeiro quadrante; o inferior esquerdo indica o conjunto de procedimentos ou operações que serão executados de acordo com as respostas das condições; e o inferior direito contém as ações tomadas em razão da avaliação das condições.

Quadro 2.5 - Quadrantes da tabela de decisão

Condições	Valores para as condições
Ações	Ações a serem tomadas

Para criar uma tabela de decisão é necessário definir as condições que podem afetar uma decisão e determinar as possíveis ações que podem ser tomadas. Em seguida, deve-se

determinar o número de valores possíveis para cada condição para que seja preenchido quando cada ação for tomada.

Tomemos como exemplo uma mercearia que deseja criar uma tabela de decisão sobre as ações a serem tomadas quando for efetuada uma compra, dependendo do valor da compra e da forma de pagamento. As condições são:

(1) Compras abaixo de R$ 100
(2) Pagamentos com cheque
(3) Pagamentos com cartão de crédito
(4) Pagamentos em dinheiro

As ações possíveis de serem tomadas para as condições são:

(1) Efetuar a venda
(2) Consultar o cheque
(3) Consultar o cartão de crédito

Usando os valores acima para construir uma tabela de decisão, deve-se definir os valores possíveis para cada condição e calcular as possibilidades. Para cada uma das condições, há apenas dois valores possíveis: Sim e Não. Como são quatro condições, com dois valores possíveis, chega-se a soma de oito possibilidades, visualizadas na Tabela 2.1.

Tabela 2.1. Condições e Ações da tabela de decisão modelo (S = Atende o critério e N = Não atende)

Compras acima de R$ 100,00	S	S	S	S	N	N	N	N
Pagamentos com cheque	S	N	N	S	S	N	N	S
Pagamentos com cartão de crédito	N	S	N	N	N	S	N	N
Pagamentos em dinheiro	N	N	S	S	N	N	S	S
Efetuar a venda			X		X	X	X	
Consultar o cheque	X			X				X
Consultar o Cartão de Crédito		X						

A partir da análise dos valores atribuídos às condições, são definidas as ações a serem tomadas.

2.1.2.5 Árvore de Decisão

Uma árvore de decisão é uma representação de uma tabela de decisão sob a forma de uma árvore. Tem a mesma utilidade da tabela, constituindo-se como uma maneira alternativa de expressar as mesmas regras que são obtidas quando se constrói a tabela (ÁRVORE DE DECISÃO, 2009).

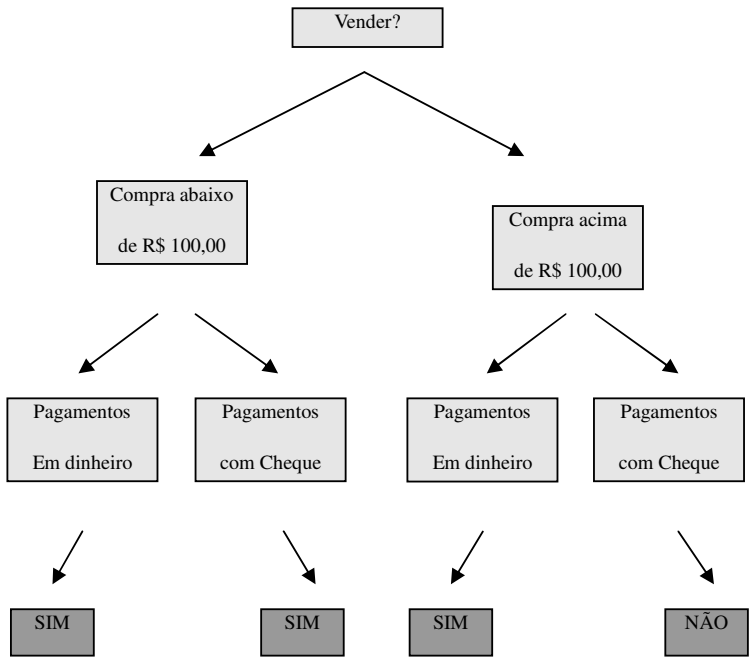
Em sua representação gráfica, a decisão (sim ou não) é representada por linhas e as questões sobre as quais se deve decidir são representadas por nós. Cada um dos ramos compostos por linhas e nós sempre termina em uma folha, indicando a consequência mais provável da sequência de decisões tomadas. Como exemplo, imagine que devemos decidir se uma venda pode ou não ser concretizada, dependendo do valor da compra e da forma de pagamento. As condições e ações a serem tomadas estão listadas na Tabela 2.2.

Tabela 2.2 - Condições e Ações da tabela de decisão modelo

Valor Compra	Pagamento Dinheiro	Pagamento Cartão	Vende
Maior que R$ 100,00	S	N	S
Maior que R$ 100,00	N	S	N
Menor que R$ 100,00	S	N	S
Menor que R$ 100,00	N	S	S

A árvore de decisão correspondente pode ser visualizada na Figura 2.7.

Figura 2.7 - Exemplo de Árvore de Decisão



As árvores de decisão, por sua representação gráfica, tornam mais fácil a visualização do problema e dos fluxos de decisão em comparação com a tabela de decisão. Este fato é válido quando consideramos um número pequeno de condições e ações; para conjuntos grandes, a árvore torna-se muito complexa e de difícil entendimento.

2.1.3 Desempenho da Classificação

Quando se avalia o desempenho de um método de classificação, são abordados, principalmente, quatro itens (OLIVEIRA, 2004):

a) Número de documentos rotulados a uma categoria c_i e corretamente classificados como pertencentes a esta categoria;
b) Número de documentos rotulados como não pertencentes a uma categoria c_i, porém incorretamente classificados como pertencentes a esta categoria – os falso-positivos;
c) Número de documentos rotulados como pertencentes a uma categoria c_i, porém incorretamente classificados como não pertencentes a esta categoria – os falso-negativos.

Baseadas nestas definições, foram determinadas importantes medidas de desempenho para mensurar a performance dos classificadores, sendo as principais:

a) **Abrangência** (*recall*) – Mede o percentual de documentos corretamente classificados na categoria c_i, considerando também os que deveriam ter sido classificados nesta categoria e não foram – os falso-negativos. (Quadro 2.6)

Quadro 2.6 - Fórmula para cálculo do abrangência

$$recall = \frac{a}{(a + c)}$$

b) **Precisão (precision) – Mede o percentual de documentos classificados corretamente em uma categoria ci , considerando também os que não deveriam ter sido classificados nesta categoria e o foram – os falso-positivos (Quadro 2.7).**

Quadro 2.7 - Fórmula para cálculo da precisão

$$precision = \frac{a}{(a + b)}$$

2.2 CLASSIFICAÇÃO DE SENTENÇAS DO TEXTO CIENTÍFICO

Como visto nas seções anteriores, um dos objetivos da mineração de texto é analisar um documento com o intuito de classificá-lo em uma determinada categoria. Neste estudo, entretanto, não é o texto como um todo que será analisado, mas segmentado em unidades menores, denominadas sentenças -que são representadas por um pedaço de texto com sentido completo. O principal motivo de utilizarmos este subconjunto do texto é o fato de que resultados de estudos são descritos em artigos científicos neste formato textual.

Estas sentenças serão representadas por vetores de termos, que são matrizes das palavras que as compõem associadas a um determinado atributo. O principal desafio é que os atributos selecionados para compor estes vetores realmente auxiliem na diferenciação dos tipos de sentenças, de forma a melhorar a precisão do classificador.

Diversos estudos propondo uma série de atributos para caracterização de sentenças vem sendo realizados nos últimos 50 anos, geralmente vinculados à atividade de extração automática de sentenças para criação de sumários de textos. As metodologias para o cálculo do valor destes atributos baseiam-se nos mais diversos paradigmas: estatístico, como a frequência das palavras-chave de Luhn (1958); semântico, como o método relacional de Skorochod'ko (1972); ou a análise do discurso, como o modelo proposto por Marcu (1998). Abordaremos, doravante, alguns atributos de sentença que são calculados com base nas estatísticas do texto.

2.2.1 Freqüência das Palavras-chave (*frequency keywords*)

Esta métrica foi proposta por Luhn (1958) com o intuito de medir o grau de importância de uma palavra na caracterização da sentença a que pertence, a partir da frequência em que aparece no texto. A hipótese é que um escritor normalmente repete certas palavras à medida que ele avança em sua argumentação ou quando redige um aspecto sobre determinado assunto. Ainda segundo o autor, é muito pouco provável que, em textos técnicos, uma

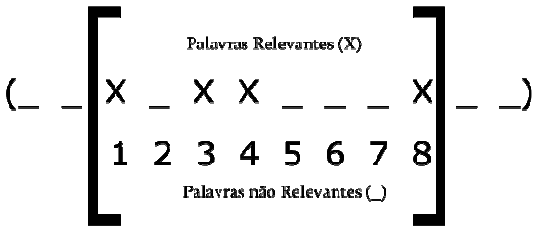
palavra seja utilizada com diferentes significados ou que se utilize diferentes palavras para representar um mesmo conceito. Se por questões estéticas o autor do estudo faz um esforço significativo em busca de sinônimos, como há quantidade limitada de palavras com sentido correlato, chega um momento em que ele fica sem alternativas e acaba por repetir as palavras-chave.

É importante observar, entretanto, que a proposição "toda palavra com alta frequência é uma palavra-chave" não é verdadeira. Palavras muito comuns, geralmente, possuem baixa importância, representando uma espécie de ruído do sistema. Palavras como artigos, preposições ou conectivos são encontradas em grande quantidade nos mais diversos textos, independente do tema que tratem, e não são importantes para a classificação destes documentos. Estas palavras, comumente conhecidas como "*stop-words*", foram abordadas na seção 2.1.1.2.

O cálculo da frequência das palavras-chave para uma sentença é derivado da análise das palavras que a compõe. São dois aspectos a serem analisados: o cálculo da frequência das palavras que compõem a sentença; e o cálculo do fator que representa a distância entre as palavras-chave. O cálculo da frequência é derivado da soma da frequência das palavras-chave encontradas na sentença (algumas das palavras mais frequentes são encontradas em conjunto umas com as outras dentro de uma sentença). Já em relação ao fator da distância entre as palavras-chave, quanto maior o número de palavras com alta frequência ocorrendo fisicamente próximas uma das outras, mais representativa é esta informação para o artigo.

Na Figura 2.8, os traços (*underline*) representam as palavras não relevantes e o "X" representa as palavras relevantes de uma sentença. O primeiro passo para o cálculo da distância entre as palavras é diminuir o tamanho da sentença para a posição das palavras-chave (relevantes) mais periféricas. A sentença reduzida é representada na figura pelas palavras no interior dos colchetes.

Figura 2.8 - Cálculo da distância das palavras de uma sentença



Neste ponto, o fator de importância é calculado elevando-se ao quadrado a quantidade de palavras relevantes (4^2). Em seguida, deve-se dividir este valor pelo número total de palavras dentro dos colchetes (8), obtendo como resultado o número 2 (16/8).

2.2.2 Método das Palavras Sinalizadoras (*cue method*)

O cálculo deste atributo baseia-se na hipótese de que a provável relevância de uma sentença é afetada pela presença de determinadas palavras, consideradas relevantes no domínio do texto (*cue words*). O método utiliza um dicionário de *cue words,* com palavras previamente selecionadas e divididas em três grupos distintos: as palavras bônus (*bonus words*), que representam as que são positivamente relevantes; as palavras stigma (*stigma words*) – que representam as negativamente relevantes; e as palavras nulas (*null words*), que são irrelevantes e devem ser ignoradas na contagem da frequência. Pesos negativos são atribuídos a cada palavra das sentenças que conste entre as palavras stigma, enquanto pesos positivos são atribuídos a cada palavra que conste na lista de palavras bônus. As palavras nulas são ignoradas pelo processamento. O peso final da sentença é formado pela média ponderada entre valores negativos e positivos de suas palavras constituintes.

Os dicionários são dependentes do contexto em que se encaixam as sentenças a serem analisadas. Em textos científicos, palavras como "método" ou "resultado" serão, muito provavelmente, altamente significativas, devendo constar no dicionário de palavras bônus. Se utilizássemos o domínio notícias de jornal, por exemplo, estas palavras não seriam relevantes.

O método permite variações, como a utilização de frases ao invés de palavras. Quando utilizadas frases, estas são tipicamente curtas e obtidas de vocabulário de documentos relacionados a determinado contexto (HUI, 2002).

Teufel e Moens (1997) definiram, a partir da observação de sentenças chave em um determinado *corpus*, um dicionário de palavras com diferentes pesos. A classe de bônus foi subdividida em três diferentes grupos, com pesos distintos, enquanto as classes de palavras nulas e de palavras stigma foram mantidas em apenas um grupo. Parte da lista definida pode ser visualizada na Tabela 2.3.

Tabela 2.3 - Dicionário de Palavras

Dicionário				
Palavras Stigma peso -1	**Palavras nulas**	**Palavras Bônus**		
		frases menos significantes – peso 1	*frases de valor intermediário – peso 2*	*frases mais significantes – peso 3*

for example
example
comments on an ...
comments on earlier drafts funded by
thankful
...
	// verbs
	am ...
	is
	…
	// articles
	the
	these
	those ...
		a central problem
		a comparison was made
		a growing need
		a limiting factor of
		…
		would be interesting
		little attention
		no current
			above show
			above shows
			elegant solution
			experimental results
			…
			what interests me here
			what interests us here
			...
				concluding
				studies to evaluate
				I have shown
				I showed that
				my goal was to
				…
				we report on our work on
				we showed

2.2.3 Método da Localização (*location method*)

O cálculo deste atributo, proposto inicialmente por Baxendale (1958), baseia-se na ideia de que: (1) sentenças que ocorrem abaixo de certos títulos são positivamente relevantes; (2) sentenças relevantes tendem a acontecer muito cedo ou muito tarde em um documento e em seus parágrafos. A primeira e a última sentença de um parágrafo ou do texto podem, por conseguinte, conter suas ideias principais, sendo selecionadas como sentenças chave do texto.

Em seu experimento, Baxendale observou que, em uma amostra de 200 parágrafos, a sentença principal encontrava-se na primeira posição em 85% dos casos, e na última posição do parágrafo em 7% dos casos. Nos 8% restantes, as informações relevantes encontravam-se entre a primeira e a última posição de um mesmo parágrafo. Apesar do valor pouco significativo, o autor considerou a inclusão da última sentença do parágrafo para manter a coesão textual, pois esta representava elo de ligação com a primeira sentença do parágrafo posterior (MARTINS, 2001).

Na operacionalização do método, cada sentença recebe um peso de acordo com a sua presença no início ou no fim do parágrafo, no início ou fim do documento, ou imediatamente abaixo de um título de seção. Em relação ao posicionamento das sentenças nos parágrafos e dos parágrafos no texto, foram definidos quatro valores, conforme Quadro 2.8.

Quadro 2.8 - Definição de peso pelo método da localização

(1)primeiro parágrafo – ordinal peso O1;

(2)último parágrafo – ordinal peso O2;

(3)primeira sentença do parágrafo – ordinal peso O3;

(4)*última sentença do parágrafo – ordinal peso O4;*

Por fim, a peso da sentença é incrementado se esta aparece abaixo de certos títulos ou subtítulos de seção como, por exemplo, “resultados” ou “conclusão”.

Segundo Inderjeet (1999), este método é dependente de contexto. A estratégia de se obter o primeiro parágrafo, por exemplo, funciona apenas em documentos do gênero notícias e não necessariamente em todos os outros. Nenhuma técnica automática foi elaborada para definir o posicionamento ótimo das sentenças em cada um dos domínios.

2.2.4 Método das Palavras-chave no Título (title-keyword method)

A base do cálculo deste atributo, proposto por Edmundson (1969), é o fato de que as palavras constantes nos títulos e subtítulos do documento e de suas seções, retiradas as “*stop words”*, são as que melhor representam o tópico principal do texto.

A operacionalização é similar ao método das palavras-chave, sendo a principal mudança o fato das palavras não mais serem selecionadas pela sua frequência no texto, mas por estarem presentes ou não nos títulos e subtítulos do texto e de suas seções.

2.2.5 Método da Palavra Temática (ThematicWord Feature)

Este método tenta identificar as palavras-chave do texto – denominadas palavras temáticas. Na teoria, sentenças contendo estas palavras temáticas conseguiriam caracterizar o documento.

Para cada palavra é calculado um peso a partir de uma função específica. Ao final, as 10 palavras com maior peso são escolhidas como temáticas. O cálculo do peso da sentença é efetivado a partir da quantidade de palavras temáticas encontradas na mesma.

Diversas são as funções utilizadas para cálculo do peso das palavras. Teufel e Moens (1997) estabeleceram o produto da frequência do termo (TF) pela frequência inversa do documento (IDF) como função de cálculo do valor deste atributo (Quadro 2.9).

Quadro 2.9 - Fórmula para Cálculo do TF-IDF

$$score(w) = f_{loc} \cdot \log\left(\frac{100 \cdot N}{f_{glob}}\right)$$

f_{loc}: *frequência da palavra w no documento*
f_{glob}: *número de documentos que possuem a palavra w*
N : *número de documentos na coleção*

Já Pardo (2003), definiu como função de cálculo a frequência inversa da sentença - TF-ISF (LAROCCA NETO, 2000). Para o cálculo da medida TF-ISF de uma dada palavra (Quadro 2.10), são utilizadas a medida da frequência da palavra na sentença – F(w), o número de palavras na sentença a que a palavra pertence – n, e o número de sentenças em que a palavra aparece – S(w).

Quadro 2.10 - Fórmula do Cálculo do TF-ISF

$$TF-ISF(w) = F(w) \cdot \frac{\log(n)}{S(w)}$$

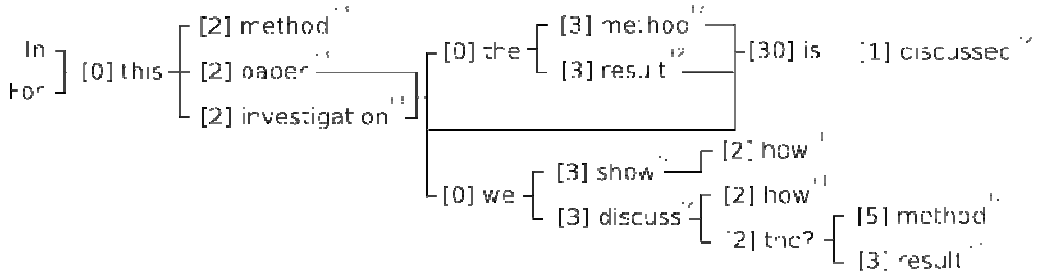
2.2.6 Frase Auto-indicativa

Alguns documentos, especialmente os discursivos, podem conter estruturas que ocorrem corriqueiramente e que, explicitamente, indicam que algo importante existe nas sentenças que a contém. As frases "*The purpose of this article is...*" e "*The results of this article show that...*" são exemplos destas estruturas, denominadas frases auto-indicativas.

Paice (1990) definiu um método de seleção destas frases auto-indicativas (*self-indicating phrases*), com base na definição de sete estruturas básicas (*templates*). Os *templates* foram identificados pelas letras A, B, C, D, E, F e G e compostos por uma sequência esperada de palavras, os "paradigmas". A partir destas estruturas são atribuídos pesos à medida que sentenças do texto se enquadravam nestes modelos pré-definidos. A busca por segmentos no texto que se adequem a uma destas estruturas forma a base para o processo de identificação das frases auto-indicativas.

A identificação das frases é um processo de busca de padrões no texto. A Figura 2.9 exibe um modelo de estrutura para indicadores do grupo B. As palavras encontradas na figura representam os paradigmas, o símbolo "?" denota que este paradigma é opcional e os pesos de cada um são mostrados em números sobrescritos (por exemplo, o valor - $^{+2}$).

Figura 2.9 - Estruturas para indicadores do grupo B (simplificado)



Fonte: Paice (1990)

Cada palavra é representante de um grupo de palavras que podem aparecer em determinado ponto da estrutura. As estruturas são utilizadas com um conjunto de grupos de palavras, sendo os paradigmas os membros chave dos grupos, conforme pode-se visualizar na Tabela 2.4.

Tabela 2.4 - Lista de paradigmas e grupos de palavra-chave

Paradigma	Grupo de Palavras-chave
Paper	Paper, article, report, review, analys-, discussion, summar-, outline, presentation, essay, stud-, survey.
Method	Method, means, model, technique, approach, process, program-, procedure, system, measure, formula, function, idea (-2)
Result	Result, outcome, finding, conclusion
Show	Show, shew, reveal, demonstrate, confirm, prove, indicate, explain, illustrate, illuminate, clarify, make clear, made clear, determine, discover, find, found, suggest (-2), propose (-2), argue (-2), see (-2).
We	We, I
How	How, why, what, whether, that.
The	Them, a, an, some, our (+2).

Fonte: Paice (1990).

Dois principais problemas são apontados neste método. O primeiro é que, apesar de ser bastante útil na maioria dos documentos, existe a possibilidade de não se encontrar nenhuma frase auto-indicativa em determinados textos. O segundo é que, com alguma frequência, frases que deveriam ser definidas como auto-indicativas podem conter em sua

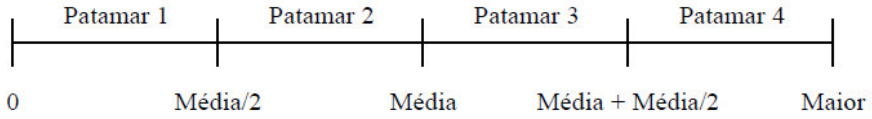
constituição palavras que não fazem parte da lista de indicadores, o que acaba por desconfigurar a instanciação de determinado padrão de estrutura pré-definido, dificultando a identificação.

2.2.7 Tamanho da Sentença (Sentence Lenght)

O Cálculo deste atributo parte da hipótese de que as sentenças mais longas, normalmente, apresentam maior conteúdo informativo sendo, portanto, mais relevantes (KUPIEC et al., 1995). Às sentenças são atribuídos pesos proporcionais ao tamanho, sendo definido um valor prévio de corte (5 palavras, por exemplo).

Na abordagem conexionista de Pardo (2003), as sentenças são enquadradas em uma escala com 4 patamares (Figura 2.10), onde são considerados o tamanho médio das sentenças do texto e o tamanho da maior sentença do texto.

Figura 2.10 - Patamares de tamanho das sentenças



2.2.8 Presença de Palavras da "Gist Sentence" na Sentença

O cálculo do peso proposto por Pardo (2002) se utiliza dos métodos de frequência das palavras (TCHEMRA, 2009) e do TS-ISF (LAROCCA NETO,. 2000) para definir a sentença com maior pontuação do texto, a qual, ao menos teoricamente, representaria a ideia principal do texto. Esta sentença é denominada *sentença-gist* do texto. A partir deste ponto são selecionadas todas as sentenças do texto que:

(1) possuam pelo menos uma das formas canônicas[4] que compõe a *sentença-gist*;

(2) possuam uma pontuação maior do que a média da pontuação de todas as sentenças do texto.

4 Para alcançar a forma canônica de uma palavra, basta reduzir os verbos para o infinitivo e os substantivos e adjetivos para o masculino singular

2.2.9 Método da Inicial Maiúscula (*UppercaseWord Feature*)

Segundo Kupiec (1995), nomes próprios são frequentemente importantes para definição das sentenças relevantes. O passo inicial é a criação de uma lista de palavras iniciadas com letra maiúscula que não iniciem sentenças, que ocorram diversas vezes e que não sejam unidades de medida ou abreviaturas. O cálculo dos pesos das sentenças é feito com base na quantidade de palavras constituintes que se encontram na lista de palavras previamente definida, sendo que as sentenças em que uma palavra da lista foi encontrada pela primeira vez tem peso dobrado em relação às outras sentenças que possuem esta mesma palavra.

2.2.10 Métodos da Linha Base (*baseline*) e da Média da Conectividade Léxica (*average lexical connectivity*)

O método da linha base e da média da conectividade léxica foram propostos por Lin (1999). O primeiro estabelece que o peso de cada sentença é baseado na sua posição no texto como um todo, sendo que a primeira sentença possui o maior valor e a última o menor valor.

O segundo método pressupõe que a sentença que compartilha muitos termos com outras sentenças é importante e, por conseguinte, define o peso da sentença com base no número de termos compartilhados entre esta e as demais, dividido pelo número total de sentenças do texto.

2.3 CONSIDERAÇÕES FINAIS

Neste capítulo apresentamos um resumo das técnicas de mineração de texto que serão aplicadas no método de sumarização de texto desenvolvido neste estudo, desde a etapa de pré-processamento do documento textual até a descrição do funcionamento dos algoritmos de classificação de texto.

Apresentamos, ainda, os indicadores que serão utilizados para medir o desempenho do algoritmo desenvolvido – abrangência e precisão, e os principais atributos de caracterização de sentença encontrados na literatura.

Capítulo 3

Análise do Artigo Científico

Este será analisada a estrutura do artigo científico com base em dois aspectos: a forma de organização do conteúdo em seções do texto; e a organização em relação ao seu modelo estilístico, que define a aparência do documento e formatação visual do texto.

De uma forma geral, podemos caracterizar estruturalmente os textos científicos sob dois aspectos: o da organização hierárquica do conteúdo, que divide o texto em seções lógicas e suas subseções, e o das normas estilísticas de escrita, que se preocupam com a aparência do documento e utilizam conceitos como margens da página, espaçamentos de linhas, tipos e formatação de letras.

Por padrão, os artigos científicos publicados em congressos, independente da área de conhecimento, seguem um modelo de submissão onde estão descritas as normas que devem ser observadas pelos autores na escrita do documento. O que se percebe, entretanto, é que a maioria dos congressos e seminários enfocam demasiadamente o aspecto estilístico do texto, em detrimento da organização hierárquica do conteúdo.

Grande parte dos autores de artigos cientificos ainda não atentaram para os benefícios advindos da utilização de um padrão para escrita do conteúdo de artigos cientificos. A padronização de seções e subseções facilita a localização da informação no documento pelos seres humanos e incrementa a performance dos algoritmos de extração da informação, tornando mais eficiente o processamento das informações pelos computadores. Com o crescimento exponencial da produção científica, a organização do conteúdo se apresenta como um auxílio considerável aos pesquisadores na tarefa de ler o conjunto de artigos publicados em sua área de interesse (DE WAARD, 2007).

3.1 ANÁLISE DAS NORMAS ESTILÍSTICAS DE ESCRITA

Como visto na seção anterior, de uma forma geral, a principal preocupação das chamadas para publicação de estudos em eventos, sobretudo em Computação, é a definição estética, o visual

do artigo. Para corroborar esta afirmação, analisamos alguns dos principais modelos utilizados na chamada de trabalhos de eventos na área, os quais:

(1)IEEE Manuscript Templates for Conference Proceedings[5];

(2)Modelo SBC[6];

(3)SMC 2009 – IEEE Xplore[7];

(4)SPRINGER[8];

(5)TRANSMAG[9];

(6)Official ACM Proceedings Format[10]

A análise do primeiro modelo, "IEEE Manuscript Templates for Conference Proceedings", exibe a diretriz adotada pelos editores, que priorizam quase que exclusivamente detalhes de formatação do texto. São encontradas determinações de como devem ser escritos os títulos ou de como deve ser estruturado o corpo do artigo, em uma ou duas colunas. O modelo da SBC segue a mesma direção do anterior e trata com primazia itens como a definição do tamanho e do tipo da letra a ser utilizado e o formato do papel. Os modelos SMC 2009 e Transmag da IEEE, o modelo de publicação da Springer e da ACM seguem as mesmas diretrizes dos outros analisados.

Nestes modelos, pouco se observa em relação a organização hierárquica do conteúdo. Neste sentido são encontradas apenas restrições superficiais relativas à parte inicial dos documentos, como a definição do título, subtítulo, autores, resumo e palavras chaves, que estão presentes em praticamente todos os modelos analisados. Definição das seções do corpo do texto, entretanto, não são contempladas.

A estrutura dos modelos analisados, com a identificação dos itens definidos e a

---

5 http://ieeec.com/portal/cms_docs_iportals/iportals/publications/pubservices/confpub/auxfiles/MSW_A4_format.doc – acessado em 18/6/2009 às 00:30

6 http://www.sbc.org.br/index.php?language=1&content=downloads&id=286 – acessado em 18/6/2009 às 00:30

7 http://www.ieee.org/publications_standards/publications/authors/authors_journals.html – acessado em 18/6/2009 às 00:30

8 http://www.springer.com/cda/content/document/cda_downloaddocument/word.zip?SGWID=0-0-45-72919-0 – acessado em 18/6/2009 às 00:30

9 http://www.ieee.org/portal/cms_docs_iportals/iportals/publications/journmag/transactions/Transmag.doc - – acessado em 18/6/2009 às 00:30

10 http://www.acm.org/sigs/publications/pubform.doc - – acessado em 18/6/2009 às 00:30

sequência em que aparecem em cada documento, é exibida na Tabela 3.1.

Tabela 3.1 - Análise de modelos para publicação de artigos

	IEEE	SBC	SMC2009	Springer	Transmag	ACM
Cabeçalho			-	-	1	-
Título	1	1	1	1	2	1
Sub-título	2	-	2	2	-	-
Autores	-	2	-	3	3	-
Afiliação	-	3	-	4	4	-
Autor + Afiliação	3	-	3	-	-	2
Resumo	4	4	4	5	5	3
Palavras-chave	5	-	5	6	6	6
Descrição da Categoria e Assunto	-	-	-	-	-	4
Termos Gerais	-	-	-	-	-	5
Rodapé	-	-	6	-	-	-

3.2 ORGANIZAÇÃO DO CONTEÚDO EM ARTIGOS CIENTÍFICOS

Enquanto os artigos em Computação padronizam poucos itens de organização do conteúdo do artigo, outras áreas de conhecimento avançam para um modelo onde seções do corpo do documento são definidas, a exemplo dos estudos em saúde. Estes trabalhos costumam apresentar seu conteúdo estruturado em quatro seções, baseadas em um modelo denominado IMRaD: Introdução, Método, Resultados e Discussão. Este modelo segue o critério estabelecido pelo Comitê Internacional dos Editores de Jornais Médicos (ICME, 2009), que afirma:

> O texto de artigos observacionais e experimentais é geralmente (mas não necessariamente) dividido em seções com os títulos Introdução, Métodos, Resultados e Discussão. A chamada "estrutura IMRAD" não é simplesmente um formato de publicação arbitrário, mas sim uma reflexão direta do processo de descoberta científica. Artigos longos podem necessitar de subtítulos dentro de algumas seções (especialmente Resultados e Discussão) para esclarecer seu conteúdo. Outros tipos de artigos, como relatos de casos, revisões e editoriais, provavelmente precisarão ser formatados de modo diferente.

O experimento desenvolvido por Teufel e Moens (2002) evidenciou a diferença de abordagem entre estas duas áreas. O resultado exposto na tabela 3.2 permite observar o quanto a estrutura do conteúdo está estabelecida na área de Cardiologia (saúde), onde quase a totalidade dos artigos segue uma padronização de divisão em seções, em contraposição à área de Lingüística Computacional, onde a preocupação com a organização do conteúdo não é tão presente.

Tabela 3.2 - Análise estatística da estrutura de artigos científicos

Linguística Computacional (80 artigos)			Cardiologia (103 artigos)		
Seção	**Frequência**		**Seção**	**Frequência**	
Introdução	63	79,00%	Introdução	103	100,00%
Conclusão	47	59,00%	Resultado	97	94,00%
Discussão	13	16,00%	Discussão	97	94,00%
Resultado	16	20,00%	Método	95	92,00%
Método	-	-	Conclusão	44	42,00%

Fonte: Teufel e Moens (2002).

Para confirmar a observação de Teufel e Moens (2002), selecionamos aleatoriamente, em bases de dados da internet, um conjunto de 33 artigos da área de computação, redigidos por 82 diferentes autores. Utilizando o modelo IMRaD como base de observação, verificamos que nenhum dos artigos selecionados possuía uma seção descrevendo o método. Já a seção de introdução estavam presentes em quase todos os artigos (94%) (Tabela 3.3).

Tabela 3.3 - Análise da estrutura de artigos em Ciência da Computação

Área: Ciência da Computação (33 artigos)	
Seção	**Seção**
Introdução	94,00%
Conclusão/Discussão	91,00%
Resultado	15,00%
Método	0,00%

3.3 POR QUE PADRONIZAR AS SEÇÕES DO CORPO DO ARTIGO

A principal vantagem de se utilizar uma estrutura padrão de seções para os artigos é a tendência que esta abordagem tem de agregar sentenças de propósito semelhante em uma mesma região do artigo. De Waard (2007) elencou características das sentenças presentes em cada uma das seções padrão do texto cientifico, que aqui são exibidas na Tabela 3.4.

Tabela 3.4 - Descritivo de Seções do artigo cientifico

Seção	Sentenças que ..
Introdução	Posicionem o presente tópico de pesquisa Posicionem a questão central da pesquisa (Provejam uma linha de como seguir)
Método	Descrevam os métodos usados
Resultados	Forneçam provas para principal argumentação (Interpretem as implicações do trabalho)
Discussão	Avaliem as alegações a luz dos resultados levantados e trabalhos relacionados (Sumarizem o trabalho Atual) (Discorram sobre os próximos passos)

É importante salientar que o fato de se utilizar as quatro seções exibidas na Tabela 3.4 como padrão de organização do artigo não implica, por exemplo, que todos os resultados do estudo estejam localizados dentro da seção resultados. O que se pretende é aglomerar a maior parte dos resultados nesta seção. Na Tabela 3.5, que tem como fonte experimento realizado por De Waard (2007), podemos observar que das 284 sentenças que indicavam resultado, aproximadamente 81% foram localizadas na seção Resultados.

Tabela 3.5 - Número de segmentos (linhas) em cada seção (colunas)

	Introduction	**Method**	**Results**	**Discussion**	**Total**
Fact	63	0	104	37	204
Problem	20	0	10	15	45
Goal	2	0	72	6	90
Method	2	All	129	6	137
Result	10	0	230	44	284
Implication	14	0	100	36	150
Hypothesis	10	-	33	26	69
Total	121	0	678	170	979

Fonte: De Waard (2007).

O experimento de Cruzes (2007), assim como o de De Waard (2007), identificou agrupamento de sentenças indicando resultado na seção de mesmo nome, conforme exibido na Tabela 3.5.

Tabela 3.6 - Distribuição dos resultados encontrados em artigos por sessão

Seção	**Resultados do Artigo**	**Resultados de Outros Artigos**
Discussão e Conclusão	32,93%	5,56%
Resultados Experimentais	56,10%	15,28%
Introdução		20,83%
Método	9,15%	8,33%
Trabalhos Relacionados		50%

Fonte: Cruzes (2007).

O agrupamento de sentenças de propósitos correlatos, em seções que possam ser identificadas automaticamente no texto, atuaria positivamente no processamento computacional destas informações.

3.4 ANOTAÇÃO DE ARTIGOS CIENTÍFICOS

Apesar da imensa maioria dos textos científicos já escritos não possuir qualquer anotação semântica do seu conteúdo, diversos pesquisadores vem desenvolvendo técnicas de anotação manual, com o intuito de prover processamento automático das informações, de forma transparente ao ser humano e com índice de erro próximo a zero.

Marcondes (2005) propõe a criação e utilização de uma ontologia para os artigos científicos, de forma que as sentenças fossem instanciadas nesta estrutura, tornando transparente o processamento do texto e incrementando as operações voltadas à extração da informação. Há ainda pesquisadores que defendem o uso da linguagem XML, a maioria para representar características específicas de modalidades científicas, como Murray-Rust (1999), que propõe a Linguagem de Marcação Química (CML – Chemical Markup Language) e, mais recentemente, a Linguagem de Marcação Científica Técnica e Média (STMML - Scientific Technical and Medical Markup Language) (MURRAY-RUST, 2002). Há ainda o trabalho de Hucka (2009), que propõe a linguagem de marcação de Sistemas Biológicos (SBML - System Biology Markup Language).

A utilização de uma marcação explícita das sentenças resolveria o problema do processamento automático dos textos científicos. Entretanto, apesar da quantidade de artigos publicados sobre o tema ter crescido nos últimos anos (o artigo de Hucka (2009) já possui mais de 500 citações[11]), o percentual de artigos anotados é insignificante em relação ao universo de trabalhos.

3.5 DISCUSSÃO

Ainda hoje, ignorando o avanço das técnicas de Mineração de Texto e recuperação da informação, grande parte dos autores demonstram preocupação apenas na formatação visual dos seus artigos, tornando-os atraentes ao trato humano, mas dificultando o processamento por computadores

Em algumas áreas, como a da Saúde, os pesquisadores já utilizam um padrão estrutural para escrita de conteúdo, melhorando os resultados do processamento computacional dos textos e tornando mais rápida a localização das informações por parte dos leitores. Já em outras áreas, como a Ciência da Computação, percebe-se que não há um padrão definido, fato que foi enunciado por Teufel e Moens (2002) e confirmado neste estudo através da seleção e análise de 33 artigos da área, conforme resultado exibido na Tabela 3.3.

Como alternativa à aglomeração de sentenças de mesmo propósito em determinadas seções do texto (padronização na escrita do conteúdo), seria possível efetuar a anotação semântica das sentenças durante a escrita do artigo, de forma a permitir o intercâmbio das

11 Segundo informações obtidas em http://esciencenews.com/articles/2010/03/02/after.5.years.free.systems.biology.markup.language.has.proven.popular

informações entre ferramentas computacionais sem necessidade de intervenção humana. Esta abordagem, entretanto, demandaria considerável esforço adicional dos pesquisadores na escrita dos seus trabalhos.

3.6 CONCLUSÃO

A falta de estrutura definida na escrita do conteúdo de artigos científicos de engenharia de software, evidenciada na análise realizada neste capítulo, age como fator dificultador no processo de classificação automática das sentenças do texto.

Em relação a ênfase dada à formatação apenas do estilo do texto, encontrada em todos os modelos analisados, o que se percebeu é que além de não auxiliar na categorização das sentenças, atua como agente dificultador em função dos ruídos gerados no texto pelas marcas de formatação.

Capítulo 4

Classificação de Sentenças

Neste capítulo serão analisados os trabalhos de Teufel (1999), Ibekwe-SanJuan (2008) e Agarwal (2009), que abordam a classificação das sentenças do texto completo do artigo científico, em relação a sua implementação e índices de abrangência e precisão na classificação de sentenças de artigos científicos. Serão, ainda, apresentados os resultados de testes realizados sobre a performance destas três técnicas, de forma a identificar a possibilidade de utilização na localização de resultados de artigos em ES.

O universo de estudos que tratam sobre classificação de sentenças do texto em relação ao seu propósito é restrito. Neste contexto, encontramos trabalhos que tratam de textos com conteúdos bastante divergentes, como o de Hachey (2005), que utiliza diversas técnicas baseadas em aprendizado de máquina para definir o status retórico das sentenças de um corpus formado por documentos da área jurídica, ou o de Khoo (2006), que propõe a classificação de sentenças de um corpus de e-mails de helpdesk em categorias relacionadas ao domínio, como instrução, questões e resposta. Há, ainda, a categoria de trabalhos com ênfase na análise de sentenças de textos científicos, que serão analisados mais detalhadamente por tratarem especificamente do tema alvo desta dissertação.

Para uma melhor avaliação, dividimos os estudos que analisam as sentenças do texto científico de acordo com o escopo de texto a ser classificado. A maior parte dos trabalhos encontrados tinha como alvo de classificação apenas as sentenças constantes no resumo (abstract) dos artigos, conquanto uma minoria atuava sobre o texto completo do artigo científico.

As pesquisas de McKnight (2003), Ruch (2007), Xu (2006) e Tbahriti (2004), utilizaram diferentes técnicas para a classificação de sentenças presentes em resumos (abstracts) não estruturados de artigos do banco de dados da área de saúde. Já os trabalhos de Teufel (1999), Ibekwe-SanJuan (2008) e Agarwal (2009), trataram da classificação do texto

completo dos artigos científicos da área de saúde e, por conseguinte, foram selecionados e avaliados nas seções subsequentes.

4.1 ARGUMENTAÇÃO BASEADA EM DISCURSO EM ARTIGOS CIENTÍFICOS

O trabalho desenvolvido por Teufel (1999), apesar de ter como objetivo promover a melhoria em um método de criação automática de resumos de artigos, foi um dos primeiros sobre a identificação automática do propósito de uma sentença em um artigo cientifico. É importante ressaltar que, para a definição das categorias do discurso, o estudo não se baseou no modelo IMRaD. As classes foram definidas pela autora segundo critérios próprios, sendo agrupadas em dois esquemas: um básico, formado por três elementos, e outro completo, que adiciona outros quatro elementos aos já existentes (Tabela 4.1). A construção de um corpus anotado composto por 80 artigos da área de linguística computacional também foi contribuição relevante deste trabalho.

Tabela 4.1 - Categorias para classificação de Sentenças do texto científico

Esquema Básico	Background	Sentenças que descrevem algum informações de contexto.	Esquema Completo
	Other	Sentenças que descrevem aspectos específicos de outras pesquisas de uma forma neutra (excluindo as sentenças definidas como CONTRAST e BASIS)	
	Own	Sentenças que descrevem qualquer dos aspectos do próprio trabalho apresentado no artigo – exceto o que é coberto pelos tópicos AIM e TEXTUAL como, por exemplo, detalhes da solução (metodologia), limitações e trabalhos futuros.	
	Aim	Sentenças que enquadram melhor o principal objetivo de pesquisa do artigo.	
	Textual	Pedaços de texto explicitamente sobre a estrutura de seções do texto do artigo	
	Contrast	Sentenças que contrapõem o trabalho atual com outros trabalhos; sentenças que indicam pontos fracos no trabalho; sentenças que afirmam que a investigação do presente trabalho nunca foi antes estudada; comparações diretas.	
	Basis	Sentenças em que o próprio trabalho usa outros trabalhos como base ou ponto de partida, ou é suportado por este outro trabalho.	

Fonte: Teufell (1999).

O estudo propõe a definição de um conjunto de características que definem diversos aspectos das sentenças (Tabela 4.2). A partir do conjunto de valores atribuídos a cada sentença, seria possível determinar em que categoria a mesma se enquadra.

Utilizando o algoritmo Naive Bayes como base do classificador automático, Teufel

apresentou resultados que corroboram a possibilidade de se automatizar o processo de anotação semântica das sentenças. A pesquisadora aponta como relevante o fato de 81% das sentenças pertencentes a categoria “OWN” terem sido classificadas corretamente (precisão) e de 91% do total de ocorrências existentes no texto terem sido localizadas (abrangência).

Tabela 4.2 - Características das Sentenças do texto científico

Tipo da Característica	Valores Possíveis
Tipo do cabeçalho da seção corrente	"introduction", "data", "result", "metbod", "conclusion" e "introduction"
Posição relativa da sentença dentro do parágrafo	Início, meio, fim
Posição relativa da sentença dentro da seção	Primeira, segunda ...
Localização relativa – o artigo é dividido em dez segmentos iguais.	1-10
A sentença contém uma citação ou o nome de um autor contido na lista de referência?	Citação, nome do autor ou nenhuma
A sentença contém uma auto citação	Sim, não
Tempo Verbal (associado ao primeiro verbo finito na sentença)	Presente, Passado, Presente Perfeito, Passado Perfeito, Fututo ou nenhum
Modal auxiliares	Presente ou não
Voz	Ativa ou passiva
Negação	Presente ou não
Tipo de ação do primeiro verbo da sentença. O verbo indica...	Emoção, argumentação, conhecimento, uma melhor solução, mudança, comparação, contraste, interesse futuro, interesse, necessidade, apresentação de algo, problema, pesquisa, algo análogo, solução, estrutura do texto, uso, "copula", "possession"
Tipo de Agente	Autor, outro ou nenhum
Type of formulaic expression ocurring in the sentence	18 diferentes tipos
A sentença contém palavras-chave definidas pela métrica TF/IDF?	Sim ou não
A sentença contém palavras que ocorrem no título ou cabeçalhos?	Sim ou não

Fonte: Adaptada de Teufell (1999).

Agrupando as informações de precisão e abrangência com a frequência da quantidade de sentenças em cada categoria no texto (Tabela 4.3.), observamos que o mais relevante resultado apontado aparenta ser consequência da grande quantidade de sentenças definidas na categoria “OWN”. Se o algoritmo de classificação definisse todas as sentenças do texto como pertencentes a esta categoria específica, ao final do experimento, obteríamos 69,4% de acerto e 100% de abrangência, resultado consideravelmente superior ao encontrado para outras categorias do estudo. É interessante ressaltar que os índices encontrados para categorias como “AIM”, “TEXTUAL” e “OTHER”, que beiram os 50%, são extremamente positivos.

Não foi possível efetuar testes sobre o método de Teufel utilizando o corpus de artigos anotados construído por Cruzes (2007), uma vez que o algoritmo não foi disponibilizado para

nossa pesquisa pelos autores do trabalho.

Tabela 4.3 - Precisão e Abrangência do algoritmo de Teufel

Categoria	Precisão	Abrangência	Frequência no texto
AIM	48,00%	56,00%	2,40%
CONTRAST	37,00%	13,00%	4,40%
TEXTUAL	55,00%	52,00%	0,90%
OWN	81,00%	91,00%	69,40%
BACKGROUND	38,00%	41,00%	5,70%
BASIS	21,00%	45,00%	1,40%
OTHER	47,00%	23,00%	15,80%

Fonte: Teufel (1999).

4.2 IDENTIFICANDO INFORMAÇÕES ESTRATÉGICAS EM ARTIGOS CIENTÍFICOS

O principal objetivo de Ibekwe-SanJuan (2008) foi o de auxiliar os usuários a identificar mais rapidamente as principais informações presentes no texto científico, a partir de elementos específicos associados às sentenças. A autora definiu uma série de gramáticas locais para anotação automática e classificação das sentenças de acordo com o seu status retórico. Para tanto, foram definidas oito categorias: objetivo, coisas novas, resultados, descobertas, hipóteses, conclusões, trabalhos relacionados, trabalhos futuros.

As gramáticas de Ibekwe-SanJuan (2008) foram desenvolvidas utilizando como ferramenta o Software UNITEX[12] e como conjunto de testes uma coleção de artigos oriundos de três diferentes áreas de conhecimento: Biologia Quantitativa, Ciências da Informação e Astronomia.

Em função das variações que podem ocorrer em diferentes níveis linguísticos da escrita (morfológico, sintático, léxico e semântico), parte dos autômatos construídos para representar os padrões léxico-sintáticos utilizaram, de forma combinada, dados destes quatro grupos citados. A Tabela 4.4. apresenta exemplos de alguns dos padrões definidos pela autora.

Utilizando a base de artigos anotados montada por Cruzes (2007), baseada em artigos da Área de Testes de Software e escritos em inglês, foi executado um experimento para testar a eficácia do autômato no reconhecimento de sentenças indicando “Resultados”. Foram utilizados 7 artigos, compostos por 2617 sentenças. Destas, o algoritmo de Ibekwe-SanJuan

12 http://www-igm.univ-mlv.fr/~unitex/

classificou 264 como “Resultados”, sendo 25 de forma correta e 239 incorretamente, gerando um índice de acerto de 9,47%.

Tabela 4.4 - Exemplo de Padrões léxico-sintáticos

Tipo de Informação	**Padrões**
Objetivo	In this_{article|paper|study|research|work}...
	We_{examine|investigate|describe|present|outline|introduce|consider}....
	DET_{motivation:|aim|goal|objective}...
Coisas Novas	Here, we propose a novel (...) approach...
	This analysis reveals...
	Emerging evidence suggests that...
	Interestingly, our results indicate that...
Hipóteses	DET_NP_{may|might}_{ADV|V_NP}...
	Our findings support the view that...
	DET_NP_can_{V|NP.}..
Trabalhos Futuros	{Further|Future|more}_{work|investigation|observation}_<verb>...
	Table

Fonte: Ibekwe-SanJuan (2008).

Ibekwe-SanJuan não explicita, em seu artigo, os índices de acerto e abrangência obtidos com o experimento, motivo pelo qual não é possível realizar uma comparação com o resultado encontrado em nosso teste. Pressupõe-se, entretanto, que por utilizar artigos de ciências experimentais, que respeitam com mais rigidez as divisões retóricas das publicações científicas, o resultado encontrado pela autora tem razoável possibilidade de ter índices de acerto superiores aos do experimento realizado neste trabalho.

4.3 CLASSIFICAÇÃO AUTOMÁTICA DE SENTENÇAS EM TEXTOS COMPLETOS DE ARTIGOS

O trabalho de Agarwal (2009) é voltado especificamente para a classificação de sentenças de textos completos de artigos científicos da área biomédica. Para tanto, utiliza-se das quatro

categorias tipicamente encontradas em textos deste gênero e constantes no padrão IMRaD. O objetivo do estudo é, através da classificação das sentenças, auxiliar no desempenho de tarefas relacionadas à mineração de textos, como extração da informação e construção automática de resumos. O estudo utiliza-se de abordagens baseadas em regras e em aprendizado de máquina para a classificação automática das sentenças. A base de dados de testes é formada por oito artigos da área biomédica, compostos por 1250 sentenças.

Pode-se dividir o estudo realizado em duas etapas: a primeira que trata da criação do corpus anotado de sentenças; e a segunda, que trata das tarefas de classificação automática. Para avaliar a qualidade da anotação das sentenças, Agarwal determinou que cada texto fosse analisado por três diferentes biólogos e estabeleceu um indicador, ao qual chamou de grau de confiança da anotação. A este indicador era atribuído, por cada biólogo, um de três valores: alto – se o responsável pela anotação estivesse certo de que a sentença pertencia a uma classe específica; médio – se ficasse incerto entre duas classes; e baixo – se a incerteza fosse entre três ou mais categorias. O percentual de concordância entre os avaliadores para sentenças definidas com grau de confiança alto foi de 88,2%, para as classificadas como introdução, 94,3%, das classificadas como método, 90,7%, das definidas como resultado, e 84,6%, das classificadas como discussão.

Como resultado da classificação automática das sentenças, parte da segunda etapa, o autor afirma ter alcançado 81,04% de abrangência e 80,73% de precisão, conforme exibido na Tabela 4.5.

Tabela 4.5 - Resultado do experimento de classificação das sentenças

Acurácia	Introdução	Método	Resultados	Discussão	Média
81,04%	A: 87,8 P: 83,7	A: 88,0 P: 82,0	A: 76,7 P: 77,5	A: 53,3 P: 75,0	A: 81,04 P: 80,73

Fonte: Agarwal (2009).

Nota: (Legenda: A – Abrangência e P - Precisão)

Uma análise aprofundada do algoritmo de Agarwal permite desvendar o seu modo de funcionamento. O método utiliza um conjunto de 2504 palavras oriundas do conjunto de sentenças utilizado no experimento realizado pelo autor. Os padrões para classificação são definidos pela combinação da presença ou ausência de elementos deste conjunto de palavras nas sentenças.

A tabela 4.6 exibe alguns dos termos oriundos do conjunto utilizado por Agarwal.

Através da análise destes números percebe-se, por exemplo, que do total de ocorrências da palavra "table", considerando o conjunto completo de sentenças, 93,02% foram em sentenças que indicavam resultado e, do total de sentenças que indicavam resultado, 14,65% eram compostas pela palavra "table". Desta forma, é provável que a utilização deste termo auxilie na classificação de sentenças que indiquem resultado.

Tabela 4.6 - Análise de termos retirados do Corpus de Agarwal

Termo	Percentual de ocorrência do termo em sentenças que indicam resultado	Percentual das sentenças que indicam resultado que possuem esta palavra
Table	93,02%	14,65%
Significantly	81,25%	9,52%
Figure	100,00%	6,23%
Showed	90,00%	6,59%
Higher	64,10%	9,16%
Observed	67,86%	6,96%
Compared to	62,07%	6,59%
Was not	64,29%	3,30%
respectively	61,54%	5,86%
Detected	66,67%	3,66%

Assim como no teste realizado com o método de Ibekwe-SanJuan (2008), utilizamos o corpus anotado de Cruzes (2007) para testar o método proposto por Agarwal. Selecionamos, neste corpus, 60 sentenças categorizadas como resultado, e efetuamos testes de classificação em ferramenta disponibilizada pelo Autor[13]. O resultado do experimento, que mostra o resultado do processo de classificação, pode ser visualizado na Tabela 4.7.

Tabela 4.7 - Classificação de Sentenças utilizando o método de Agarwal

	Discussão	Método	Background	Resultado
Quantidade	4	4	37	15
Porcentagem	6,67%	6,67%	61,67%	25,00%

A análise dos dados mostra que apenas 25% dos resultados foram classificados corretamente, valor bastante distinto do resultado de 77,5% encontrado pelo autor para esta

13 Disponibilizado em http://wood.ims.uwm.edu/full_text_classifier, acessado em 10/11/2009 às 16:30

mesma classe (Tabela 4.5).

De modo a identificar a causa da divergência dos resultados encontrados entre os testes de Agarwal e o realizado neste trabalho, resolveu-se não apenas efetuar um teste de alto nível na ferramente disponibilizada, mas avaliar detalhadamente o funcionamento do método proposto, identificando os termos presentes no conjunto utilizado e a sua importância na classificação de cada tipo de sentença.

O primeiro passo foi reduzir a dimensionalidade do conjunto de palavras utilizado por Agarwal. Desta forma, os 2504 termos inicialmente listados por Agarwal foram reduzidos para 289, após a eliminação das palavras pouco representativas – aquelas com menos de 1% do total de ocorrências em relação ao total de palavras do conjunto.

A segunda etapa consistiu em calcular a frequência de cada um destes termos nas sentenças que indicavam resultados e as que não representavam resultados, de forma a definir o grau de importância de cada termo na classificação da sentença.

A análise da presença ou não dos termos em cada tipo de sentença mostrou que dos 289 termos inicialmente selecionados, 139 possuíam entre 0 e 15% das ocorrências em sentenças classificadas como resultado; 78 entre 25,01% e 50%; 50 termos possuíam entre 50,01% e 75% e 22 termos tinham de 75,01% à 100% de suas ocorrências concentradas nos resultados. Alguns dos termos e seus percentuais podem ser observados na Tabela 4.8.

Tabela 4.8 - Importância dos termos na definição da categoria das sentenças

Termo	**Percentual de ocorrência do termo em sentenças que indicam resultado**	**Percentual das sentenças que indicam resultado que possuem esta palavra**
Higher	50,00%	10,00%
Seen	66,67%	10,00%
Tend	50,00%	5,00%
Sized	100,00%	5,00%
Susceptible	100,00%	5,00%
Possibly	100,00%	5,00%
Compared	60,00%	6,67%
More	50,00%	18,18%
Much	80,00%	7,27%
Figure	30,00%	16,36%

A etapa final consistiu em gerar um conjunto de termos composto por 1116 sentenças, utilizando artigos do corpus de Cruzes como base. O objetivo foi o de comparar os termos que caracterizavam resultados presentes da lista de Agarwal com os desta lista montada do corpus de Cruzes e observar se eram coincidentes. O resultado demonstrou que apenas 10% das palavras dos conjuntos era coincidentes.

Em tese, foi possível perceber que termos ligados à comparação, probabilidade e números são úteis no auxílio à classificação das sentenças que indicam resultado, independente do contexto dos artigos. Por outro lado, a maior parte dos termos aparentou ser dependente de contexto. Houve termos que, no corpus de Agarwall, apresentavam todas as suas ocorrências em sentenças que indicavam resultados e no de Cruzes tinham 70% de suas ocorrências concentradas em sentenças que não indicavam resultados.

O que se percebeu com a análise do Corpus de Cruzes é que a informação diretamente ligada ao contexto do artigo e capaz de prestar considerável auxílio na diferenciação entre categorias se perdeu no estudo de Agarwal, sobretudo em função do pesquisador tratar a lista de termos encontrados em função do "conjunto" de artigos e não individualmente.

Através da análise individualizada de três artigos do Corpus de Cruzes, foram identificados os substantivos com maior frequência em cada um dos estudos. Os resultados demonstraram que:

a) Elevado percentual dos substantivos que possuíam alta frequência em um artigo era pouco frequente nos demais, o que acabava por eliminá-los ou reduzir sua importância na lista final de termos;

b) Estes substantivos eram relevantes apenas para o artigo onde foi frequente, causando ruído caso fosse utilizado em outros artigos;

c) Embora boa parte destas palavras, em termos absolutos, se localizavam em sentenças que não indicavam resultados, em termos relativos (levando em consideração cada grupo isoladamente), estavam presentes em grande parte dos resultados.

Observando a tabela 4.9, percebe-se, por exemplo, que o termo "Defects" tem, do seu total de ocorrências, 36,11% em sentenças que indicam resultado e, por conseguinte, 63,89% em outras sentenças. Entretanto, dentre os resultados, 47,27% possuem "defects" em sua composição, enquanto apenas 11,89% das outras sentenças contam com este termo entre suas

palavras constituintes. O interessante a ressaltar é que a palavra “defect” só é importante no contexto do Artigo 1. Utilizada nos Artigos 2 ou 3 não teria qualquer relevância.

Tabela 4.9 - Frequência de termos em artigos retirados do Corpus de Cruzes

Artigo 1				Artigo 2				Artigo 3			
Termo	**Frequência nas sentenças que indicam resultado**	**Percentual dos resultados que possuem o termo**	**Percentual das demais sentenças que possuem o termo**	**Termo**	**Frequência nas sentenças que indicam resultado**	**Percentual dos resultados que possuem o termo**	**Percentual das demais sentenças que possuem o termo**	**Termo**	**Frequência nas sentenças que indicam resultado**	**Percentual dos resultados que possuem o termo**	**Percentual das demais sentenças que possuem o termo**
Defects	36,11%	47,27%	11,89%	Change(s)	12,62%	86,67%	24,66%	Error	12,27%	100,00%	52,38%
Project	30,91%	30,91%	9,82%	Man-hours	75,00%	60,00%	0,82%	Modules	13,79%	100,00%	45,79%
Interface	80,00%	29,09%	1,03%	Distributions	30,77%	26,67%	2,47%	Percent	35,00%	35,00%	65,00%
External	85,71%	21,82%	0,52%	Errors	5,88%	20,00%	13,15%	Effort	35,00%	35,00%	65,00%
projects	20,00%	18,18%	10,34%	Average	75,00%	20,00%	0,27%	Specifications	42,86%	30,00%	2,93%

4.4 CONCLUSÃO

Este capítulo teve como principal objetivo analisar a performance dos algoritmos e métodos propostos por Teufell (1999), Ibekwe-SanJuan (2008) e Agarwal (2009), selecionados por tratarem de tópico correlato a este trabalho. Para o primeiro não foi possível efetuar testes de performance, já que não tivemos acesso ao algoritmo utilizado no experimento original. Os demais foram testados com base em informações enviadas pelos autores dos estudos.

O teste do autômato de Ibekwe-SanJuan foi realizado utilizando modelo encaminhado pela autora e, para o conjunto de treinamento e teste fornecido, gerou um índice de acerto de 9,47%. Já o algoritmo de Agarwal, foi testado a partir de um software disponibilizado pelo autor e alcançou índices de acerto um pouco superiores ao estudo de Ibkwe-SanJuan, apesar de ainda insatisfatórios em função da precisão de 25% na classificação dos resultados e de 6,67% na classificação dos métodos.

O principal aspecto observado em ambos os testes foi a baixa performance dos classificadores e, no caso de Agarwal, a divergência dos resultados encontrados neste estudo em relação aos estudos originais, fato provavelmente ocasionado em função do uso de diferentes conjuntos de treinamento e teste. Conquanto os autores em seus experimentos originais utilizaram artigos da área biomédica, que possuem um alto nível de organização do conteúdo, em nosso experimento utilizamos um conjunto de artigos da área de Engenharia de

Software oriundos do corpus de Cruzes (2009), artigos que não costumam seguir um padrão de organização do conteúdo.

A escolha de efetuar os testes destes métodos com artigos da área de Engenharia de Software, um domínio diferente do utilizado nos testes originais pelos autores, foi tomada em razão de que o nosso estudo busca o desenvolvimento de um método que é voltado para esta área específica. Precisávamos descobrir como se comportavam estes métodos atuando sobre o domínio dos artigos de Engenharia de Software.

Em relação à contribuição destes estudos para a definição do método proposto nesta dissertação, destacamos os seguintes pontos:

a) Do estudo de Teufell (1999) utilizamos a metodologia utilizada para análise das sentenças. Observamos que o conjunto de atributos definidos para caracterização de sentenças pode evidenciar atributos não apenas das classes definidas por Teufel em seu estudo, mas também para as classes do modelo IMRaD;

b) Os métodos apresentados nos trabalhos de Ibekwe-SanJuan (2008) e Agarwal (2009) forneceram a inspiração para a definição de uma série de atributos de sentença, utilizados no método proposto.

Capítulo 5

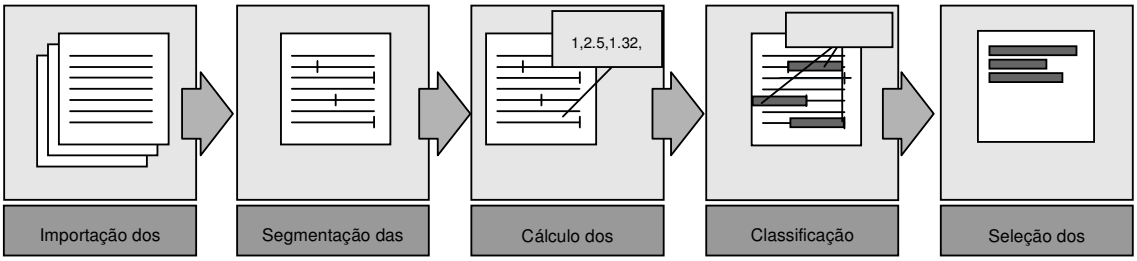
Método Proposto

Neste capítulo descreveremos o método proposto neste estudo, explicitando os caminhos adotados e a justificativa para a escolha.

O método proposto neste estudo tem como principal objetivo automatizar parte do processo de Revisão Sistemática, auxiliando os pesquisadores e reduzindo a quantidade de esforço a ser empreendido na construção do estudo. O algoritmo é construído com base em técnicas de mineração de texto, com a definição de estratégias de classificação que utiliza uma série de atributos previamente definidos.

De uma forma geral, o método utilizado pode ser dividido em cinco atividades, que vão desde a importação do texto dos artigos até a seleção dos parágrafos que irão compor o documento final (Figura 5.1).

Figura 5.1 - Esquema do método de sumarização do texto



5.1 IMPORTAÇÃO DOS ARTIGOS E SEGMENTAÇÃO DAS SENTENÇAS

A primeira atividade é a de importação dos artigos, e consiste em converter os documentos de entrada para o formato de texto plano. Nesta etapa são removidas marcas de formatação e caracteres indesejáveis.

A próxima etapa consiste na extração das sentenças, que são tratadas neste estudo como pedaços de texto com significado completo. Esta atividade é realizada com a utilização

de um algoritmo baseado em reconhecimento de padrões denominado YASD[14] (Yet Another Sentence Detector) que considera a utilização de caracteres de pontuação, utilização de maiúsculas, nomes próprios, siglas e posicionamento de partículas gramaticais no texto. Parte da heurística utilizada pelo algoritmo pode ser visualizada no Quadro 5.1.

Quadro 5.1 - Heurística do algoritmo de detecção de Sentenças

```
function DECIDE_ORTHOGRAPHIC (TOKEN):
        Se TOKEN tem a pprimeira letra em maiúscula:
                Se TOKEN já ocorreu com  primeira letra em minúscula:
                        Se TOKEN nunca ocorre com a primeira letra em maiúscula no
interior da sentença
                                Return divisor_de_sentença
                        Senão
                                Return indeciso
                Senão
                        Return indeciso
        Senão Se TOKEN tem a primeira letra em minúscula
                (...)
```

5.2 CÁLCULO DOS ATRIBUTOS

O núcleo do algoritmo se encontra na etapa de Cálculo dos Atributos, atividade onde são definidos os valores de características para cada uma das sentenças do texto de entrada. São calculados os seguintes atributos, já comentados no Capítulo 2:

a) **Freqüência das palavras-chave**: o cálculo do atributo segue o modelo de Luhn (1958), sendo composto pela soma das frequências das palavras da sentença (exceto *stop words*) multiplicado pelo cálculo da distância entre as palavras.

b) **Palavras sinalizadoras (cue method)** – a partir da utilização dos conjuntos de palavras definidos por Teufel e Moens (1997), atribuiu-se a cada palavra constante na lista de palavras "stigma" o peso -1, na lista de palavras " bônus A" foi atribuído peso 1, na lista "bônus B" atribuiu-se peso 2 e na "bônus C" peso 3. Ao final, o valor atribuído à sentença é a soma dos pesos das suas palavras constituintes.

c) **Setor no Parágrafo** – foi utilizado método proposto por Teufel e Moens, onde é considerada a localização da sentença no parágrafo a que pertence. O total de sentenças é dividido por 4, às presentes no primeiro quarto é atribuído valor 0, e, sequencialmente, são atribuídos os valores 1, 2 e 3 aos demais quartos.

14 https://sites.google.com/site/carlossillajr/files/yasd.zip?attredirects=0

d) **TF-IDF** - cálculo do produto da frequência do termo -TF, pela frequência inversa do documento – IDF. O método de cálculo é semelhante ao apresentado por Salton (1988).

e) **TF-ISF** – a frequência inversa da sentença é definida com base na frequência da palavra na sentença – F(w), no número de palavras na sentença a que a palavra pertence – n, e no número de sentenças em que a palavra aparece – S(w) (Quadro 5.2).

Quadro 5.2 - Fórmula de Cálculo do TF-ISF

$$TF - ISF\,(w) = F(w) \times \frac{\log n}{S(w)}$$

Fonte: Larocca Netoe outros (2000).

f) **Tamanho da Sentença** – o atributo é representado pela quantidade de palavras que compõem a sentença.

g) **Presença de palavras da "gist sentence" na sentença** – o cálculo é semelhante ao apresentado por Pardo (2002) .

h) **Média da conectividade léxica** - o peso é definido com base no número de termos compartilhados por uma sentença com as demais sentenças do texto, dividido pelo número total de sentenças.

i) **Posição da Sentença na seção** – após a identificação de cada título de seção, numera-se a posição de cada sentença sequencialmente, iniciando-se com 1.

j) **Posição da Sentença em relação aos 10 segmentos** – variação do modelo definido por Teufel e Moens (1997). O texto é dividido em 10 partes iguais e à cada seção é atribuída uma letra, de A à J. As sentenças presentes em cada uma das seções terá como valor de atributo a letra que identifica a seção.

k) **Tempo verbal da sentença** – o atributo é definido através da análise da primeira ocorrência de verbo na sentença. Os valores possíveis são: presente na terceira pessoa do singular; presente com exceção da terceira pessoa do singular; infinitivo; passado particípio; presente particípio ou gerúndio; e sentença sem verbo.

l) **Citações** – se a sentença contém uma citação, valor 1, se não contém, valor 0.

m) **Cabeçalho** – indica se uma sentença se encontra em uma seção que possui um cabeçalho relevante. A definição da relevância do cabeçalho é dada se o seu título possuir uma das palavras cabeçalho, as quais: “result”, “introduction”, “conclusion”, “data”, “implementation” e “discussion”. O valor do atributo para cada sentença é a palavra cabeçalho relacionada ao título da seção em que se encontra.

n) **Frases auto-indicativas** – a abordagem utilizada é a proposta por Ibekwe-SanJuan, e utiliza a sua gramática de reconhecimento de frases que indicam resultados. As sentenças selecionadas pelo autômato terão peso 1, as não selecionadas, peso 0.

o) **Presença de Comparação** – verifica se existem partículas indicando comparação nas sentenças, como “er”, “est”, “more” … “than”. Se encontradas, atribui valor “Sim” à sentença, caso contrário, atribui o valor “Não”.

p) **Presença de Número** – verifica, através dos rótulos sintáticos atribuídos às palavras das sentenças, se existe palavra classificada como número na sentença. Se existe, é atribuído o valor “Sim”, caso contrário, é atribuído o valor “Não”.

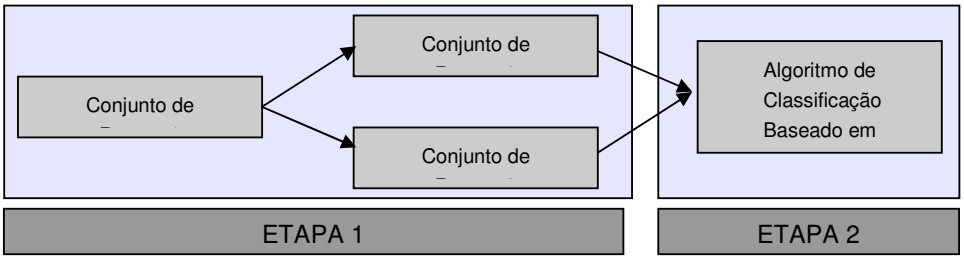
q) **Indicador de porcentagem** – se for encontrado o símbolo “%”, a palavra “percent”, ou variação, dentre os termos da sentença, é atribuído o valor “Sim”, caso contrário, é atribuído o valor “Não”.

5.3 CLASSIFICAÇÃO

De posse dos conjuntos de treinamento e testes, formados por sentenças já definidas e com os atributos calculados, é o momento de enviar os dados para o algoritmo de classificação. Cabe a este algoritmo classificar as sentenças que representam resultados e as que não representam.

A classificação é realizada em duas etapas, a primeira utiliza um algoritmo de classificação baseado em regras preestabelecidas, e a segunda um algoritmo de aprendizado de máquina, como pode ser observado na Figura 5.2.

Figura 5.2 - Esquema do método classificação



5.3.1 Classificação Baseada em Regras

O processo de definição do conjunto de regras de classificação se baseou na análise da distribuição de valores de cada um dos atributos de sentença. Todos os dados levantados para as sentença foram analisados em forma de gráfico. Como resultado, buscou-se identificar atributos que estivessem causando ruído na classificação por não auxiliarem na diferenciação das categorias. Buscou-se, ainda, descobrir os atributos que exibiam grande divergência de valores entre as sentenças *resultado* e as *não resultado*, apresentando-se com atributos chave de diferenciação entre as categorias.

A análise detalhada dos atributos de sentença é apresentada na seção 5.3.1.1. Os conjuntos de regras definidos com base nestes dados são exibidos nos quadros 5.3, 5.4 e 5.5.

Quadro 5.3 - Conjunto de Regras 1

- Se a sentença não possuir verbo, deve ser definida como "não resultado";
- Se a sentença, excluídas as "stop words", possuir tamanho menor ou igual a 3, deve ser definida como "não resultado";
- As sentenças que possuem valor inferior ou igual a 600 para o atributo "frequência das palavras-chave" devem ser definidas como "não resultado";
- As sentenças que iniciam um parágrafo devem sem definidas como "não resultado".
- As sentenças que possuem valor de TF-IDF inferior ou igual a 33,47 devem ser definidas como "não resultado";
- As sentenças que possuem valor de TF-ISF igual a "0" devem ser definidas como "não resultado";
- As sentenças com valor de Conectividade Léxica inferior ou igual a 0,01 devem ser definidas como "não resultado";
- As sentenças que possuem valor "A, B, C, D ou J" para o atributo "posição em relação aos 10 segmentos" devem ser definidas como "não resultado";
- As sentenças que possuem citação devem ser definidas como "não resultado";
- As sentenças pertencentes às seções: "introduction", "data" e "Implementation", devem ser definidas como "não resultado";

Quadro 5.4 - Conjunto de Regras 2

- As sentenças do conjunto resultante (após processamento do conjunto de regras 1) que possuírem uma outra sentença que seja pertencente ao conjunto e esteja posicionada imediatamente antes ou imediatamente após à sentença em análise "e" possuam um número em sua composição devem ser classificadas como "resultado";
- As sentenças com indicação de porcentagem devem ser classificadas como "resultado";

Quadro 5.5 - Conjunto de Regras 3

- As sentenças do conjunto resultante (após processamento do conjunto de regras 1) que possuírem uma outra sentença que seja pertencente ao conjunto e esteja posicionada imediatamente antes ou imediatamente após à sentença em análise "e" possuam um número em sua composição "e" possuam indicativo de comparação devem ser classificadas como "resultado" ;
- As sentenças com indicação de porcentagem devem ser classificadas como "resultado";

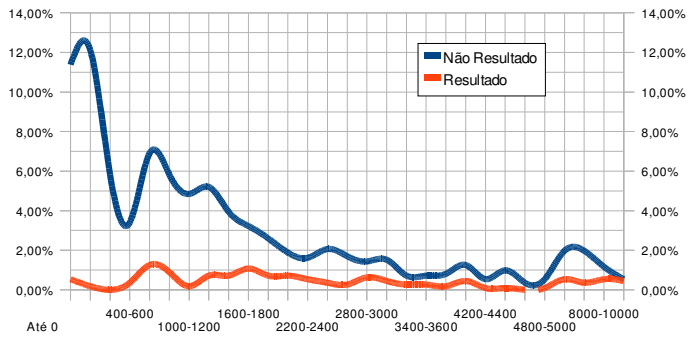
5.3.1.1 Análise dos Atributos da Sentença

A definição das regras elencadas nos quadros 5.3, 5.4 e 5.5 baseou-se na análise da distribuição de valores de cada um dos atributos, considerando-se a categoria a que pertencia cada sentença. Os dados foram planilhados e plotados em gráficos de fácil compreensão, conforme pode ser visualizado na descrição dos atributos, elencada a seguir:

a) Frequência das Palavras-chave

A análise do gráfico 5.1 aponta que há uma concentração de sentenças que não indicam resultado na faixa de valores entre 0 e 600, situação que não é compartilhada pelas sentenças que indicam resultado.

Gráfico 5.1 - Distribuição dos valores da frequência das palavras-chave nas sentenças resultado e não resultados em relação ao total de sentenças do texto.

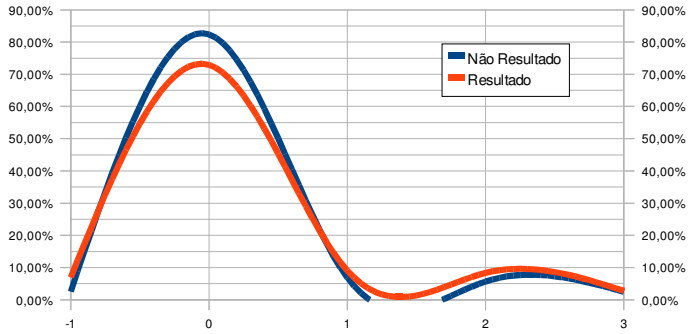


O experimento mostrou que aproximadamente 34% das sentenças do texto possuem valor entre 0 e 600, para o atributo frequência das palavras-chave. Deste total, aproximadamente 33% são sentenças que não se apresentam como resultado e apenas 1% são formados por sentenças que indicam resultado. Este valor de 1% do total de sentenças representam aproximadamente 8% dos total de resultados existentes.

b) Palavras sinalizadoras (cue method)

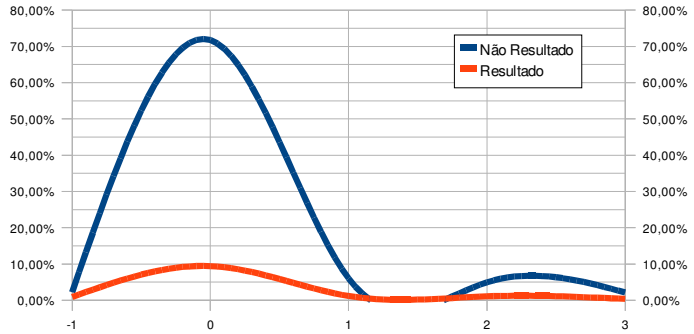
A distribuição dos valores deste atributo pelos cinco patamares é semelhante entre resultados e não resultados, conforme se observa no gráfico 5.2. Aproximadamente 72% dos resultados e 82% dos não resultados existentes possuem valor 0 para este atributo. A curva se mantém semelhante entre as duas categorias nos demais patamares, como se pode observar nos 5% dos resultados e 4% dos não resultados que possuem valor 2.

Gráfico 5.2 - Distribuição dos valores das palavras sinalizadoras entre as sentenças resultado e não resultados, relativo a cada categoria, isoladamente



É interessante ressaltar que a análise foi realizada, considerando-se a distribuição dos valores do atributo isoladamente em cada categoria. A análise do gráfico 5.3. permite concluir que o conjunto de não resultados é muito mais representativo quantitativamente do que o de resultados, independente da distribuição relativa dos valores do atributo.

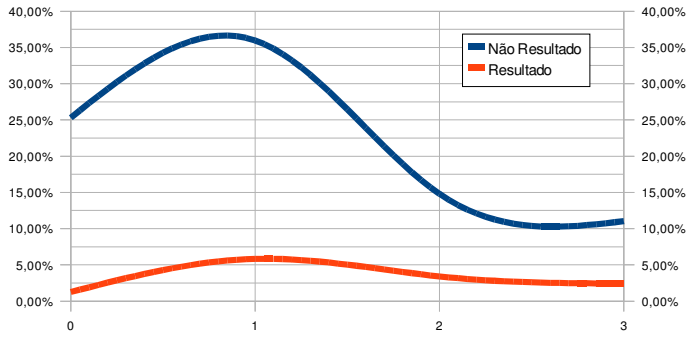
Gráfico 5.3 - Distribuição dos valores das palavras sinalizadoras, das sentenças resultado e não resultados, em relação ao total de sentenças



c) Setor no Parágrafo

A distribuição dos valores é realizada entre os quatro patamares previamente definidos – 0, 1, 2 e 3, e exibida no gráfico 5.4. É interessante ressaltar que não há uma distribuição equivalente entre os quatro patamares, já que nem todos os parágrafos do texto possuem uma quantidade de sentenças igual a quatro ou múltiplo de quatro.

Gráfico 5.4 - Distribuição dos valores do atributo "Setor no Parágrafo" entre as sentenças do corpus

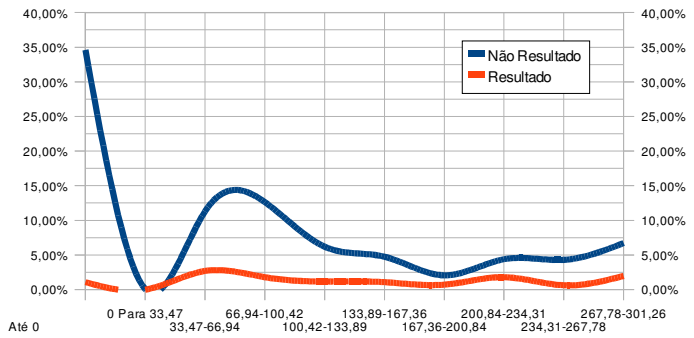


Através da análise dos valores do atributo, pôde-se observar que 26,55% do total de sentenças do texto se encontram no patamar 0. Destas, 1,26% são classificadas como resultados e 25,29% como não resultados. O percentual de 1,26% das sentenças representa 9,72% do total de resultados do conjunto, de forma que podemos concluir que 90,28% dos resultados se encontram nas seções 1, 2 e 3.

d) *TF-IDF*

A distribuição dos valores do produto da frequência do termo pela frequência inversa do documento (TF-IDF) é exibida no gráfico 5.5, onde pode-se observar uma concentração de não resultados na parte inicial, situação que não se observa nas sentenças da outra categoria.

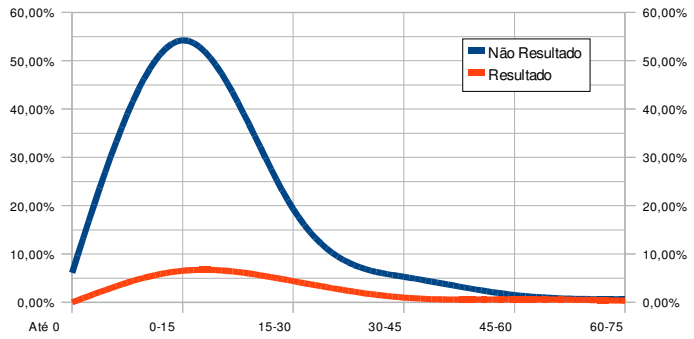
Gráfico 5.5 - Distribuição dos valores do atributo "TF-IDF" entre as sentenças do corpus



e) TF-ISF

O resultado resumido do cálculo da frequência inversa da sentença para as duas categorias é apresentado no gráfico 5.6.

Gráfico 5.6 - Distribuição dos valores do atributo "TF-ISF" entre as sentenças do corpus

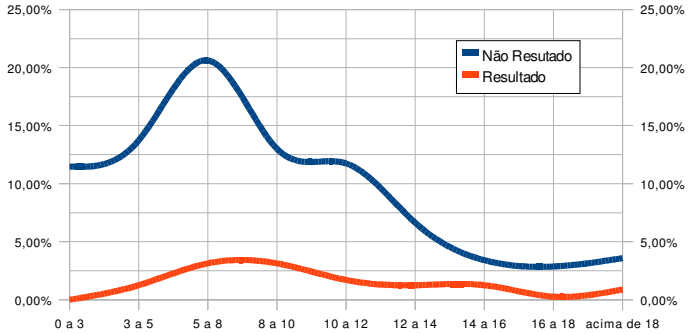


À exceção do valor inicial, onde não há presença de resultados e aproximadamente 8% de não resultados, os números encontrados apontam para uma distribuição similar entre as duas categorias, com crescimento e declínio do gráfico em pontos semelhantes.

f) Tamanho da Sentença

O gráfico 5.7 apresenta a evolução do valor do atributo tamanho da sentença entre as sentenças do conjunto, separadas em função de duas categorias em análise.

Gráfico 5.7 - Distribuição dos valores do atributo "Tamanho da Sentença" entre as sentenças do corpus

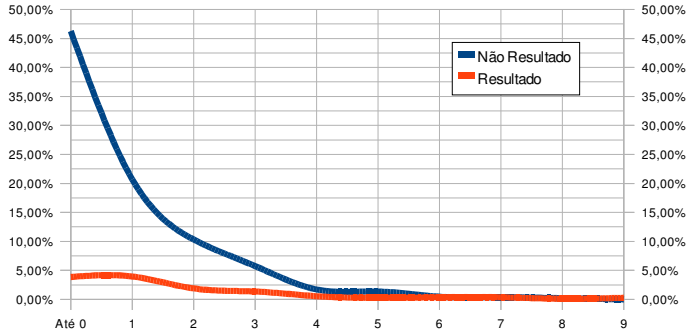


Assim como no TF-ISF, observa-se uma divergência entre a distribuição dos valores de resultados e não resultados, sobretudo na parte inicial do gráfico. Os dados obtidos através do experimento demonstraram que aproximadamente 13% das sentenças não resultados possuíam valor entre 0 e 3 para este atributo, conquanto nenhuma das sentenças resultados se enquadrava neste intervalo.

g) Presença de palavras da "gist sentence" na sentença

O agrupamento dos valores resultantes do cálculo deste atributo pode ser visualizado no gráfico 5.8

Gráfico 5.8 - Distribuição dos valores do atributo "Gist Sentence" entre as sentenças do corpus



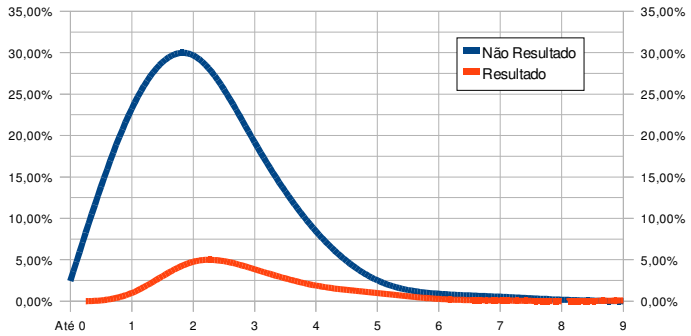
Apesar da aparente divergência de valores em algumas das faixas definidas, o que se percebe, de uma forma geral, é uma grande disparidade no início da série, mas uma evolução similar dos valores ao longo do gráfico.

Aproximadamente 53% dos resultados e 30% dos não resultados possuem valor igual a "zero" para este atributo, o que representa, respectivamente, 46% e 4% do total de sentenças do conjunto.

h) Média da conectividade léxica

Através da análise deste atributo, observando-se o gráfico 5.9, percebe-se uma provável zona de diferenciação entre sentenças que indicam resultado e as demais sentenças, na faixa que vai de 0 à 1.

Gráfico 5.9 - Distribuição dos valores do atributo "Média da conectividade léxica" entre as sentenças do corpus

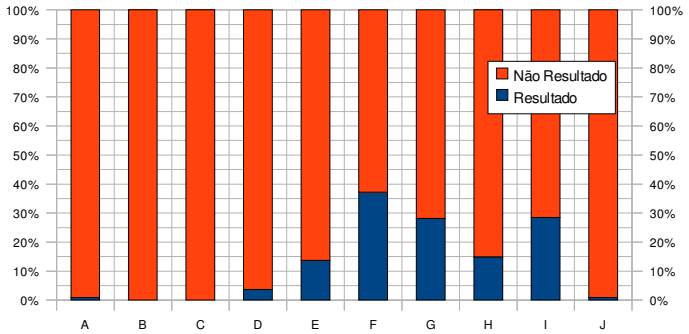


O experimento revelou que 29,45% dos não resultados e 7,64% dos resultados concentram-se nesta faixa de valores.

i) Posição da Sentença em relação aos 10 segmentos

O que se percebe, com a análise do gráfico 5.10, é que há uma concentração de sentenças que indicam resultado do meio para o final do texto, especificamente nas seções E, F, G, H e I.

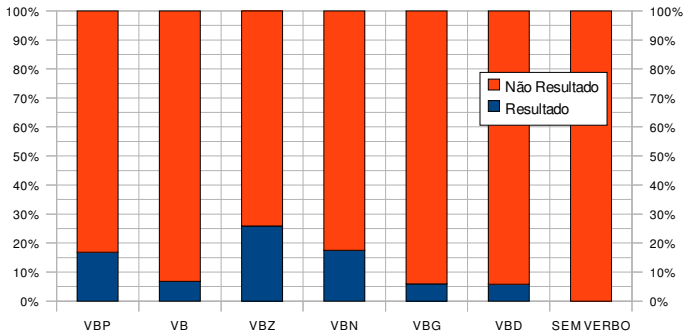
Gráfico 5.10 - Distribuição dos valores do atributo "Média da conectividade léxica" entre as sentenças



j) Tempo verbal da sentença

A análise linguística das sentenças dos artigos teve como resultado os valores exibidos no gráfico 5.11, onde se evidencia que a categoria resultado engloba sentenças que sempre possuem um verbo em sua composição.

Gráfico 5.11 - Distribuição dos valores do atributo "Tempo verbal da sentença" entre as sentenças



Os verbos se dividem nas seguintes categorias, a depender do tempo verbal:

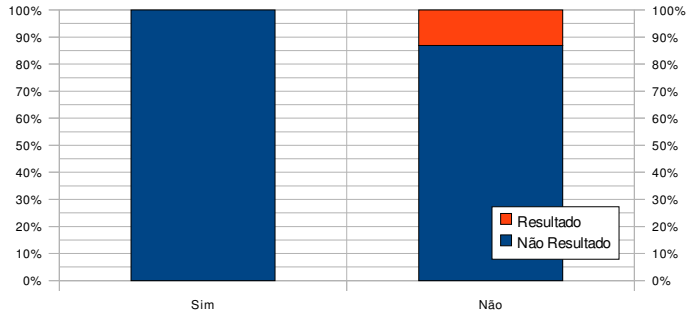
a) VBZ – verbo no presente e na terceira pessoa do singular
b) VBP – verbo no presente à exceção da terceira pessoa do singular

c) VB – verbo no infinitivo
d) VBN – verbo no passado particípio
e) VBG – verbo no presente particípio ou gerúndio
f) VBD – verbo no passado

k) Citação

O que se observa no gráfico 5.12 é que nenhum dos resultados encontrados é uma citação, de forma que este atributo torna-se importante para diferenciação das categorias de sentenças.

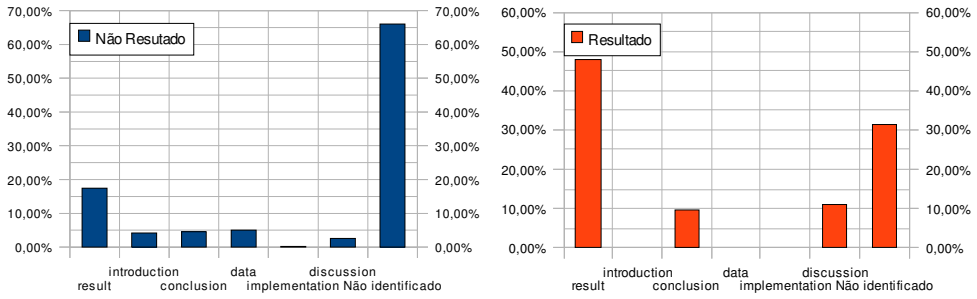
Gráfico 5.12 - Distribuição das citações entre resultados e não resultados



l) Cabeçalho

O gráfico 5.13 apresenta a distribuição das duas categorias de sentença entre as seções do texto que possuem, em seu título, as palavras: “result”, “introduction”, “conclusion”, “data”, “implementation” e “discussion”. A análise dos dados permite observar que 50% dos resultados se encontram na seção de resultados, confirmando a informação levantada no trabalho de Cruzes (2007).

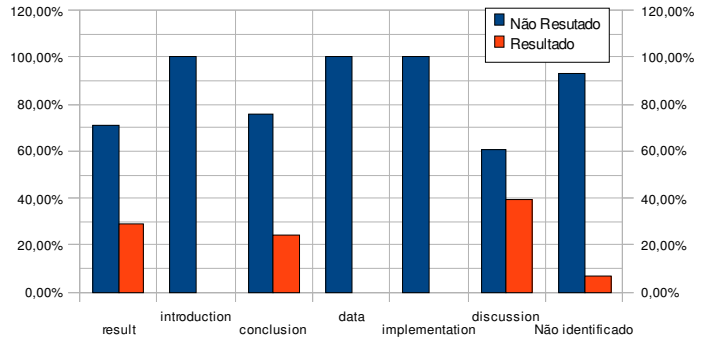
Gráfico 5.13 - Distribuição dos resultados e não resultados entre as diferentes seções do texto



O que dificulta a caracterização desta seção como representativa na classificação dos resultados é a diferença global entre a quantidade de resultados e não

resultados. Apesar de menos de 20% dos não resultados se enquadrarem na seção de resultados, estas sentenças representam quase 80% do total que compõe a seção, conforme pode ser visualizado no gráfico 5.14.

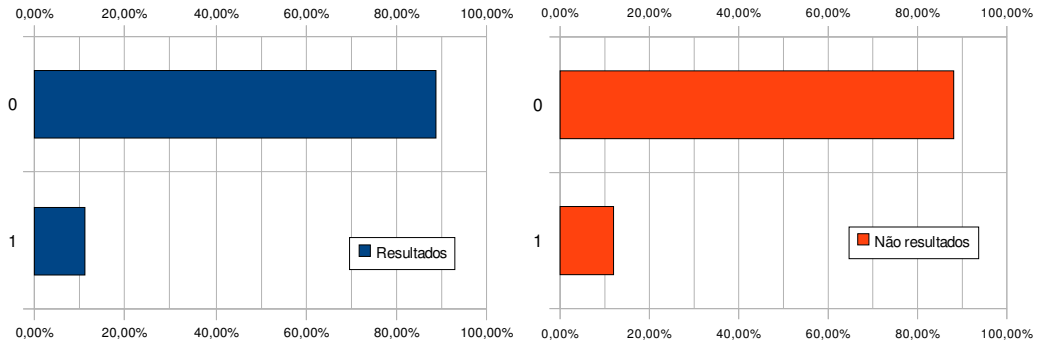
Gráfico 5.14. - Divisão das sentenças pertencentes a cada uma das seções entre as duas categorias em análise



m) Frases auto-indicativas

A análise do Gráfico 5.15 evidencia que há grande semelhança na distribuição dos valores deste atributo entre as sentenças que indicam resultado e as demais, de modo que as frases auto indicativas pouco contribuem para a classificação automática das sentenças, agindo, sobretudo, como ruído ao processo. Os valores possíveis são: 0 - caso não encontre frase auto indicativa; 1 – caso encontre frase auto indicativa.

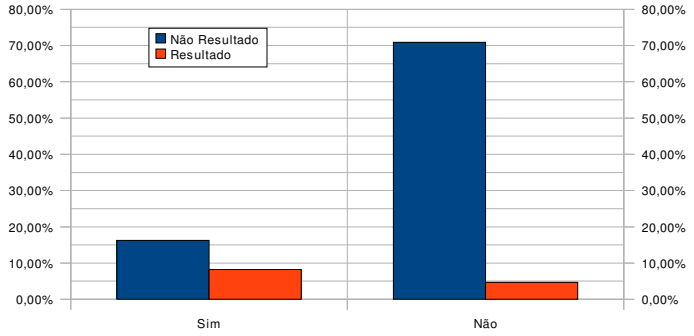
Gráfico 5.15 - Distribuição dos resultados e dos não resultados entre os possíveis valores das Frase auto indicativas



n) Presença de Comparação

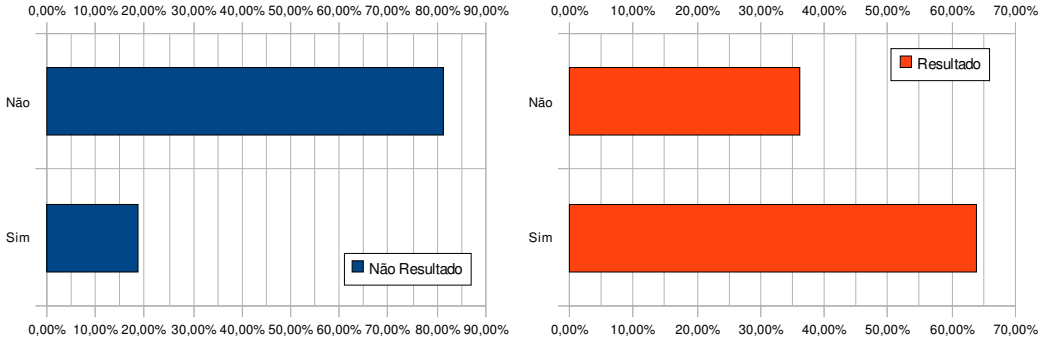
O que se observa no gráfico 5.16 é a grande quantidade de sentenças resultado que possuem termos indicadores de comparação. Mesmo com o universo dos resultados sendo aproximadamente dez vezes menor do que os de não resultados, das 24% sentenças do texto que possuem comparação, 16% são não resultados e 8% são resultados.

Gráfico 5.16 - Presença de comparação nas sentenças do texto



Este resultado se evidencia na análise do gráfico 5.17, onde se percebe que quase 60% das sentenças que representam resultado possuem algum tipo de comparação.

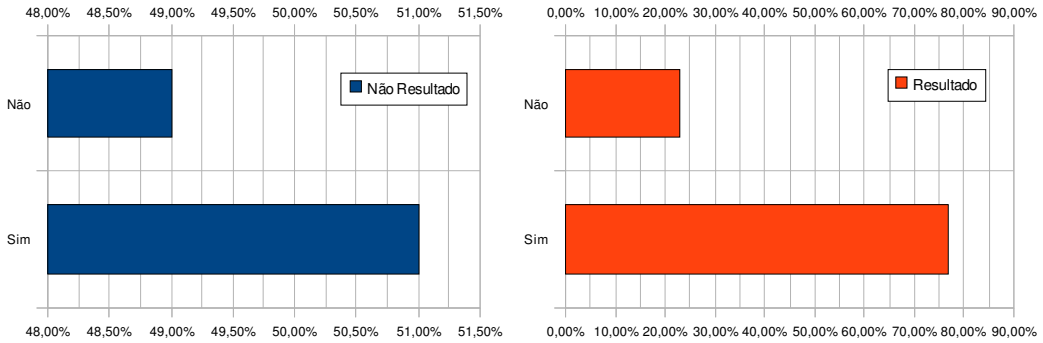
Gráfico 5.17 - Distribuição dos resultados e não resultados em relação à presenção ou não de comparação



o) Presença de Número

Este trabalho analisou o percentual das sentenças que possuíam algum termo com valor numérico em sua composição, independente do tipo de representação. O resultado pode ser visualizado no gráfico 5.18.

Gráfico 5.18 - Análise da presença de número nas sentenças

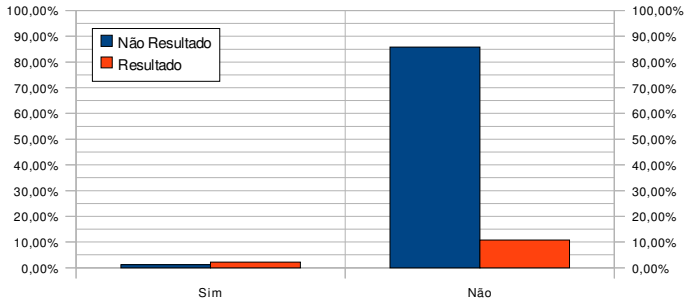


A análise do gráfico 5.18 demonstra que aproximadamente 75% dos resultados possuem algum tipo de número em sua composição, valor expressivo numa análise relativa, mas irrelevante quando se considera todo o universo de sentenças, vez que o valor encontrado para os não resultados representam, de forma absoluta, uma quantidade consideravelmente superior de sentenças.

p) Indicador de porcentagem

O experimento buscou por indicadores de porcentagem entre as sentenças do texto, como o símbolo "%", ou a palavra "percent". Os valores encontrados para cada tipo de sentença pode ser visualizado no gráfico 5.19.

Gráfico 5.19 - Análise da presença de indicador de porcentagem nas sentenças



É interessante observar que, em termos absolutos, do total de sentenças que possuem indicador de porcentagem, a maior parte é de resultados, mesmo esta categoria representando menos de 10% do total das sentenças do conjunto. De todas as sentenças existentes, 3,41% apresentavam este indicador. Deste valor, 1,26% eram enquadrados como não resultados e 2,15% como resultados.

5.3.2 Classificação Baseada Em aprendizado de Máquina

Seis diferentes algoritmos de classificação foram testados: Nayve Bayes, Tree J 4.8, Tabelas de Regras de Decisão (TRD), Máquinas de Vetor de Suporte (SVM), Vizinho mais Próximo (IBK) e um algoritmo baseado em Redes Neurais (Multlayer Perceptron - MP).

Para o teste, a definição do conjunto de testes e classificação apresentou duas variações. A primeira, com 30% do conjunto de documentos sendo utilizado como conjunto de treinamento, e a segunda, com 50%. O percentual restante das sentenças (70% e 50% do total, respectivamente) passou a compor o conjunto a ser classificado. O resultado do experimento pode ser visualizado na Tabela 5.1.

Considerando-se apenas a classificação das sentenças indicando resultado, o experimento realizado obteve como melhores índices uma abrangência (recall) de 39,7%, utilizando o método do vizinho mais próximo (IBK), e uma taxa de precisão de 55,6%, utilizando-se o algoritmo T J.48. É importante ressaltar, entretanto, que esta precisão foi obtida para uma taxa de recall de apenas 7,9%, e o recall de 39,7% foi encontrado para uma precisão de apenas 28,1%, tornando, por conseguinte, os valores encontrados pouco significantes.

Tabela 5.1 - Resultado da classificação das sentenças

Algoritmo		N. Bayes	N. Bayes	T. J.48	T. J.48	TRD	TRD	SVM	SVM	MP	MP	IBK	IBK
Tamanho Conjunto Treinamento		50,00%	30,00%	30,00%	50,00%	30,00%	50,00%	30,00%	50,00%	30,00%	50,00%	30,00%	50,00%
Instâncias classificadas de forma correta		88,68%	89,36%	92,05%	91,56%	91,92%	91,92%	91,03%	91,92%	89,49%	90,48%	86,92%	88,15%
Instâncias classificadas de forma incorreta		11,32%	10,64%	7,95%	8,44%	8,08%	8,08%	89,74%	8,08%	10,51%	9,52%	13,08%	11,85%
Resultados	**Precisão**	24,20%	26,20%	55,60%	25,00%	0,00%	0,00%	31,60%	0,00%	32,70%	36,70%	28,10%	28,60%
	Recall	17,80%	17,50%	7,90%	2,20%	0,00%	0,00%	9,50%	0,00%	28,60%	24,40%	39,70%	31,10%
Não Resultados	**Precisão**	92,90%	93,00%	92,50%	92,00%	91,90%	91,90%	31,60%	91,90%	93,80%	93,50%	94,50%	93,90%
	Recall	95,10%	95,70%	99,40%	99,40%	100,00%	100,00%	9,50%	84,50%	94,80%	96,30%	91,10%	93,20%
Média (Resultados e Não Resultados)	**Precision**	87,40%	87,60%	89,50%	86,60%	84,50%	84,50%	87,60%	84,50%	88,90%	89,00%	89,10%	88,60%
	Recall	88,90%	89,40%	92,10%	91,60%	91,90%	91,90%	91,00%	91,90%	89,50%	90,50%	86,90%	88,20%

Os resultados encontrados apontam, ainda, que, considerando a classificação dos dois tipos de sentença, a média de acerto e recall encontrada para qualquer combinação de algoritmo e conjunto de treinamento ficou acima dos 80%. Apesar de aparentemente satisfatório, este resultado não é relevante para este trabalho, pois o fato da categoria "não resultados" possuir 9 vezes mais elementos do que a categoria "resultados" distorce o valor final médio de precisão e recall encontrados no experimento. Tomando como base, por exemplo, o teste realizado utilizando o algoritmo Máquina de Vetores de Suporte, associado a um conjunto de treinamento formado por 50% do total de sentenças do corpus, encontramos um "valor médio" para precisão de 84,5 % e de 91,9% para o recall. Nestas mesmas condições, considerando-se apenas os valores para classificação das sentenças da categoria "resultado", encontramos um valor de 0% para precisão e recall.

Observando os resultados encontrados e considerando que devemos priorizar a abrangência em detrimento da previsão, selecionamos o método do vizinho mais próximo (IBK) como algoritmo de aprendizado de máquina. O principal motivo de se priorizar a abrangência é o fato de que a falta de um resultado causa muito mais prejuízos ao estudo do que o tempo que será gasto a mais pelo pesquisador em ler uma sentença que não é um resultado.

5.4 SELEÇÃO DOS PARÁGRAFOS

Como entrada para esta etapa são passadas as sentenças devidamente categorizadas e identificadas no texto como resultados ou como não resultados, em suas posições originais no texto. A partir desta informação, são selecionados os parágrafos do texto que possuem pelo menos uma sentença classificada como resultado, o conjunto destes parágrafos formará o sumário do texto.

5.5 CONCLUSÃO

Este capítulo apresentou e descreveu os detalhes de implementação do método de sumarização de texto, principal objetivo deste trabalho. O conjunto de técnicas, a sequencia de atividades, os detalhes de implementação, a seleção das técnicas mais apropriadas para a classificação e, sobretudo, a análise detalhada dos atributos de sentença, se apresentam como a principal contribuição deste estudo.

Para chegar ao conjunto final do método, fez-se necessário testar dezessete atributos de sentença e mais de cinco diferentes algoritmos de classificação de texto, que culminou na

proposta inovadora de classificação em duas etapas, integrando duas diferentes heurísticas de classificação, uma baseada em regras e outra em aprendizado de máquina.

Capítulo 6

Método Proposto

Neste capítulo será apresentado o protótipo de ferramenta desenvolvido com o objetivo de avaliar o modelo proposto no capítulo anterior.

Com o objetivo de avaliar o método proposto, fez-se necessária a construção de um protótipo de ferramenta para análise automática das sentenças. Dentre as funcionalidades da ferramenta, estão:

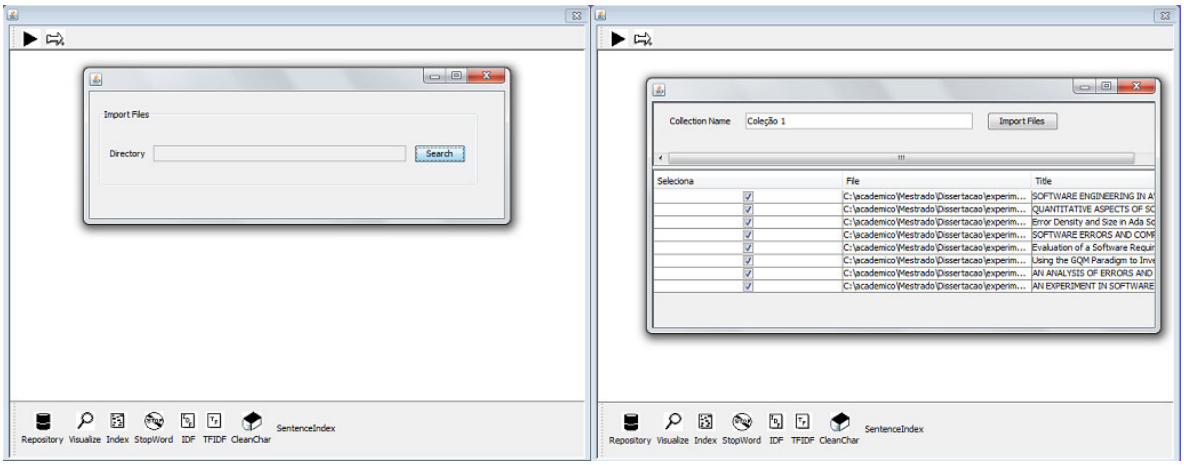
a) Importação de artigos e indexação dos textos em função das palavras e sentenças;
b) Possibilidade de aplicação automática de operações de mineração de texto, a exemplo da remoção de "Stop Words" e realização de "Stemming";
c) Permitir a marcação (anotação) manual das sentenças;
d) Efetuar o cálculo dos atributos de sentença;
e) Gerar uma arquivo no formato ARFF[15] a ser processado pela ferramenta Weka.

6.1 DESCRIÇÃO DAS FUNCIONALIDADES

A primeira funcionalidade construída foi a de importação dos artigos científicos e seu agrupamento em coleções de textos, conforme pode ser visualizado na Figura 6.1. O usuário seleciona o diretório onde os arquivos a serem importados estão armazenados e a ferramenta efetua todo o processamento do texto, sendo válidos apenas os arquivos com a extensão ".txt" e ".pdf". Nesta etapa os arquivos no formato proprietário ".pdf" são automaticamente convertidos para um formato que permite a edição, e as marcas de formatação, como negrito, itálico e correlatos, são removidas.

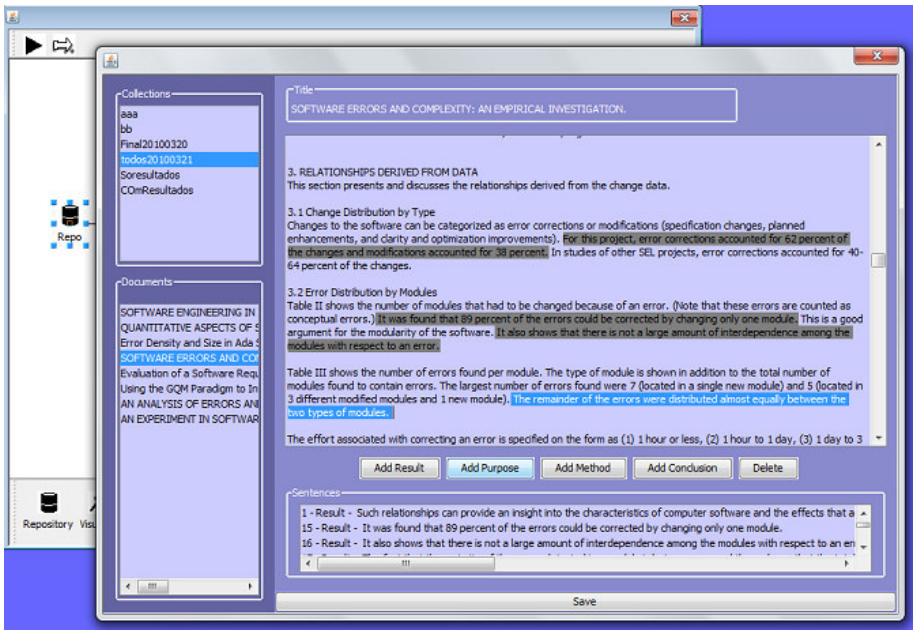
15 ARFF – tipo de arquivo em formato ASCII que descreve uma lista de instância a partir de seus atributos. É o modelo padrão de arquivo da ferramenta WEKA - http://www.cs.waikato.ac.nz/~ml/weka/arff.html

Figura 6.1 - Importação de Artigos e Criação de Coleções



Com os artigos já disponíveis na Ferramenta, a próxima funcionalidade é a de anotação manual das sentenças de forma a permitir a construção do conjunto de testes e a verificação automática da precisão da classificação. (Figura 6.2.)

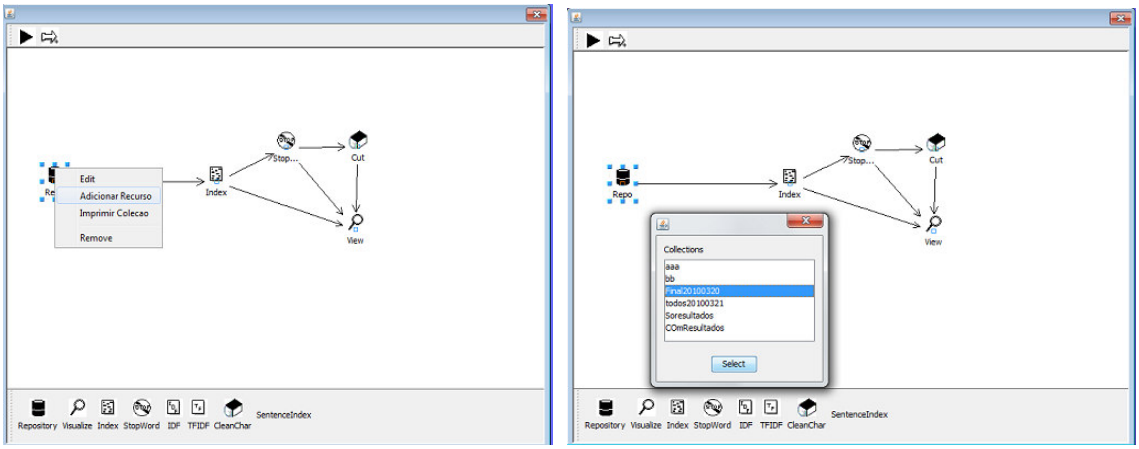
Figura 6.2 - Edição e Anotação de Artigos



Nesta etapa, a ferramenta permite que as sentenças do texto de cada um dos artigos importados sejam selecionadas e rotuladas em diferentes categorias. Como citado anteriormente, a anotação das sentenças (identificação manual da categoria em que se enquadra) é fundamental apenas quando a ferramenta for utilizada em modo de teste, ou seja, com o intuito de aferir, automaticamente, a precisão e abrangência do processo de classificação. Em modo de produção, não é necessário realizar esta etapa.

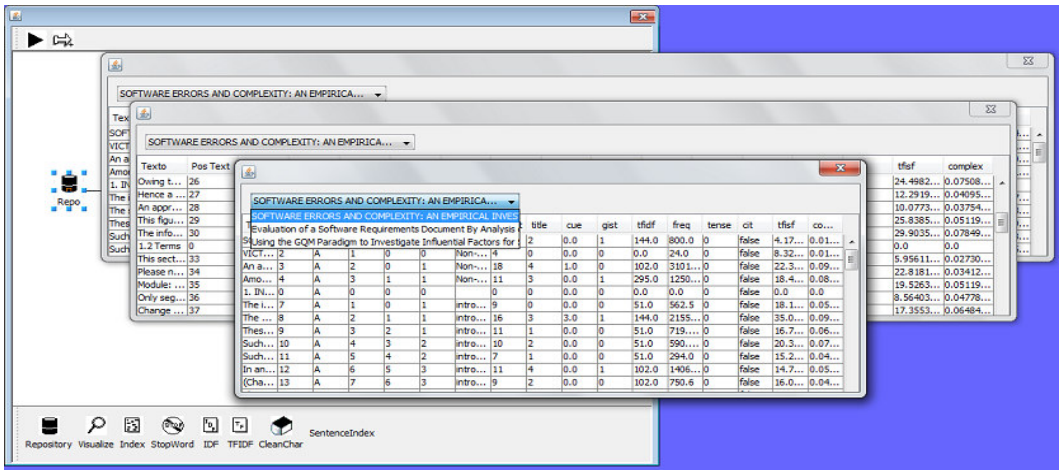
Com os conjuntos já definidos inicia-se, de fato, a etapa de processamento automático dos textos. A ferramenta permite a utilização de funcionalidades da mineração de texto, como a remoção de "StopWords", utilização de "Stemming" e remoção de caracteres indesejados. O modo como o texto será processado e por quais etapas deve passar é definido de acordo com o desejo do usuário, a partir da criação de um fluxo de trabalho, conforme pode ser visualizado na Figura 6.3.

Figura 6.3 - Construção do fluxo de trabalho e execução dos testes



Diversos fluxos podem ser construídos a partir de um único inicio, sendo possível visualizar o resultado de cada etapa do processamento e efetuar ajustes em tempo real. Este resultado do processamento, etapa a etapa, pode ser visualizado no formato de tabulação dos dados, como exibido na Figura 6.4.

Figura 6.4 - Tabulação e Visualização dos Resultados



Já o resultado final permite a visualização e exportação em formato ARFF, de forma a possibilitar a análise automática através da ferramenta Weka (Quadro 6.1). Esta funcionalidade é fundamental para que se possa dar continuidade ao processo e efetuar a etapa de classificação das sentenças.

Quadro 6.1 - Resultado resumido do processamento em formato ARFF

```
@ATTRIBUTE AbsolutLocation {A,B,C,D,E,F,G,H,I,J}
@ATTRIBUTE ComplexidadeLexica NUMERIC
@ATTRIBUTE CueMethod NUMERIC
(…)
@ATTRIBUTE HeadLine {introduction, implementation, example, conclusion, result, evaluation, solution, experiment, discussion, method, problems, "related work", data, "further work", "problem statement"}
(...)
@ATTRIBUTE Result {true,false}
@DATA
A,0.02736318407960199,0.0,1080.888888888889,0,2,?,0,1,0,0,0.0,4.288254020520038,false,false,false,false,,false,false,fal,(...)
```

6.2 CONCLUSÃO

São encontradas ferramentas de mineração de texto onde são efetuadas as operações básicas de remoção de stop words, redução de dimensionalidade e cálculo de atributos difundidos, como a frequência do termo (TF) ou seu derivado pela frequência inversa do documento TF-IDF.

O principal diferencial deste protótipo em relação às demais ferramentas encontradas diz respeito, em primeiro lugar, ao paradigma de mineração visual, que permite que o usuário

defina o seu fluxo de trabalho, visualize cada uma das técnicas aplicadas e acompanhe a saída de cada uma das etapas, permitindo a realização de ajustes das variáveis dependentes em tempo real e a percepção do impacto imediato nos resultados. O outro diferencial é o fato de possuir uma quantidade maior de atributos dos termos e permitir que o usuário possa não apenas implementar um novo atributo como criar novos elementos para utilizar no fluxo de trabalho, a partir da extensão dos modelos existentes.

Capítulo 7

Experimento

Este capítulo descreve o experimento executado para validar o método proposto. Aqui são explicitadas as variáveis, questões, hipóteses e são descritos os procedimentos para realização do procedimento..

No capítulo anterior, descrevemos detalhadamente o método proposto neste estudo para categorização de sentenças e sumarização de texto com base nos resultados dos estudos selecionados. Após as definições, faz-se necessário avaliar, com base em aspectos objetivos, se os resultados da aplicação do método coadunam com a sua proposta e quais os índices de abrangência e precisão que apresentam.

Desta forma, com o intuito de avaliar o método proposto, desenhamos um experimento controlado, com a definição das questões a serem observadas, das métricas a serem utilizadas, das variáveis, recursos necessários e fluxo de execução. Em seguida, executamos o experimento, capturamos e discorremos sobre os resultados encontrados, conforme será apresentado nas seções subsequentes.

7.1 DEFINIÇÃO DO EXPERIMENTO

Com o objetivo principal de avaliar a precisão, abrangência e efetividade da utilização do método de identificação automática de sentenças proposto neste trabalho, fez-se necessária a montagem e execução de um experimento.

Este experimento foi conduzido em ambiente acadêmico, com o objetivo de analisar a técnica proposta e com o intuito de avaliar a sua efetividade no que diz respeito a identificação de sentenças que indicam resultados em artigos científicos. Para isto, será levado em consideração a precisão na identificação das sentenças e a quantidade de sentenças localizadas em relação ao total existente.

7.1.1 Questões

São três principais questões relacionadas ao experimento, as quais:

a) Pode um método de identificação automática de resultados em artigos científicos obter melhor abrangência na classificação das sentenças do que os 50% obtidos por alunos no experimento de Cruzes (2007)?

b) O isolamento manual das variáveis utilizadas na classificação das sentenças e a análise de sua distribuição pode auxiliar na remoção de ruído e melhorar o processo de classificação?

c) É possível reduzir a quantidade de texto dos artigos sem que se reduza proporcionalmente a quantidade de resultados no texto final? (proporcionalmente, se reduzirmos em 30% um texto, tomando como base que a distribuição dos resultados é uniforme, perderíamos 30% dos resultados).

7.1.2 Métricas

As métricas adotadas neste experimento são:

a) Abrangência (recall) – quantidade de sentenças corretamente identificadas em relação ao total de sentenças existentes em uma determinada categoria (total localizado / total existente).

b) Precisão (precision) – quantidade de sentenças corretamente classificados em função do total de sentenças classificadas como pertencentes a uma determinada categoria (corretamente classificados / corretamente classificados + falso positivos).

7.1.3 Seleção de Variáveis

A lista de variáveis independentes é composta por:

a) Cálculo dos Atributos da Sentença

b) Tamanho do Conjunto de Treinamento

c) Algoritmo utilizado na classificação

A lista de variáveis dependentes é composta por:

a) Precisão

b) Abrangência

7.1.4 Recursos

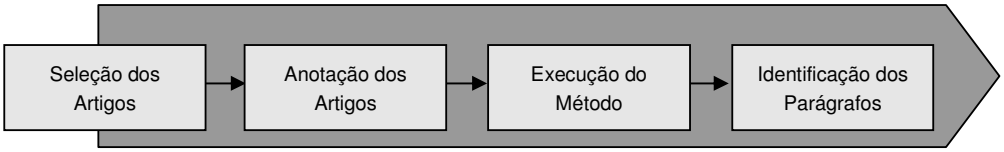
Para o desenvolvimento do experimento, serão necessários os seguintes recursos:

- Software: Ferramenta para Pré-processamento de documentos textuais (construída neste estudo), Ferramenta de Mineração de Texto Weka.

7.2 PLANEJAMENTO DO EXPERIMENTO

São três as principais etapas para execução do experimento. A lista e ordem de execução podem ser visualizados na Figura 7.1.

Figura 7.1 - Diagrama de Fluxo do Experimento



A primeira etapa contempla a seleção dos artigos que irão compor o conjunto de testes e o de classificação, oriundos do Corpus de Cruzes (2007). Após a escolha dos artigos, que devem ser da área de Engenharia de Software, o próximo passo é definir em qual categoria cada uma das sentença se enquadra. Neste estudo, lidamos com duas categorias: “resultados” e “não resultados”. O texto deve ser marcado com a categorias das sentenças de modo que o processo de treinamento e de validação da classificação sejam automáticos.

A terceira etapa, e principal, é a de execução do método. Nesta fase os atributos de cada uma das sentenças são calculados utilizando a ferramenta Textum, descrita em capítulo anterior, e o texto é exportado para um formato compatível com a ferramenta de classificação Weka, que deve ser configurada com o algoritmo de classificação do vizinho mais próximo e percentual de conjunto de treinamento de 30%. O conjunto de entrada é enviado ao classificador e, após execução do processo, são definidas as categorias das sentenças. Estes valores foram definidos em estudo preliminar que efetuou testes de abrangência e precisão em 5 artigos capturados do Corpus de Cruzes (2007) e retirados, por conseguinte, do conjunto de documentos a ser utilizado neste experimento. Os testes prévios utilizaram os seguintes valores para as variáveis dependentes:

d) Tamanho do Conjunto de Treinamento: 30% ou 50% do conjunto de entrada.

e) Algoritmos utilizados na classificação: Nayve Bayes, Tree J 4.8, Tabelas de Regras de Decisão (TRD), Máquinas de Vetor de Suporte (SVM), Vizinho mais Próximo (IBK) e um algoritmo baseado em Redes Neurais (Multlayer Perceptron - MP).

A última etapa consiste em selecionar no texto original os parágrafos que contém sentenças classificadas como "resultado", de modo a compor o texto final, resumido.

7.2 EXECUÇÃO DO EXPERIMENTO

Um dos subprodutos do trabalho de Cruzes (2007) foi a construção de um pequeno Corpus de artigos de Engenharia de Software, cujas sentenças foram classificadas manualmente, por pesquisadores, como pertencentes a categoria de "resultados" ou "não resultados". Em função desta característica, elegemos este corpus para selecionar os artigos a serem utilizados no experimento. É interessante ressaltar que os artigos deste corpus utilizados para definição das regras utilizadas no algoritmo foram excluídos e não compõe o conjunto utilizado no experimento. A lista dos artigos escolhidos encontra-se no Anexo I.

Toda a etapa de anotação e parte da etapa de execução do método foram realizadas utilizando-se uma ferramenta desenvolvida neste estudo – denominada *Textum*, e descrita no Capítulo 6. Atividades como importação dos artigos e conversão para o formato de texto editável, anotação das sentenças, cálculo dos atributos, a etapa de classificação baseada em regras e a geração do arquivo compatível com a ferramenta de classificação Weka também foram executadas na ferramenta *Textum*.

A parte final do experimento consistiu em fornecer o arquivo de entrada à Ferramenta Weka, definir os atributos de classificação conforme especificados no planejamento do experimento e executar o classificador. De posse das sentenças selecionadas como "resultado", foram identificados os parágrafos que as continham e definido o texto final. A última atividade foi a de quantificar a redução do texto final em relação ao inicial e os percentuais de acerto e abrangência na classificação.

7.3 RESULTADOS DO EXPERIMENTO

De modo a facilitar o entendimento do processo, separamos os resultados encontrados para cada uma das etapas de classificação: a classificação por regras e a classificação por aprendizado de máquina.

O resultado da classificação com base em regras foi capaz de definir a classificação de 74% do total de sentenças do texto com uma precisão de 82%. É interessante observar, ainda, que 27% das sentenças de classificação definidas como "resultado" foram incorretamente classificadas nesta etapa, deixando para a próxima etapa um percentual máximo de abrangência de 73%.

O conjunto textual restante, correspondente a 26% das sentenças do conjunto inicial, foi, então, levado à classificação automática na ferramenta Weka, utilizando-se o algoritmo vizinho mais próximo. Os resultados da classificação podem ser observados na Tabela 7.1.

Tabela 7.1 - Classificação das sentenças após aplicação do conjunto de regras 1, 2 e 3.

	Conjunto de regras 1 + Conjunto de regras 2	Conjunto de regras 1 + Conjunto de regras 3
Acerto	**47,00%**	**56,50%**
Recall	**60,00%**	**41,20%**

Em uma primeira análise, o resultado encontrado na tabela 7.1 parece ser insatisfatório. Vale ressaltar, entretanto, que no experimento descrito por Cruzes (2007), alunos analisaram manualmente os artigos e conseguiram capturar apenas 50% dos resultados do texto, devendo ser este o patamar de comparação deste método.

Antes de analisarmos os resultados da seleção dos parágrafos, convém comentar um fato importante observado no resultado final da classificação das sentenças. O que se percebeu após a classificação foi um conjunto final no qual as sentenças constantes eram extremamente semelhantes, seja em relação às características analisadas pelo computador ou até mesmo em análise efetuada por humanos. Isso se comprova através da análise das sentenças, que demonstrou que os elementos da categoria *não resultados* presentes no conjunto possuem palavras chave, números e informações similares às encontradas na

categoria de *resultados.* No entanto, com o objetivo distinto de explanar uma conclusão ou explicitar como foi feita a coleta de determinado dado para se chegar ao resultado. De certa forma, estas sentenças classificadas incorretamente forneciam o contexto aos resutados apresentados.

Voltando aos resultados da seleção dos parágrafos, a utilização do método proposto apresentou uma taxa de acerto de 74% e de recall de 72%, sendo que o conjunto textual formado pelos parágrafos selecionados representaram, aproximadamente, 20% do texto, ou seja, uma considerável redução na quantidade de informação a ser lida pelo pesquisador com consequente redução do tempo e custo para realização do trabalho.

7.4 CONCLUSÃO

Os resultados do experimento evidenciaram taxas de abrangência e precisão na localização de sentenças que indicam resultados do artigo científico próximas às descritas por Cruzes (2007) em experimento realizado com alunos da Universidade de Maryland. O principal diferencial diz respeito ao tempo gasto para realização do procedimento, enquanto o processamento computacional se posiciona na casa de minutos, a leitura dos textos por humanos possui ordem de grandeza em horas. Não é possível estabelecer um valor exato de tempo, já que o processamento computacional é dependente do equipamento que se está utilizando para a execução do método e, assim como o humano, são dependentes do tamanho do conjunto de entrada (artigos).

Levando em consideração o método de seleção automática dos parágrafos, com base nos resultados encontrados, podemos estimar um tempo 80% menor para a localização dos resultados em comparação à seleção por humanos, levando em conta igual redução no tamanho do conjunto original de artigos. Esta constatação necessita, entretanto, de comprovação através de estudo complementar.

Capítulo 8

Resultados e Discussão

Este capítulo agrupa os resultados dos experimentos realizados neste estudo. Contém, ainda, informações relevantes do processo de execução e ajuste do método proposto para classificação das sentenças como "resultado" ou "não resultado", e identificação dos parágrafos relevantes do texto científico.

O conjunto de procedimentos para definição do método de sumarização de texto, objetivo principal deste estudo, mostrou o quanto é complexa a atividade de processar e analisar dados em linguagem natural, corroborando fato que já é consenso entre os pesquisadores da área de Recuperação da Informação.

A dificuldade se agrava quando lidamos com informações que são ambíguas até mesmo para os próprios seres humanos. Cruzes (2007) realizou um experimento em que fornecia alguns artigos a grupos de alunos para que estes identificassem no texto as sentenças que representavam resultados dos artigos. O resultado demonstrou que, em média, apenas 50% do total de sentenças existentes nos textos foi localizada corretamente.

Agarwal (2010) realizou um experimento em que era necessário definir a classificação das sentenças dos artigos. Para tanto, forneceu o mesmo texto a três diferentes especialistas neste tipo de identificação e, mesmo entre estudiosos da área, houve divergência na definição da categoria de parte das sentenças, conflito resolvido apenas com a interação entre os três membros do experimento.

É neste contexto que propomos um método para classificação automática de sentenças do texto científico, identificação dos parágrafos e sumarização do texto, baseado na análise da estrutura do artigo, na definição de atributos de sentença e na observação de métodos similares.

8.1 ANÁLISE DA ESTRUTURA DO ARTIGO EM ENGENHARIA DE SOFTWARE

O levantamento realizado comprovou que, de uma forma geral, os artigos científicos na área

de saúde tem seu conteúdo estruturado com base na separação do texto em quatro seções básicas: Introdução, Método, Resultados e Discussão. Esta prática se consolidou a partir do momento em que grande parte dos mais conceituados jornais científicos passaram a dar preferência a textos escritos neste formato, explicitamente recomendado pelo Comitê Internacional de Editores de Jornais Médicos. Estudo realizado por Teufel e Moens (2002) confirma a padronização dos estudos em Saúde. A análise de 108 artigos da área de cardiologia demonstrou que 92% dos textos possuíam as quatro seções.

Quando lidamos com artigos de outras áreas, entretanto, a situação é oposta. Ainda no estudo de Teufel, a pesquisadora analisou 80 artigos de Linguística Computacional e observou que apenas 20% apresentavam seção de resultados, 16% apresentavam seção de discussão e nenhum dos textos analisados possuía uma seção descrevendo o método.

Com o intuito de avaliar a organização dos artigos da área de Engenharia de Software, foco desta dissertação, selecionamos 33 artigos para análise estrutural. O resultado encontrado foi semelhante ao apresentado por Teufel para os artigos de Linguística Computacional, com apenas 15% dos artigos apresentando uma seção de resultados e nenhum apresentando uma seção de método.

Em complemento a esta análise, fizemos um levantamento de alguns dos principais modelos de chamada para publicação de artigos de congressos e jornais na área de Engenharia de Software. A observação dos modelos demonstrou que estes não trazem exigências formais em relação a modelagem do conteúdo e sua organização em seções. O que se percebe, apenas, são imposições de ordem estilística que priorizam a apresentação visual dos artigos.

A utilização de um padrão de organização do conteúdo é um facilitador para a localização das informações desejadas não apenas para os leitores dos artigos, mas para algoritmos de recuperação da informação. Seguir ou não este padrão pode representar uma diferença substancial nos resultados da aplicação de métodos automáticos de mineração de texto. O fato dos artigos de Engenharia de Software, por padrão, não seguirem um modelo de organização do conteúdo é mais um fator dificultador na construção de um método de classificação de sentenças neste tipo de texto.

8.2 ANÁLISE DOS MÉTODOS DE CLASSIFICAÇÃO DE SENTENÇAS

Este estudo analisou três métodos de classificação de sentenças em artigos científicos. É

importante ressaltar que as técnicas analisadas tem um escopo um pouco mais amplo do que o definido nesta dissertação, já que classifica as sentenças em mais de duas categorias, conquanto trabalhamos, nesta pesquisa, apenas com as categorias “resultado” e “não resultado”.

Em relação ao método proposto por Teufel (1999), observamos um pequeno desvio no resultado declarado para as sentenças classificadas na categoria “OWN” em função da quantidade muito superior de sentenças desta categoria em comparação a quantidade das demais. Desta forma, acatamos a precisão do método pelo resultado encontrado para as demais categorias, por volta de 50% de acerto. Vale ressaltar que estes resultados foram encontrados para um conjunto de artigos seguindo o padrão IMRaD, já que não foi possível testar o método com artigos de Engenharia de Software, pois não tivemos acesso à ferramenta utilizada pela autora e nem a uma descrição suficientemente detalhada que nos permitisse a implementação do método.

O estudo de Ibekwe-SanJuan (2008) não apresentou resultados estatísticos de validação para o método descrito pela autora. Um experimento realizado com ferramenta disponibilizada por Ibekwe-SanJuan tendo como entrada artigos de Engenharia de Software apresentaram precisão inferior a 10% na classificação das sentenças indicando “resultados”.

O último estudo analisado, de Agarwal (2009), apresentou precisão superior a 75% na classificação das categorias e abrangência superior a 75% na localização dos resultados. Estes resultados, entretanto, foram conseguidos com a utilização de artigos da área de saúde, estruturados conforme o modelo IMRaD. Experimento realizado utilizando artigos de Engenharia de Software como entrada para o algoritmo proposto obteve uma precisão de 25% na identificação dos resultados.

O que se percebe é que os métodos analisados não apresentaram resultados satisfatórios quando utilizados na classificação de sentenças em artigos científicos que não seguem padrões para estruturação do conteúdo. Os valores encontrados ficam muito aquém dos levantados no experimento de Cruzes (2007) com seres humanos, o que inviabiliza a utilização destes algoritmos para realização desta tarefa em ambiente não acadêmico.

8.3 DEFINIÇÃO DO MÉTODO DE SUMARIZAÇÃO DE TEXTO

A base do método proposto é a classificação das sentenças em duas categorias base: "resultados" e "não resultados". A idéia foi a de construir, como núcleo do método, um algoritmo similar ao de Teufel (1999), com base na definição de atributos de sentença e sua

utilização na classificação utilizando uma das técnicas de aprendizado de máquina. Os resultados iniciais não foram animadores, já que os índices de precisão se estabilizaram em pouco menos da metade do encontrado pelos alunos participantes do experimento de Cruzes (2007).

Uma análise detalhada da distribuição dos valores dos atributos em gráficos representou o início do processo de ajuste do método. A identificação e eliminação de atributos que geravam ruídos associado a criação de uma primeira etapa de classificação, baseada em regras, mais do que duplicou a performance do classificador elevando-a a índice superior ao de referência, do já citado experimento de Cruzes (2007), de 50% de acerto.

É interessante ressaltar, ainda, que a validação do método foi realizada através de um experimento controlado. De forma a dar transparência e facilitar a análise, elencamos os diversos parâmetros utilizados, definimos e documentamos as variantes de análise e a sequência de procedimentos utilizada tanto na montagem quanto na execução do experimento.

8.3.1 Ferramenta para pré-processamento textual

É comum entre os pesquisadores da área de recuperação da informação a utilização de ferramentas que atuam tanto na classificação automática quando na preparação dos mais diferentes dados. A grande preocupação, entretanto, ocorre com o primeiro tipo de ferramenta, já que a etapa de pré-processamento é algo secundário, de preparação para a etapa principal, que é a classificação. Neste trabalho, enfatizamos o processo de preparação dos dados de modo a reduzir, ao máximo, possíveis ruídos gerados nesta etapa e que influenciam negativamente o processo de classificação.

Como subproduto do processo de validação do método, fez-se necessária a construção de um protótipo de ferramenta para pré-processamento visual de documentos textuais, cuja especificação encontra-se no Apêndice A. As atividades desde a importação dos artigos, formatação de caracteres, remoção de "stop words", redução de dimensionalidade e diversas outras atividades relacionadas à mineração de texto foram implementadas de forma visual. Com um simples clicar do mouse pode-se criar um novo workflow, inserir diversas origens de dados, ajustar os mais variados parâmetros e observar a saída de cada uma das etapas.

O principal diferencial desta ferramenta em relação às demais produzidas é o conceito da montagem visual do workflow. A possibilidade de se criar dois, três, quatro caminhos

simultâneos com parâmetros e objetos diferentes, comparar a saída entre estes e identificar o melhor ajuste, apenas clicando e arrastando objetos é algo que em muito facilita o trabalho do pesquisador. A utilização desta ferramenta – denominada de "Textum", nos fez economizar tempo considerável no ajuste do método.

Capítulo 9 Conclusão

Este capítulo apresenta as conclusões sobre os resultados deste estudo e aponta para as linhas a serem seguidas em trabalhos futuros.

Nas últimas décadas, tem crescido vertiginosamente a quantidade de revisões sistemáticas produzidas. Esta técnica, que até então era utilizada apenas em trabalhos da área de saúde, avançou para outras áreas, a exemplo da Engenharia de Software. A grande quantidade de tempo e esforço gasto na construção de uma revisão sistemática tem agido como um dificultador a uma maior expansão e utilização desta técnica.

Ferramentas e métodos computacionais vêm sendo densenvolvidos com o intuito de tornar mais ágil o trabalho dos pesquisadores que constroem estas revisões, como o trabalho de Malheiros (2007), que propõe uma ferramenta para auxiliar na etapa da revisão em que são selecionados os artigos que irão compor o estudo final, ou o de Silva Rocha (2009), cuja ferramenta construída auxilia na localização automática de informações de contexto no texto científico.

O objetivo deste estudo é a definição de um método, com construção de um protótipo de ferramenta, para auxiliar os pesquisadores na localização das sentenças que indicam resultado no texto científico, um dos elementos base para a construção da revisão e um dos que mais ocupa o tempo dos pesquisadores.

Dos três estudos que tratavam da anotação automática da semântica das sentenças, analisados no capítulo 3, nenhum obteve resultados satisfatórios quando o conjunto de entrada era formado por artigos que não seguiam o modelo IMRaD, principal problema dos artigos de outras áreas que não a de saúde, a exemplo da Engenharia de Software.

Conquanto as demais técnicas testadas se restringiram a um nível de acerto inferior a 30%, para um conjunto de artigos oriundos da ES, o método proposto neste estudo superou os 47% de acerto para um nível de recall superior a 60%. Este resultado não ficou muito distante do encontrado no experimento realizado por Cruzes (2007), com a utilização de seres humanos não especialistas para identificar manualmente os resultados. O grande diferencial se

dá em relação ao tempo gasto para realização da atividade, já que o processamento computacional durou cerca de dois minutos, consideravelmente inferior ao encontrado na pesquisa com seres humanos.

Os índices encontrados, entretanto, ainda estavam abaixo do esperado para um algoritmo de recuperação da informação. Percebeu-se, ainda, que as sentenças selecionadas, para serem compreendidas pelos pesquisadores, precisariam de informações de contexto, presente nas sentenças adjacentes. Desta forma, ao invés de marcar no texto apenas as sentenças indicando resultados, resolveu-se que os parágrafos que continham as sentenças resultado seriam selecionados. O método proposto permitiu um nível de acerto de 74% e um recall de 72% na identificação dos parágrafos compostos por resultados.

A seleção apenas dos parágrafos reduziu para 1/5 a quantidade de texto a ser analisado pelo pesquisador o que, em tese, ensejaria uma redução de 80% no tempo gasto na análise do artigo. É necessário, entretanto, realizar em um estudo futuro, um experimento para comprovar o valor encontrado.

O método desenvolvido neste artigo teve como foco artigos científicos na área de Engenharia de Software. A aplicação em artigos de outras áreas, sobretudo a de saúde, não foi testada por não possuirmos um conjunto anotado de artigos, condição essencial para execução do experimento. Se o comportamento do método proposto for similar ao dos outros métodos analisados, pode-se inferir que os resultados da aplicação do método neste novo contexto serão ainda melhores do que os obtidos com o processamento de artigos de Engenharia de Software.

Como trabalho futuro, faz-se necessário, ainda, construir um corpus de artigos científicos com anotação dos tipos de sentença mais robusto. A criação e compartilhamento desta base de dados contribuiria substancialmente para o avanço dos estudos em recuperação da informação em artigos. O conjunto utilizado neste trabalho, oriundo do estudo de Cruzes (2007), possuía quantidade reduzida de artigos o que pode ter influenciado negativamente no resultado da classificação.

R Referências

ABRAHÃO, P. R. C. **Modelagem e Implementação de um léxico semântico para o português**. 1995. Dissertação (Mestrado) - Pontifícia Universidade Católica do Rio Grande do Sul, Instituto de Informática, Porto Alegre, 1995

AGARWAL, S.; Yu, H. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. In: AMIA SUMMIT ON TRANSLATIONAL BIOINFORMATICS. **Proceedings,** 2009.

ÁRVORE de decisão. In: WIKIPÉDIA, a enciclopédia livre. Flórida: Wikimedia Foundation, 2011. Disponível em: <http://pt.wikipedia.org/w/index.php?title=%C3%81rvore_de_decis%C3%A3o&oldid=26382926>. Acesso em: 4 set. 2009.

ATALLAH, A. N.; CASTRO A. A. **Revisão sistemática e metanálises, em evidências para melhores decisões clínicas**. São Paulo. Lemos Editorial 1998.

BASILI, Victor R. The past, present, and future of experimental software engineering. **Journal of the Brazilian Computer Society (JBCS),** dec., 2006.

BAXENDALE, P. B. Machine-made index for technical literature: an experiment. **IBM Journal of Research and Development**, v.2, n.4, p.354–361, 1958.

BIOLCHINI, J.et al. **Systematic review in software engineering:** relevance and utility, **Technical Report ES67905.** Rio de Janeiro: PESC – COPPE/UFRJ, 2005.

CRUZES, D. S. **Análise secundária de estudos experimentais em engenharia de software**. 2007. Tese (Doutorado em Engenharia Elétrica)- Faculdade de Engenharia Elétrica e Computação, Universidade de Campinas, 2007.

DE WAARD, A. A pragmatic structure for research articles. In: INTERNATIONAL CONFERENCE ON THE PRAGMATIC WEB, 2. 2007. **Anais…** 2007.

EDMUNDSON, H. P. New methods in automatic extracting. **Journal of the Association for Computing Machinery,** v. 16, n.2, p. 264-285, apr. 1969.

ENGENHARIA DE SOFTWARE. In: WIKIPÉDIA, a enciclopédia livre. Flórida: Wikimedia Foundation, 2011. Disponível em: <http://pt.wikipedia.org/w/index.php?title=Engenharia_de_software&oldid=19544477>. Acesso em: 10 abr. 2010.

FRAKES, W. ; BAEZA-YATES, R. **Information retrieval, data structures and algorithms**. New York: Prentice Hall, Englewood Cliffs, N.J., 1992.

HACHEY, B.; Grover, C. Sequence modelling for sentence classification in a legal summarisation system. In: ACM symposium on Applied computing, 2005. Santa Fe, New Mexico, **Proceedings…** anta Fe, New Mexico2005.

HALL, M. et al. The WEKA data mining software: an update**. SIGKDD Explorations**, v.1, n.1, 2009.

HUCKA, M. et al.**System biology markup language (SBML) level 1:** structures and facilities for basic model definitions. Disponível em <http://www.sbml.org/specifications/sbml-level-1/version-2/sbml-level-1-v2.pdf>. Acesso em: 10 maio 2009.

HUI, B. **Developing and evaluating a document visualization system for information management**. Master Thesis. Toronto : University of Toronto, 2002

IBEKWE-SANJUAN, F.; CHEN, C.; PINHO, R. Identifying strategic information from scientific articles through sentence classification**.** In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION CONFERENCE (LREC-08), 6., 2008. Marrakesh, Morocco. **Proceedings** Marrakesh, Morocco, 2008.

ICME. **Uniform requirements for manuscripts submitted to biomedical journals**: writing and editing Biomedical publications. Disponível em: <http://www.icmje.org>. Acesso em: 18 jun. 2009.

JOACHIMS, T. Text categorization with support vector machines: Learning with many relevants features. In: EUROPEAN CONFERENCE ON MACHINE LEARNING (ECML'98), 10., 1998. Chemnitz, Germany. **Proceedings…**, Chemnitz, Germany, 1998.

KHOO, A.; MAROM, Y.; ALBRECHT, D. Experiments with sentence classification. In: AUSTRALASIAN LANGUAGE TECHNOLOGY WORKSHOP. 2006. **Proceedings…** 2006.

KITCHENHAM, B. **Procedures for undertaking systematic reviews. Tech. Report TR/SE-0401.** Australia: Department of Computer Science, Keele University, Australia Ltd., 2004.

KUPIEC, J.; PEDERSEN, J.; Chen, F. A trainable document summarizer. **Proc. SIGIR, ACM Press,** 68–73, 1995.

LACERDA, W. S.; BRAGA, A. P. Experimento de um classificador de padrões baseado na Regra Naive de Bayes. In: SECICOM, 6., 2004. Lavras. **Anais…** Lavras, 2004

LAROCCA NETO, J. et al. Generating text summaries through the relative importance of topics. In: INTERNATIONAL JOINT CONFERENCE IBERAMIA/SBIA. 2000. Atibaia, SP. Proceedings… Atibaia, SP, 2000.

LIN, C. Training a selection function for extraction**.** In: ANNUAL INTERNATIONAL ACM CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 18., 1999. **Proceedings…** 1999.

LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal of Research and Development**, n.2, p.157-165, 1958

MALHEIROS, V. D. et al. A visual text mining approach for systematic reviews. In: INTERNATIONAL SYMPOSIUM. INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT, 2007. ESEM 2007. 2007, Madrid, Espanha. Proceedings of the First Washington, DC, EUA : IEEE Computer Society, 2007. p. 245-254, 2007.

MARCONDES, C. H; MENDONÇA, M. A. R; MALHEIROS, L. A estrutura dos elementos de metodologia científica no textos de artigos de periódicos eletrônicos em Ciências da Saúde. In: INTERNATIONAL CONFERENCE ON HEALTH INFORMATION AND LIBRARIES, 9., 2005. Salvador. **Proceedings…** Salvador, 2005.

MARCU, D. Improving summarization through rhetorical parsing tuning. In: WORKSHOP ON VERY LARGE CORPORA, 6., 1998. Montreal, Canadá. **Anais...** Montreal, Canadá,1998.

MARTINS, C. B. et al. **Introdução à sumarização automática**. Relatório Técnico RT-DC 002/2001. São Carlos-SP: Departamento de Computação, Universidade Federal de São Carlos, 2001

MARTINS JUNIOR, J. **Classificação de páginas na internet**. 2003. Dissertação. (Mestrado em Ciências da Computação e Matemática Computacional)- Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, São Carlos, 2003.

MCKNIGHT, L.; SRINIVASAN, P. Categorization of Sentence Types in Medical Abstracts. In: AMIA SYMPOSIUM, 2003. **Anais…** 2003.

MURRAY-RUST, P; RZEPA H. S. Chemical Markup, XML and the worldwide web. I: basic principles. **Journal of Chemical Information and Computer Science**, 1999.

MURRAY-RUST, P; RZEPA H. S. STMML. A markup language for Scientific, Technical and Medical Publishing. **Data Science Journal,** 2002.

OLIVEIRA, G. L. **Categorização automática de documentos por múltiplos classificadores Naive Bayes**. 2004. 157 f. Dissertação. (Mestrado em Redes de Computadores, área de concentração Gestão do Conhecimento)- Universidade Salvador, Salvador, 2004.

OLIVEIRA, F. A. D. **Processamento de linguagem natural:** princípios básicos e a implementação de um analisador sintático de sentenças da língua portuguesa. Disponível em:

<http://www.inf.ufrgs.br/gppd/disc/cmp135/trabs/992/Parser/parser.html>. Acesso em: 23 jan.2010.

PAICE, C. D. Another stemmer. **SIGIR Forum**, v.24, n.3, p.56-61, 1990.

PARDO, T. A. S. **Gistsumm:** um sumarizador automático baseado na idéia principal de textos. São Paulo: Núcleo Interinstitucional de Linguística Computacional, 2002. (Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional).

PARDO, T. A. S; RINO, L .H. M.; NUNES, M. G. V. NeuralSumm: uma abordagem conexionista para a sumarização automática de textos. In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL – ENIA, 4., 2003. Campinas-SP, **Anais...** Campinas-SP, 2003

PEARSON, K. **Report on certain enteric fever inoculation statistics**. BMJ 3:1243-1246, 1904.

PETROSINO, A. **Lead authors of cochrane reviews:** survey results. Report to the Campbell Collaboration, University of Pennsylvania, Cambridge, MA, 1999.

PORTER, M.F. An algorithm for suffix stripping**. Program,** v. 14, n.3, 130-137, 1980.

RUCH, P. A et al. **Using discourse analysis to improve text categorization in MEDLINE**. Medinfo, 2007.

SALTON, G. ; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information Processing & Management,** v.24, n.5, p. 513–523, 1988.

SEBASTIANI, F. **Machine learning in automated text categorization**. Technical Report IEI B4-31-12-99, Instituto di Elaborazione della Informazione, Consiglio Nazionale delle Ricerche, Pisa, Itália, 1999.

SILVA, L.K. Avaliação tecnológica e análise custo-efetividade em saúde: a incorporação de tecnologias e a produção de diretrizes clínicas para o SUS**. Ciência e Saúde Coletiva,** v. 8, n.2), p.501-520, 2003.

SILVA ROCHA, M. C. **Contextextractor:** uma ferramenta de apoio para a extração de informações de contexto de artigos de engenharia de software experimental. 2009. Dissertação. (Mestrado Profissional em Sistemas e Computação)-Universidade Salvador, 2009.

SKOROKHOD'KO, E. F. Adaptive method of automatic abstracting and indexing. In: IFIP CONGRESS. 1972. Ljubljana, Yugoslavia **Proceedings…** . North Holland, 1972.

STATSOFT. **Naive bayes classifier:** introduction overview. Disponível em: <http://www.statsoft.com/textbook/stnaiveb.html> Acesso em: 14 ago.2009.

TAN, A.H. Text mining: the state of the art and the challenges. In: PAKDD'99 WORKSHOP ON KNOWLEDGE DISOCOVERY FROM ADVANCED DATABASES. 1999. Beijing. **Proceedings…** Beijing, 1999.

TBAHRITI, I. et al.. Using argumentation to retrieve articles with similar citations: an inquiry into improving related articles search in the MEDLINE digital library. **International Journal of Medical Informatics**, v.75, n.6, p.488–495, 2006.

TCHEMRA, A. H. **Tabela de decisão adaptativa na tomada de decisões multicritério**. 2009. 172 p. Tese (Doutorado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais, São Paulo, 2009.

TEUFEL, S. ; MOENS, M. Sentence extraction as a classification task. In: WORKSHOP 'INTELLIGENT AND SCALABLE TEXT SUMMARIZATION, ACL/EACL, jul. 1997.

TEUFEL, S.; MOENS, M. Discourse-level argumentation in scientific articles: human and automatic annotation. In: TOWARDS standards and tools for discourse tagging. ACL 1999 Workshop, 1999.

TEUFEL, S.; MOENS, M**.** Summarising scientific articles: experiments with relevance and rhetorical status. In: **Journal Computational Linguistics - Summarization archive,** v.28 n. 4, dec. 2002.

TORRES, J.A.S. **Proposta de arquitetura para construção de perfis de usuário através da mineração da Web**. 2005. Monografia (Graduação em Análise de Sistemas) - Universidade do Estado da Bahia, Salvador, 2005.

XU, R. et al. Combining text classification and hidden markov modeling techniques for structuring randomized clinical trial abstracts. In: ANNUAL SYMPOSIUM OF AMIA.2006. **Proceedings…** 2006.

YANG Y. ; LIU X. A re-examination of text categorization methods. In: ACM INT. CONF. ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (SIGIR'99), 22., 1999. **Proceedings…** Berkeley, CA, 1999.

YANG, Y.; PEDERSEN J.P. A Comparative study on feature selection in text categorization. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING (ICML'97), 14., 19997. **Proceedings…,** 1997.

Anexo A

Anexo A - Lista dos Artigos Utilizados nesta Dissertação oriundos do Corpus de Cruzes (2007)

- V.R. Basili and B. Perricone, *Software errors and complexity: An empirical investigation*, Communications of the ACM 27(1) (1984) 42–52.
- V. Basili and D. Weiss, "Evaluation of a Software Requirements Document By Analysis of Change Data " in Proc. 5th Int. Conf. Software Eng., Mar 1981, pp 314-323.
- Mashiko, Y., and Basili, V. R. (1997) Using the GQM Paradigm to Investigate Influential Factors for Software Process Improvement. Journal of Systems and Software, 36, 17-32
- Endres, A. AN ANALYSIS OF ERRORS AND THEIR CAUSES IN SYSTEM PROGRAMS, Proceedings of the international conference on Reliable software, Los Angeles, California, Pages: 327 – 336, 1975.
- Elovitz, H. S. AN EXPERIMENT IN SOFTWARE ENGINEERING: THE ARCHITECTURE RESEARCH FACILITY AS A CASE STUDY
- Bell, T. E., and T. A. Thayer. Software requirements: Are they really a problem? International Conference on Software Engineering, Proceedings of the 2nd international conference on Software engineering, San Francisco, California, 1976.
- Dniestrowski, A., Guillaume, J. M., Mortier, R. Software Engineering in Avionics Applications. Proceedings of the 3rd International Conference on Software Engineering, May 10-12, 1978, Atlanta, Georgia, USA. IEEE Computer Society, Catalog No.78CH1317-7C
- POTIER, D.; ALBIN, J.L.; FERREOL, R.; BILODEAU, A. EXPERIMENTS WITH COMPUTER SOFTWARE COMPLEXITY AND RELIABILITY. Proceeding ICSE '82 Proceedings of the 6th international conference on Software engineering, 1982
- Raymond, J. R. QUANTITATIVE ASPECTS OF SOFTWARE VALIDATON. Proceeding ICSE '82 Proceedings of the 6th international conference on Software engineering, 1982
- Withrow, C. Error Density and Size in Ada Software. Journal IEEE Software archive Volume 7 Issue 1, January 1990
- Mohagheghi, P. e Conradi, R. An Empirical Study of Software Change: Origin, Acceptance Rate, and Functionality vs. Quality Attributes. Proceedings of the 2004 International Symposium on Empirical Software Engineering (ISESE), 2004.
- Chaumun, M. A.; Kabaili, H.; Keller, R. K.; e Lustman, F. A Change Impact Model for Changeability Assessment in Object-Oriented Software Systems. Third European Conference on Software Maintenance and Reengineering, 2005.
- WEYUKER, E. J. The Cost of Data Flow Testing: An Empirical Study. Journal

IEEE Transactions on Software Engineering archive Volume 16 Issue 2, February 1990

- Bieman, J. M. and Schultz, J. L. An empirical evaluation (and specification) of the all-du-paths testing criterion. Software Engineering Journal,Volume 7 Issue 1, Jan. 1992
- Frankl, P. G. and Iakounenko, O. Further Empirical Studies of Test Effectiveness. Proceeding SIGSOFT '98/FSE-6 Proceedings of the 6th ACM SIGSOFT international symposium on Foundations of software engineering, 1998
- Myers, G. J. A Controlled Experiment in Program Testing and Code Walkthroughs/Inspections. Magazine Communications of the ACM, Volume 21 Issue 9, Sept. 1978
- Kamsties, E. and Lott, C. M. An Empirical Evaluation of Three Defect-Detection Techniques. Proceedings of the 5th European Software Engineering Conference, 1995