



**Universidade de Brasília**  
Faculdade de Ciência da Informação - FCI  
Programa de Pós-Graduação em Ciência da Informação

**B2:**  
**Um Sistema para Indexação e Agrupamento  
de Artigos Científicos em Português  
Brasileiro Utilizando Computação  
Evolucionária**

Alexandre Ribeiro Afonso

**Brasília-DF**

**2013**

**B2:**  
Um Sistema para Indexação e Agrupamento de Artigos  
Científicos em Português Brasileiro Utilizando Computação  
Evolucionária

Alexandre Ribeiro Afonso

Tese apresentada à banca examinadora como requisito parcial à obtenção do Título de Doutor em Ciência da Informação pelo Programa de Pós-Graduação em Ciência da Informação da Faculdade de Ciência da Informação da Universidade de Brasília.

Orientador: Prof. Dr. Cláudio Gottschalg Duque

**Brasília-DF**

**2013**



### FOLHA DE APROVAÇÃO

**Título:** "B2: Um Sistema para Indexação e Agrupamento de Artigos Científicos em Português Brasileiro Utilizando Computação Evolucionária"

**Autor (a):** Alexandre Ribeiro Afonso


**Área de concentração:** Transferência da Informação


**Linha de pesquisa:** Arquitetura da Informação

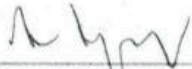
Tese submetida à Comissão Examinadora designada pelo Colegiado do Programa de Pós-graduação em Ciência da Informação da Faculdade em Ciência da Informação da Universidade de Brasília como requisito parcial para obtenção do título de **Doutor** em Ciência da Informação.

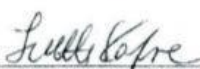
Tese aprovada em: 11 de novembro de 2013.

  
\_\_\_\_\_  
**Prof. Dr. Cláudio Gottschalg-Duque**  
Presidente (UnB/PPGCINF)

  
\_\_\_\_\_  
**Prof. Dr. Leonardo da Cunha Brito**  
Membro Interno (UFG)

  
\_\_\_\_\_  
**Prof. Dr. René Gottlieb Strehler**  
Membro Externo (UnB/LET)

  
\_\_\_\_\_  
**Prof. Dr. André Porto Ancona Lopez**  
Membro Interno (UnB/PPGCINF)

  
\_\_\_\_\_  
**Profª. Drª. Ivette Kafure Muñoz**  
Membro Interno (UnB/PPGCINF)

\_\_\_\_\_  
**Prof. Dr. Jorge Henrique Cabral Fernandes**  
Suplente (UnB/PPGCINF)

*Dedico este trabalho à minha família, especialmente aos meus pais Wanda e Victor. Meu pai nos deixou em maio de 2013, quando eu estava já no fim do trabalho de doutorado, após lutar bravamente para permanecer conosco. Um homem corajoso e verdadeiro que tantas outras lutas enfrentou de maneira destemida. Seu maior sonho era casar e ter filhos, sonho realizado por 42 anos ao lado da minha mãe, mulher companheira e amorosa. Ambos são raro exemplo de cumplicidade, amor e dedicação aos filhos. Levo comigo todos os ensinamentos, as doces lembranças do nosso lar (com minhas irmãs e sobrinhos) e principalmente o amor incondicional que me ofertaram. Minha família é um presente da vida, amo vocês eternamente.*

## Agradecimentos

Nesta jornada acadêmica, passei por várias áreas do conhecimento em cursos complementares e de pós-graduação: Ciência da Computação, Matemática, Engenharia, Linguística, Música e, finalmente, a Ciência da Informação. A busca por algo que despertasse o interesse me levou a conhecer um pouco das diversas áreas e nada melhor para estimular a criatividade que ter uma noção múltipla do conhecimento humano e da vida.

Aliás, criatividade é algo que preciso exercer, e apesar da escolha inicial em ciências exatas, meu espírito não se contentaria apenas com elas. Neste sentido, a *informação*, seus aspectos multidisciplinares e as diversas visões do tema me encantaram, logo nas primeiras leituras que fiz sobre a área. Uma ciência de vários paradigmas, de facetas múltiplas, tão universal, com tantas possibilidades de caminhos a seguir, descobrir, podendo ser tão humana, podendo ser tão tecnológica e podendo ser cibernética, só desenvolveu ainda mais essa busca pela multiplicidade que já é nata em minha pessoa. Por me compreender, agradeço ao meu orientador Cláudio, pois neste sentido, da criação, sou livre, e ele deixou-me a oportunidade para que pudesse trilhar meu caminho altamente pessoal e criativo durante a pesquisa, porém, reorientando quando necessário. Sem essa liberdade, talvez, não tivesse conseguido desenvolver um estudo significativo.

Manter a originalidade foi algo trabalhoso e este pode ser apontado como o principal atributo da pesquisa realizada. Foi criada, implementada e experimentada uma técnica computacional sob o paradigma da Computação Evolucionária para indexação e agrupamento de textos, codificada em cerca de 4500 linhas de código. Espera-se de um trabalho de doutorado a inovação, e procurei ser o mais original possível. Inclusive toda a arte de fundo da tela principal do protótipo foi criada por mim (mesmo não sendo um grande desenhista) e mesmo os ícones de cada tela foram projetados um a um, para manter a linha da originalidade proposta no trabalho. A Computação Evolucionária é uma técnica de vanguarda bioinspirada na Teoria da Evolução de Charles Darwin, sendo um dos trabalhos que evidencia a Teoria da Evolução a pesquisa realizada por Bernard Kettlewell sobre a seleção natural das mariposas da espécie *Biston betularia*, daí utilizar as mariposas como inspiração ao título do trabalho (B2) e a utilização da imagem como ilustração em algumas situações.

Quanto às pessoas que interagi neste tempo, recordo-me dos professores Suzana e Tarcísio, responsáveis pelas atraentes primeiras disciplinas do curso de pós-graduação em Ciência da Informação, onde começaram as intrigantes questões sobre os fundamentos da

Ciência da Informação e que me levaram a horas de reflexão. Agradeço à coordenação do programa de pós-graduação e agradeço à CAPES que permitiram o recebimento da bolsa, inicialmente, programa coordenado pelo professor André Porto e agora pela professora Lílian Álvares, agradecimentos também à secretária Martha, todos auxiliando e prestativos quando necessário.

Agradeço aos familiares Ramiro e tias Berenice, Míriam, Neusa, minhas irmãs Alessandra e Imara, meu cunhado Alexandre, primo José Augusto, e todos presentes no momento difícil de 2013. Meus sobrinhos, a alegria da casa. Aos meus amigos todos, agradeço as conversas e momentos de distração, em ambiente real ou virtual.

Agradeço também ao orientador de mestrado Leonardo Brito o suporte na época, e ao professor Oto Vale e à professora Suelí Aguiar as oportunidades anteriores. Agradeço ao professor René Gottlieb a participação na prévia banca de qualificação deste trabalho. Também agradeço à banca de avaliação deste trabalho o aceite do convite.

Meus agradecimentos também ao NILC e à *Linguateca*, entidades que disponibilizaram on-line e gratuitamente os recursos em Linguística Computacional os quais têm sido utilizados por mim durante a pesquisa, também à ANCIB a qual permite que os artigos em Ciência da Informação produzidos possam ser publicados no ENANCIB, evento muito bem organizado.

Obrigado.

## Espírito Livre



Voa coração feito mariposa  
Liberta-te do teu casulo e voa  
Aventura-te como ser da noite  
Corre riscos, mas descobre!

O que te aguarda?  
A luz mortal da lamparina?  
A beleza da flor?

Quem sabe teu futuro?  
Pertence a Deus?  
Pertence ao homem?

Voa coração feito mariposa  
Ainda que sem direção exata  
Leva tua sede de descoberta ao sopro  
E a pupa cansada ao movimento

Voa para o futuro  
pois tua certeza única...  
é a LIBERDADE.

## Resumo

Nesta tese é apresentado um estudo estatístico sobre o agrupamento automático de artigos científicos escritos em português do Brasil, são propostos novos métodos de indexação e agrupamento de textos com o objetivo futuro de desenvolver um software para indexar e agrupar textos por área de conhecimento. Foram testadas três classes conhecidas de termos simples para representar (indexar) os textos de entrada a agrupar: (substantivos), (substantivos e adjetivos), (substantivos, adjetivos e verbos) e também foram desenvolvidas três novas classes de termos compostos para representação (indexação) dos textos: classes de termos mais complexos, onde um termo pode ser composto pela junção de substantivos, adjetivos e preposições. Durante a fase de agrupamento textual dos experimentos foram testados os algoritmos de agrupamento: *Expectation-Maximization (EM)*, *X-Means*, um *Algoritmo Evolucionário de Agrupamento Convencional* e, ainda, um novo *Algoritmo Evolucionário de Agrupamento Proposto* cujo diferencial é trabalhar em duas etapas de processamento: uma etapa para localização do agrupamento subótimo genérico e outra etapa para melhorar tal solução. Adicionalmente, o novo algoritmo permite ao usuário definir a formação de mais grupos ou menos grupos no resultado de agrupamento. Os algoritmos de indexação e agrupamento propostos foram codificados e implementados em um protótipo denominado *B2*, no entanto, para testar os algoritmos de agrupamento *EM* e *X-Means* foi utilizado o pacote de mineração de dados *WEKA*. Quatro corpora de artigos científicos, diferentes entre si por guardarem artigos de áreas científicas distintas, foram reunidos para testar as combinações de indexação e algoritmo de agrupamento propostas. Melhores resultados de agrupamento (por área de conhecimento dos artigos) foram obtidos utilizando termos compostos na indexação, ao invés do uso de termos simples, quando combinados com o uso do novo *Algoritmo Evolucionário de Agrupamento Proposto*, porém, para obter grupos bem formados, um número excessivo de grupos é gerado pelo protótipo, consumindo alto tempo de computação para executar tais novos métodos, em um computador pessoal convencional do ano de 2012. Pode-se concluir que o problema de agrupar automaticamente artigos científicos em suas áreas originais é uma tarefa complexa. Logo, acredita-se que os métodos de indexação e agrupamento desenvolvidos possam ser aprimorados para utilização futura em situações específicas, onde a fragmentação e geração adicional de grupos além do esperado não seja um problema maior.

**Palavras-chave:** Indexação Automática, Agrupamento Automático de Textos, Linguística Computacional, Algoritmos Evolucionários, Mineração de Textos, Artigos Científicos.



## Abstract

This thesis presents an empirical study about automated text clustering for scientific articles written in Brazilian Portuguese. We tested three already known classes of simple terms for representing (or indexing) the input texts: (nouns), (nouns and adjectives) and (nouns, adjectives and verbs); we also developed three new classes of composed terms for text representation (or indexing): the new classes consist of more complex terms, where a complex term could be composed by the joint of nouns, adjectives and prepositions. Our final goal is to develop new software for text indexing and clustering. During the clustering stage of the experiments we tested the Expectation-Maximization (EM) Clustering Algorithm, the X-Means Clustering Algorithm, the Conventional Clustering Evolutionary Algorithm and, finally, we also proposed a new Two Phase Clustering Evolutionary Algorithm which works in two phases, the first phase finds the sub-optimal text clustering and the second one improves the result found by the first phase. The Two Phase Clustering Evolutionary Algorithm also permits the user to define whether the system should create a high number or a low number of clusters. The new indexing and clustering algorithmic strategies presented were implemented in a prototype named *B2*, but for testing the EM and X-Means algorithms we used the known WEKA data mining package. Four different scientific corpora having different sets of scientific topics were assembled and applied for testing the combinations of indexing and clustering methods. Although considerable better results were achieved when indexing with the classes of composed terms combined with the new Two Phase Clustering Evolutionary Algorithm, a considerable higher number of clusters was generated and a considerable additional time was consumed when running the new system over a 2012 conventional personal computer. We conclude that the problem of clustering scientific articles in their original topics is a complex task. Good results of clustering correctness were achieved by the new methods but producing many fragmented additional clusters as output, so, in the future, the methods can be improved and applied in specific situations where the fragmentation and additional production of clusters are not a major problem.

**Keywords:** Automated Text Indexing, Automated Text Clustering, Computational Linguistics, Evolutionary Algorithms, Text Mining, Scientific Articles.

## **Lista de Abreviaturas e Siglas**

**ACO:** Ant Colony Optimization

**AE:** Algoritmos Evolucionários

**ARIA:** Adaptive Radius Immune Algorithm

**CDD:** Classificação Decimal de Dewey

**CDU:** Classificação Decimal Universal

**CE:** Computação Evolucionária

**CSPA:** Cluster-based Similarity Partitioning Algorithm

**EC:** Evolutionary Computing

**EM:** Expectation-Maximization Algorithm

**ENANCIB:** Encontro Nacional de Pesquisa em Ciência da Informação

**FID:** Institute International de Documentation

**IFLA:** International Federation of Library Associations

**ILS:** Interated Local Search

**JDK:** Java Development Kit

**KWIC:** Key-Word in Context

**LSI:** Latent Semantic Indexing

**MVGA:** Modified Variable String Length Genetic Algorithm

**ND:** Número de Desvios

**NILC:** Núcleo Interinstitucional de Linguística Computacional

**NMF:** Non-negative Matrix Factorization

**OC:** Organização do Conhecimento

**OI:** Organização da Informação

**PB:** Português Brasileiro

**PE:** Português Europeu

**PLN:** Processamento de Linguagem (Língua) Natural

**PLSA:** Probabilistic Latent Semantic Analysis

**PROPOR:** Conferência Internacional sobre o Processamento Computacional do Português

**RC:** Representação do Conhecimento

**RI:** Recuperação da Informação/ Representação da Informação

**SA:** Simulated Annealing

**semantic-SOM:** Semantic Self-Organizing Map

**SKM:** Simple K-Means

**SOC:** Sistemas de Organização do Conhecimento

**SRI:** Sistema de Recuperação da Informação

**STIL:** Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana

**TC:** Termos Compostos

**TF.IDF:** Term Frequency  $\times$  Inverse Document Frequency

**TF:** Term Frequency

**TS:** Tabu Search

**UFOD:** Union Française des Organismes de Documentation

**UNICLASS:** United Classification for the Construction Industry

**VCI:** Vetor de Características Individual

**VGA:** Variable String Length Genetic Algorithm

**VSM-BoW:** Vector Space Model – Bag of Words

## Lista de Figuras

Figura 1 - Níveis de complexidade do processo de classificação de textos .....	53
Figura 2 - Passos do processo de agrupamento de textos .....	54
Figura 3a - Agrupamento hierárquico para a área de Computação Natural.....	55
Figura 3b - Agrupamento por particionamento não hierárquico .....	56
Figura 3c - Agrupamento nebuloso .....	56
Figura 4 - Mariposas <i>Biston betularia</i> .....	66
Figura 5 - Fluxograma dos módulos de um <i>Algoritmo Evolucionário Convencional</i> .....	71
Figura 6 - Funcionalidades do sistema <i>B2</i> a partir do menu principal .....	80
Figura 7 - Representação convencional de textos (Índices VSM) .....	86
Figura 8 - Representação de textos para o <i>Algoritmo Evolucionário de Agrupamento Proposto</i> (índices VCI).88	
Figura 9 - Representação dos cromossomos na Computação Evolucionária .....	89
Figura 10 - Representação dos cromossomos na implementação proposta .....	90
Figura 11 - Representação do cromossomo de tamanho fixo na implementação proposta.....	91
Figura 12 - Representação dos genes de um cromossomo nas implementações .....	92
Figura 13 - Esquema do <i>Algoritmo Evolucionário de Agrupamento Proposto</i> .....	99
Figura 14 - Exemplo de mutação no cromossomo na implementação proposta .....	103
Figura 15 - Exemplo de <i>crossover</i> entre dois cromossomos na implementação proposta .....	104
Figura 16a - Exemplo de ciclo de evolução para a etapa 1 na implementação proposta .....	106
Figura 16b - Exemplo de ciclo de evolução para a etapa 2 na implementação proposta .....	107
Figura 17 - Esquema do processo iterativo de evolução convencional .....	109
Figura 18 - Modelo dos experimentos realizados na pesquisa .....	113
Figura 19 - Gráfico com melhores resultados de corretude dos experimentos .....	137

## Lista de Tabelas

Tabela 1A – Experimento 1A .....	129
Tabela 1B – Experimento 1B .....	129
Tabela 1C – Experimento 1C .....	129
Tabela 1D – Experimento 1D .....	129
Tabela 1E – Experimento 1E .....	130
Tabela 1F – Experimento 1F .....	130
Tabela 2A – Experimento 2A .....	131
Tabela 2B – Experimento 2B .....	131
Tabela 2C – Experimento 2C .....	131
Tabela 2D – Experimento 2D .....	131
Tabela 2E – Experimento 2E .....	132
Tabela 2F – Experimento 2F .....	132
Tabela 3A – Experimento 3A .....	133
Tabela 3B – Experimento 3B .....	133
Tabela 3C – Experimento 3C .....	133
Tabela 3D – Experimento 3D .....	133
Tabela 3E – Experimento 3E .....	134
Tabela 3F – Experimento 3F .....	134
Tabela 4A – Experimento 4A .....	135
Tabela 4B – Experimento 4B .....	135
Tabela 4C – Experimento 4C .....	135
Tabela 4D – Experimento 4D .....	135
Tabela 4E – Experimento 4E .....	136
Tabela 4F – Experimento 4F .....	136

## Lista de Pseudocódigos

Listagem 1 - Algoritmo <i>K-Means</i> original .....	59
Listagem 2 - Algoritmo de agrupamento <i>X-Means</i> .....	60
Listagem 2 - Algoritmo de agrupamento <i>X-Means</i> (continuação).....	61
Listagem 3 - Algoritmo de agrupamento <i>Expectation-Maximization</i> .....	63
Listagem 4 - Algoritmo <i>Cross-Validation</i> .....	64
Listagem 5 - Índice de avaliação <i>Davies-Bouldin</i> .....	93
Listagem 6a - Função de avaliação para calcular a similaridade entre dois vetores (VCI) .....	95
Listagem 6b - Cálculo do peso (aptidão) de um gene (grupo) na implementação proposta .....	97
Listagem 6c - Cálculo do peso (aptidão) de um cromossomo na implementação proposta .....	97

## Lista de Quadros

Quadro 1 - Configurações da primeira etapa do <i>Algoritmo Evolucionário Proposto</i> .....	123
Quadro 2 - Configurações da segunda etapa do <i>Algoritmo Evolucionário Proposto</i> .....	123
Quadro 3 - Valores <i>default</i> para taxa de mutação com número de textos maior que 180 unidades.....	124
Quadro 4 - Valores <i>default</i> para taxa de mutação com número de textos igual a 180 unidades .....	124
Quadro 5 - Valores <i>default</i> para taxa de mutação com número de textos igual a 100 ou 120 unidades.....	125
Quadro 6 - Valores <i>default</i> para taxa de mutação com número de textos igual a 60 unidades .....	125
Quadro 7 - Configurações para o Algoritmo <i>EM (WEKA)</i> .....	126
Quadro 8 - Configurações para o Algoritmo <i>X-Means (WEKA)</i> .....	126
Quadro 9 - Configurações para o <i>Algoritmo Evolucionário de Agrupamento Convencional</i> .....	127

# Sumário

<b>1 Introdução</b> .....	18
1.1 Apresentação.....	18
1.2 Problemática e Metodologia.....	21
1.3 Justificativa.....	22
1.4 Objetivo geral .....	24
1.5 Objetivos específicos.....	24
1.6 Organização da tese .....	25
<b>2 Indexação e Classificação: Revisão da Literatura</b> .....	26
2.1 Apresentação .....	26
2.2 Motivações para um estudo específico sobre o português do Brasil.....	26
2.3 Sistemas de Organização da Informação e do Conhecimento.....	28
2.3.1 Tipos de Classificação Bibliográfica .....	32
2.4 A Classificação na Arquivologia e Museologia .....	34
2.5 Considerações sobre a classificação na Biblioteconomia, Arquivologia e Museologia.....	38
2.6 Indexação manual .....	39
2.7 Indexação automática .....	43
2.8 Indexação automática para o português do Brasil .....	47
2.9 Classificação automática de textos sob o enfoque da Inteligência Computacional .....	51
2.9.1 O algoritmo de agrupamento <i>X-Means</i> .....	58
2.9.2 O algoritmo de agrupamento <i>Expectation-Maximization (EM)</i> .....	62
2.9.3 O algoritmo de agrupamento baseado em <i>Computação Evolucionária</i> .....	64
2.10 Estudos em Agrupamento Automático de Textos para o português do Brasil.....	74
<b>3 Inovações Propostas para Indexação e Agrupamento de Textos</b> .....	79
3.1 Apresentação .....	79
3.2 Arquitetura do sistema <i>B2</i> .....	79



3.3 Indexação (ou representação) por termos compostos.....	82
3.4 Inovações nos algoritmos de agrupamento.....	85
3.5 Características do <i>Algoritmo Evolucionário de Agrupamento Proposto</i> .....	86
3.5.1 Estruturas de dados para os índices (ou representações) de textos .....	86
3.5.2 A estrutura dos cromossomos .....	88
3.5.3 A função de avaliação .....	91
3.5.4 O processo iterativo de evolução.....	98
3.5.5 Procedimento de seleção para o <i>Algoritmo Evolucionário de Agrupamento Proposto</i> .....	102
3.5.6 Procedimentos de mutação e <i>crossover</i> na implementação proposta .....	103
3.6 Características do <i>Algoritmo Evolucionário de Agrupamento Convencional</i> .....	109
3.6.1 O processo iterativo de evolução.....	109
3.6.2 Representação dos textos.....	110
3.6.3 A estrutura dos cromossomos .....	110
3.6.4 A função de avaliação .....	110
3.6.5 Critério de parada .....	110
3.6.6 Procedimento de seleção.....	110
3.6.7 Procedimentos de mutação e <i>crossover</i> .....	111
<b>4 Metodologia e Descrição dos Experimentos</b> .....	<b>112</b>
4.1 Apresentação .....	112
4.2 Descrição das variáveis independentes.....	113
4.2.1 Seleção do corpus .....	113
4.2.2 Seleção por classes de palavras ou termos.....	115
4.2.3 Algoritmos de filtragem .....	115
4.2.4 Algoritmos de agrupamento .....	117
4.3 Descrição das variáveis dependentes.....	118
4.3.1 Métricas de corretude e tempo .....	118
<b>5 Resultados e Análise de Resultados</b> .....	<b>121</b>
5.1 Configurações <i>default</i> para os experimentos.....	121
5.1.1 Configurações dos corpora .....	121
5.1.2 Configurações de indexação por tipo de termo.....	121
5.1.3 Configurações do <i>Algoritmo Evolucionário de Agrupamento Proposto</i> .....	122
5.1.4 Configurações para o Algoritmo <i>EM (WEKA)</i> .....	126

5.1.5 Configurações para o Algoritmo <i>X-Means</i> (WEKA) .....	126
5.1.6 Configurações para o <i>Algoritmo Evolucionário de Agrupamento Convencional</i> .....	127
5.2 Resultados.....	128
5.3 Análise de resultados .....	137
5.3.1 Sobre a eficácia e eficiência do <i>Algoritmo Evolucionário de Agrupamento Proposto</i> .....	137
5.3.2 Sobre o impacto da forma de indexação nos algoritmos de agrupamento.....	139
5.3.3 Sobre o impacto da natureza do corpus nos algoritmos de agrupamento .....	139
<b>6 Conclusão e Trabalhos Futuros</b> .....	<b>141</b>
<b>7 Referências Bibliográficas</b> .....	<b>145</b>
Apêndice A: Lista de Periódicos Utilizados nos Corpora da Pesquisa .....	153
Apêndice B: Descrição de Opções para os Algoritmos <i>EM</i> e <i>X-Means</i> no <i>Weka</i> .....	155
Apêndice C: Lista de Publicações Realizadas durante a Pesquisa .....	157

# 1 INTRODUÇÃO

## 1.1 Apresentação

A elaboração de técnicas e instrumentos que suportam a recuperação de registros produzidos pela humanidade é uma tarefa exercida desde as mais antigas bibliotecas. Na atualidade, utilizando a nomenclatura “Recuperação da Informação”, tal campo de estudos encontra-se em expansão, com diversas pesquisas tecnologicamente motivadas e com diferentes enfoques, seja ele: social, computacional, linguístico, cognitivo ou multidisciplinar.

É complexo definir exatamente o que é e qual a abrangência científica, os limites e fronteiras, da especialidade Recuperação da Informação (RI). Tradicionalmente, a RI é colocada como subárea da Ciência da Informação: segundo Saracevic (1996), a Ciência da Informação surge após o período da Segunda Guerra Mundial, numa tentativa de controlar a explosão informacional proveniente de pesquisas científicas e tecnológicas e da necessidade em criar técnicas, ferramentas, métodos para a eficiente e eficaz distribuição informacional, ou mesmo para compreender e teorizar sobre o ciclo de produção e consumo da informação registrada pela sociedade. Ainda, segundo este autor, o termo Recuperação da Informação foi cunhado por Calvin Moores em 1951, destacando que ele "engloba os aspectos intelectuais da descrição de informações e suas especificidades para a busca, além de quaisquer sistemas, técnicas ou máquinas empregadas para o desempenho da operação".

Apesar da definição concisa, existem visões diversificadas para o tema. A Ciência e Engenharia de Computação têm tido uma participação notável nas pesquisas em Recuperação da Informação, abrigando o enfoque tecnológico, no desenvolvimento de métodos computacionais e na construção de softwares para organizar e recuperar documentos ou registros digitais (FERNEDA, 2003). Ao mesmo tempo, as áreas de Biblioteconomia, Documentação, Ciência Cognitiva, Sociologia da Ciência, Comunicação, Linguística, Lógica, Psicologia, Estatística e Economia fazem interface com a Ciência da Informação, o que demonstra o caráter interdisciplinar desta ciência (ORTEGA, 2004), e logo, os estudos em Recuperação da Informação, como especialidade da Ciência da Informação, herdarão essa característica multi e interdisciplinar.

Não é encontrada, na literatura em Ciência da Informação, uma definição única para o objeto de estudos: *informação*, ou seja, os autores por vezes divergem ao tentarem

desenvolver definições sobre o objeto de estudos e, como consequência, não há uma definição consensual para os vários termos derivados (Ciência da Informação, Informática, Arquitetura da Informação, Recuperação da Informação, Documento, Dado, Conhecimento, etc.). Tão complexo quanto definir o objeto de estudo desta ciência pelos autores da Ciência da Informação, é relacionar as visões das áreas, subáreas e especialidades de estudo, sendo difícil dizer exatamente até que ponto a Recuperação da Informação da Biblioteconomia, Arquivologia, Museologia, Ciência da Informação, Ciência e Engenharia da Computação, Ciência Cognitiva e outras se assemelham e se diferenciam.

A dificuldade em estabelecer limites e definições para a Ciência da Informação (e logo, para a Recuperação da Informação) é apontada por diversos autores. Araújo (2010) destaca a existência de um problema concreto: o compartilhamento, por parte da Biblioteconomia, da Arquivologia e da Museologia, de um mesmo espaço institucional (ou um mesmo departamento nas universidades brasileiras) e a escassez de produção científica que possibilite que esse compartilhamento seja harmonioso e produtivo, ou seja, não existem investigações científicas suficientes que apontem ou limitem a forma de união dos campos de estudo. A dificuldade de integração prática é exemplificada por Araújo, Marques e Vanz (2011), os quais destacam a dificuldade de integralizar um tronco curricular comum (pertencente à Ciência da Informação) entre os cursos de graduação em Biblioteconomia, Arquivologia e Museologia nas universidades brasileiras. Ou seja, estas ciências estudam o objeto *informação* em diferentes perspectivas.

Neste trabalho, o tópico de interesse é a Recuperação da Informação sob o olhar da Ciência da Informação, tomando especificamente a definição citada anteriormente e referenciada em Saracevic (1996) para a ciência, junto à definição de Calvin Moores (1951) para a prática da Recuperação da Informação, onde a tecnologia tem um papel importante ao auxiliar o usuário na busca dos registros. Na perspectiva de Saracevic, a Ciência da Informação é vista como ciência interdisciplinar e métodos e teorias de outras áreas do conhecimento, como a Linguística, Biblioteconomia, Ciência Cognitiva e a Ciência da Computação são utilizadas para atingir os objetivos de organização e recuperação de registros. Saracevic (1996) ainda afirma que “os problemas básicos de se compreender a informação e a comunicação, suas manifestações, o comportamento informativo humano, (...), incluindo as tentativas de ajustes tecnológicos, não podem ser resolvidos no âmbito de uma única disciplina” (SARACEVIC, 1996, p. 48). Portanto, a Recuperação da Informação aqui apresentada é de natureza interdisciplinar e socialmente motivada, no sentido que a prática

científica é ativada ao identificar e procurar satisfazer as necessidades informacionais dos usuários, daí ser considerada uma ciência social aplicada. Entende-se, portanto, que o computador, os algoritmos e estruturas de dados são instrumentos imprescindíveis no processo de recuperação da informação, mas o entendimento das necessidades informacionais do usuário, seu comportamento ao recuperar a informação e a escolha da maneira de organizar a informação produzida pela sociedade também influenciará significativamente na eficácia e eficiência do processo de recuperação. Logo, os entendimentos dos aspectos sociais, comportamentais, linguísticos e tecnológicos são contribuintes para levar o usuário ao sucesso no processo de recuperação da informação, e não opostos.

Cunha (2008) observa que a biblioteca convencional e a digital permanecem com a mesma função, a de disponibilizar informação e promover conhecimento. O que muda é o instrumento que elas utilizam para levar a informação ao usuário. Com as novas tecnologias a biblioteca digital conseguiu reunir os materiais em um só lugar, em vários tipos de formato em um único suporte, o digital. É possível preservar obras raras, frágeis e únicas, podendo ser recuperadas em qualquer lugar que possua acesso a *internet* em qualquer hora do dia por vários usuários simultaneamente.

Dentre as diversas possibilidades de pesquisa relacionadas à Recuperação da Informação, pode-se citar a Classificação da Informação. Classificam-se documentos textuais em classes e subclasses com o intuito de facilitar a recuperação da informação registrada por parte do usuário. A Teoria da Classificação é tradicionalmente conhecida da velha arte da Biblioteconomia, a qual elabora diversos métodos de classificação para efetivar a recuperação da informação por parte dos usuários. A Classificação Decimal Universal (CDU), a Classificação Decimal de Dewey (CDD) e a Classificação Facetada são Sistemas de Classificação Bibliográfica utilizados nas bibliotecas institucionais por décadas para auxiliar o processo de localização do material impresso por área de conhecimento.

É notável, porém, que com o estabelecimento do aparato tecnológico, outras necessidades e maneiras de trabalho estão surgindo, e classificar não necessariamente significa utilizar um desses métodos tradicionais, ainda que eles possam ser utilizados como métodos eficazes no meio digital. A classificação automática também tem sido vista como prática auxiliar do processo de recuperação da informação, ou a prática em “colocar documentos em classes” de forma automática, utilizando um software (geralmente aplicando métodos de Inteligência Artificial) para tal objetivo (MARKOV; LAROSE, 2007). Portanto,

pode-se afirmar que o escopo da pesquisa aqui relatada faz parte do campo de estudos da Recuperação da Informação, considerado um campo interdisciplinar, e especificamente, o ponto de estudos é a classificação por mecanismos computacionais, de forma automatizada, com o desenvolvimento de algoritmos, representação de dados e softwares que aprimoram a classificação de textos, visando uma recuperação da informação facilitada, ágil e correta, do ponto de vista do usuário.

Segundo Markov e Larouse (2007), a Classificação Automática ainda poderia ser dividida em duas vertentes metodológicas: a Categorização Automática, quando o sistema classificador recebe do usuário algum tipo de conhecimento formal externo (ontologia, tesouro, vocabulário controlado ou pré-treinamento) cedido ao sistema de forma supervisionada para efetivar a classificação; e outro ramo seria o Agrupamento Automático, uma tarefa mais complexa, onde somente os textos de entrada são utilizados pelo sistema de classificação sem nenhum conhecimento formal prévio cedido, ou seja, o sistema computacional deveria aprender de maneira não supervisionada, com o próprio corpus a classificar, a identificar as classes de textos existentes e efetuar o agrupamento.

## 1.2 Problemática e Metodologia

Na pesquisa aqui descrita, primeiramente, foram avaliados algoritmos de agrupamento de textos aplicados a artigos científicos escritos em português brasileiro, onde tais algoritmos são provenientes da atual tecnologia da informação, são implementados em softwares conhecidos, de livre acesso, e foram previamente testados em diversas pesquisas para outras línguas. Esses algoritmos já testados e descritos na literatura são: um *Algoritmo Evolucionário Convencional* de agrupamento, o conhecido algoritmo *Expectation-Maximization (EM)* e o algoritmo de agrupamento *X-Means*, uma extensão do conhecido algoritmo *K-Means* para quando o número de grupos  $K$  a gerar não é previamente conhecido pelo usuário. Na segunda etapa da pesquisa, é proposto um novo algoritmo de agrupamento de textos para artigos científicos escritos em português do Brasil. Tal algoritmo também utiliza técnicas da subárea da Inteligência Computacional denominada Computação Evolucionária.

Sobre o pré-processamento dos textos a agrupar, os artigos científicos para testes com os três algoritmos de agrupamento conhecidos e o novo algoritmo proposto são indexados utilizando termos simples e termos compostos, sendo tal indexação com termos compostos ainda não descrita na literatura. Também, foram testados quatro corpora diferentes para as

diferentes combinações possíveis de técnicas de indexação e algoritmos de agrupamento. Os quatro corpora de teste utilizados contêm combinações de áreas científicas diferentes, pois a meta, inclusive, era verificar se combinações diferentes de áreas científicas nos corpora poderiam alterar os resultados de agrupamento. Para especificamente testar as inovações (indexação por termos compostos e o *Algoritmo de Agrupamento Evolucionário Proposto*) foi codificado um sistema computacional, que ainda se encontra em versão beta, denominado *B2*. Para testar os algoritmos convencionais de agrupamento *Expectation-Maximization* e *X-Means* foi utilizado o pacote de mineração de dados *WEKA*<sup>1</sup>.

A pesquisa procurou a elaboração de uma técnica computacional aprimorada para agrupamento de textos, e para demonstrar se essa melhoria é alcançada, uma comparação entre os métodos de indexação e agrupamento de textos já conhecidos e os novos métodos propostos é realizada. Utilizaram-se medidas estatísticas conhecidas, citadas por cientistas da área, para a comparação dos algoritmos em termos de eficácia (qualidade do agrupamento textual) e eficiência (tempo de computação) no processo de agrupamento de textos científicos em português do Brasil. Como são testados quatro corpora diferentes, seis formas de indexação diferentes e quatro algoritmos de agrupamento, ter-se-ia uma possibilidade combinatória de 96 (noventa e seis) testes possíveis a verificar, porém, o algoritmo de agrupamento proposto pode ser configurado de duas maneiras distintas, logo, têm-se o equivalente a cinco algoritmos de agrupamento a verificar, totalizando 120 (cento e vinte) testes experimentais.

### 1.3 Justificativa

A tarefa de organizar (indexar, classificar, catalogar) a informação registrada para facilitar a recuperação por usuários da informação, apesar de estar ativa na Biblioteconomia/Arquivologia por décadas, tem exigido novas experiências científicas e tecnológicas para a recuperação da informação em meio digital, uma vez que a *World Wide Web* e a atual tecnologia da informação trazem uma nova perspectiva metodológica para o armazenamento e a localização de registros em bibliotecas e repositórios digitais (BARRETO, 2008).

Neste sentido, quando se trata especificamente da recuperação da informação textual em meio digital, o fator “língua” acaba por exigir novas pesquisas e experimentações na área,

---

<sup>1</sup> [www.cs.waikato.ac.nz/ml/weka/](http://www.cs.waikato.ac.nz/ml/weka/)

considerando que muitas técnicas de Representação do Conhecimento, Classificação e Agrupamento de Textos, Técnicas de Sumarização, Tradução e Busca por Assunto/Tema são na maioria das vezes dependentes do conhecimento científico da língua, para que ocorram em fluxo eficiente e eficaz no atendimento às necessidades de informação dos usuários. As línguas possuem características peculiares que as diferenciam e as terminologias das áreas de conhecimento são concebidas nacionalmente, com características originais e herdando as peculiaridades sociolinguísticas de cada nação (BIDERMAN, 2006).

Tal dependência da língua significa que o sucesso na construção de técnicas, métodos e ferramentas ou sistemas para recuperação da informação textual, e a forma de testar esses métodos em meio digital, estão altamente ligados à expressão linguística e terminológica dos textos a processar. Métodos que são implementados em software para recuperar ou organizar a informação registrada dependem de boas escolhas algorítmicas/computacionais e técnicas bibliométricas ou de manipulação de arquivos implementadas neste software, mas o conhecimento científico da língua para o sucesso do sistema de organização e recuperação desta massa de textos também é vital em muitas aplicações de recuperação da informação (DUQUE, 2005).

Se o conhecimento linguístico é importante para a construção de técnicas, métodos e sistemas para a organização e recuperação da informação textual, tem-se, portanto, uma nova gama de ideias e experimentos a serem implementados e testados pelas ciências que lidam com informação registrada, uma vez que um método de classificação automática de textos para o inglês pode não produzir os mesmos resultados para o português, e isso abre uma ampla lista de escolhas nas experimentações científicas.

Logo, a justificativa deste trabalho, do ponto de vista científico-social, é contribuir para o desenvolvimento dos estudos em classificação da informação textual aplicada ao português do Brasil, uma vez que a quantidade de informação registrada produzida na língua é ampla (seja jornalística, comercial, científica, etc.) e há necessidade de meios específicos para sua organização e consequente recuperação em meio digital.

Há uma preocupação social evidente: a construção e o teste de métodos computacionais que serão utilizados por usuários que buscam a informação registrada nesta língua, especificamente a utilizada no Brasil, com suas peculiaridades culturais e terminológicas presentes.



## 1.4 Objetivo geral

Propor e implementar um método de indexação por termos compostos e um Algoritmo Evolucionário para o agrupamento automático de artigos científicos em português do Brasil.

## 1.5 Objetivos específicos

- a) investigar os algoritmos de agrupamento automático de textos aplicados a outras línguas, aplicando-os a artigos científicos escritos em português do Brasil;
- b) propor a aplicação das técnicas de uma das subáreas da Inteligência Computacional denominada Computação Evolucionária, de forma inédita, onde o desenvolvimento e implementação de um novo algoritmo para agrupamento de textos é efetuada. Procura-se verificar e comparar formalmente a eficiência (tempo) e eficácia (agrupamento correto) dos algoritmos de agrupamento conhecidos e o proposto;
- c) verificar formalmente se a indexação prévia dos textos, antes do agrupamento, com termos compostos, traz melhorias para o agrupamento em relação ao uso de termos simples, para todos os algoritmos de agrupamento estudados;
- d) aplicar diversos corpora de testes com textos de diferentes áreas do conhecimento, observando formalmente se grupos de artigos científicos diferentes produzem resultados melhores ou piores de agrupamento, para textos em português do Brasil.

## 1.6 Organização da tese

Este trabalho está dividido nas seguintes etapas: no capítulo 2 é realizada a contextualização da pesquisa na Ciência da Informação, a problemática é apresentada em detalhes e relacionada aos objetivos desta ciência. Neste capítulo, também são apresentados trabalhos relacionados ao tema: indexação e classificação de textos, especificamente para o português brasileiro, e as referências bibliográficas relevantes são enunciadas. No capítulo 3 são apresentadas as inovações em indexação e classificação propostas, com a descrição das técnicas utilizadas e implementadas no sistema *B2*. No capítulo 4, descreve-se a metodologia utilizada nos experimentos para verificar se melhorias foram alcançadas com o novo sistema *B2*, também, descrevem-se as etapas dos experimentos e o método estatístico para avaliação de resultados. No capítulo 5, os 120 (cento e vinte) resultados dos experimentos realizados em agrupamento automático de textos são apresentados, analisados e comparados. No capítulo 6 é elaborada uma pré-conclusão sobre os resultados de eficiência e eficácia obtidos nos experimentos, também há considerações gerais sobre a pesquisa realizada e perspectivas futuras.

## **2 INDEXAÇÃO E CLASSIFICAÇÃO: REVISÃO DA LITERATURA**

### **2.1 Apresentação**

O objetivo da pesquisa aqui descrita leva ao estudo e a proposta de Sistemas Inteligentes para o agrupamento de textos científicos por área temática, porém, os conhecidos Sistemas de Classificação Bibliográfica e as Linguagens Documentárias, por décadas, têm sido amplamente utilizados nas bibliotecas institucionais com o intuito de facilitar a recuperação da informação em suporte impresso e têm sido adaptados para a utilização em meio digital. Os dois assuntos não são interdependentes, mas inter-relacionados. Assim, uma breve (mas significativa) descrição da teoria e das técnicas convencionais de classificação bibliográfica e arquivística é realizada, com o objetivo de contextualizar a pesquisa aqui descrita, historicamente. A indexação tem uma relação próxima com o desenvolvimento de sistemas inteligentes de classificação, uma vez que a classificação automática pede uma prévia indexação automática. Adicionalmente, no Brasil, a indexação automática já é explorada desde o início da década de oitenta. São abordadas, neste capítulo, as técnicas de indexação e classificação convencionais, com breve perspectiva histórica das metodologias mais amplamente estudadas. Nestas abordagens, procura-se a valorização dos autores brasileiros, uma vez que a pesquisa realizada é voltada para textos escritos em português do Brasil e para as características da comunicação científica brasileira. A revisão da literatura atual, amplamente relacionada ao tema, e as propostas dos autores para a automação classificatória destinada ao uso em bibliotecas ou repositórios digitais é realizada em seguida.

### **2.2 Motivações para um estudo específico sobre o português do Brasil**

Na literatura relacionada a este estudo, é possível encontrar alguns artigos científicos que tratam da influência das características da língua no processo de organização e recuperação automática da informação textual. A seguir, o relato sucinto de tais estudos.

Rossel e Velupillai (2005) investigaram o impacto do uso de sintagmas ao representar textos utilizando o modelo de representação vetorial (*vector space model*) para fins de agrupamento de textos em sueco, em diferentes situações. Foi utilizado um corpus científico da área médica e outro corpus de textos jornalísticos. Como resultado, a pesquisa mostrou que o uso de sintagmas não trás melhorias para o agrupamento quando comparado ao uso de

termos simples, e ainda, os resultados diferem significativamente entre os dois tipos de textos. Os autores avaliam que os resultados obtidos são diferentes daqueles apresentados por Hammouda e Kamel (2004), o que ocorre, presumidamente, pelo menos parcialmente, devido às diferenças entre o inglês e o sueco. Os compostos cristalizados (*solid compounds*) do sueco frequentemente correspondem a sintagmas do inglês.

Weiss e Stefanowski (2003) descrevem um estudo similar para o polonês, na tarefa de agrupamento de textos no contexto de um sistema experimental denominado *Carrot*. O algoritmo em consideração, denominado *Suffix Tree Clustering*, tem sido reconhecido como eficaz quando aplicado ao inglês. Quando aplicado ao polonês, é percebida uma fragilidade na experimentação, onde falhas do algoritmo ocorrem devido às características dos dados de entrada. Os autores indicam que as características dos grupos textuais produzidos (número, valor), ao contrário do inglês, são dependentes da fase de pré-processamento de tais textos. Os autores também afirmam que as propriedades de uma língua severamente influenciam o desempenho do algoritmo.

Para a língua croata, Basic, Berecek e Cvitas (2005) argumentam que os sistemas e algoritmos para processamento de textos para o inglês são bem desenvolvidos, o que não é o caso para a língua croata. O texto explora a aplicação dos sistemas existentes e avalia os parâmetros ótimos para a língua croata. Uma dos problemas básicos, que tornam o processamento desta língua difícil, é a complexidade morfológica, onde a diferença é particularmente óbvia quando comparada ao inglês.

O português do Brasil apresenta também características peculiares que exigem experimentações específicas nas tarefas de processamento de textos. Recentemente, alguns autores evidenciam tais peculiaridades. Por exemplo, Silva (2006) relata um paralelo entre textos jornalísticos do domínio do futebol e do vestuário escritos em português brasileiro e português europeu, analisando as diferenças terminológicas entre os dois países a partir de artigos escritos nas décadas de: 50, 70, 90-2000. O autor aponta uma diferença significativa entre os termos utilizados pelas duas nações nos dois domínios, além disso, verifica que os textos brasileiros apresentaram uma maior abertura ao uso de termos estrangeiros.

Seguindo a mesma ideia de contraste entre as duas variantes, Biderman (2001) também descreve as diferenças linguísticas entre as duas nações. Justifica que o léxico registra o conhecimento do mundo de uma comunidade linguística através da palavra, mais ainda, o léxico tem papel fundamental na estrutura e funcionamento da língua porque refere

os conceitos linguísticos e extralinguísticos da cultura e da sociedade, por essa razão são bem grandes as diferenças lexicais entre o português brasileiro (PB) e europeu (PE). Afirma que o Brasil possui uma alta biodiversidade (considere, por exemplo, a fauna e a flora amazônica) onde muitos nomes vieram de línguas indígenas que habitavam e habitam tal região e tais nomes não têm correspondentes no português europeu. Há ainda, a influência de africanos (da época da escravidão) e de imigrantes. Todas essas influências criaram um léxico próprio no Brasil que difere do léxico europeu.

Tais evidências sobre a influência das características linguísticas e culturais nos sistemas de organização e recuperação da informação e as evidências de divergências entre as variantes do português justificam a necessidade de estudos e experimentos específicos para o português brasileiro. Tais estudos servem para avaliar o processamento computacional da comunicação científica, jornalística, empresarial, jurídica em nível nacional, visando aprimoramentos em tais sistemas computacionais.

### **2.3 Sistemas de Organização da Informação e do Conhecimento**

Ao explorar os temas indexação e classificação textual, além de uma justificativa sobre a influência das características linguísticas para o processo de organização e recuperação da informação, é necessário rever os conceitos teóricos de Organização da Informação (OI) e Organização do Conhecimento (OC), no contexto desta pesquisa e no contexto da Ciência da Informação, já que tais conceitos são base teórica da pesquisa. Apesar de não existir um consenso sobre o significado dos dois termos (informação e conhecimento) entre os autores da Ciência da Informação, foram adotadas como base as definições propostas por Bräscher e Café (2008) que contrastam as diferenças entre OI e OC. Em princípio, as autoras descrevem as características que Fogl (1979) atribui aos conceitos de informação e conhecimento:

- a) conhecimento é o resultado da cognição (processo de reflexão das leis e das propriedades de objetos e fenômenos da realidade objetiva na consciência humana);
- b) conhecimento é o conteúdo ideal da consciência humana;
- c) informação é uma forma material da existência do conhecimento;
- d) informação é um item definitivo do conhecimento expresso por meio da linguagem natural ou outros sistemas de signos percebidos pelos órgãos e sentidos;
- e) informação existe e exerce sua função social por meio de um suporte físico;
- f) informação existe objetivamente fora da consciência individual e independente

dela, desde o momento de sua origem.

O objetivo do processo de organização da informação é possibilitar o acesso ao conhecimento contido na informação. Esse objetivo pode ser detalhado com base nos ajustes propostos por Svenonius (2000) aos objetivos bibliográficos definidos pela *International Federation of Library Associations* (IFLA), a saber:

- a) localizar entidades, em arquivo ou base de dados como resultado de uma busca por meio de atributos e relacionamentos entre as entidades;
- b) identificar uma entidade, isto é, confirmar que a entidade descrita em um registro corresponde à entidade desejada ou distinguir entre duas ou mais entidades com características similares;
- c) selecionar uma entidade que é apropriada às necessidades dos usuários;
- d) adquirir ou obter acesso à entidade descrita;
- e) navegar numa base de dados, isto é, encontrar obras relacionadas à determinada obra por meio de generalização, associação, agregação; encontrar atributos relacionados por equivalência, associação e hierarquia.

As autoras descrevem que no contexto da OI (Organização da Informação) e da RI (Representação da Informação), temos como objeto os registros de informação. Estamos, portanto, no mundo dos objetos físicos, distinto do mundo da cognição, ou das ideias, cuja unidade elementar é o conceito. A cognição, como afirma Fogl (1979, p.22), "é o processo de reflexão das leis e das propriedades de objetos e fenômenos da realidade objetiva na consciência humana". Ainda segundo o autor, o resultado da cognição é o conhecimento e não a informação, quando se refere à Organização do Conhecimento (OC) e à Representação do Conhecimento (RC), estamos no mundo dos conceitos e não naquele dos registros de informação. Tais sistemas de organização do conhecimento são elementos-chave das bibliotecas, museus e arquivos, pois são mecanismos de organização da informação, ou seja, a recuperação da informação por parte do usuário é altamente apoiada por tais sistemas.

Observa-se então, a existência de dois tipos distintos de processos de organização, um que se aplica às ocorrências individuais de objetos informacionais - o processo de organização da informação (OI), e outro que se aplica a unidades do pensamento (conceitos) - o processo de organização do conhecimento (OC). A OI compreende, também, a organização de um conjunto de objetos informacionais para arranjá-los sistematicamente em coleções, neste caso, temos a organização da informação em bibliotecas, museus, arquivos, tanto tradicionais

quanto eletrônicos. A organização do conhecimento, por sua vez, visa à construção de modelos de mundo que se constituem em abstrações da realidade. Esses dois processos produzem, conseqüentemente, dois tipos distintos de representação: a representação da informação (RI), compreendida como o conjunto de atributos que representa determinado objeto informacional e que é obtido pelos processos de descrição física e de conteúdo, e a representação do conhecimento (RC), que se constitui numa estrutura conceitual que representa modelos de mundo.

Considerando tal proposta, os índices e resumos seriam então representações da informação, pois representam um documento físico, individualmente, facilitando a sua recuperação. As representações do conhecimento seriam feitas por diferentes Sistemas de Organização do Conhecimento (SOC) que são sistemas conceituais que representam um domínio (por exemplo, um domínio científico) pela sistematização dos conceitos e das relações semânticas que se estabelecem entre eles, tais como os tesouros e as ontologias. Nota-se que os SOC podem ser utilizados para auxiliar a Organização da Informação (OI) através da construção de Representações da Informação (RI), como índices e resumos.

Sobre o mesmo tema, Tristão (2004) afirma que Sistemas de Organização do Conhecimento (SOC) incluem a variedade de esquemas que organizam, gerenciam e recuperam a informação. Existem desde os tempos remotos e estão presentes em todas as áreas do conhecimento humano, dos simples aos mais complexos. Esses sistemas abrangem classificação, tesouro, ontologia, assim como os conhecidos glossários e dicionários, específicos a cada área e, em sua maioria, ligados a bibliotecas e outras organizações de gerenciamento da informação visando a organizar, recuperar e disseminar a informação. Dessa maneira, adotam-se como definições dos termos que se seguem, os mais referenciados na literatura:

- a) Classificação: conjunto de conceitos organizados sistematicamente de acordo com os critérios ou características escolhidas;
- b) Tesouro: definido como um vocabulário de termos relacionados genérica e semanticamente sobre determinada área de conhecimento;
- c) Ontologia: especificação formal e explícita de uma conceitualização compartilhada, em que:
  - **Conceitualização** se refere a um modelo de fenômeno abstrato no mundo por ter identificado os conceitos relevantes daquele fenômeno,

- **Explícito** significa que o tipo dos conceitos usados e as restrições no seu uso são definidos explicitamente,
- **Formal** se refere ao fato de que a ontologia deveria ser lida pela máquina,
- **Compartilhado** reflete que ontologia deveria capturar conhecimento consensual aceito pelas comunidades.

Assim, os sistemas de classificação, as ontologias, as taxonomias e os tesouros são linguagens documentárias, ou seja, são sistemas artificiais de signos normalizados que permitem representação mais fácil e efetiva do conteúdo documental, com o objetivo de recuperar manual ou automaticamente a informação que o usuário solicita. Entende-se que as linguagens documentárias é que farão a comunicação entre a linguagem natural dos usuários e a unidade de informação, elas são utilizadas para representar o conteúdo dos documentos, por isso alguns autores as definem como sistemas simbólicos instituídos, que visam a facilitar a comunicação.

Especificamente, sobre os Sistemas de Classificação, de acordo com Piedade (1977), classificar é dividir em grupos ou classes, segundo as diferenças e semelhanças; é dispor os conceitos segundo suas semelhanças e diferenças, em certos números de grupos metodicamente distribuídos.

Existem diversas formas de classificação, o homem, por viver em sociedade, atribui classes no convívio social, classificando objetos, seres e pessoas de diversas formas: humanos do sexo masculino ou feminino, militares ou civis; classifica hierarquicamente as pessoas com quem convive, lhes atribuindo níveis de respeitabilidade, como: pais, parentes, amigos, colegas de trabalho, ou desconhecidos; classifica os seres como vivos ou não vivos e animados ou não animados, entre várias outras formas de classificação. A classificação é essencial ao homem, para que este possa compreender o mundo a sua volta, se organizar e executar as tarefas da melhor forma possível. A classificação da informação física segue a mesma vertente: classificam-se informações registradas para facilitar sua localização e, logo, poder agregar a informação obtida ao conhecimento e utilizá-lo de forma estratégica.

Historicamente, a classificação das áreas de conhecimento é um problema que vem sendo estudado e desenvolvido há séculos, com diversas propostas sobre a melhor forma de realizar tal classificação. Sobre o surgimento da classificação bibliográfica e das áreas de conhecimento, Dahlberg (1976) coloca:



No início, a sistematização do conhecimento não era feita de maneira esquemática como a conhecemos hoje. Até 1491 não era hábito elaborar sistemas para a classificação das ciências como um fim em si mesmo. Provavelmente só após 1491, quando o humanista e poeta italiano Angelo Poliziano publicou seu "Panepistemon" - um plano destinado não a ser o esboço de um texto, mas a mostrar esquematicamente as relações entre as ciências ou áreas do conhecimento - é que realmente foi iniciado o "movimento" de elaboração de sistemas de classificação. Após Poliziano, muitos outros tentaram a mesma coisa, nenhum deles tão conhecido como Francis Bacon que, cerca de cem anos depois, em 1605 para sermos exatos, publicou um plano de classificação das ciências em seu trabalho "De dignitate et augmentis scientiarum". Contudo, esta arte não foi chamada de "classificação" até quase duzentos anos mais tarde, por volta do fim do século XVIII. Somente a partir dessa época é que temos evidências, especialmente através das bibliografias de C. W. Shields, R. Flint e B. C. Richardson, de que o termo "classificação" foi utilizado em títulos de livros, relacionado com a apresentação de um plano para a classificação das ciências e dos livros. No século XIX especialmente, a elaboração de tais planos tornou-se um hobby para cada filósofo, bem como para alguns cientistas - por exemplo, o físico A.-M. Ampère- e até para um homem de estado como T. G. Masaryk, presidente da Tcheco-Eslováquia (1886). A inspiração decorrente desses trabalhos filosóficos também influenciou os bibliotecários no sentido de construírem continuamente novos sistemas para a organização do conteúdo de suas coleções de livros (DAHLBERG, 1978).

Segundo a finalidade a que se destina, a classificação pode ser filosófica ou bibliográfica, a primeira foi criada pelos filósofos com a finalidade de definir, esquematizar e hierarquizar o conhecimento, preocupados com a ordem das ciências ou a ordem das coisas. As classificações bibliográficas são sistemas destinados à classificação de documentos nas estantes, em catálogos, em bibliografias, etc.

### 2.3.1 Tipos de Classificação Bibliográfica

Tristão (2004) descreve os tipos de classificação:

#### *Classificações Especializadas e Gerais*

Uma classificação denomina-se por especializada, se tiver por objetivo um assunto em particular, como, por exemplo, o sistema de classificação da *United Classification for the Construction Industry (Uniclass)*, direcionado à indústria da construção, ou geral, se pretende cobrir o universo mais complexo da informação, como, por exemplo, à área de Ciência da Informação, a Classificação Decimal Universal (CDU).

#### *Classificações Analíticas e Documentais*

Uma classificação denomina-se analítica quando pretende sistematizar fenômenos físicos e providencia uma base para a sua explicação e entendimento. Também se denominam por classificações científicas ou taxonomias, como exemplo, a classificação do reino animal. Uma classificação designa-se como documental, quando a sua utilização pressupõe a

classificação de documentos ou outros tipos de informação, com o objetivo principal de facilitar a localização dessa informação, como exemplo, a Classificação Decimal Dewey (CDD), bastante utilizada em bibliotecas.

#### *Classificações Enumerativas*

São classificações que prescrevem um universo de conhecimento subdividido em classes sucessivamente menores que incluem todas as possíveis classes compostas (relações sintáticas). Essas classes são organizadas de forma a apresentar suas relações hierárquicas. Apresenta-se em listagem exaustiva de termos, organizados em classes e subclasses. Este tipo de classificação é limitativo, uma vez que coloca dificuldades à inserção de novos termos. A ordem predefinida para os termos em cada classe apenas permite a introdução de novos termos de forma sequencial. Relativamente à notação, por exemplo, de produtos, os dígitos de reserva necessários para a introdução de novos produtos são de difícil previsão, podendo tornar a notação muito extensa.

#### *Classificações por facetas*

Desenvolvida por Shiyali Ramamrita Ranganathan na década de 1930, atualmente tem sido largamente discutida na academia como uma solução para a organização do conhecimento, em decorrência de suas potencialidades de acompanhar as mudanças e a evolução do conhecimento. Muitos termos e expressões têm surgido, mas retratam nada mais do que a classificação facetada que, segundo Ranganathan (1967), conceitua o conhecimento “como a totalidade das ideias conservadas pelo ser humano” por meio da observação das coisas, fatos e processos do mundo que o cerca.

Os Sistemas de Classificação Bibliográfica utilizam conjuntos simbólicos para representar os assuntos contidos nas fontes de informação. É uma linguagem formal, controlada, que permite a conexão entre o usuário da informação e a informação desejada. Logo, utilizando essa interface comunicativa, é possível converter uma consulta em linguagem natural (português, por exemplo) do usuário para a linguagem controlada e localizar a informação na estante de uma biblioteca. Verifica-se que tal sistema procura aumentar a eficácia (acesso ao material) e eficiência (tempo) na recuperação da informação desejada pelo usuário dentre centenas ou milhares de documentos existentes, ainda que estes sistemas de classificação bibliográfica sejam amplamente dependentes da intervenção humana no casamento da informação catalogada no repositório e da consulta do usuário. Observa-se

que nada impediria, porém, a aplicação de tais Sistemas de Classificação tradicionais numa biblioteca ou repositório digital, para auxiliar a busca da informação registrada.

Segundo Carlan (2010), não existe um esquema de classificação do conhecimento sobre o qual todos concordem. Um Sistema de Organização do Conhecimento (SOC) pode ser significativo e vantajoso para uma cultura, uma coleção ou um domínio e para outros pode não ser. Apesar da multiplicidade de maneiras para organizar o conhecimento, Hodge (2000) aponta algumas características comuns dos SOC usadas em organização de bibliotecas digitais:

- a) os SOC impõem uma visão particular do mundo, de uma coleção e de itens;
- b) a mesma entidade pode ser caracterizada de diferentes maneiras, dependendo do SOC que é usado e;
- c) deve haver identificação suficiente entre o conceito expresso no SOC e o objeto do mundo real, ao qual aquele conceito se refere. Pois assim, quando uma pessoa procura algo sobre determinado objeto, o SOC deve ser capaz de conectar o conceito do objeto com sua respectiva representação no sistema.

Os SOC tradicionais, como as classificações e tesouros, têm sido utilizados também para organizar recursos digitais na Internet. Com a Web Semântica, as ferramentas para desenvolvimento de SOC estão se popularizando, devido à necessidade de compartilhamento com uso de padrões orientados ontologicamente.

Sobre uma Teoria da Classificação, Carlan (2010, p. 66) considera a existência de uma diversidade de visões sobre o tema classificação entre os autores, porém, afirma que há elementos essenciais que caracterizam o processo de classificar, que é a formação metódica e sistemática de grupos onde se estabelecem critérios para a divisão. Segundo a descrição da autora, a classificação é, provavelmente, o método mais simples de descobrir ordem na múltipla e confusa diversidade da natureza. É usada como instrumento de representação do conhecimento com a finalidade de organizar e recuperar informações.

## **2.4 A Classificação na Arquivologia e Museologia**

De acordo com Siqueira (2011), ao contrário da Museologia e da Biblioteconomia que também lidam com a informação registrada e até compartilham procedimentos técnicos, nota-se que a informação arquivística é originária de fontes únicas, que se organizam segundo sua

proveniência para fins de prova, principalmente utilizadas em contextos jurídicos e administrativos.

Na Arquivologia, a classificação como método empregado para facilitar a localização de documentos tem outras formas de ação. Logicamente, se os objetos de estudo: informação e documento são vistos de maneiras diferentes entre a Biblioteconomia, Arquivologia e Museologia, logo, as técnicas de classificação com tais objetos seriam também diferentes.

As diferenças se evidenciam inclusive nos aspectos terminológicos no campo teórico da classificação; sobre a Arquivologia, Golçalves (1998) afirma que:

No meio arquivístico brasileiro, foi consagrada a distinção entre “classificação” e “arranjo”. De acordo com tal distinção, a “classificação” corresponderia às operações técnicas destinadas a organizar a documentação de caráter **corrente**, a partir da análise das funções e atividades do organismo produtor de arquivos. Por seu turno, o “arranjo” englobaria as operações técnicas destinadas a organizar a documentação de caráter **permanente** (GONÇALVES, 1998, p.11).

A autora também enfatiza que a criação das classes no processo de classificação depende totalmente do produtor dos documentos de arquivo: são criadas categorias, **classes** genéricas, que dizem respeito às funções/atividades detectadas (estejam elas configuradas ou não em estruturas específicas, como departamentos, divisões, etc.).

Além do procedimento de classificação, existe a ordenação: quanto à **ordenação**, seu objetivo básico é facilitar e agilizar a consulta aos documentos, pois, mesmo no que se refere a uma mesma atividade, e em relação a um mesmo tipo documental, os documentos atingem um volume significativo. A adoção de um ou mais critérios de ordenação para uma série documental permite evitar, em princípio, que, para a localização de um único documento, seja necessária a consulta de dezenas ou centenas de outros.

O procedimento técnico de classificação alcança, portanto, os tipos documentais (identifica-os e articula-os entre si), mas considera, sobretudo, a forma e as razões que determinaram sua existência (como e por quê foram produzidos). Já a ordenação aborda os tipos documentais especialmente do ponto de vista das consultas que lhes forem feitas. Cabe à ordenação definir a melhor maneira de dispor fisicamente as notas de empenho (numericamente?), os extratos bancários (cronologicamente?) e todos os demais tipos documentais.

Uma questão importante em um plano de classificação de documentos de arquivo é o critério de classificação: um critério **funcional** (classes correspondendo estritamente a funções) ou **estrutural** (classes correspondendo a “estruturas” - setores, divisões, departamentos). A opção pela classificação “estrutural” é, tradicionalmente, mais aceita e adotada. Apresenta, porém, inconvenientes - quando não há estruturas que digam respeito à totalidade das funções e atividades do organismo; quando, eventualmente, as estruturas existentes são confusas, misturando indevidamente funções; quando as estruturas sofrem alterações constantes. De modo geral, e salvaguardadas as exceções de praxe, a opção pela classificação estritamente “funcional”, apesar de menos frequente e tecnicamente mais complexa, costuma atender melhor as exigências da classificação arquivística. No entanto, cabe ao profissional de arquivo examinar cada situação e decidir pelo que se apresenta como tecnicamente mais correto.

Em relação à trajetória temporal da prática de classificação em arquivos, Sousa (2006) faz a seguinte consideração sobre a interdisciplinaridade na questão da classificação:

As reflexões sobre a classificação de documentos arquivísticos na literatura apresentam alguns aspectos comuns. O primeiro deles é que essa operação intelectual não agregou em suas concepções e nos seus fundamentos as contribuições da classificação vindas da Filosofia e, posteriormente, da Teoria da Classificação. A teoria do conceito, que estabelece as várias relações possíveis entre os conceitos, é desconhecida pela teoria arquivística. Os requisitos e os princípios desenvolvidos nessas áreas quando aparecem é de forma muito tímida. Observou-se, apenas, nos trabalhos de Schellenberg alguma influência desses conhecimentos no processo classificatório em Arquivística.

Isso demonstra, de certa forma, a falta de comunicação da Arquivística com outras áreas do conhecimento, que podem contribuir para o desenvolvimento de um arcabouço teórico-metodológico próprio da disciplina, levando em consideração as especificidades do objeto de estudo. Esteban Navarro (1995, p. 67), analisando a relação da Arquivística com as outras áreas da documentação (Biblioteconomia e Documentação), percebe que essa ausência de diálogo ocorre, também, pela falta de interesse das outras disciplinas em conhecer e compreender as peculiaridades do trabalho realizado nos arquivos.

Esteban Navarro propõe, de uma maneira pioneira, a articulação da gestão documental dentro da área da representação e organização do conhecimento. Não estamos defendendo a “importação” sem critérios de conhecimentos de outras disciplinas, mas, no caso específico da classificação, a construção de um saber interdisciplinar que confira ao processo classificatório um fundamento teórico-metodológico. Acreditamos na possibilidade de construção de esquemas de classificação baseados nos princípios próprios da área e de seu objeto de estudo e nos conceitos e requisitos da classificação desenvolvidos pela Filosofia, pela Teoria da Classificação e, recentemente, pela Teoria do Conceito. O processo classificatório em Arquivística resente dessa ausência e a prática é testemunha desse fato (SOUSA, 2006, p.137).

Sobre o uso de vocabulários controlados na organização da informação arquivística, Aguiar (2008) afirma que do ponto de vista do ciclo documentário arquivístico (produção, registro, organização, disseminação, recuperação e assimilação) pode-se dizer que o uso de uma linguagem normalizada de acordo com o contexto e a cultura organizacional de uma instituição é um dos fatores determinantes para garantir a dinâmica e a totalidade do ciclo documentário. A linguagem normalizada e comum a toda a organização teria o papel de garantir a comunicação compartilhada entre departamentos, por exemplo, evitando a ambiguidade da linguagem natural e a imposição de uma linguagem dos setores produtores de informação (jurídica, fiscal-financeira, administrativa, etc.). A normalização com base terminológica, raras vezes, ocorre somente na fase permanente dos arquivos, no momento da elaboração de um quadro ou plano de classificação.

Em relação à similaridade da atividade de Classificação na Biblioteconomia, Arquivologia e Museologia, Siqueira (2011) afirma que nas propostas de definição do termo classificação nos domínios das três ciências, há grande proximidade terminológica, já que mesmo utilizando expressões linguísticas distintas, elas reportam a significados análogos. Mesmo que a Arquivologia denomine as principais características do termo classificação como “ação intelectual” e “disposição físico-material”; a Biblioteconomia prefira as designações de “operação intelectual” e “operação material”; e a Museologia opte por “aspecto intelectual” e “aspecto físico”, temos nos três domínios uma representação comum, ou seja, o termo possui uma faceta de caráter cognitivo e outra de natureza material.

Quanto ao caráter cognitivo é comum nos três domínios o estabelecimento de classes que reúnam documentos com atributos comuns, ação que se diferencia entre as três ciências no que tangem aos critérios para a sua elaboração. No âmbito da Arquivologia, o elemento chave que definirá a organização sistemática é a função/atividade desempenhada pela entidade ou pessoa produtora do fundo, ou seja, é o caráter orgânico-funcional que delimita as classes. Já na Biblioteconomia, as classes são definidas pelo assunto, também denominada como produto de representação temática. Na Museologia, consideram-se os critérios da natureza da coleção, a função do objeto e a perspectiva do sujeito ou da civilização produtora do artefato. Em relação ao aspecto físico-material, também notamos uma proximidade funcional desse caráter nos três domínios. Nos três domínios, observa-se que a classificação no âmbito material tem o objetivo de auxiliar na localização, acesso e recuperação de documentos, ações que se vinculam indiretamente à ação intelectual, pois a disposição física está atrelada à organização intelectual.

O índice e o catálogo podem aparecer nos três domínios, no entanto, não só seu conteúdo, como sua estrutura pode ser totalmente distinta. O catálogo, por exemplo, considerado um instrumento de pesquisa pode ser elaborado segundo critérios temáticos, cronológicos, onomásticos ou geográficos. Mesmo se considerarmos um tipo, o temático, por exemplo, observamos que ele terá uma estrutura específica para cada instituição. No arquivo se consideram as séries arquivísticas, tendo como pressuposto o caráter da proveniência dos fundos; na biblioteca formada por uma coleção com obras múltiplas se evidenciará o assunto; já o museu com obras únicas, exige-se maior detalhamento de caracteres, o que justifica um catálogo *raisonné*<sup>2</sup>, por exemplo, usado na sobreposição de características temáticas com dados bibliográficos do artista.

## **2.5 Considerações sobre a classificação na Biblioteconomia, Arquivologia e Museologia**

Como apresentado nos parágrafos anteriores (e como era de se esperar, já que os conceitos básicos de informação, conhecimento, documento e outros conceitos são estabelecidos em diferentes perspectivas) há diversas visões para o tema, a teoria e prática da classificação. Cada área ou especialidade do conhecimento (Filosofia, Biblioteconomia, Arquivologia, Museologia, Ciência da Computação e Ciência da Informação) tratam a classificação utilizando terminologias e metodologias particulares, ainda que existam semelhanças e empréstimos no desenvolvimento da teoria ou processo classificatório.

Sobre a contribuição destas visões citadas anteriormente para a atividade de classificação (como teoria ou processo) em relação ao trabalho aqui descrito, observa-se que os Sistemas de Organização do Conhecimento (SOC) e as definições da Teoria da Classificação da Biblioteconomia e Ciência da Informação estão altamente relacionadas à representação do conhecimento, ou a uma linguagem de representação de conceitos ou termos e suas relações semânticas, com vistas a auxiliar o humano na organização da informação ou documentos, tradicionalmente em formato impresso. Porém, a perspectiva do trabalho aqui descrito é diferente: neste contexto, o enfoque e o trabalho de pesquisa maior é destinado ao desenvolvimento de sistemas de classificação automática, aplicando as técnicas de

---

<sup>2</sup> Um catálogo *raisonné* é uma listagem abrangente, com anotações de todas as obras conhecidas do artista, quer em uma determinada mídia ou todas as mídias. Como exemplo, pode-se citar o catálogo *raisonné* on-line de Picasso em <http://onlinepicassoraisonne.com/>. Acesso em 04/10/2013.

Inteligência Computacional, para a construção dos sistemas de classificação, em oposição aos tradicionais sistemas de classificação de conceitos ou termos já utilizados há tempos em bibliotecas.

Apesar dos assuntos serem distintos, pois há o objetivo de construção do mecanismo computacional para automatizar a classificação de registros digitais, eles têm em comum o objetivo geral que é colocar registros textuais em classes específicas para facilitar a recuperação e, neste sentido, os sistemas computacionais se igualam a todos os outros descritos em termos de função social. Além disso, os sistemas tradicionais e a automação classificatória se inter-relacionam e se complementam em diversas etapas: a indexação automática claramente procura simular uma indexação humana, como será visto adiante; também, as linguagens documentárias ou sistemas de organização do conhecimento (SOC), como os tesouros e as ontologias, podem ser utilizados em conjunto com os sistemas computacionais inteligentes para auxiliar a classificação automática.

Portanto, nas seguintes seções, visando diferenciar as abordagens apresentadas e para diferenciar essas duas vertentes, utilizaremos o termo *Sistema de Classificação Automática de Textos* para referenciar a classificação usando o mecanismo computacional e *Sistema de Organização da Informação* e *Sistema de Organização do Conhecimento* para os tradicionais sistemas de organização e representação de documentos e conceitos.

## **2.6 Indexação manual**

A indexação por termos ou palavras-chave tem sido utilizada há tempos nas tradicionais bibliotecas institucionais, cujo resultado é representar os registros textuais com uma linguagem controlada ou conceitos que representam de forma sucinta um texto maior e, desta forma, auxiliar o processo de recuperação da informação. Silva e Fujita (2004) descrevem alguns pontos históricos importantes sobre o desenvolvimento dos métodos de indexação (inicialmente, métodos manuais), como descrito abaixo:

A indexação surgiu com a atividade de elaboração de índices. Gomes e Gusmão (1983, p.12) afirmam que o índice, como um instrumento de armazenagem e recuperação da informação, tem sua origem a partir do momento em que o homem passou a se preocupar em tornar acessível a informação registrada em um documento e para isso resolve ordená-la de alguma forma. A forma mais antiga de armazenagem de informação de que se tem conhecimento foi encontrada nas tábuas de argila produzidas pela extinta Mesopotâmia no



século II A.C. Nelas foi grafada uma espécie de resumo dos livros antigos considerada como forma de representação condensada do conteúdo informacional que dava acesso ao assunto dos livros (WITTY, 1973).

No histórico da indexação, Collinson (1971) indica que o primeiro tipo de indexação existente era baseado na memória. Textos célebres, como as grandes epopeias, por exemplo, eram transmitidos oralmente. Depois disso, os primeiros índices de que se têm notícia eram arrançados pela primeira sentença de cada parágrafo. Na Biblioteca de Alexandria, organizada pela classificação de Calímaco, seu catálogo era arrançado em ordem alfabética de autores e subordinados a assuntos mais gerais. Várias obras, principalmente as histórias e peças dos grandes dramaturgos da época, eram condensadas.

Para Kobashi (1994) a documentação como é praticada hoje, nasceu no século XVII com a edição de *Le Journal des Sçavans* publicado em Paris no ano de 1665. Tratava-se de um periódico semanal que trazia os resumos dos trabalhos científicos, filosóficos e artísticos. Esse periódico deu origem a uma série de outros posteriores de mesma natureza que surgiram na Europa. Nos séculos seguintes, XVIII e XIX, aconteceu o crescimento com mais intensidade de periódicos referenciais que atualmente encontram-se no formato eletrônico denominados base de dados. Até o surgimento da imprensa, os índices eram a única forma de acesso aos livros encontrados nas bibliotecas dos mosteiros, a partir do registro dos títulos dos livros. A partir de então, houve um significativo aumento da literatura que impulsionou o aparecimento de várias listas com diferentes finalidades. Konrad Gesner elaborou um repertório geral e europeu – o *Bibliotheca Universalis* – no qual relacionava cerca de 12 mil títulos de todos os livros latinos, gregos e hebraicos de seu conhecimento. Mais tarde foi publicado o índice alfabético de assunto do referido repertório, cujo nome era *Pandectarum sive partitionum uníversalium, libri XXI*.

Segundo Chaumier (1971) foi em 1931 que a palavra “documentação” começou a ser usada. Os organismos criados pelas atividades de Documentação foram o *Institute International de Documentation* (FID) e a *Union Française des Organismes de Documentation* (UFOD). Os principais instrumentos de organização documentária criados foram os sistemas de classificação bibliográfica com destaque para a Classificação Decimal Universal (CDU), os estudos para criação de sistemas classificatórios realizados em 1929 e 1933 por H.G. Bliss e R.S. Ranganathan, a criação da Classificação Decimal de Dewey (CDD), além dos repertórios documentais que incluíam as bibliografias, códigos de

abreviaturas dos títulos de periódicos e catálogos bibliográficos. É importante ressaltar que La Fontaine e Paul Otlet introduziram o princípio de pré-coordenação, ao criarem na CDU o uso de dois pontos (:) para relacionar duas classes de assunto.

Apesar de atualmente presenciarmos uma evolução ainda maior da indexação, é importante considerarmos a importância do índice enquanto ferramenta de busca. Robredo (1994, p.202) classifica o índice em dois sentidos: no sentido tradicional e no amplo. No primeiro sentido afirma ser uma listagem alfabética ou sistemática de tópicos que indicam a existência e localização de cada um deles num documento ou em uma coleção de documentos. No segundo sentido, “... um conjunto ordenado de códigos representativos de assuntos, tópicos ou conceitos (por exemplo, códigos de classificação, grafismos diversos, incluindo palavras ou frases), os quais podem servir como critérios de busca relacionando com alguma chave de acesso que permita localizar os documentos – ou suas partes ou representações – relativas a cada assunto”. O autor complementa dizendo que o índice é o mais significativo instrumento para recuperação da informação, sendo definido como uma ‘chave’ condensada que dá acesso à informação contida nos documentos, ou como uma ponte entre o conteúdo de um acervo de informação e os usuários (ROBREDO, 1994, p.244).

Examinando a literatura da área, entende-se que a indexação pode ser metodologicamente realizada de diversas maneiras possíveis, e diferentes técnicas ou métodos para a obtenção dos termos-chave que melhor identificam o texto escrito poderiam ser utilizados. Considerando uma divisão ampla da área, é possível dividir as possibilidades de indexação em três enfoques: manual (humana), automática ou híbrida; para a indexação automática, ainda poderíamos tê-la realizada com enfoque estatístico ou linguístico, ou ambos (híbrido).

Sobre a indexação manual, Fujita (2003) descreve o processo como sendo formado por duas etapas: a *analítica*, em que é realizada a compreensão do texto como um todo, a identificação e a seleção de conceitos válidos para a indexação, e a segunda etapa o *estágio de tradução*, que consiste na representação de conceitos por termos de uma linguagem de indexação:

- a) determinação do assunto: estabelecimento dos conceitos tratados num documento;
- b) representação de conceitos por termos de uma linguagem de indexação: a tradução dos conceitos nos termos da linguagem de indexação, geralmente utilizando para isso uma linguagem documentária, tal como o tesauro.

O primeiro estágio, a análise de assunto, é subdividido em outros três estágios:

- a) compreensão do conteúdo do documento;
- b) identificação dos conceitos que representam este conteúdo;
- c) seleção dos conceitos válidos para recuperação

A autora ainda destaca que a análise de assunto é uma etapa importante do trabalho do indexador. Tem como objetivo identificar e selecionar os conceitos que representam a essência de um documento. O processo de identificação de conceitos envolve certo grau de complexidade por exigir do indexador o uso de metodologia adequada para garantir bons resultados na recuperação, o que pressupõe o conhecimento de abordagens sistematizadas ao texto. Além disso, pela análise de literatura, a identificação de conceitos depende da tematicidade do texto e está atrelada à leitura do indexador e às suas concepções de análise de assunto adquiridas pela sua formação, objetivos e políticas de indexação.

Ainda sobre a fase analítica, Fidel (1994) afirma que a determinação da tematicidade de um documento, ou sobre o que ele trata, pode ser algo subjetivo. Do ponto de vista prático, fortes evidências sobre a natureza subjetiva do processo de indexação têm sido verificadas em testes de consistência entre indexadores. Muitos desses testes têm mostrado um indicativo de baixa concordância entre os indexadores sobre como indexar um documento, e que eles frequentemente indexam um mesmo documento de forma diferente em ocasiões diferentes.

Sobre a segunda fase do processo de indexação, a representação de conceitos por termos de uma linguagem de indexação (ou a fase de tradução), Fidel (1994) afirma que a experiência dos indexadores aponta um conjunto de considerações a serem observadas ao se estabelecer uma política de indexação:

- a) *Fonte dos termos de índice*: qual fonte vocabular pode o indexador utilizar para a seleção dos termos? A política limita o indexador ao uso do tesouro ou termos da linguagem natural podem ser utilizados?
- b) *Especificidade*: o quão específico deve o indexador ser ao traduzir um conceito em termos de índice, ou seja, deveria o termo do índice se limitar ao conceito observado, ou deveria ser ampliado?
- c) *Pesos*: o indexador pode estabelecer um peso para um termo no índice de um documento? Conceitos mais importantes ou centrais receberiam um peso maior no processo de indexação.

- d) *Precisão*: o indexador deveria indexar considerando termos relacionados? Ou, como proceder quando um termo não possui descritores equivalentes?
- e) *Nível de combinação de termos*: o indexador deve atribuir termos elementares ao índice ou utilizar combinação de termos? Por exemplo, o termo “educação em saúde”, deve ser inserido no índice como “educação em saúde”, ou “educação” e “saúde” como termos separados?
- f) *Linguagem do Usuário*: o indexador deveria utilizar uma linguagem próxima ao do usuário do índice? Algumas políticas de indexação apontam o usuário pretendido, se eles são profissionais da área ou o público leigo.

Algumas políticas de indexação também enfocam a análise de conteúdo:

- a) *Exaustividade*: quão compreensível ou exaustivo deveria ser a descrição do índice, quais dos vários aspectos de um conteúdo deverão ser representados no índice? Por exemplo, um artigo descrevendo um projeto na área de educação em saúde deveria também conter os termos “mulheres” e “americanos asiáticos” se a maioria dos participantes no projeto são “mulheres americanas asiáticas”?
- b) *Material indexável*: qual parte intelectual de um documento o indexador deveria considerar na representação do conteúdo? “resultados negativos”, “implicações”, “usos possíveis”, “sugestões” e “pesquisas futuras” deveriam participar no processo de escolha de termos?

## 2.7 Indexação automática

A indexação automática utiliza os computadores para selecionar os termos que melhor representam um documento textual, sem intervenção humana, visando uma posterior recuperação de documentos eficiente e eficaz. Também, há relatos de estudos em indexação híbrida, com a intervenção de máquinas e agentes humanos. Apesar da automação passar a ideia de modernidade em relação à indexação manual, os estudos em indexação automática têm sido realizados já desde a década de 1960 do século XX.

De acordo com Silva e Fujita (2004), foi na década de 1960 que o índice KWIC (*Key-Word in Context* ou Indexação pela Palavra-Chave no Contexto) apareceu propagando um novo método de indexação: a indexação pela palavra. Representa a primeira aplicação de indexação automática de documentos técnicos, tendo por base as palavras significativas dos títulos. A história da indexação evidencia que a sua criação foi atribuída a William Frederick

Poole que em 1882 com a publicação de “Poole’s Index”, criou um índice dando entrada do assunto pela palavra-chave do título dos artigos desse periódico. É atribuída a Poole a criação do índice KWIC (BORKO; BERNIER, 1978, p.8). O índice KWIC é caracterizado pelo uso da linguagem natural, conseqüentemente não há controle de termos significativos e os sinônimos não são identificados.

Ward (1996) indica as vantagens e desvantagens do uso do indexador automático.

Desvantagens de um indexador automático:

- a) funciona somente em documentos separadamente;
- b) não consegue fazer relações entre os textos ou entre um texto e uma visão de mundo;
- c) fica amarrado ao vocabulário e à gramática usada no documento indexado;
- d) não consegue lidar com dados gráficos;
- e) não consegue lidar com línguas estrangeiras;
- f) não consegue avaliar textos;
- g) não consegue criar relações intertextuais;
- h) só consegue indexar o que está explícito, não consegue indexar o que está implícito;
- i) não é capaz de imitar o questionamento, a resposta humana a um texto, o que acrescenta valor à indexação;
- j) requer constante aprimoramento para manter-se em dia com os novos desenvolvimentos;
- k) não consegue catalogar ou classificar.

Vantagens de um índice automático:

- a) leitura instantânea de todo texto;
- b) diz-se que é mais coerente do que um indexador humano;
- c) não é tendencioso.

A discussão sobre a melhor forma de indexação (ou a mais produtiva) entre a indexação manual e automática também tem sido um tópico recorrente na área. Viera (1988) afirma que ambas as técnicas são consideradas eficientes. Em alguns casos há maior aceitação da indexação automática, em outros, da manual. Depende das línguas, das áreas do

conhecimento em que foram aplicadas e das fontes de informação utilizadas na extração do termo que expressará o assunto do documento.

Segundo a literatura da área (descrita a seguir), os métodos para extração automática de termos e formação do índice variam em três possibilidades metodológicas: indexação linguística, estatística ou híbrida.

Uma técnica de indexação automática e estatística bastante utilizada e conhecida é baseada na frequência dos termos do texto. As leis bibliométricas de Zipf e Goffman podem ser utilizadas para tal indexação. No trabalho de Guedes (1994), dentre os estudos de indexação automática, são de interesse os estudos bibliométricos, fundamentados na frequência de ocorrência das palavras, principalmente, nas leis de Zipf e Ponto T de Goffman. Zipf observou que, em um texto suficientemente longo, o produto da ordem de série ( $r$ ) de uma palavra (dada pela frequência de ocorrência em ordem decrescente) pela sua frequência de ocorrência ( $f$ ) era aproximadamente constante. Enunciou, então, que

$$r \cdot f = c \text{ (equação 1)}$$

expressão que ficou conhecida como Primeira Lei de Zipf. A Segunda Lei de Zipf enuncia que, em um texto, várias palavras de baixa frequência de ocorrência (alta ordem de série) aparecem o mesmo número de vezes. Booth, ao modificá-la, a representa matematicamente por:

$$\frac{l_1}{l_n} = \frac{n(n+1)}{2} \text{ (equação 2)}$$

onde  $l_1$  é o número de palavras que têm frequência 1, e  $l_n$ , o número de palavras que têm frequência  $n$ .

Os comportamentos, inteiramente distintos, da primeira e segunda Lei de Zipf definem as duas extremidades da lista de distribuição de palavras de um texto. Assim, é razoável esperar uma região crítica, na qual há a transição do comportamento das palavras de baixa frequência para as de alta frequência. Para se chegar a essa região de transição, a expressão da 2ª Lei de Zipf teria de fornecer o comportamento típico das palavras de alta frequência, isto é, o número de palavras que têm frequência  $n$  tenderia a 1.

Substituindo-se, na expressão da 2ª Lei de Zipf (equação 2),  $l_n$  por 1, obtém-se:

$$\frac{l_1}{1} = \frac{n(n+1)}{2} \text{ (equação 3)}$$

ou ainda, rearranjando:

$$n^2 + n - 2l_1 = 0 \text{ (equação 4)}$$

cujas raízes são:

$$n = \frac{-1 \pm \sqrt{1+8l_1}}{2} \text{ (equação 5)}$$

sendo utilizado o resultado positivo apenas.

Ao valor de  $n$  assim determinado dá-se o nome de Ponto de Transição de Goffman (T). O Ponto de Transição de Goffman determina a vizinhança onde, de acordo com Goffman, devem estar incluídas as palavras de maior conteúdo semântico e, portanto, aquelas que seriam usadas para a indexação de um texto em questão. Esta linha de raciocínio representa um passo importante na busca de um critério de indexação automática. Segundo Guedes (1994), vários estudos posteriores se baseiam em tais leis bibliométricas para a indexação de textos em áreas diversas.

Sobre a indexação automática utilizando métodos de análise linguística, Narukawa, Gil-Leiva e Fujita (2009) afirmam que tal forma de indexação surge como uma tentativa de resolver os problemas da indexação baseada em seleção estatística de palavras, obviamente tais métodos estatísticos são superficiais, pois não consideram, por exemplo, relações sinonímias entre termos e a existência de ambiguidades. Nesse sentido, Leiva (1999, p. 82) explica que a partir do início dos anos sessenta associam-se as técnicas de Processamento da Linguagem Natural (PLN) que consiste no estudo e análise dos aspectos linguísticos de um texto mediante a utilização de programas informáticos – e a automatização da indexação. Os estudos linguísticos avançaram em direção à compreensão da estrutura textual, suas relações e seu significado.

Segundo Leiva (2008, p. 339), os primeiros analisadores linguísticos surgiram na década de 1960 para o processamento automático de informação. Os avanços e melhorias produzidas nestes sistemas têm permitido utilizá-los para a recuperação da informação, extração de informação, classificação, indexação e resumos de documentos ou para o

reconhecimento automático da fala. Estes analisadores linguísticos, coincidindo com os níveis de linguagem, se dedicam ao tratamento das palavras (analisador morfológico), ao tratamento das orações (analisador sintático) e ao tratamento das palavras e orações segundo o contexto em que se encontram para conhecer seu significado (analisadores semânticos), e coloca-se também, a interpretação dos enunciados levando em consideração o contexto de uso, o estilo e a prática social (analisadores pragmático-discursivos).

Na atualidade, segundo Guimarães (2000), verificam-se os métodos mistos ou híbridos de indexação automática que reúnem aportes da estatística, da linguística textual e ainda utilizam tesouros como instrumento de controle de vocabulário, auxiliando e contribuindo para eliminar problemas como a sinonímia e a identificação de funções sintáticas dos termos, proporcionando benefícios à revocação na recuperação da informação.

As últimas tendências da automatização da indexação é denominada de indexação inteligente, por Mendez Rodríguez e Moreira González (1999). Explicam que esse tipo de indexação está voltado ao acesso direto de documentos por meio do processamento linguístico automático e uso de linguagem natural combinando outras técnicas como análise estatística ou a ponderação dos termos. Esses sistemas buscam interfaces inteligentes para que o usuário possa utilizar a linguagem natural como linguagem de intercâmbio de conhecimento e é atribuída ao computador a competência linguística e/ou cognitiva, tendo não só bases linguísticas, mas também bases de conhecimento.

## **2.8 Indexação automática para o português do Brasil**

Para o português do Brasil, é possível encontrar trabalhos publicados sobre indexação automática, considerando o impacto das características da língua no processo de indexação, já há trinta anos.

Andreewski e Ruas (1983), por exemplo, descrevem em seu artigo a indexação automática utilizando o processamento de documentos em linguagem natural, que é obtido com o auxílio de métodos linguísticos combinados com métodos estatísticos permitindo uma indexação ponderada. A título ilustrativo, o autor descreve em linhas gerais um sistema de indexação desse gênero denominado SPIRIT, o qual foi desenvolvido originalmente para o idioma francês. No texto, são tratados aspectos essenciais de sua adaptação à língua portuguesa.



Outro estudo sobre indexação estatística utilizando o Ponto T de Goffman é realizado por Mamfrim (1991) utilizando textos científicos em português do Brasil. A autora (para os estudos da época) conclui que o uso desta técnica estatística se mostrou viável no processo de identificação de termos-chave, onde a região de texto apontada realmente revela termos que identificam a temática do texto. A autora ainda afirma que os resultados obtidos para a Fórmula de Transição de Goffman são mantidos, sem maiores problemas, para o português.

Desde então, o tema tem evoluído ao longo dos anos com propostas diversas para a indexação tanto nos paradigmas estatísticos quanto linguísticos ou híbridos. Algo importante a destacar, é que para tal processamento linguístico ser eficaz, uma série de recursos linguístico-computacionais é necessária para a língua a ser processada. Tesouros, dicionários, ontologias, etiquetadores morfosintáticos, analisadores sintáticos, *stemmers*, extratores sintagmáticos, corpora e estudos sobre a estrutura e as características do português brasileiro são recursos necessários para a construção de índices eficazes e representativos (e para obter resultados eficientes e eficazes de classificação e recuperação da informação, em geral).

O Núcleo Interinstitucional de Linguística Computacional (NILC) foi criado no Brasil em 1993 com o objetivo justamente de gerar tais recursos, visando ampliar os estudos relacionados ao processamento do português brasileiro que necessitam de tais recursos previamente construídos. O NILC tem se empenhado em garantir a geração e manutenção de tais recursos (NUNES; ALUÍSIO; PARDO, 2010).

Sobre indexação e recuperação da informação utilizando processamento linguístico voltado ao português brasileiro, Duque (2005) propõe o desenvolvimento de um Sistema de Recuperação da Informação (SRI) que utiliza teorias da Linguística Computacional e Ontologia denominado SiRILiCO. Presumiu que um SRI elaborado desta forma poderia ser efetivamente mais eficiente que os sistemas da época, no quesito qualidade de resposta, uma vez que a geração de índices a partir de conceitos estruturados (uma ontologia) é permitida, empregando-se técnicas de Linguística Computacional. Uma ontologia foi criada automaticamente a partir dos conceitos encontrados nos textos da coleção de testes e armazenada. Essa ontologia, obtida através dos conceitos extraídos da análise proposicional (FREDERIKSEN, 1975) dos textos da coleção, serve de base para a geração do índice da coleção. De acordo com as suas classes, é possível identificar quais os conceitos relevantes para a coleção e em que textos eles se encontram. O Modelo SiRILiCO, apesar dos problemas de ruído apresentados pelo protótipo, apresentou resultados superiores aos

resultados apresentados pelo Modelo Vetorial para a coleção em questão. A ideia de utilizar conhecimento de ciências cognitivas para indexar uma coleção de documentos eletrônicos através de frases com conteúdo semântico (proposições) mostrou-se promissora.

Sobre a comparação entre indexação manual e automática Araújo Júnior e Tarapanoff (2006) tratam da comparação entre a indexação manual e a ferramenta de mineração de textos, por meio da análise do índice de precisão de resposta no processo de busca e recuperação da informação. O estudo de caso escolhido para o desenvolvimento da pesquisa foi o Centro de Referência e Informação em Habitação (Infohab), cuja base de dados sobre habitação, saneamento e urbanização foi indexada de forma manual por bibliotecários da Caixa Econômica Federal, com base em uma lista de palavras-chave. Houve o desenvolvimento de um protótipo cujos itens bibliográficos correspondem às teses e dissertações contidas no Infohab, o que permitiu a aplicação do software BR/Search para a execução da mineração de textos. As pesquisas no Infohab e no protótipo foram realizadas a partir da demanda de especialistas da Caixa nos assuntos contidos na base. Os autores concluem que não há ganhos significativos na precisão ao se aplicar a ferramenta de mineração de textos em relação à indexação manual.

Um tópico, na década de 90 e 2000, que foi investigado por alguns pesquisadores brasileiros foi a indexação linguística utilizando sintagmas nominais (SNs) a partir de textos em português brasileiro. Kuramoto (1995) apresenta um sistema de auxílio à recuperação da informação utilizando SNs, como uma alternativa ao uso de termos simples que podem levar a uma recuperação de informação ambígua. O trabalho de Kuramoto (1995) utiliza uma extração manual para SNs, simulando uma extração automática. Essa escolha ocorreu pelo fato de na época ainda não existir um extrator para SNs.

Souza (2006) apresenta a indexação também utilizando SNs, porém diferentemente do trabalho apresentado por Kuramoto (1995), é utilizado um extrator de SNs descrito em (GASPERIN; 2003). No projeto de Kuramoto buscava-se apresentar uma maquete de um SRI baseado em sintagmas nominais, o objetivo do trabalho de Souza (2006) foi propor uma metodologia de auxílio à indexação automática, utilizando uma metodologia aplicada sobre os sintagmas nominais extraídos automaticamente a partir de textos digitalizados em língua portuguesa.

Câmara Júnior (2007) apresenta uma ferramenta para a indexação de acórdãos jurídicos, no escopo do Direito Penal, em português brasileiro, utilizando sintagmas nominais

e um tesouro de jurisprudência do Superior Tribunal Federal de Justiça, visando uma recuperação de tais documentos mais efetiva. O autor conclui que a indexação automática proposta equivale à indexação manual para o contexto analisado. Algumas pequenas diferenças de precisão e revocação a favor da indexação manual são alcançadas, para alguns parâmetros de pesquisa, mas de maneira geral, os resultados (de recuperação da informação) são bastante semelhantes. Alerta, porém, que os bons resultados dependem também da representação do conhecimento, tal como o tesouro utilizado.

Borges, Maculan e Lima (2008) apresentam o planejamento de um sistema de indexação sintático-semântico para o português do Brasil para textos de teses e dissertações. As autoras utilizaram um *parser* denominado *Tropes* associado a uma taxonomia da área de Ciência da Informação para auxiliar o *parser* no processo de escolha dos termos de índice para a área de Ciência da Informação.

Maia (2008) apresenta um sistema para extração de SNs voltado para o português do Brasil denominado OGMA. O OGMA é uma ferramenta para análise de textos, cálculo da similaridade entre documentos e extração de sintagmas nominais. O aplicativo foi desenvolvido com a ferramenta *Visual Studio.NET* em linguagem C#. O OGMA realiza também a identificação da classe do sintagma nominal, bem como o cálculo da pontuação do mesmo como descritor de forma automática. Para realizar a extração de sintagmas nominais o OGMA faz uso de um léxico da língua portuguesa construído a partir do vocabulário utilizado pelo dicionário *BR.ISPELL* e uma lista de 475 palavras irrelevantes criada tendo como base a gramática de Tufano (1990).

Ainda utilizando o sistema OGMA, o trabalho de Correa, Miranda, Lima e Silva (2011) descreve o uso de SNs na indexação e recuperação de teses e dissertações por meio de sintagmas nominais. Os autores concluem que o uso de SNs como itens de índice são melhores que o uso de termos simples como descritores por resolverem o problema da polissemia. O processo de extração de sintagmas nominais através do OGMA teve diferentes desempenhos para cada programa de pós-graduação, sendo obtido melhor desempenho (melhor índice de precisão) para resumos de Direito, seguidos dos de Computação e Nutrição. Esta diferença de desempenho pode em parte ser explicada pela diferente natureza dos termos técnicos presentes nos resumos. Conclui que embora existam limitações nas ferramentas disponíveis, a aplicação de métodos automatizados de extração e indexação por sintagmas nominais mostra-se promissora, pois os sintagmas nominais se configuram como melhores

descritores e pontos de acesso aos documentos, eliminando os problemas causados pela sinonímia e a polissemia das palavras isoladas.

Como observado nos parágrafos acima, a escolha do método de indexação depende de uma série de fatores em relação a seu uso. Um índice a ser consultado por humanos no processo de recuperação da informação pode diferir de um índice a ser utilizado por máquinas. No caso da classificação automática, observa-se que a indexação e a classificação automática caminham juntas. Isso ocorre, pelo fato de ser improvável classificar textos automaticamente em classes ou grupos sem alguma representação reduzida e selecionada de termos-chave a partir de um coleção ou corpus de testes. Em relação aos índices utilizados para classificação automática de textos, deve-se atentar para o fato que o formato do índice é específico: os itens descritores de cada índice, originados de cada texto, são colhidos por um processo automático (linguístico, estatístico ou híbrido), passam por um processo de redução (pois índices muito longos, com vários termos, atrasam o processo de classificação) e, ainda, são adicionados pesos numéricos aos termos do índice, os quais tentam revelar ao programa de computador classificador o quanto eles, os termos selecionados no índice, contribuem semanticamente para a identificação do grupo, ou tema, a que pertencem.

Na seção seguinte, as principais técnicas de agrupamento de textos, levando em consideração a participação prévia do processo de indexação (ou pré-processamento, ou representação textual, ou extração de características, como também é chamado) são descritas.

## **2.9 Classificação automática de textos sob o enfoque da Inteligência Computacional**

O que diferencia a Mineração de Dados e a Mineração de Textos é que a primeira atividade trabalha com a descoberta de informações em fontes estruturadas, como, por exemplo, os conhecidos bancos de dados relacionais empresariais. Ou seja, os dados estão formatados em estruturas de dados (preenchendo tabelas e campos no banco de dados) e não na forma textual em linguagem natural. A informação em texto está na *Web* e nos arquivos digitais em diferentes formatos ou fontes de informação: emails, publicações eletrônicas de bibliotecas digitais, textos de redes sociais, blogs, jornais informativos, memorandos e vários outros formatos; os dados também podem estar em formato híbrido ou semiestruturado (HAN; KAMBER; PEI, 2006). Várias tarefas de mineração de textos podem ser realizadas sobre tais fontes de informação, estas incluem: categorização e agrupamento de documentos,

extração de informações, análise de associações, análise de sentimentos, análise de tendências, entre outras.

As tarefas de classificação de documentos textuais como a categorização e o agrupamento referem-se às ações de colocar tais documentos em classes, onde a classificação se dá de forma automática ou semiautomática, motivada por algum critério temático, tal como a classificação textual por área científica, por assunto, por estilo de escrita, por expressão emotiva, etc. A diferença entre a categorização de textos e o agrupamento de textos é que a categorização utiliza alguma fonte externa de conhecimento formalizado e legível em máquina, tal como os SOC descritos na seção 2.3 anterior (tesauros, ontologias, taxonomias e vocabulários controlados em geral) ou ainda, algum treinamento prévio do sistema realizado de forma supervisionada. Observa-se que, neste caso, ao utilizar fontes de conhecimento externas para treinar ou auxiliar o sistema de categorização, é necessário que o usuário do sistema conheça o número e o conteúdo temático das classes para que o treinamento ou inserção do conhecimento possa ser realizado de forma supervisionada.

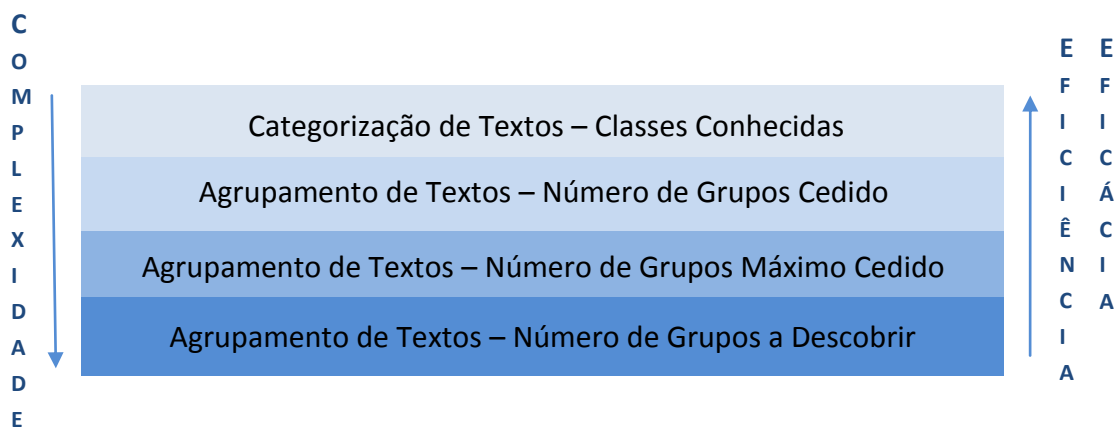
Quando é desejado agrupar os textos sem qualquer conhecimento prévio dos textos a serem reunidos (os temas dos grupos são incógnitos) utilizam-se os sistemas de classificação do tipo agrupamento, e neste caso, o usuário não treina o sistema, ou insere conhecimento formalizado, sendo que a reunião dos textos em grupos é feita utilizando-se somente as características, as palavras-chave, termos ou sintagmas dos textos de entrada. Em algumas situações é necessário descobrir o número de grupos automaticamente, pois o usuário pode não ter esse dado, o que é o caso mais comum, porém, tal tarefa se torna mais complexa para ser automatizada. Como os sistemas de categorização trabalham de forma supervisionada no treinamento classificatório, a tarefa de agrupar textos acaba, experimentalmente, levando a resultados com mais erros que a categorização de textos.

A classificação poderia também ser semisupervisionada, por exemplo, sem intervenção humana somente até certo ponto de execução do sistema agrupador.

É notável, portanto, que os sistemas de classificação de textos mais comuns podem ser vistos de acordo com o nível de complexidade de suas tarefas: o nível menos complexo seria relativo aos sistemas de categorização que utilizam algum conhecimento formal como auxílio para classificação, o segundo nível seria representado pelos sistemas de agrupamento contendo o número de classes possíveis a classificar cedido pelo usuário, em seguida, os sistemas de agrupamento que têm como entrada cedida pelo usuário o número máximo de

grupos que podem ser criados, e por último, os sistemas que procuram descobrir o número exato e correto de grupos a partir dos textos de entrada e, então, classificar os textos de entrada nos grupos criados. É visto que quanto mais complexa a tarefa de classificação, provavelmente, maior será o tempo de computação tomado e maior será a taxa de erros do processo classificatório, ou seja, mais baixa será a eficiência e a eficácia do sistema classificador, como ilustra a figura a seguir:

**Figura 1 - Níveis de complexidade do processo de classificação de textos**



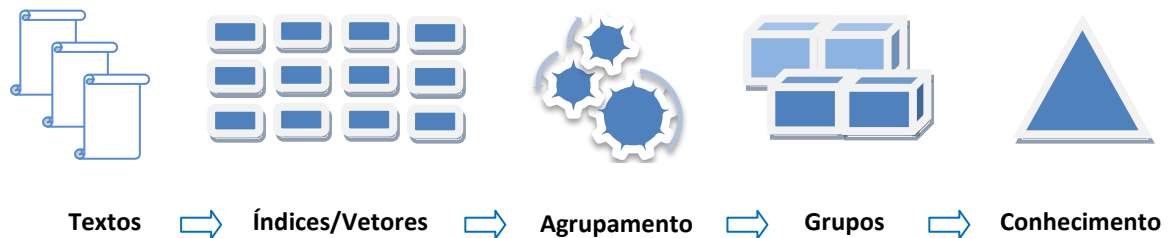
Fonte: elaborado pelo autor, 2012

Uma característica dos algoritmos de agrupamento de textos que geralmente não existe no agrupamento de dados estruturados é que seus resultados dependem fortemente do pré-processamento do texto, ou seja, do método como a extração de características (ou indexação) é realizada e os textos são apresentados ao algoritmo de agrupamento. Métodos diferentes de escolha de termos e pesagem para a formação de bons índices representativos têm sido propostos na literatura, inclusive as diversas formas de indexação automática descritas na seção anterior (linguística, estatística ou híbrida) podem ser utilizadas previamente ao processo de agrupamento, e de acordo com a escolha, os resultados de tempo (eficiência) e correteza de classificação (eficácia) podem ser alterados. A indexação ou seleção de características para aplicação em agrupamento automático, porém, tem diferenças consideráveis da simples seleção de palavras-chave. Os índices para agrupamento geralmente contêm termos de todo o corpus de entrada sendo, por vezes, chamado *vetor de características* para cada texto. Tais vetores geralmente são compactados, contendo os termos principais do corpus, a sua dimensão é diminuída e os termos mais significativos são

selecionados por algum algoritmo: isso faria o tempo de computação do agrupamento diminuir e poderia gerar menos erros de agrupamento. Um peso numérico para cada termo do vetor de características também pode ser gerado para ser utilizado durante a execução do algoritmo de agrupamento. Várias propostas, tanto para a melhor seleção de termos, redução da dimensão do vetor de características e pesagem são descritas na literatura e também, como especificado anteriormente, podem depender da língua utilizada, caso um abordagem linguística seja escolhida na indexação. Aggarwal e Zhai (2012) descrevem alguns desses procedimentos de vanguarda na extração e redução de características, entre elas: *Latent Semantic Indexing* (LSI), *Probabilistic Latent Semantic Analysis* (PLSA), e *Non-negative Matrix Factorization* (NMF).

A figura 2 abaixo ilustra o processo de agrupamento desde a seleção de textos até a análise de resultados e extração de conhecimento, por parte do usuário.

**Figura 2 - Passos do processo de agrupamento de textos**



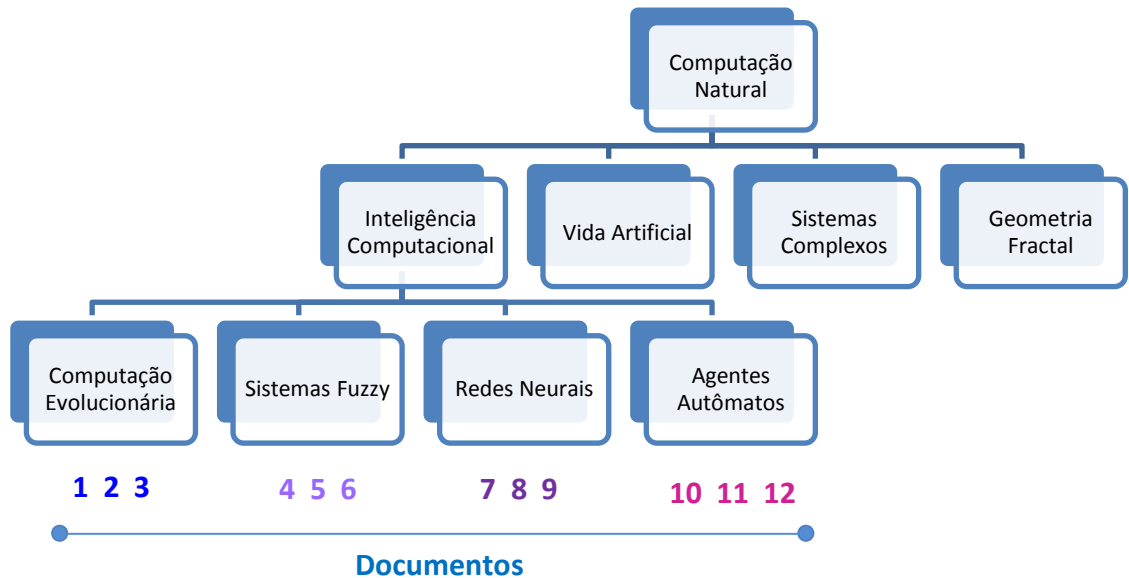
Fonte: elaborado pelo autor, 2012

Assim como a indexação/extração de características, o agrupamento automático de textos pode ser efetuado de diversas maneiras distintas e com diversos algoritmos distintos, cada um com propósitos específicos na organização da informação e utilizando estratégias matemáticas e algorítmicas diferenciadas para tais objetivos. Halkidi, Batistakis e Vazirgiannis (2001) descrevem vários paradigmas para algoritmos de agrupamento, entre eles:

*Algoritmos de Agrupamento Hierárquico:* tais algoritmos reúnem os textos em subgrupos e grupos formando hierarquias, de forma análoga a uma taxonomia, onde os nós folhas da árvore de agrupamento gerada (chamada dendrograma) possuem apenas grupos

contendo um texto e os nós internos vão sendo formados por grupos mais concentrados que englobam os grupos de textos das hierarquias inferiores.

**Figura 3a - Agrupamento hierárquico para a área de Computação Natural**

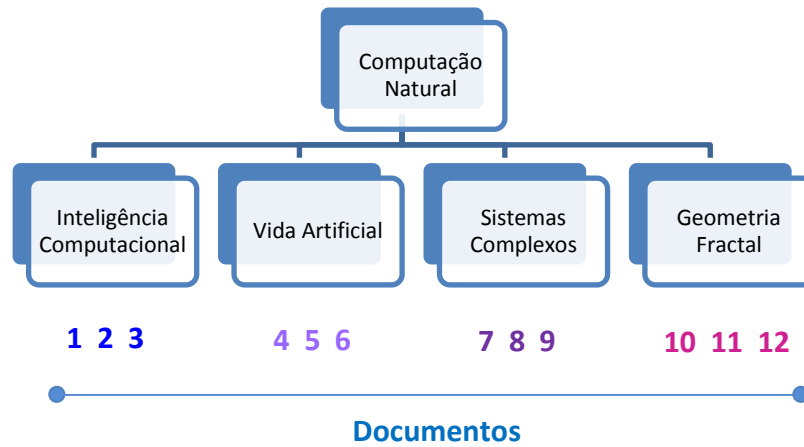


Fonte: elaborado pelo autor, 2012

*Algoritmos de Agrupamento por Particionamento Baseados em Distância:* neste processo, os documentos são particionados em grupos e o objetivo é que os grupos cada vez mais tenham característica diferentes entre si, ou seja, gerando o agrupamento por dissimilaridade entre os grupos. Neste caso, não há formação de hierarquias, há apenas um nível hierárquico.



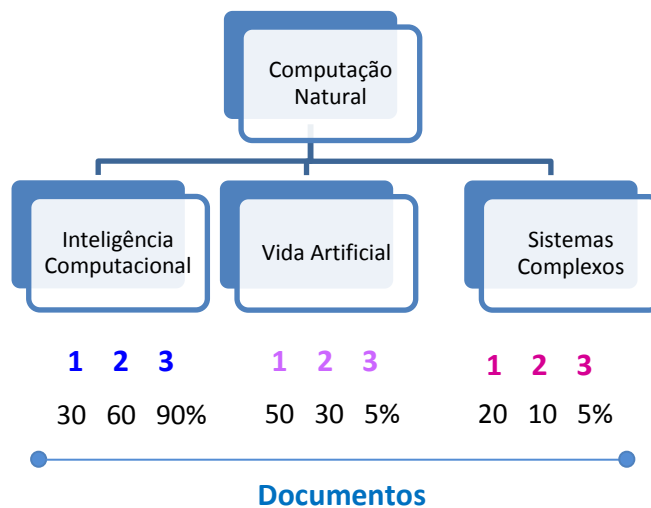
**Figura 3b - Agrupamento por particionamento não hierárquico**



Fonte: elaborado pelo autor, 2012

*Algoritmos de Agrupamento Probabilístico ou Nebuloso:* os algoritmos agrupam o documento pela “probabilidade dele pertencer a um determinado grupo”, ou seja, a classificação é não determinística, cada documento recebe do sistema uma taxa de probabilidade ou percentual de pertencer a cada um dos grupos existentes.

**Figura 3c - Agrupamento nebuloso**



Fonte: elaborado pelo autor, 2012

Embora os algoritmos por particionamento sejam mais rápidos, os resultados de agrupamento podem não ser tão eficazes, pois o resultado depende de escolhas iniciais de

particionamento que afetam o resultado final. Os métodos híbridos foram então criados, visando encontrar um meio termo entre melhoria da eficiência e eficácia no agrupamento de textos.

Outras diversas possibilidades de agrupamento podem existir de acordo com o formato dos dados a classificar e a maneira de criação dos grupos, e a escolha vai depender dos objetivos de agrupamento do usuário. Nesta pesquisa, trabalha-se com os *Algoritmos de Agrupamento por Particionamento Baseados em Distância*, especificamente: os Algoritmos Evolucionários e o algoritmo *X-Means* (que é uma variante do conhecido algoritmo *K-Means*, mas com a especificação por parte do usuário de um número máximo de grupos a criar) e também foi testado um algoritmo de agrupamento probabilístico (o *Expectation Maximization-EM*, com número de grupos a descobrir). A razão para esta escolha foi que todos esses três algoritmos estão nos dois últimos níveis de complexidade, como mostrado na figura 1 anterior, que é o alvo da pesquisa, ou seja, ou os algoritmos pedem o número máximo de grupos a criar ou tentam descobrir ou aproximar sozinhos o número correto de grupos originalmente existente no corpus de teste. Outra razão para a escolha, é pelo fato dos algoritmos *EM* e *X-Means* serem clássicos, com várias citações na literatura, e inclusive estão implementados no software de livre acesso para *Data Mining* e *Text Mining WEKA*, software bastante utilizado em pesquisas sobre *Data Mining* e *Text Mining*. Como o novo algoritmo proposto utiliza Computação Evolucionária para agrupamento, também o Algoritmo Evolucionário para agrupamento utilizado na forma convencional é testado para que comparações de melhoria, dentro do paradigma Computação Evolucionária, possam ser verificadas. O *Algoritmo Evolucionário de Agrupamento Convencional* pede ao usuário a especificação de um número máximo de grupos a criar, já o *Algoritmo Evolucionário de Agrupamento Proposto* tenta descobrir o número de grupos.

Algo notável é que o agrupamento de um texto em vários grupos (pluriclassificação) não é considerado no trabalho desenvolvido: o objetivo primário é avaliar o agrupamento determinístico (um texto somente em um grupo). O algoritmo *EM* é nebuloso e realiza a pluriclassificação (figura 3c), mas neste trabalho, para este algoritmo, foi considerado apenas o agrupamento com máxima probabilidade de um texto para um único grupo.

A escolha de trabalhar essas classes de algoritmos se dá também pelo fato que o conjunto de artigos científicos do corpus a classificar pode vir sem um número especificado de textos, ou seja, não se saberia, numa situação real de uso, se os artigos a classificar são

artigos de uma única área científica ou de 20 áreas científicas, portanto, esse enfoque escolhido é mais próximo à realidade. O resultado ideal seria que os algoritmos conseguissem deduzir o número correto de grupos (sendo cada grupo uma área científica) e agrupasse os artigos corretamente, porém, a situação ideal é ainda longe da realidade: os algoritmos convencionais dificilmente conseguem descobrir exatamente o número correto de grupos e quando isso ocorre de forma aproximada a taxa de acertos no agrupamento geralmente é baixa, sendo que, muitas vezes, para a casa de centenas de artigos a classificar, o tempo pode ser de horas de execução dependendo do algoritmo escolhido e da quantidade de textos.

Pelo fato da tarefa ser complexa e de tais sistemas serem os ideais necessários é que surgem as pesquisas na área, sempre objetivando melhorias nos mecanismos de agrupamento de textos. A hipótese aqui estabelecida, é que o novo algoritmo proposto, juntamente com uma nova forma de escolha de termos (indexação) proposta gere taxas de corretude de agrupamento maiores e taxas de consumo de tempo mais baixas, não é almejado resolver o problema, pois talvez o agrupamento perfeito ou quase perfeito, comparável à classificação humana, nem seja alcançável computacionalmente. Neste trabalho, verifica-se uma hipótese e abre-se uma nova possibilidade de estudos para um problema que pode ainda ser potencialmente investigado futuramente e que se encontra em passos iniciais de estudos.

Os algoritmos descritos a seguir são os algoritmos de agrupamento já conhecidos e testados nos experimentos, cuja forma de execução será comparada ao novo *Algoritmo Evolucionário de Agrupamento Proposto* que será descrito em detalhes no próximo capítulo.

### 2.9.1 O algoritmo de agrupamento *X-Means*

Este algoritmo deriva do conhecido algoritmo *K-Means*, a diferença é que o *X-Means* procura descobrir o número de grupos dentro de um intervalo entre  $[2, k]$ ,  $k$  é o número máximo de grupos dado pelo usuário como entrada na execução. O algoritmo *K-Means*, ao contrário, pede como entrada o número exato de grupos  $k$  a criar como entrada. Por tal fato, o trabalho matemático do algoritmo *X-Means* é maior devido à maior complexidade do problema.

O algoritmo *K-Means* possui diversas extensões e modificações, a ideia original foi proposta por Hugo Steinhaus em 1957. Recentemente, Vattani (2011) discute e apresenta melhorias na complexidade temporal do algoritmo. O *K-Means* particiona um conjunto  $X$  de  $n$  pontos em  $\mathbf{R}^d$  em  $k$  grupos.  $X$  é semeado com um conjunto inicial de  $k$  grupos (*clusters*) e seus

centroides em  $\mathbf{R}^d$ . Todo ponto do conjunto  $X$  é associado ao centroide mais próximo matematicamente a ele. O nome *K-Means* (*K-Médias*) refere-se ao fato que a nova posição de um centroide é computada como o centro da massa (ou o ponto médio) dos pontos associados a tal centroide. A listagem, a seguir, ilustra formalmente o algoritmo original em quatro passos (VATTANI, 2011).

### Listagem 1 - Algoritmo *K-Means* original

0. Arbitrariamente, escolha  $k$  centroides iniciais  $c_1, c_2, \dots, c_k$ .
1. Para cada  $1 \leq i \leq k$ , estabeleça o grupo  $C_i$  como o conjunto de pontos em  $X$  que são mais pertos de  $c_i$  do que qualquer outro  $c_j$  com  $j \neq i$ .
2. Para cada  $1 \leq i \leq k$ , estabeleça  $c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$ , como o centroide dos pontos em  $C_i$ .
3. Repita os passos 1 e 2 até os grupos  $C_i$  e os centroides  $c_i$  não se modificarem mais. O particionamento de  $X$  é o conjunto de grupos  $C_1, C_2, \dots, C_k$ .

Fonte: adaptado de Vattani, 2011

O *X-Means* é uma versão modificada e mais complexa do *K-Means*. A letra “X” no nome “*X-Means*” indica justamente que um número  $k$  de grupos não é conhecido, ou seja, não é uma constante  $k$  dada pelo usuário, como indica o nome “*K-Means*”, e sim, uma incógnita entre  $[2, k]$  a calcular. Dentre as versões para o *X-Means*, uma é descrita por Ishioka (2005) e também há o algoritmo descrito por Pelleg e Moore (2000). A versão apresentada a seguir, descrita por Ishioka (2005), procura encontrar o melhor valor de grupos para o conjunto a agrupar, e esta versão do algoritmo ainda utiliza o *K-Means* original como módulo de apoio. É descrito, formalmente, o algoritmo *X-Means* em onze passos:

## Listagem 2 - Algoritmo de agrupamento *X-Means*

**Passo 0 :** Prepare os dados  $p$ -dimensionais, onde o número de dados é  $n$ .

**Passo 1:** Estabeleça um número inicial de grupos  $k_0$  (o número *default* é 2), o qual deve ser suficientemente pequeno.

**Passo 2:** Aplique o algoritmo K-Médias (*K-Means*) para todos os dados de entrada, fazendo  $k=k_0$ . Chamemos os grupos divididos:  $C_1, C_2, \dots, C_{k_0}$ .

**Passo 3:** Repita o procedimento seguinte, do passo 4 ao passo 9, estabelecendo:  $i = 1, 2, \dots, k_0$ .

**Passo 4:** Para um grupo de  $C_i$  aplique o K-Médias, estabelecendo  $k=2$ . Os grupos são chamados:  $C_i^{(1)}, C_i^{(2)}$ .

**Passo 5:** Assumimos a seguinte distribuição normal  $p$ -dimensional para o dado  $x_i$  contido em  $C_i$ :

$$f(\theta_i, x) = (2\pi)^{-p/2} |V_i|^{-1/2} \times \exp \left[ -\frac{1}{2} (x - \mu_i)^t V_i^{-1} (x - \mu_i) \right],$$

então, calculamos o BIC (Bayesian Information Criterion – Critério de Informação Bayesiano):

$$BIC = -2 \log L(\hat{\theta}_i; x_i \in C_i) + 2p \log n_i,$$

onde  $\hat{\theta}_i = [\hat{\mu}_i, \hat{V}_i]$  é a máxima verossimilhança estimada da distribuição normal  $p$ -dimensional;  $\mu_i$  é o vetor-médio  $p$ -dimensional e  $V_i$  é a matriz variância-covariância  $p \times p$ -dimensional; o número total de parâmetros é  $2p$ .  $x_i$  é o dado  $p$ -dimensional contido em  $C_i$ ;  $n_i$  é o número de elementos contidos em  $C_i$ .  $L$  é a função de verossimilhança que indica  $L(.) = \prod f(.)$ .

**Passo 6:** É assumida a distribuição normal  $p$ -dimensional com os parâmetros  $\theta_i^{(1)}$ ,  $\theta_i^{(2)}$  para  $C_i^{(1)}$ ,  $C_i^{(2)}$  respectivamente; A função densidade de probabilidade para este modelo de duas divisões será:

$$g(\theta_i^{(1)}, \theta_i^{(2)}; x) = \alpha_i [f(\theta_i^{(1)}; x)]^{\delta_i} [f(\theta_i^{(2)}; x)]^{1-\delta_i}, \quad (1) \text{ onde}$$

$$\delta_i = \begin{cases} 1, & \text{Se } x \text{ está incluído em } C_i^{(1)}, \\ 0, & \text{Se } x \text{ está incluído em } C_i^{(2)}; \end{cases}$$

$x_i$  será incluído tanto em  $C_i^{(1)}$  ou  $C_i^{(2)}$ ;  $\alpha_i$  é uma constante que faz a equação (1) ser uma função densidade de probabilidade; isto é:

### Listagem 2 - Algoritmo de agrupamento *X-Means* (continuação)

$$\alpha_i = 1 / \int [f(\theta_i^{(1)}; x_i)]^{\delta_i} [f(\theta_i^{(2)}; x_i)]^{1-\delta_i} dx,$$

( $1/2 \leq \alpha_i \leq 1$ ). Se for desejado um valor exato, pode-se utilizar integração numérica  $p$ -dimensional. Mas isso requer significativa computação. Assim,  $\alpha_i$  é aproximado como se segue:

$$\alpha_i = 0,5/K(\beta_i), \text{ onde } \beta_i = \sqrt{\frac{\|\mu_1 - \mu_2\|^2}{|V_1| + |V_2|}}.$$

$K(\cdot)$  permanece como a probabilidade mais baixa da distribuição normal. Quando é estabelecido  $\beta_i=0,1,2,3$ ,  $\alpha_i$  torna-se  $0,5/0,500 = 1$ ;  $0,5/0,841 = 0,59$ ;  $0,5/0,977 = 0,51$  e  $0,5/0,998 = 0,50$ , respectivamente. O BIC' para este modelo é:

$$BIC' = -2 \log L'(\hat{\theta}'_i; x_i \in C_i) + 4p \log n_i,$$

onde  $\hat{\theta}'_i = [\hat{\theta}_i^{(1)}, \hat{\theta}_i^{(2)}]$  é a verossimilhança máxima estimada de duas distribuições normais  $p$ -dimensionais. Já que existem dois parâmetros de média e variância para cada variável  $p$ , o número total de parâmetros se torna  $2 \times 2p = 4p$ .  $L'$  é a função de verossimilhança a qual indica  $L'(\cdot) = \prod g(\cdot)$ .

**Passo 7:** Se  $BIC > BIC'$ , o modelo dividido em dois é preferido, e a divisão é continuada; é estabelecido:  $C_i \leftarrow C_i^{(1)}$ . E para  $C_i^{(2)}$  coloca-se os dados  $p$ -dimensionais, os centroides, a verossimilhança logarítmica e o BIC em uma estrutura de dados pilha. Retorne ao passo 4.

**Passo 8:** Se  $BIC \leq BIC'$ , os grupos não são mais divididos. Extraia da pilha os dados que foram armazenados no passo 7 e faça  $C_i \leftarrow C_i^{(2)}$ . Retorne ao passo 4. Se a pilha estiver vazia vá ao passo 9.

**Passo 9:** o procedimento de duas divisões para  $C_i$  está completo. Os grupos sofrem uma renumeração para que eles se tornem únicos em  $C_i$ .

**Passo 10:** o procedimento de duas divisões para um  $k_0$  inicial está completa. Todos os grupos são renumerados para identificação para que se tornem únicos.

**Passo 11:** Devolva o número de identificação do grupo para o qual cada elemento está associado, o centroide de cada grupo, a verossimilhança logarítmica e o número de elementos em cada grupo.

### 2.9.2 O algoritmo de agrupamento *Expectation-Maximization (EM)*

O algoritmo *EM* descrito por McLachlan e Krishnan (2007) é um método de agrupamento não supervisionado e tem base nos Modelos de Mistura. Funciona em uma abordagem iterativa, subótima, que tenta encontrar os parâmetros da distribuição de probabilidade que tem a máxima probabilidade de seus atributos.

Geralmente, utilizam-se Modelos de Mistura Gaussianas, ou seja, o conjunto de dados de entrada a agrupar  $x_i$  são representados por distribuições de probabilidade gaussianas, o número de grupos  $K$  será o número destas distribuições de probabilidade. Inicialmente, os parâmetros  $\theta_j = \langle p_j, \mu_j, \Sigma_j \rangle$  (probabilidade do grupo (ou classe)  $p_j$ , média dos valores do grupo  $\mu_j$ , e cálculo da matriz de covariância  $\Sigma_j$  do grupo) de cada distribuição são escolhidos aleatoriamente. O algoritmo funciona então em dois passos, no passo *E* são calculados os valores de probabilidade de cada dado  $x_i$  pertencer ao grupo  $w_j$ , para todo grupo existente. No passo *M* são recalculados os valores dos parâmetros das distribuições  $\langle p_j, \mu_j, \Sigma_j \rangle$ . Esses passos são executados até que um critério de parada ocorra, podendo ser um número de iterações máximo ou até que não ocorram mudanças nos valores dos parâmetros. Um ótimo local é alcançado. A listagem a seguir ilustra o algoritmo *Expectation-Maximization* básico, algumas estratégias são propostas na literatura para sua melhoria, por exemplo, para que o número de grupos  $K$  seja calculado automaticamente.

### Listagem 3 - Algoritmo de agrupamento *Expectation-Maximization*

1. Propor um conjunto de valores iniciais para os parâmetros: a probabilidade de ocorrência de cada classe  $p_j$ , a média de valores do grupo  $\mu_j$  e sua matriz de covariância  $\Sigma_j$ . Ou seja:

$$\theta_j = \langle p_j, \mu_j, \Sigma_j \rangle, j = 1 \dots K$$

2. Repita os passos até a convergência:

a) Passo-E: Para cada  $i=1, \dots, m$  e cada  $j=1, \dots, K$ , calcule a probabilidade de  $x_i$  ser projetado a partir da distribuição de classe  $j$ :

$$w_{ij} = p(x_i | p_j, \mu_j, \Sigma_j) \propto p_j p(x_i | \mu_j, \Sigma_j)$$

b) Passo-M: Atualizar os parâmetros do modelo para maximizar a verossimilhança dos dados:

$$p_j = \frac{1}{m} \sum_{i=1}^m w_{ij} \quad \mu_j = \frac{\sum_{i=1}^m w_{ij} x_i}{\sum_{i=1}^m w_{ij}}$$

$$\Sigma_j = \frac{\sum_{i=1}^m w_{ij} (x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^m w_{ij}}$$

Fonte: adaptado do material do site Machine Learning<sup>3</sup>, 2006

<sup>3</sup> <http://www.facweb.iitkgp.ernet.in/~sudeshna/courses/ML06/ml-lecture19.pdf>. Acesso em 04/10/2013.



Uma das formas de estimar o número correto de grupos ao executar o algoritmo *EM* é utilizar uma técnica de partição dos dados de teste  $x_i$  denominada *Cross-Validation*. Tal técnica é inclusive implementada no software *WEKA* e pode ser acionada pelo usuário, caso este deseje que o número de grupos seja deduzido. A seguir, um exemplo da técnica em cinco passos, para mais detalhes consultar (SMYTH, 1998):

#### **Listagem 4 - Algoritmo *Cross-Validation***

1. O número de grupos é estabelecido em 1.
2. O conjunto de treinamento é dividido randomicamente em 10 partes.
3. O algoritmo *EM* é executado por 10 vezes usando as 10 partições.
4. A média das verossimilhanças  $l = \sum_{i=1}^N \log(\sum_{j=1}^K p_j p(x_i|\mu_j, \Sigma_j))$  é calculada para as 10 partições.
5. Se a média das verossimilhanças  $l$  aumentar, então o número de grupos é aumentado em 1 e o programa continua no passo 2.

Fonte: adaptado de Smyth, 1998

### **2.9.3 O algoritmo de agrupamento baseado em *Computação Evolucionária***

A partir de 1960, a observação centrada na relação subjacente entre a otimização e a evolução biológica levou ao desenvolvimento de um importante paradigma da Inteligência Computacional: a Computação Evolucionária (CE), a qual objetiva realizar buscas complexas e otimizações (DAS; ABRAHAM; KONAR, 2009).

Segundo Scriptor (2006), a Computação Evolucionária faz parte de uma série de novos algoritmos de aproximação (algoritmos empregados para encontrar soluções aproximadas em problemas de otimização) que emergiu nos últimos 20 anos e que tentam combinar métodos heurísticos básicos dentro de esquemas de nível maior (ou seja, heurísticas específicas dentro de heurísticas genéricas) destinados a eficientemente e eficazmente explorar um espaço de busca. Segundo este autor, o termo meta-heurística foi, inicialmente,

utilizado por Glover em 1986 (GLOVER, 1986) e é melhor descrito como sendo uma estratégia iterativa do processo de busca que guia tal processo sobre o espaço de busca na tentativa de encontrar a solução ótima. Esta classe de algoritmos inclui, mas não está restrita a: Otimização por Colônia de Formigas (*Ant Colony Optimization* - ACO), Computação Evolucionária (*Evolutionary Computing* - EC), Busca Local Interativa (*Iterated Local Search* - ILS), *Simulated Annealing* (SA) e a Busca Tabu (*Tabu Search* - TS).

Destas Meta-Heurísticas, tem-se a Computação Evolucionária, inspirada na teoria da seleção natural das espécies, descrita pelo biólogo Charles Darwin, sendo utilizada para resolver problemas de busca de soluções onde tal espaço de busca é extenso, onde seria inviável a busca em todo o espaço de forma exaustiva, pois o consumo de tempo de computação seria grande demais ou indeterminado. Poderia ser utilizada, por exemplo, numa otimização numérica, se o objetivo fosse encontrar os valores de uma variável  $x \in R^n$ , para maximizar uma função  $f(x)$ ,  $f: R^n \rightarrow R$ , tais Algoritmos Evolucionários buscariam as soluções no espaço de busca  $R^n$ . Ou ainda, poderia ser também utilizada a Computação Evolucionária para o caso de problemas de otimização combinatória, onde o número de combinações possíveis a verificar como solução seja muito alto, entre diversos outros tipos de problemas envolvendo otimização.

Na pesquisa aqui descrita, o problema de agrupar textos é de otimização combinatória, pois o objetivo é encontrar a melhor combinação de documentos dentro de cada grupo, ou seja, agrupar os documentos com temas afins (no caso, artigos científicos de uma mesma área científica). Se o algoritmo executado fosse verificar todas as combinações possíveis de agrupamento ter-se-ia um espaço de verificação muito grande à medida que o número de documentos de entrada crescesse, logo, esta meta-heurística procura achar a solução ótima (os melhores agrupamentos) em tempo reduzido, em relação à computação exaustiva que verificaria todas as possibilidades.

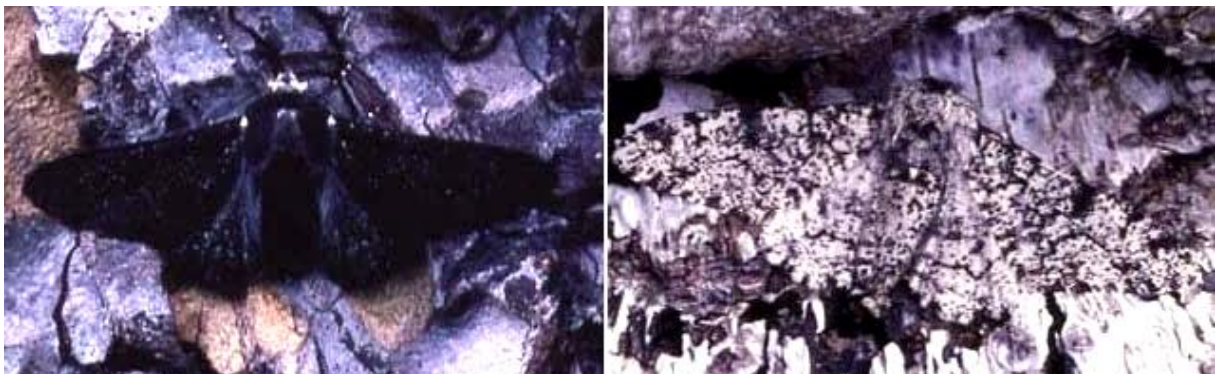
Para se ter uma ideia de como os grupos crescem de forma combinatória, à medida que o número de elementos a agrupar aumenta, considere um conjunto de cinco símbolos {+, \*, \$, &, %} eles podem ser combinados criando 31 grupos diferentes de 1 a 5 elementos, já um conjunto de 25 símbolos pode ter  $3,335 \times 10^7$  grupos formados, onde os grupos têm de 1 a 25 elementos. Algoritmos que verificam todos os grupos possíveis para encontrar o melhor agrupamento de textos (onde os grupos teriam textos mais similares) seriam, portanto, lentos ou a solução seria não alcançável para situações reais de agrupamento.

Logo, a diminuição de tempo ocorrerá porque um Algoritmo Evolucionário irá melhorar as soluções candidatas de agrupamento ao longo do tempo (de forma evolutiva), ao invés de verificar todas as combinações possíveis ao procurar a melhor solução de agrupamento. Esta é a inteligência do mecanismo evolucionário: ir melhorando uma solução de agrupamento ao invés de procurar todas as possíveis: uma a uma. Essa “racionalidade” também é encontrada na natureza, a seleção natural das espécies permite que os seres vivos melhorem ao longo dos milhares de anos, pois as mais aptas espécies sobreviverão e se reproduzirão e as menos aptas acabam desaparecendo.

A Computação Evolucionária é um ramo da Inteligência Computacional que tem por base os mecanismos evolutivos encontrados na natureza. Esses mecanismos estão diretamente relacionados com a teoria da evolução de Darwin, onde ele afirma que a vida na Terra é o resultado de um processo de seleção, feito pelo meio ambiente, em que somente os mais aptos e adaptados possuirão chances de sobreviver e, conseqüentemente, reproduzir-se.

Uma discussão conhecida sobre a seleção natural das espécies que ilustra como o mecanismo artificial bioinspirado trabalha, é o caso do processo evolutivo das mariposas da espécie *Biston betularia*, figura 4 a seguir.

**Figura 4 - Mariposas *Biston betularia***



Fonte: Homepage da FSC – Field Studies Council<sup>4</sup>

---

<sup>4</sup> <http://www.field-studies-council.org/urbaneco/urbaneco/introduction/evolution.htm>. Acesso em 04/10/2013.

Segundo Pazzo (2004), antes da revolução industrial na Grã-Bretanha, a forma mais coletada destas mariposas era a clara, salpicada. A forma melânica, escura, foi identificada pela primeira vez em 1848, perto de Manchester, e aumentou em frequência até constituir mais de 90% da população de áreas poluídas em meados do século 20. Em áreas despoluídas, a forma clara ainda era comum. A partir dos anos 1970, entretanto, em decorrência de práticas conservacionistas e consequente diminuição da poluição, a frequência das formas melânicas diminuíram drasticamente, de cerca de 95% até menos de 10% em meados dos anos 90.

Desde 1890, vários trabalhos tentam explicar os fenômenos envolvidos no aumento da frequência da forma melânica, como: efeito da cor sobre a eficiência térmica, indução das formas melânicas por efeitos diretos da poluição, entre outros diversos fatores atuando sozinhos ou em conjunto. Em meados dos anos 50, Kettlewell explicou a mudança na frequência pela ação da caça visual por pássaros. A forma melânica ficava melhor camuflada no tronco de árvores em regiões poluídas, onde a fuligem matou o líquen. Por outro lado, as mariposas salpicadas ficavam melhor camufladas em áreas despoluídas. Um estudo de L. M. Cook conclui que no melanismo industrial de *Biston betularia*, tanto o aumento original e a recente diminuição na frequência das formas melânicas são notáveis exemplos de mudança genética natural, intimamente relacionada com a mudança do meio ambiente. Como a evolução é definida pela mudança na frequência das características herdadas ao longo do tempo, e a frequência da forma melânica da mariposa *Biston betularia* (cujos padrões de coloração são regidos por leis Mendelianas) aumentou e agora diminuiu em decorrência das leis antipoluição, isto é prova de evolução. Além disso, a velocidade e direção das mudanças podem ser explicadas apenas através da seleção natural, sendo assim, prova da evolução Darwiniana.

Recentemente, o trabalho de Cook (2012) discute e reafirma tal verificação sobre o processo de seleção natural das mariposas por pássaros predatórios, observado inicialmente por Kettlewell. Segundo o autor, os dados colhidos em sua pesquisa fornecem a evidência direta para implicar a camuflagem e predação das aves como a explicação primordial para o aumento e diminuição do melanismo em mariposas.

A Computação Evolucionária procura justamente simular esse processo de evolução e seleção natural, dada a pressão existente no meio-ambiente. Para tal, trabalha com as três seguintes ideias básicas: criação de uma população de soluções, criação de uma função de

avaliação para as soluções propostas e a criação dos operadores de seleção, recombinação e mutação (EIBEN; SMITH, 2010).

***a) A criação de uma população de soluções***

É necessário que haja uma população inicial de possíveis soluções para o problema proposto. Essa população pode ser gerada aleatoriamente ou através de alguma técnica. Cada indivíduo dessa população terá registrado em si os parâmetros que descrevem a sua respectiva solução para o problema. No caso do problema de agrupamento de textos, essa população inicial seria diversas combinações aleatórias de documentos em grupos, criadas sem critério de avaliação prévio.

***b) A criação de uma função de avaliação***

A função de avaliação terá o trabalho de julgar a aptidão de cada indivíduo da população. Ela não precisará deter o conhecimento de como encontrar a solução do problema, somente precisará julgar a qualidade da solução que está sendo apresentada por aquele indivíduo. A função é definida através da codificação das soluções para o problema para que se possa avaliar se o indivíduo está ou não apto. No caso do agrupamento textual, a função de avaliação seria capaz de avaliar cada solução de agrupamento de uma população, quanto melhor a solução de agrupamento de um indivíduo da população (textos mais similares dentro do mesmo grupo) melhor a pontuação da solução (uma nota de zero a cem, por exemplo).

***c) A criação dos operadores: seleção, recombinação (ou crossover) e mutação***

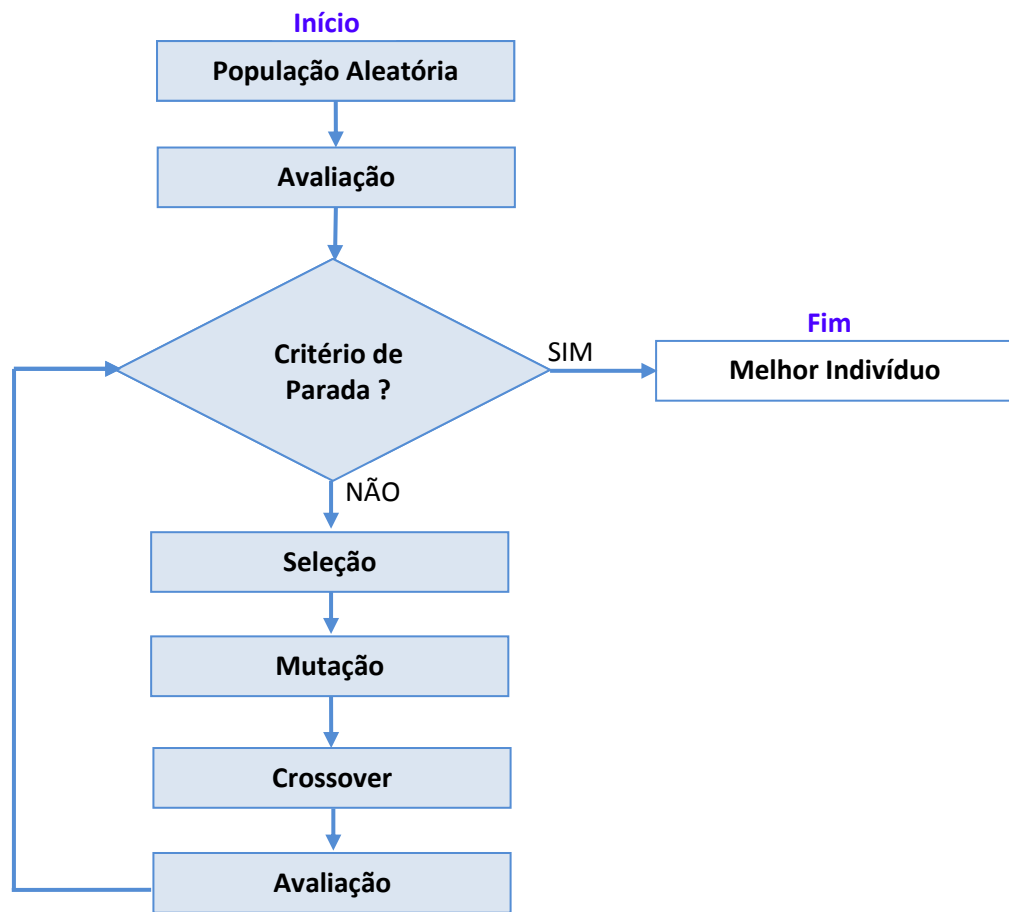
É necessário gerar novas populações para que se encontre o mais apto. Essas são chamadas de gerações e são obtidas através da aplicação de três operadores: *seleção*, *recombinação (ou crossover)* e *mutação*. Na seleção escolhem-se os indivíduos mais aptos (ou melhor pontuados) para gerarem descendentes. Já a reprodução pode ocorrer de duas formas: um indivíduo gera a descendência ou um par de indivíduos geram a descendência. O primeiro caso simula uma reprodução assexuada e o segundo uma reprodução sexuada. É importante destacar que os descendentes serão diferentes de seus antecedentes. Na recombinação ocorre a troca de material genético entre o par de antecedentes definindo assim a carga genética dos descendentes. Por tal reprodução, cada descendente desse par herdará uma parte do material genético de seus antecedentes. Na mutação ocorrem mudanças no material genético de um indivíduo para gerar outro indivíduo, logo, o novo indivíduo poderá estar mais apto dentro daquela população. A simulação da passagem de gerações ocorre na repetição deste ciclo, ou seja, a cada iteração.

No caso do agrupamento de textos, a mutação poderia ser simplesmente mudar os documentos de grupo dentro de uma solução de agrupamento. A recombinação (ou *crossover*) poderia ser: juntar duas diferentes partes de duas soluções de agrupamento (que seriam indivíduos da população) e criar uma terceira solução de agrupamento (um novo indivíduo) a partir dessa junção.

Observa-se que tal Computação Evolucionária consiste numa máquina “aprendente”, otimizada, baseada nos moldes dos mecanismos de evolução biológica e seleção natural. De maneira análoga, o mecanismo computacional evolucionário imita a natureza como no caso das mariposas *Biston betularia*. Ou seja, uma pressão seletiva externa é aplicada às gerações desde a geração inicial, para que somente as soluções de agrupamento ideais sobrevivam e se reproduzam, melhorando assim, de forma adaptativa, a geração seguinte. Os pássaros (predadores) que pressionam a seleção da população de mariposas seriam representados pelo módulo de *seleção*, a reprodução das mariposas pelos módulos de *crossover* e *mutação* e a *avaliação* seria comparável à cor definida às duas classes de mariposas; mariposas de cores claras teriam baixa pontuação, pois seriam mais facilmente vistas pelo predador, por exemplo.

O procedimento seguinte ilustra um Algoritmo Evolucionário genérico e seus passos na busca de uma solução otimizada.

Figura 5 - Fluxograma dos módulos de um *Algoritmo Evolucionário Convencional*



Fonte: elaborado pelo autor, 2012



Dentro desta estrutura (descrita na figura 5 anterior), existem diversas variantes dos Algoritmos Evolucionários (AE) e muitos sistemas híbridos incorporam várias características desse paradigma. Entretanto, todas as estruturas dessas variantes possuem métodos evolutivos muito semelhantes. Segundo Souza, Neves Jr. e Lopes (2006), os cinco principais paradigmas dos Algoritmos Evolucionários são:

- a) Algoritmos Genéticos: Técnica de busca baseada na teoria de evolução de Darwin. Desenvolvida originalmente por John Henry Holland em 1975, modela a seleção natural e o processo da evolução das espécies. Eles podem ser considerados um processo de pesquisa, ao determinar os melhores indivíduos no espaço de busca de todos os possíveis indivíduos. Resumidamente, compreende a evolução de uma população de inteiros binários, os quais são submetidos a transformações unitárias e binárias genéricas e a um processo de seleção.
- b) Programação Evolutiva: Consiste na evolução de população com máquinas de estados finitos submetendo-as a transformações unitárias.
- c) Estratégias de Evolução: Trata-se de evoluir uma população de números reais que codificam as possíveis soluções de um problema numérico, onde a seleção está implícita.
- d) Sistemas de Classificadores: São sistemas capazes de perceber e classificar os acontecimentos em seu ambiente e reagir a eles apropriadamente.
- e) Programação Genética: Técnica que utiliza a metodologia da Computação Evolucionária não para solucionar o problema, mas sim para obter os melhores procedimentos possíveis para sua resolução.

Os AE apresentam vantagens e desvantagens em relação aos métodos tradicionais de busca e otimização (COELHO, 2003).

Entre as *vantagens* dos AE tem-se:

- a) não existe a necessidade de assumir-se características do espaço do problema;
- b) vastamente aplicável (algoritmos de propósito geral);
- c) baixo custo de desenvolvimento e aplicação;
- d) facilidade de incorporar outros métodos;
- e) pode ser executado interativamente e possibilita a acomodação de soluções propostas pelo usuário no procedimento de otimização.

Entre as *desvantagens* dos AE deve-se mencionar que:

- a) não garantem uma solução ótima;
- b) podem necessitar de sintonia de alguns parâmetros inerentes à metodologia evolutiva adotada;
- c) frequentemente apresenta alto custo computacional.

O trabalho de Song e Park (2006), sobre agrupamento de textos, representa bem como o enfoque evolucionário pode ser utilizado no agrupamento de textos. Para uma revisão geral dos métodos de agrupamento utilizando Computação Evolucionária e seus paradigmas, suas similaridades e diferenças, veja o artigo de revisão geral da área (HRUSCHKA; CAMPELLO; FREITAS; CARVALHO, 2009).

Song e Park (2006) propõem o *MVGA (Modified Variable String Length Genetic Algorithm)* o qual utiliza uma nova forma de representar os cromossomos de um Algoritmo Genético, ou seja, a forma de representar os documentos nos cromossomos para o tratamento algorítmico é inovadora em relação ao seu antecessor, o sistema *VGA (Variable String Length Genetic Algorithm)*. As funções de seleção, mutação e *crossover* também foram melhoradas, visando aumentar a pressão de seleção de indivíduos e aumentar a diversidade populacional. Porém, a função de avaliação utilizada é o índice *Davies-Bouldin* apresentado em 1979, podendo ser uma forma cara de avaliação em termos de eficiência (tempo) de agrupamento.

Sobre resultados, utilizando as técnicas tradicionais de medição de resultados (Precisão, Revocação e a Medida-F) da área de agrupamento e um corpus de teste tradicional para o inglês (a base de documentos *Reuter-21578*), o autor afirma o aprimoramento do algoritmo anterior com a melhoria da Precisão em 19,5 pontos percentuais, da Revocação em 2,5 pontos percentuais, e da Medida-F em 12,1 pontos percentuais.

No capítulo seguinte serão listadas as inovações pretendidas nas fases do novo *Algoritmo Evolucionário Proposto*. É feito um paralelo com o modelo de Algoritmo Evolucionário genérico da figura 5 descrita anteriormente e a nova proposta, para cada etapa do novo Algoritmo Evolucionário.

## 2.10 Estudos em Agrupamento Automático de Textos para o português do Brasil

O número de trabalhos em Linguística Computacional e Processamento de Linguagem (Língua) Natural são considerados reduzidos no Brasil, se comparado a outros países da Europa e América do Norte. Porém, nos últimos anos, tem havido um crescimento significativo da área no Brasil, principalmente no meio acadêmico, com pesquisas e publicações na área. Essa redução ocorre também pela falta de interação entre as áreas de Linguística e Ciência da Computação no Brasil para resolução de problemas de PLN (DIAS-DA-SILVA, 2006).

Seguindo tal tendência, o mesmo ocorre para a área de Recuperação da Informação na Ciência da Informação que está fortemente ligada ao processamento linguístico, incluindo as subáreas de Categorização e Agrupamento de Textos. No caso de trabalhos em Agrupamento de Textos voltado para a língua portuguesa do Brasil, sendo este o alvo de estudo principal deste trabalho, foram encontradas algumas publicações relevantes, cujos objetivos se relacionam com os objetivos deste trabalho, tais publicações são descritas a seguir.

Maia e Souza (2010) descrevem o estudo que mais se aproxima a esta pesquisa. É descrita pelos autores uma experimentação sobre agrupamento não hierárquico de textos e outra em categorização não hierárquica de textos em português brasileiro, sendo os mesmos corpora utilizados nas duas situações. A pesquisa procurou verificar qual a melhor forma de representação linguística de textos na fase de pré-processamento, antes das realizações de classificação em si. Para tais formas de representação textual, os autores testaram sintagmas nominais e termos simples, verificando se o uso de sintagmas nominais seria mais produtivo para a construção de grupos ou categorias com menos erros de classificação. O algoritmo de agrupamento utilizado foi o *Simple K-Means* (SKM) e a técnica de categorização com treinamento prévio foi a *Naive-Bayes*. Os corpora descritos por Maia e Souza (2010) são jornalísticos e científicos. O corpus jornalístico contém 160 textos de quatro seções (Informática, Veículos, Mundo e Turismo), o corpus científico é formado por 50 textos das subáreas da Ciência da Informação (Estudos Históricos e Epistemológicos da Informação; Organização do Conhecimento e Representação da Informação; Mediação, Circulação e Uso da Informação; Gestão de Unidades de Informação; Política, Ética e Economia da Informação). Sobre os melhores resultados de agrupamento, os autores descrevem 44% de corretude de agrupamento utilizando termos simples para o corpus científico e 81% de corretude de agrupamento utilizando os sintagmas nominais do corpus jornalístico. Os autores

concluem que o uso de sintagmas nominais não é tão melhor que o uso de termos simples para a tarefa de agrupamento, já que a porcentagem de acertos é similar para as duas formas de representação linguística, contudo o pré-processo de extração de sintagmas nominais é mais demorado que o processo de extração de termos simples. Sobre a diferença de resultados de corretude no processo de agrupamento e no processo de categorização, os autores afirmam que 8 pontos percentuais de diferença entre o agrupamento e a categorização para o corpus científico e 10 pontos percentuais de diferença para o corpus jornalístico foram alcançados. Os processos de classificação com treinamento supervisionado do tipo categorização foram os melhores considerando a taxa de corretude de classificação.

DaSilva, Vieira, Osório e Quaresma (2004) propõem e avaliam o uso de informação linguística na fase de pré-processamento nas tarefas de mineração de textos (agrupamento não hierárquico e categorização não hierárquica). Eles apresentam diversos experimentos comparando as suas propostas para seleção de termos baseadas em conhecimento linguístico com técnicas usuais aplicadas no campo de agrupamento/categorização. O estudo mostra que o uso de informações linguísticas, como classe de palavras (*part-of-speech information*), para identificar os termos dos textos, é útil durante a fase de pré-processamento, antes da categorização e do agrupamento textual, como alternativa para o simples uso de radicalização (extração dos radicais dos termos) e a retirada de palavras sem peso semântico antes da classificação (preposições, artigos, interjeições, etc.).

Seno e Nunes (2008) apresentam alguns experimentos na área de detecção e agrupamento não hierárquico de frases similares em textos escritos em português brasileiro. Eles propõem um esquema de avaliação baseado em um método de agrupamento incremental e não supervisionado, o qual é combinado com métricas de similaridade estatística para medir a distância semântica entre sentenças. Experimentos mostram que este método é robusto mesmo para tratar pequenos conjuntos de dados. O método atingiu 86% e 93% de Medida-F e Purismo, respectivamente, e 0.037 de Entropia para o melhor caso. A Medida-F, o Purismo e a Entropia são medidas tradicionais usadas na área de RI, inclusive na categorização e agrupamento, para medir a qualidade dos grupos gerados em termos de acertos no agrupamento.

Schiessl e Brascher (2012) analisam um Serviço de Atendimento ao Consumidor de uma instituição financeira que centraliza, em forma textual, os questionamentos, as reclamações, os elogios e as sugestões, verbais ou escritas, de clientes. Discute a

complexidade da informação armazenada em linguagem natural e oferece alternativa para extração de conhecimento de bases textuais com a criação de agrupamentos e modelo de classificação automática de textos para agilizar a tarefa realizada atualmente por pessoas. Apresenta uma revisão de literatura que mostra a Descoberta de Conhecimento em Texto como uma extensão da Descoberta de Conhecimento em Dados que utiliza técnicas do Processamento de Linguagem Natural para adequar o texto ao formato apropriado para a mineração de dados e destaca a importância do processo na Ciência da Informação. Aplica a Descoberta de Conhecimento em Texto na base do Serviço de Atendimento ao Consumidor com objetivo de criar automaticamente agrupamentos de documentos para posterior criação de um modelo categorizador automático dos novos documentos recebidos diariamente. Essas etapas são validadas por especialistas de domínio que atestam a qualidade dos agrupamentos e do modelo. Finalmente, geram-se indicadores de desempenho do grau de satisfação do cliente referente a produtos e serviços oferecidos que subsidiam a gestão na política de atendimento.

Nassif (2011) aplica uma série de algoritmos (*K-Means*, *K-Medoids*, *Single Link*, *Complete Link*, *Average Link* e *CSPA*) para agrupamento de documentos e analisa suas potencialidades na prática de análise pericial de computadores apreendidos durante investigações policiais. Aplica tais técnicas em cinco bases de textos provenientes de investigações reais. Adicionalmente, utiliza dois índices de validade relativos (*Silhueta* e sua versão simplificada) para estimar automaticamente o número de grupos. Conclui que os algoritmos hierárquicos *Complete Link* e *Average Link* possuem os melhores resultados, e os algoritmos particionais *K-Means*, *K-Medoids* atingem resultados similares quando inicializados adequadamente.

Furquim (2011) estuda a aplicação de técnicas de aprendizado de máquina (agrupamento e categorização) à pesquisa de jurisprudência, no âmbito do processo judicial eletrônico. Discute e implementa alternativas para o agrupamento dos documentos da jurisprudência, gerando automaticamente classes que servem ao posterior processo de categorização dos documentos anexados ao processo jurídico. O algoritmo *TClus* de Aggarwal, Gates e Yu é selecionado para desenvolvimento de exemplo de uso, com propostas de alteração no descarte de documentos e grupos, e passando a incluir a divisão de grupos. A proposta ainda introduz um paradigma "*bag of terms and law references*" em lugar do "*bag of words*", quando utiliza, na geração dos atributos, os tesouros do Senado Federal e da Justiça Federal para detectar termos jurídicos nos documentos e expressões regulares para detectar referências legislativas. Contribuições deste estudo: confirmação da possibilidade de uso do

aprendizado de máquina na pesquisa jurisprudencial, evolução do algoritmo *TClus* ao eliminar os descartes de documentos e grupos e ao implementar a divisão de grupos, proposta de novo paradigma “*bag of terms and law references*”, através de prototipação do processo proposto com exemplo de uso e avaliações automáticas na fase de *clustering*, e por especialista humano na fase de categorização.

Nogueira, Camargo e Rezende (2009) descrevem o tratamento da imprecisão e incerteza na identificação de documentos similares em português para subáreas da Inteligência Artificial. Apresenta os resultados estatísticos da análise de comportamento de dois algoritmos em tais tarefas: *Fuzzy C-Means* e *Expectation-Maximization* que consideram a possibilidade dos documentos pertencerem a mais de um grupo tópico ou classe, com diferentes graus.

Bezerra, Barra, Ferreira e Von Zuben (2008) descrevem um novo algoritmo de agrupamento cujo nome é ARIA (*Adaptive Radius Immune Algorithm*). Os autores adaptaram um algoritmo inspirado em imunidade, projetado originalmente para o agrupamento baseado em densidade. O algoritmo efetua um agrupamento hierárquico de textos e usa uma medida de similaridade entre textos denominada “Distância Correlativa”, ao invés da clássica “Distância Euclidiana”. A principal vantagem do agrupamento baseado em densidade é que ele é capaz de detectar grupos de formato arbitrários ao invés de formatos pré-definidos, e tal característica melhora a identificação de grupos naturais de documentos. Tal algoritmo trabalha juntamente com uma rede neural do tipo *Semantic Self-Organizing Map* (*semantic-SOM*), para extração de características, e adaptada para trabalhar com textos jornalísticos em português brasileiro.

Pires (2008) apresenta um estudo sobre a utilização e desempenho do algoritmo de agrupamento incremental e hierárquico, em bases de documentos conhecidas e que já foram utilizadas com outros algoritmos. A etapa de pré-processamento dos dados foi realizada com a utilização da plataforma “Biguá”, aplicada para os idiomas português e inglês e para os formatos de documentos em (.txt) e (.pdf). Foram utilizadas três bases como estudos de caso, e para todas foram utilizados três métodos de similaridade. O agrupamento incremental foi executado com várias taxas para poder avaliar o grau de coesão dos grupos formados. Como resultado do algoritmo incremental e hierárquico, é gerada uma árvore contendo nove níveis representando os grupos formados e suas relações.

Existem outros estudos recentes sobre agrupamento e mineração de textos para o português do Brasil utilizando técnicas diversas para outras fontes de informação que não jornais informativos ou documentos científicos: páginas web, e-mails, registros médicos, tais como são os trabalhos de Palazzo (2006), Bastos (2006), Lopes (2009) e Pinheiro (2009).

## 3 INOVAÇÕES PROPOSTAS PARA INDEXAÇÃO E AGRUPAMENTO DE TEXTOS

### 3.1 Apresentação

Nesta pesquisa, as inovações ocorreram em todos os níveis dos experimentos realizados, considerando que os algoritmos de agrupamento conhecidos e implementados no pacote *WEKA* (*X-Means* e *EM*) e o *Algoritmo Evolucionário de Agrupamento Convencional* não foram amplamente testados para o português do Brasil, e especificamente para agrupamento de artigos científicos. Entretanto, pode-se afirmar que a contribuição maior desta pesquisa seja na indexação (representação dos textos) utilizando termos compostos e, também, na proposta de um novo Algoritmo Evolucionário para agrupamento de artigos científicos. Tais inovações poderão, inclusive, ser testadas para outros idiomas e outras fontes de informação (emails, memorandos empresariais, mensagens em redes sociais, textos jornalísticos e jurisprudências) futuramente.

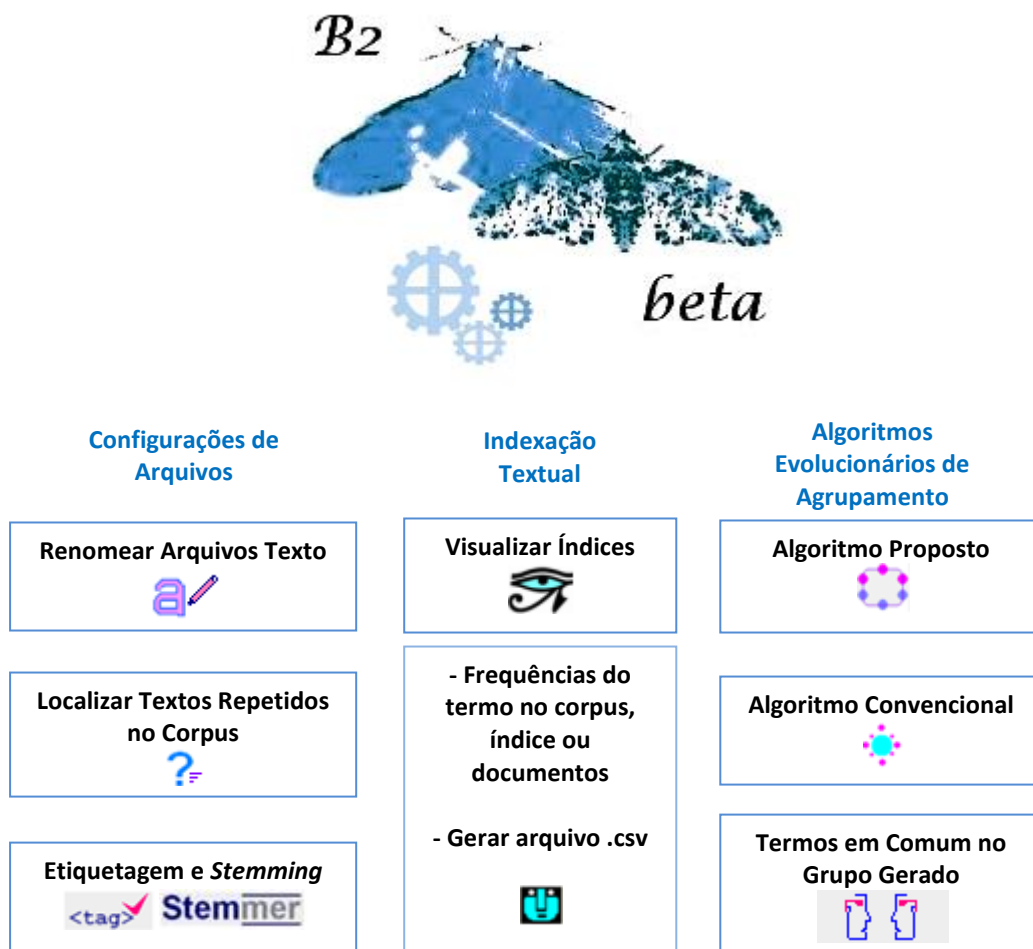
### 3.2 Arquitetura do sistema B2

O protótipo do sistema *B2* implementado contém os algoritmos de indexação por termos simples e compostos e os Algoritmos Evolucionários: convencional e o proposto. Os algoritmos *EM* e *X-Means* são testados utilizando o pacote de mineração de dados e textos *WEKA*. Para tal tarefa, o *B2* é capaz de gerar os índices de representação de textos para seu uso interno e também para o formato (.csv) que é o formato de índice utilizado pelo *WEKA*. O sistema *B2* inclui uma série de funcionalidades em cerca de 4500 linhas de código de programa escritas até o momento: para realizar a indexação, ele chama e executa o programa de etiquetagem (*MXPOST tagger*) e o de radicalização (*stemmer*), já existentes e produzidos pelo NILC, auxiliares nas tarefas de extração de termos simples e compostos. Ainda, o protótipo *B2* possui alguns módulos auxiliares os quais permitem visualizar os índices textuais gerados por ele e visualizar a frequência de termos: nos índices, nos documentos, e no corpus, caso deseje o usuário analisar tais dados. Também, há um módulo para renomear arquivos em série, com o objetivo de identificar a área científica dos arquivos texto do corpus, o que facilita ao usuário saber no fim do processo se tais arquivos foram realmente agrupados corretamente. Há mais um módulo para localizar a repetição de textos no corpus de teste, e um módulo que lista os termos em comum entre os textos de um grupo gerado, após o processo de agrupamento.



O nome *B2* faz uma referência ao importante e conhecido estudo sobre evolução das espécies com as duas formas das mariposas *Biston betularia*. O sistema *B2*, porém, ainda é um protótipo e foi elaborado com o intuito de testar os algoritmos de indexação e agrupamento, uma quantidade de trabalho considerável é necessária ainda para que este se torne um software de uso público. “Protótipo” quer dizer que o software funciona em situações específicas e controladas. A figura, a seguir, ilustra os módulos e submódulos no protótipo do sistema *B2*:

**Figura 6 - Funcionalidades do sistema *B2* a partir do menu principal**



Fonte: elaborado pelo autor, 2012

O novo algoritmo de agrupamento proposto e implementado no sistema *B2*, utilizando Computação Evolucionária, surgiu pela observação dos resultados dos algoritmos tradicionais. Os algoritmos de agrupamento em execução foram observados nos experimentos (descritos adiante) e foi constatado que os algoritmos clássicos *EM* e *X-Means* têm baixos

resultados de acerto no agrupamento. Os experimentos indicam, portanto, que, particularmente, os algoritmos *X-Means* e *EM* talvez não sejam as melhores escolhas para o trabalho com textos (ou dados não estruturados), e acredita-se que isso ocorra também pelo fato dos vetores de representação textual (ou índices) serem extensos, com alta dimensão, ao contrário dos índices de representação para dados estruturados. Tornando o problema mais complexo, um termo existente em um vetor de representação pode ter significados terminológicos completamente diferentes nos textos de origem do corpus, fenômeno linguístico conhecido como polissemia. O uso destes longos vetores de representação, associado ao tradicional conceito de centroide utilizado em cada grupo de textos durante a execução do agrupamento e as constantes operações para cálculo da distância entre vetores de representação e grupos (tal como a Distância Euclidiana) podem ser a causa para baixos resultados de acerto no agrupamento considerando os três algoritmos convencionais citados. Os Algoritmos Evolucionários de agrupamento de textos já propostos na literatura são uma tentativa para melhorar a eficácia (corretude) no agrupamento dos algoritmos clássicos mais antigos, porém, o conceito de centroide e a alta dimensionalidade dos vetores de representação, ainda, são conceitos e métodos mantidos em tais implementações, como observado, por exemplo, no algoritmo de Song e Park (2006). Também, além da pouca eficácia, nos experimentos realizados, verifica-se que o tradicional uso do conceito de centroide e a alta dimensionalidade dos vetores de representação acompanhando os Algoritmos Evolucionários convencionais de agrupamento resultam em processos de agrupamento consideravelmente lentos.

Com a meta de contrapor as características das técnicas tradicionais nos três algoritmos: *EM*, *X-Means* e *Algoritmo Evolucionário de Agrupamento Convencional*, é proposto um novo conceito de representação dos textos que usa termos compostos ao invés de termos simples nos vetores de representação, levando em consideração as formas mais comuns de construção de termos compostos em português brasileiro. Ainda, no algoritmo de agrupamento proposto, é abandonada a abordagem de agrupamento baseada em centroide e utilizam-se vetores individuais para cada texto e não para o todo corpus, ou seja, com dimensão reduzida. Outra inovação é não exigir do usuário entrar com um número máximo de grupos a gerar, como acontece no algoritmo *X-Means* e na implementação evolucionária convencional de Song e Park (2006). Nos experimentos realizados, outra observação diferenciada é a verificação do efeito de combinação de diferentes áreas científicas nos

corpora de teste: foi verificado se diferentes áreas científicas combinadas nos corpora a agrupar causam aumento ou diminuição na taxa de acertos de agrupamento por um algoritmo.

A seguir, descreve-se a indexação e o algoritmo de agrupamento propostos.

### 3.3 Indexação (ou representação) por termos compostos

Neste trabalho, para a representação dos textos, utilizam-se três conjuntos de termos simples (apenas substantivos (S)<sup>5</sup>; substantivos e adjetivos (S,A); substantivos, adjetivos e verbos (S,A,V)), sendo esta uma forma de indexação já conhecida e testada para o português brasileiro em outros trabalhos citados no capítulo anterior.

Os seguintes padrões para termos compostos, descritos abaixo, estão representados como expressões regulares e são buscados para cada texto através do módulo de indexação do sistema *B2*, na tentativa de localizar os termos compostos dos textos. Leia o símbolo de barra vertical “|” a seguir como união e a justaposição como concatenação. O módulo de extração de termos seleciona um termo composto de um texto se ele se enquadrar em uma das formas de representação descritas abaixo, onde *S* indica substantivo, *A* indica adjetivo, *Prep.* indica preposição e *Cont\_Prep.* indica contração de preposição com artigo:

- a) **forma de representação de termos compostos 1:** substantivo seguido de substantivo, ou substantivo seguido de adjetivo, ou adjetivo seguido de adjetivo, ou adjetivo seguido de substantivo. Formalmente: (SS|SA|AA|AS);
- b) **forma de representação de termos compostos 2:** (SS|SA|AA|AS); ou também (substantivo ou adjetivo) seguido de (preposição ou contração de preposição com artigo) seguido de (substantivo ou adjetivo): ((S|A) (Prep|Cont\_Prep) (S|A)). Formalmente: ((SS|SA|AA|AS) | ((S|A) (Prep|Cont\_Prep) (S|A)));
- c) **forma de representação de termos compostos 3:** (SS|SA|AA|AS); ou também só substantivos (S). Formalmente: ((SS|SA|AA|AS) | S);

A razão de permitir os padrões do tipo (AA), (SS), ((A) (Prep|Cont\_Prep) (A)) numa representação textual ocorre devido a três fatores:

---

<sup>5</sup> O termo “substantivo” e o seu símbolo representante (S) utilizados neste texto substitui o termo “nome” e símbolo (N) empregados pelo etiquetador (*POS-Tagger*) *MXPOST* utilizado e descrito em Aires (2000). Esta substituição foi adotada para não confundir os leitores com pouca proximidade à área de Linguística Computacional.

Primeiramente, não consideramos o símbolo de hífen nos textos, esse símbolo e os sinais de pontuação são retirados do texto antes da etiquetagem, logo, termos como *sofá-cama* se tornam disjuntos: “*sofá*” e “*cama*”, antes da etiquetagem, tornando-se dois substantivos separados. Por esta razão, ao localizar termos compostos, padrões do tipo SS são considerados.

Outro motivo seria que segundo Souza-e-Silva e Koch (2011, p. 63) embora as gramáticas do português tragam que adjetivos e substantivos sejam categorias distintas a flutuação categorial entre elas é grande. Muito dos nomes podem ser funcionalmente substantivos (termos determinados) ou adjetivos (termos determinantes) dependendo do contexto como são usados. Por exemplo, no enunciado *um diplomata mexicano* o primeiro vocábulo é substantivo (*diplomata*) e o outro adjetivo (*mexicano*), já no enunciado *um mexicano diplomata* o inverso ocorre, ou seja, o primeiro vocábulo *mexicano* é o substantivo e *diplomata* o adjetivo. Logo, padrões do tipo “adjetivo seguido de adjetivo”, ou (AA), são considerados como termos compostos válidos, pois existe a flutuação categorial que deve ser considerada e não é possível prever como o etiquetador efetua tal classificação morfossintática.

Um terceiro motivo é considerar o sistema de etiquetagem como passível de erro, o software que executa a etiquetagem é treinado utilizando um corpus limitado de textos que não abrange todas as terminologias existentes provenientes das diversas áreas do conhecimento, logo, a possibilidade de troca ao associar termos técnicos que são obviamente substantivos a adjetivos (e a associação contrária) existe.

Sobre o tratamento dos textos antes do processo de indexação, deve-se observar também que foram retirados todos os sinais de pontuação existentes nos textos, pois estes, se deixados, como estão juntos às palavras do texto, podem comprometer de forma significativa o desempenho dos algoritmos de agrupamento (por esta razão ter que retirar os símbolos de hífen). Sobre a consideração do novo acordo ortográfico entre países lusófonos, durante os experimentos, foi considerado que como o cumprimento do acordo só é obrigatório a partir do ano 2016 e os softwares de etiquetagem e *stemming* utilizados não consideram tais mudanças, o acordo não foi tomado como relevante na experimentação e coleta de resultados. Claramente, o uso do trema em alguns textos do corpus e o não uso em outros pode afetar o resultado dos algoritmos de agrupamento (não se sabe exatamente até que nível), porém, o objetivo dos experimentos científicos é justamente acompanhar a realidade social corrente da

produção científica nacional e, na realidade atual, tanto para as ferramentas computacionais quanto para a ortografia dos textos científicos, é encontrado um misto de atualizações e não atualizações.

O módulo de indexação elaborado para extração de termos trabalha para localizar as três formas de representação de termos compostos descritas anteriormente, da seguinte maneira: o usuário escolhe como entrada uma das três formas para representar os textos do corpus a agrupar e o módulo localiza e extrai os padrões quando são identificados em um texto. O exemplo, abaixo, ilustra a forma de representação de termos compostos 1 (SS|SA|AA|AS), se escolhida pelo usuário:

- a) SS: *grande relevância, força diminuição;*
- b) SA: *educação física; atrofia muscular; musculatura esquelética;*
- c) AS: *esquelético atrofiado; considerável número; grande relevância;*
- d) AA: *muscular esquelética; muscular taxas; gênica vias.*

É possível observar claramente que tal tipo de extração, apenas verificando a existência de padrões sequenciais do tipo (SS|AS|AA|AS) leva a erros de extração de termos compostos. Abaixo, listamos as frases originais de onde foram retirados os exemplo de extração *a), b), c), d)* acima, abaixo listados em negrito:

- a) ... observaram resultados positivos e de **grande relevância** clínica...;
- b) ... perda de **força, diminuição** da autonomia...;
- c) ... embora a compreensão do mecanismo de turnover protéico-**muscular** (**taxas** de síntese e degradação) ter sido descrito na década de 70...;
- d) ... principalmente no tocante à biologia molecular, proporcionam-nos razoável compreensão em termos de expressão **gênica, vias** de sinalização celular, etc. ...

Os casos acima (*b, c, d*) em negrito, são erros claros de localização de termos compostos reais, uma vez que todos os elementos lexicais individuais compondo os supostos termos compostos são delimitados por sinais de pontuação.

Tais erros na tentativa do módulo de indexação ao localizar os termos reais dos textos podem ser minimizados pela próxima etapa (etapa 3) do mecanismo de pré-processamento dos textos, antes do agrupamento: a aplicação dos “Algoritmos de Filtragem”. Nesta etapa, ocorre a seleção dos termos por frequência nos textos do corpus de entrada. Nesta pesquisa, consideramos que somente os termos compostos que aparecem com uma frequência igual ou

maior que 4 (quatro) documentos poderiam permanecer no índice de representação do texto, isso acaba filtrando e retirando grande quantidade de termos irreais extraídos e exemplificados acima. Por exemplo, o termo errôneo escolhido e exemplificado acima “gênica vias” dificilmente vai aparecer novamente no corpus e em quatro documentos, logo, somente os termos reais que repetem em outros documentos acabam ficando no índice. O problema para o sistema de agrupamento, porém, pode ser ainda a existência de padrões como a sequencia “grande relevância” que não são termos científicos, mas são padrões válidos que podem ter ocorrências consideráveis em documentos do corpus e causar confusões ao algoritmo de agrupamento. A exigência de determinada frequência dos termos compostos em documentos para que o termo pudesse ser incluído ao índice, também foi utilizada para termos simples (substantivos, verbos e adjetivos), porém, o valor para termos simples foi de pelo menos 5 aparições em documentos diferentes, pelo fato de suas frequências serem maiores que os termos compostos nos corpora de teste.

### **3.4 Inovações nos algoritmos de agrupamento**

Sobre as inovações na etapa onde ocorre a aplicação dos algoritmos de agrupamento, pode-se afirmar que até o atual momento da pesquisa, não foram encontradas aplicações de Computação Evolucionária no agrupamento de textos, especificamente para o português brasileiro. Algumas aplicações para o inglês existem e a implementação descrita se baseia nas características destas implementações, porém, reafirma-se que a ideia é a construção de um novo modelo de Algoritmo Evolucionário que inclusive poderá ser testado em outras línguas, futuramente.

No método evolucionário proposto, a maneira de representar os cromossomos, a operação de *crossover*, a técnica para realizar a mutação e a forma de avaliar os indivíduos são inovadoras ou aplicadas diferentemente. O algoritmo proposto é executado em duas fases evolutivas ao invés de uma, e ainda, um limite máximo de grupos gerados como saída (algo que seria requisitado ao usuário pelo sistema agrupador) não é exigido, ou seja, o sistema procura encontrar o número de grupos corretos e realizar o agrupamento sobre estes grupos.

A razão destas modificações é justificada pela hipótese de se conseguir um algoritmo mais veloz e que produz resultados melhores (em termos de correteude no agrupamento) quando utilizando tais possibilidades originais. A construção de algoritmos de agrupamento, que procuram calcular o número correto de grupos, além de agrupar os dados, é considerado

um tópico de pesquisa bastante atual e inovador na área de meta-heurísticas, uma vez que poucos trabalhos têm sido desenvolvidos neste sentido (DAS; ABRAHAM; KONAR, 2009).

### 3.5 Características do *Algoritmo Evolucionário de Agrupamento Proposto*

#### 3.5.1 Estruturas de dados para os índices (ou representações) de textos

Os três algoritmos de agrupamento apresentados no capítulo anterior e utilizados para testes nesta pesquisa: o algoritmo *EM*, o algoritmo *X-Means* e o *Algoritmo Evolucionário de Agrupamento Convencional* trabalham com um vetor de características (VSM – BoW) (*Vector Space Model – Bag of Words*) que é um tipo de índice representante do texto a classificar, cujo formato é único para todos os textos do corpus de entrada. Isso significa que, para cada texto do corpus, tal vetor de características sempre contém os mesmos termos provenientes de todo o corpus a classificar, e para cada termo é inserido o peso *TF.IDF* calculado para cada texto, ou seja, os termos do índice VSM são os mesmos para todos os textos do corpus, mas o que varia é o valor do peso *TF.IDF* de um termo para um texto específico. A figura, a seguir, ilustra tais vetores para cada texto do corpus.

**Figura 7 - Representação convencional de textos (Índices VSM)**

	Univers_N	Clinic_N	Induz_VERB	Calci_N	Muscul_ADJ
<b>EdFísica1.txt</b>	$3,82 \times 10^{-4}$	$3,82 \times 10^{-4}$	$3,82 \times 10^{-4}$	$2,82 \times 10^{-3}$	$2,00 \times 10^{-3}$
<b>EdFísica2.txt</b>	$3,82 \times 10^{-4}$	$3,82 \times 10^{-4}$	$3,82 \times 10^{-4}$	$1,82 \times 10^{-3}$	$1,00 \times 10^{-3}$
<b>EdFísica3.txt</b>	$3,82 \times 10^{-4}$	$3,82 \times 10^{-4}$	$3,82 \times 10^{-4}$	$1,82 \times 10^{-4}$	$2,82 \times 10^{-4}$
<b>Farmácia1.txt</b>	$2,82 \times 10^{-4}$	$2,82 \times 10^{-3}$	$3,82 \times 10^{-4}$	$1,82 \times 10^{-4}$	$3,82 \times 10^{-4}$
<b>Farmácia2.txt</b>	$2,82 \times 10^{-4}$	$1,82 \times 10^{-3}$	$4,82 \times 10^{-4}$	$3,82 \times 10^{-4}$	$1,82 \times 10^{-4}$

Fonte: elaborado pelo autor, 2012

A figura 7 ilustra a representação tradicional na forma de matriz, cada linha da matriz é a representação de um texto com seus pesos *TF.IDF*, a primeira linha contém todos os termos do corpus a classificar já etiquetados pelo *MXPOST* e processados pelo *Stemmer*.

Um inconveniente desta representação é que ela resulta em vetores muito longos, com alta dimensionalidade para cada texto. Se, por exemplo, cada texto for representado por 25 termos e existirem 100 documentos no corpus a classificar, logo, cada vetor de cada texto terá  $(25 \times 100 = 2500)$  posições, supondo que 500 termos sejam repetidos teremos 2000 termos. Se executarmos algum algoritmo de filtragem para diminuir essa dimensão (como por exemplo, a usada nesta pesquisa, a de exigir que o termo apareça em pelo menos 4 documentos para que ele seja inserido nos vetores de características) e supondo que o número de termos a serem inseridos caia a metade, ainda teremos uma representação vetorial de 1000 termos, para somente 100 documentos a agrupar.

Para o *Algoritmo Evolucionário de Agrupamento Proposto*, não são utilizados vetores de características contendo todos os elementos do corpus a agrupar, ao contrário, cada documento é representado por um Vetor de Características Individual (VCI), contendo apenas os termos do seu texto e o peso *TF.IDF* do termo. A filtragem, exigindo a ocorrência do termo em pelo menos 4 documentos para que o termo seja inserido no índice é mantida para evitar inserção de ruídos no vetor de representação (termos com baixa frequência no corpus e baixo peso semântico). A vantagem esperada desta abordagem para os vetores de características seria diminuir o tempo de computação no agrupamento, já que os cálculos com vetores de representação VSM consomem muito tempo ao verificar se dois vetores são similares ou não no agrupamento. A figura 8, a seguir, ilustra a representação individual utilizada.



**Figura 8 - Representação de textos para o *Algoritmo Evolucionário de Agrupamento Proposto* (índices VCI)**

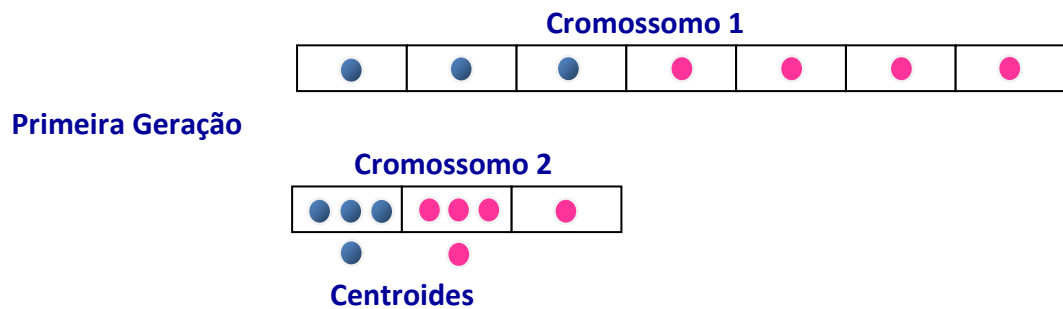
EdFísica1.txt		Farmácia1.txt	
Univers_N	$3,82 \times 10^{-4}$	Plant_N	$2,82 \times 10^{-3}$
Clinic_N	$3,82 \times 10^{-3}$	Clinic_N	$2,82 \times 10^{-3}$
Calci_N	$3,82 \times 10^{-4}$	Calci_N	$3,82 \times 10^{-4}$
Induz_Verb	$2,00 \times 10^{-4}$	Medic_N	$1,82 \times 10^{-3}$
Muscul_ADJ	$2,00 \times 10^{-3}$	Muscul_ADJ	$3,82 \times 10^{-4}$

Fonte: elaborado pelo autor, 2012

### 3.5.2 A estrutura dos cromossomos

Na implementação do Algoritmo Evolucionário de Song e Park (2006) e em diversas implementações para agrupamento utilizando Algoritmos Evolucionários (HRUSCHKA; CAMPELLO; FREITAS; CARVALHO, 2009) o usuário, geralmente, deve inserir um número mínimo de grupos (geralmente de valor 2) e um número máximo de grupos  $k$ . Outra característica, existente em diversas implementações, é a presença do centroide em cada grupo. O centroide é um vetor médio calculado sendo o valor médio dos outros vetores de representação do grupo, quando o grupo contém somente um vetor de representação o centroide é o próprio vetor. Os algoritmos *EM*, *X-Means* e o *Algoritmo Evolucionário de Agrupamento Convencional* testados na pesquisa também trabalham com o conceito de vetor centroide. No trabalho de Song e Park (2006), outra característica notável é a possibilidade de existência de cromossomos de tamanho diferentes, a figura, a seguir, ilustra tais características convencionais.

**Figura 9 - Representação dos cromossomos na Computação Evolucionária**



Fonte: elaborado pelo autor, 2012

Na implementação proposta, optou-se pela representação genotípica e fenotípica de forma idêntica, ou seja, um cromossomo representa um indivíduo que é um agrupamento de vetores textuais, logo, no cromossomo não existem codificações em número binários. Cada gene de um cromossomo representa um grupo de textos e estes textos são representados por vetores de características individuais (VCI), como os mostrados na figura 8 anterior. Em termos de programa e estrutura de dados, cada vetor (VCI) é representado por um número inteiro único dentro do gene do cromossomo, como ilustra a figura 10 a seguir.

### Figura 10 - Representação dos cromossomos na implementação proposta

Representação do Cromossomo e do Vetor de Representação Individual (VCI) para textos

#### Cromossomo 1

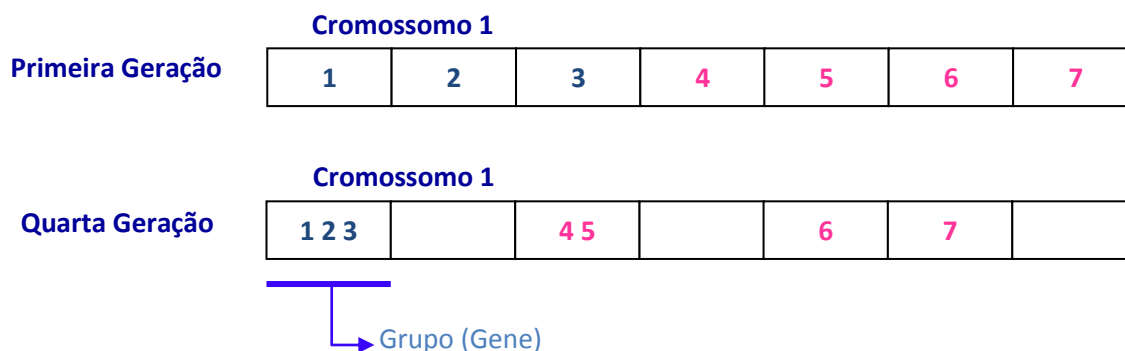
1	2	3	4	5	6	7
1 EdFísica1.txt		4 Farmácia1.txt				
Univers_N	$3,82\text{Ex}10^{-4}$	Plant_N	$2,82\text{Ex}10^{-3}$	Clinic_N	$2,82\text{Ex}10^{-3}$	Calci_N
Clinic_N	$3,82\text{Ex}10^{-3}$	Calci_N	$3,82\text{Ex}10^{-4}$	Medic_N	$1,82\text{Ex}10^{-3}$	Muscul_ADJ
Calci_N	$3,82\text{Ex}10^{-4}$	Muscul_ADJ	$2,00\text{Ex}10^{-3}$			
Induz_Verb	$2,00\text{Ex}10^{-4}$					
Muscul_ADJ	$2,00\text{Ex}10^{-3}$					

Fonte: elaborado pelo autor, 2012

Também, na implementação proposta, um número máximo  $k$  de grupos a gerar definido pelo usuário não é exigido, todo cromossomo é uma estrutura de dados *array* (ou matriz de linha unitária), cujo número máximo de elementos é fixo em  $n$ , onde  $n$  é o número de textos a classificar. Cada elemento do cromossomo é um grupo (que pode também ser chamado de gene na nomenclatura dos Algoritmos Evolucionários), mas caso o número de grupos ao final do processo de agrupamento seja menor que  $n$ , os elementos do *array* que sobram ficam vazios. A figura, a seguir, ilustra a representação no algoritmo proposto, todos os cromossomos têm o mesmo tamanho (dimensão  $n$ ) que é fixo durante toda a execução do algoritmo, as estruturas dos grupos (genes) do cromossomo é que mudam no decorrer da evolução.

### Figura 11 - Representação do cromossomo de tamanho fixo na implementação proposta

Cada número representa um vetor de características (VCI) de um texto, o tamanho de todos os cromossomos (*array*) é o mesmo e é fixo durante toda a evolução.



Fonte: elaborado pelo autor, 2012

A vantagem dessa representação é que não seria necessário trabalhar com cromossomos de tamanho variável, facilitando a representação e agilizando as operações de mutação e *crossover* nos cromossomos.

#### 3.5.3 A função de avaliação

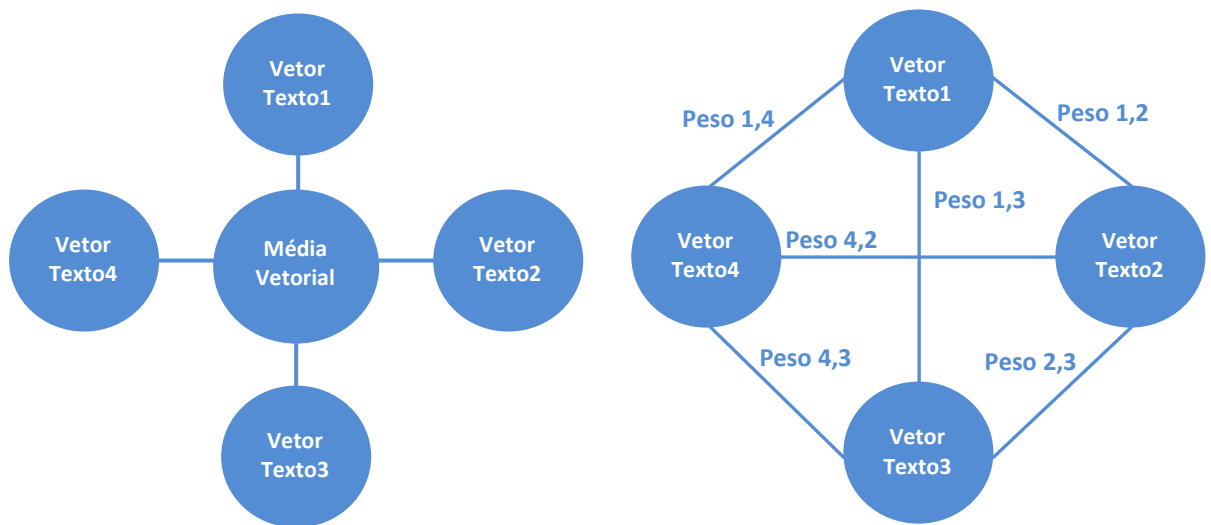
Como especificado, outra característica da implementação proposta é o não uso de centroides. O uso de centroides é uma causa da queda de eficiência no Algoritmo Evolucionário no cálculo da pontuação do cromossomo, pela função de avaliação (*fitness*). É observado nas implementações convencionais de agrupamento que toda vez que um grupo (gene) é modificado no cromossomo, o centroide deve ser recalculado. Como a representação dos textos é, na maioria das vezes, feita por longos vetores de características VSM, a consequência é que o tempo de computação torna-se extenso no cálculo da pontuação do cromossomo pela função de avaliação. Além disso, as distâncias entre os centroides de cada gene devem ser recalculadas pela função de avaliação durante cada ciclo de evolução visando a sua maximização, o que consome ainda mais tempo. Existe também o recálculo da distância entre os elementos do gene e seu centroide a cada ciclo de evolução, que deve ser minimizada no processo evolucionário, e mais tempo é exigido.

A figura 12, a seguir, ilustra a diferença: à esquerda, a média vetorial representa o centroide que é o valor médio atribuído ao gene de um cromossomo. À direita, a

representação proposta, onde para cada dupla possível de vetores de características individuais (VCI) em um gene é atribuído um peso, o peso do gene é a soma dos pesos entre todas as duplas existentes no gene.

**Figura 12 - Representação dos genes de um cromossomo nas implementações**

Convencional, à esquerda, e na implementação proposta, à direita.



Fonte: elaborado pelo autor, 2012

Existem diversas maneiras de representação dos cromossomos e funções de avaliação para Algoritmos Evolucionários. O índice *Davies-Bouldin* é comumente usado para avaliar soluções de agrupamento (os cromossomos) em Algoritmos Evolucionários (HRUSCHKA; CAMPELLO; FREITAS; CARVALHO, 2009), inclusive Song e Park (2006) utilizam tal índice como meio de avaliação dos cromossomos gerados. O índice *Davies-Bouldin* trabalha com a computação de vetores e centroides, exatamente da forma citada anteriormente: recalculando o centroide em cada gene após quaisquer substituições de vetor interna, maximizando a distância entre os centroides de cada gene e minimizando a distância entre cada vetor de representação e seu centroide dentro de um gene. O cálculo do índice *Davies-Bouldin* (DAVIES; BOULDIN, 1979), (SONG; PARK, 2006) é formalmente descrito a seguir:

### Listagem 5 - Índice de avaliação *Davies-Bouldin*

Utilizado em algoritmos de agrupamento convencionais para avaliar soluções de agrupamento.

Inicialmente, uma medida de dispersão  $S$  é calculada para cada grupo (gene)  $C_i$  do cromossomo, contendo  $|C_i|$  elementos.  $S$  retorna um valor real, que é a média das Distâncias Euclidianas ( $q=1$ ) de cada vetor de representação  $x$  do grupo em relação ao seu centroide  $z_i$ :

$$S_{i,q} = \left( \frac{1}{|C_i|} \sum_{x \in C_i} \{\|x - z_i\|_2^q\} \right)^{\frac{1}{q}}$$

O centroide é a media dos elementos de  $C_i$ :

$$z_i = \frac{\sum_{x \in C_i} x}{n_i}, \text{ } n_i \text{ é o número de pontos em } C_i.$$

Em seguida, é computado para cada grupo (gene)  $i$  :

$$d_{ij,t} = \|z_i - z_j\|_t, \text{ distância de ordem } t \text{ entre centroides.}$$

$$R_{i,qt} = \max_{j, j \neq i} \left\{ \frac{S_{i,q} + S_{j,q}}{d_{ij,t}} \right\}$$

E então, o valor do índice *Davies-Bouldin*:

$$DB = \frac{1}{K} \sum_{i=1}^k R_{i,qt}$$

Na implementação de Song e Park (2006), a minimização da função de avaliação  $F$  para os cromossomos é almejada nas evoluções:

$$F = 1/DB$$

Fonte: adaptado de Song e Park, 2006

A computação exigida no índice de avaliação *Davies-Bouldin* é massiva. A solução para diminuir o tempo de computação, então, pode ser reduzir o número de cromossomos nas

gerações, ou diminuir o número de gerações no processo de evolução, porém, para aumentar tal eficiência a eficácia poderá ser sacrificada e boas soluções de agrupamento podem não ser alcançadas.

Como especificado na figura 12 anterior, para tentar reduzir o tempo de avaliação dos cromossomos e gerar um processo de busca de agrupamento mais eficiente, o procedimento de avaliação proposto compara os elementos (vetores de representação) de um gene entre si dois a dois e calcula o peso da dupla, sem construir centroides, e calcula a aptidão (*fitness*) do gene do cromossomo por uma função *soma de pesos*. Desta maneira, os vetores de representação (VCI) do grupo são comparados dois a dois, verificando os termos iguais entre eles. A ideia é que quando os termos forem iguais entre os dois vetores de representação textual (VCI) uma *pontuação de recompensa* é cedida ao grupo, caso contrário uma *pontuação de punição* é imposta. O pseudocódigo, a seguir, descreve a estratégia:

### Listagem 6a - Função de avaliação para calcular a similaridade entre dois vetores (VCI)

```

Função   Calcula_Similaridade_Entre_Dois_Vetores (vetor(i), vetor(j),
Tipo_de_Termo, P)

// A variável peso é o valor acumulado de recompensa por termos iguais
// em dois vetores

peso=0
NumeroDeTermosDiferentes=0

Para cada termo em vetor (i) faça
  Para cada termo em vetor (j) faça
    Se vetor(i).termo = vetor(j).termo então
      peso = (peso * 2) + vetor(i).termo.(Tf.Idf)+
      vetor(j).termo.(Tf.Idf)
    SeNão
      NumeroDeTermosDiferentes ++
    FimSe
  FimPara
FimPara

// Posteriormente um valor de punição é atribuído ao peso acima
// calculado com base
// no número de termos diferentes entre os dois vetores.
// O cálculo para termos compostos, por exemplo(SS, AS, AA, AS) e
// termos simples (S, V, A) é diferente.

Se Tipo_de_Termo = "Termo Composto" então
  peso = peso - ((NumeroDeTermosDiferentes) / (3000 * P))
SeNão
  peso = peso - ((NumeroDeTermosDiferentes) / (1500 * P))
FimSe

// Armazena o peso final calculado entre os dois vetores do gene
// em uma matriz (variável global) na posição i,j

Matriz (i,j) = peso
Retorne (peso)
FimFunção

```

Fonte: elaborado pelo autor, 2012

No código acima, na fase de recompensa, apenas é somado os valores *TF.IDF* dos termos iguais entre dois vetores (VCI) de um gene, com o dobro do peso já acumulado toda vez que termos iguais são encontrados. Na fase de punição, subtraí-se do peso da fase de recompensa o número de termos diferentes dividido por uma constante multiplicada por *P*. Esse valor *P* é dado pelo usuário: se ele for alto a punição será branda e pouco será tirado do peso calculado na fase de recompensa, mas se o valor de *P* for baixo, muito será tirado do



peso e a punição será alta (o valor de  $P$  varia entre 1 e 9). Observe que o valor de  $P$  não é um número mínimo ou máximo de grupos, mas se refere ao nível de punição e recompensa da função de avaliação e é definido pelo usuário.

O efeito do valor de  $P$  dado pelo usuário na avaliação do cromossomo será a quantidade de erros e de grupos que serão retornados na solução final. Quanto mais baixo o valor de  $P$ , menos erros ocorrem e mais grupos são gerados, porém, quanto mais alto o valor de  $P$  menos grupos ocorrem e mais erros de agrupamento surgem. Para algumas combinações de áreas científicas no corpus de entrada, entretanto, aumentar o valor de  $P$  visando a diminuição do número de grupos, não necessariamente implica em aumentar o número de erros de agrupamento de forma significativa. Isso quer dizer que o efeito do valor de punição e recompensa  $P$  no número de erros e grupos depende também da natureza do corpus de entrada.

Para calcular o valor de pontuação do grupo (gene) chama-se a função `Calcula_Similaridade_Entre_Dois_Vetores` descrita anteriormente para todos os elementos do grupo (gene) a avaliar. Executa-se a função para os vetores do grupo, dois a dois, até que não haja pares de vetores a comparar no grupo, e então, os valores dessas chamadas são somados para obter o valor de peso ou pontuação do grupo (ou gene) todo (figura 12, à direita).

Observe que os resultados de peso calculados pela função `Calcula_Similaridade_Entre_Dois_Vetores` são armazenado em uma matriz no final da execução, logo, toda vez que for necessário comparar e calcular a similaridade de dois vetores, se esse cálculo já tiver sido realizado, basta extrair da matriz o resultado da comparação e seu peso, não sendo necessário chamar a função novamente. Isso pode ser feito em qualquer fase da evolução para quaisquer grupos e qualquer cromossomo, em qualquer geração do processo evolucionário. A matriz é uma variável global e pode ser acessada em qualquer tempo do processo, por qualquer módulo. A tendência, portanto, é que as gerações precisem, cada vez menos, do cálculo de semelhança entre vetores (VCI) e apenas será necessário consultar a matriz para saber o cálculo da similaridade entre dois vetores (VCI) em um alto estágio de evolução, o uso dessa memória agiliza ainda mais a evolução e obtenção do agrupamento final.

### Listagem 6b - Cálculo do peso (aptidão) de um gene (grupo) na implementação proposta

```

Função Calcula_Peso_Grupo (Grupo, Tipo_de_Termo, P)

somaGrupo = 0

Para cada vetor (i) em Grupo faça
  Para cada vetor (j) em Grupo faça
    Se (vetor (i) <> vetor (j)) então
      Se (Matriz (i,j) <> vazio) então
        somaGrupo = somaGrupo + Matriz (i,j)
      SeNãoSe (Matriz(j,i) <> vazio)
        somaGrupo = somaGrupo + Matriz (j,i)
      SeNão
        somaGrupo = somaGrupo + Calcula_Similaridade_Entre_Dois_Vetores
          (vetor(i), vetor(j), Tipo_de_Termo, P)
      FimSe
    FimSe
  FimPara
FimPara

Retorne (somaGrupo)
FimFunção

```

Fonte: elaborado pelo autor, 2012

O calculo de avaliação final do cromossomo é dado pela soma de valor de cada gene (grupo):

### Listagem 6c - Cálculo do peso (aptidão) de um cromossomo na implementação proposta

```

Função Calcula_Peso_Cromossomo (Tipo_de_Termo, P)

somaCromossomo = 0
Para cada Grupo em Cromossomo faça
  Se (Grupo <> vazio) então
    somaCromossomo = somaCromossomo + Calcula_Peso_Grupo(Grupo,
      Tipo_de_Termo, P)
  FimSe
FimPara

Retorne (somaCromossomo)
FimFunção

```

Fonte: elaborado pelo autor, 2012

Para cada cromossomo de cada população o valor de avaliação (*fitness* ou aptidão) é calculado pela função `Calcula_Peso_Cromossomo` acima, que utiliza as funções `Calcula_Peso_Grupo` e `Calcula_Similaridade_Entre_Dois_Vetores` descritas anteriormente em pseudocódigo.

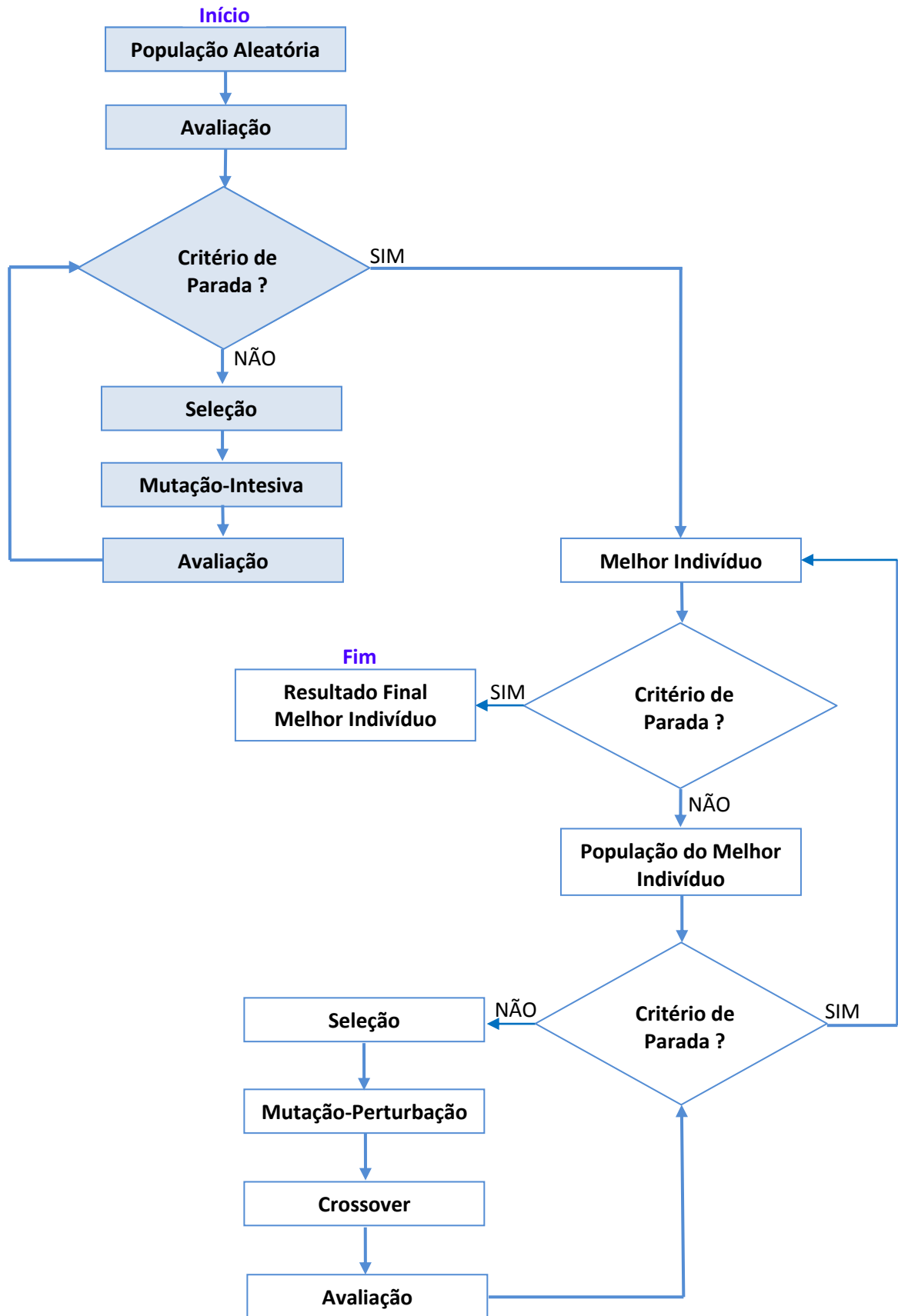
Em suma, a característica principal da função de avaliação utilizada para calcular a aptidão do cromossomo (*fitness*) é a ausência de centroides e cálculos de distâncias, junto à utilização da matriz que guarda o valor de comparação de vetores como memória auxiliar do processo evolucionário, ou seja, é a oposição ao uso do conceito de centroide com o índice *Davies-Bouldin*. Outra característica é a não necessidade de estabelecer um valor numérico máximo de grupos, ao invés disso, um valor de nível de recompensa e punição  $P$  é estabelecido pelo usuário, evitando ter que estabelecer um limite exato para número máximo de grupos a criar. Também, a representação dos textos não utiliza vetores que representam todos os termos do corpus: vetores (VSM – BoW), mas cada texto tem seu índice ou vetor individual (VCI), e somente contendo termos do próprio texto e seus pesos *TF.IDF*. Essa última característica da representação utilizada leva a índices (ou vetores de representação) menores e com menos cálculos necessários para obter a pontuação do cromossomo. Observe ainda, que o novo cálculo para aptidão de cromossomos se baseia na comparação de termos iguais entre índices de termos, ao contrário de uma computação baseada em médias de distâncias entre vetores numéricos de pesos de todo o corpus.

Devido a tais características, é notável que o tempo seja diminuído consideravelmente no cálculo da função de avaliação (*fitness* ou aptidão) para um cromossomo e, logo, pode-se aumentar o número de comparações entre vetores, o número de cromossomos da população e número de gerações, mantendo ainda um tempo bem melhor de resposta e, conseqüentemente, melhorando a qualidade do resultado evolucionário (o indivíduo final, resultante da evolução).

### 3.5.4 O processo iterativo de evolução

O método evolucionário proposto ocorre em duas etapas, diferente do *Algoritmo Evolucionário Convencional* da figura 5 ([página 71](#)) do capítulo anterior. Na primeira etapa há o processo de evolução somente com mutação intensiva nos cromossomos e na segunda etapa ocorre a otimização da melhor solução (ou indivíduo melhor pontuado) obtida na primeira etapa. A segunda etapa ocorre recursivamente até o critério de parada ser acionado, a figura, a seguir, esquematiza o processo:

Figura 13 - Esquema do Algoritmo Evolucionário de Agrupamento Proposto



### *Funcionamento da Etapa 1:*

No algoritmo proposto, visto na figura 13 anterior, na etapa 1, representada com preenchimento azul no fluxograma, um número máximo e fixo de cromossomos são gerados na primeira geração, aleatoriamente. Os cromossomos são gerados numa matriz representando a população inicial, cada um em uma linha da matriz, e o usuário pode estabelecer o tamanho (número de cromossomos) da população. Para gerar a população inicial de forma mais diversa, já que a geração inicial influencia todo o processo evolucionário considerando eficiência e eficácia de busca, metade da população é gerada inserindo os inteiros representativos dos textos (representações VCI) em posições totalmente aleatórias no cromossomo, a outra metade da população inicial é gerada somente nas posições que são múltiplos da raiz quadrada do número de textos a agrupar. Por exemplo, se o número de textos a agrupar for 100, serão inseridos os inteiros representativos dos índices textuais nas posições 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 dos cromossomos. Caso a raiz gere um número fracionado, ocorre o arredondamento para o inteiro mais próximo. O objetivo desta inicialização é gerar, também, grupos iniciais mais concentrados, com mais elementos.

Após a geração inicial ser criada, um número  $G$  de gerações é produzido (ciclos da etapa 1). O número de gerações  $G$  é estabelecido pelo usuário. O usuário pode escolher parar o algoritmo após um número de gerações  $n$  sem melhoria na pontuação do melhor cromossomo.

Também foi utilizada uma abordagem elitista no processo evolucionário: em toda geração, os cromossomos que não são inativados pelo processo de seleção vão para a próxima geração, os cromossomos que continuam ativos sofrem mutação, porém, não existe nesta abordagem a criação de novos cromossomos nas gerações posteriores, ou seja, os próprios cromossomos pais são alterados. Após a seleção e a mutação intensiva os indivíduos que sobrevivem na seleção vão para a geração seguinte, mas se o valor de avaliação do novo cromossomo (indivíduo) modificado por mutação for pior que o antecessor (o original), ele é trocado por este antecessor (o original).

Observe, portanto, que o número de cromossomos apenas cai na evolução das gerações e nunca aumenta. A razão para tal escolha é controlar o tempo gasto e aumentar a eficiência do algoritmo ainda mais. Como os descendentes da geração atual que são piores que seus antecessores da geração prévia são excluídos, o mecanismo acaba forçando a

convergência da geração inicial para uma melhor geração ou uma geração igual à geração anterior, não permitindo diminuição na pontuação dos cromossomos entre as gerações.

#### *Funcionamento da Etapa 2:*

A etapa 2 do processo evolucionário consiste em aperfeiçoar a solução obtida da etapa 1. Ou seja, perturbá-la para que possa ser melhorada ainda mais.

Para isso, inspira-se no conceito de clonagem para otimizar o indivíduo final gerado da etapa 1. A estratégia é gerar uma população inteira contendo apenas clones do melhor indivíduo final da etapa 1, onde o número de cromossomos (indivíduos) é igual ao número de cromossomos da primeira população da etapa 1, definida inicialmente pelo usuário. Estes clones, inicialmente, têm todos a estrutura cromossômica idêntica e posteriormente entram num processo de mutação e *crossover* contínuo, cujo objetivo é evoluir e atingir uma melhoria genética em termos de pontuação (*fitness* ou aptidão), caso isso seja possível.

Durante a evolução, na segunda etapa, as operações de seleção, mutação e *crossover* ocorrem sobre tal população de indivíduos, até atingir um número definido de gerações ou até não ocorrer melhoria dos indivíduos durante certo número de gerações, esses dois números são definidos pelo usuário. Após tal ciclo de gerações sobre a população de clones, uma nova geração clonada é criada: novamente selecionando da última geração da etapa 2 o melhor indivíduo da população de clones evoluída, e em seguida, de forma repetida, gerando uma nova população somente com o melhor indivíduo e reativando o processo evolucionário da etapa 2. Este processo global da etapa 2, de seleção do clone, geração da população idêntica e ativação do ciclo evolucionário, tal como mostra a figura 13 anterior, ocorre também por um período determinado pelo usuário.

Observe que o *crossover* só é utilizado na etapa 2 do mecanismo evolucionário, esse não uso do *crossover* aumenta a rapidez de convergência na etapa 1. Também, foi observado que o uso do *crossover* na segunda etapa do processo evolucionário diminui consideravelmente a taxa de erros de agrupamento da solução final. O *crossover* foi retirado da etapa 1 justamente para diminuir o tempo de computação, em seu lugar, a mutação intensiva (que é mais rápida) foi utilizada.

### 3.5.5 Procedimento de seleção para o *Algoritmo Evolucionário de Agrupamento Proposto*

O procedimento de seleção de cromossomos que produziu melhores resultados de evolução é baseado no valor do melhor cromossomo da geração corrente. Todos os cromossomos que ficam abaixo de 70 % do valor de pontuação (*fitness*) do melhor cromossomo da população corrente são inativados, não passam para a geração seguinte e não sofrem mutação ou recombinação, eles são mortos e desligados do processo evolucionário até o final do processo (parada de execução). É aberta uma exceção para cromossomos cujos valores variem entre 45% e 55% do valor do melhor cromossomo da população corrente, o objetivo desta estratégia é aumentar a variabilidade genética de gerações vindouras, por isso não são excluídos totalmente os cromossomos menos pontuados do processo. Este procedimento de seleção trabalha desta maneira descrita para as duas etapas 1 e 2.

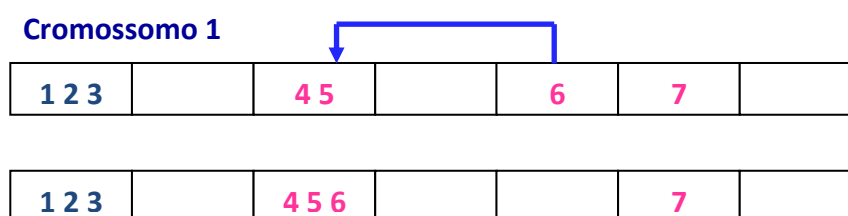
Na implementação proposta, também é possível ativar a seleção de cromossomos somente após certo número de gerações, tanto na primeira quanto na segunda etapa, de acordo com a opção do usuário.

### 3.5.6 Procedimentos de mutação e *crossover* na implementação proposta

#### *Mutação*

O procedimento de mutação nas duas etapas consiste apenas na troca de elementos (VCI) entre genes, ou seja, deslocamos o inteiro representantes do índice VCI para um outro gene qualquer, como ilustra a figura 14 a seguir. Tanto o gene de origem como de destino são escolhidos aleatoriamente. A mutação ocorre para todos os cromossomos ativos e não elitizados da população corrente, nas duas etapas do processo evolucionário.

**Figura 14 - Exemplo de mutação no cromossomo na implementação proposta**



Fonte: elaborado pelo autor, 2012

Nesta implementação proposta, na etapa 1 do ciclo evolucionário, o procedimento mutação é intensivo, ou seja, ocorre muitas vezes entre os genes de um mesmo cromossomo. O número de mutações pode variar por ciclo de evolução: nas primeiras gerações a taxa de mutação em um único cromossomo é maior que nas evoluções perto do número máximo de evoluções definido pelo usuário, logo, essa taxa vai caindo até que nas gerações finais haja apenas um ou dois procedimento de mutação por cromossomo. Tal estratégia tem o objetivo de causar alta variação nos cromossomos das gerações iniciais, onde a população é maior e longe da solução, e pouca variação nos cromossomos das gerações finais que estão mais próximas da solução. A escolha de tais taxas de mutação depende do número de textos a classificar, ou seja, quanto maior o número de textos, maior a taxa de mutação no decorrer da evolução em todos os níveis. Por exemplo, para 120 textos o número de mutações em cada cromossomo é:

2 (Após 80% das gerações)

2 (Após 50% das gerações e antes de 80% )

4 (Após 30% das gerações e antes de 50%)



5 (Após o início das gerações e antes de 30%)

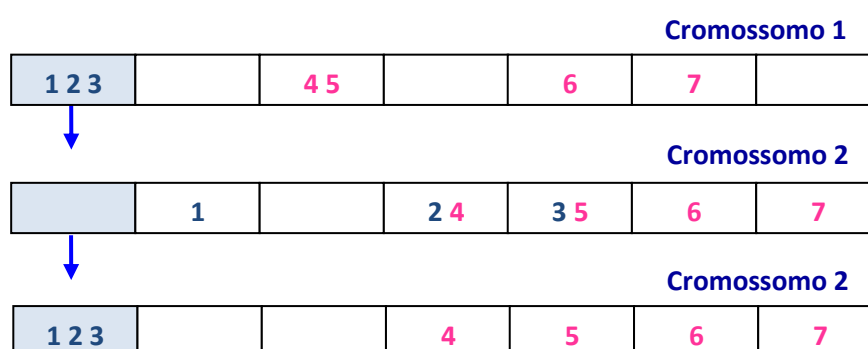
Na segunda etapa, a mutação tem o objetivo de perturbar as soluções:

1 (Durante toda a otimização)

### *Recombinação ou Crossover*

A recombinação ou *crossover* é efetuada obtendo-se o melhor gene (grupo) de um cromossomo aleatório da população corrente, e então, é inserido tal gene em um outro cromossomo escolhido aleatoriamente. Diferentemente da mutação, que opera no nível dos elementos internos do gene, a operação de *crossover* ocorre no nível dos genes. O objetivo de tal operação é propagar o melhor gene do cromossomo para outros cromossomos da população. O número de operações de *crossover* em uma população depende do tamanho da população, quanto maior esse número maior será a repetição do procedimento em uma população. Nos experimentos realizados foi colocado o número de recombinações por geração como 1/3 do tamanho da população. A figura 15, a seguir, ilustra o processo de recombinação ou *crossover*.

**Figura 15 - Exemplo de *crossover* entre dois cromossomos na implementação proposta**



Fonte: elaborado pelo autor, 2012

Na figura acima, o melhor cromossomo (cromossomo 1) da população fornece ao cromossomo 2, o primeiro gene, pelo fato de ser o melhor gene pontuado. O cromossomo 2 recebe o gene do primeiro cromossomo também na posição do gene 1 e para não haver

repetições de um mesmo vetor de representação (VCI) em outros genes, as repetições de números inteiros são retiradas dos outros grupos.

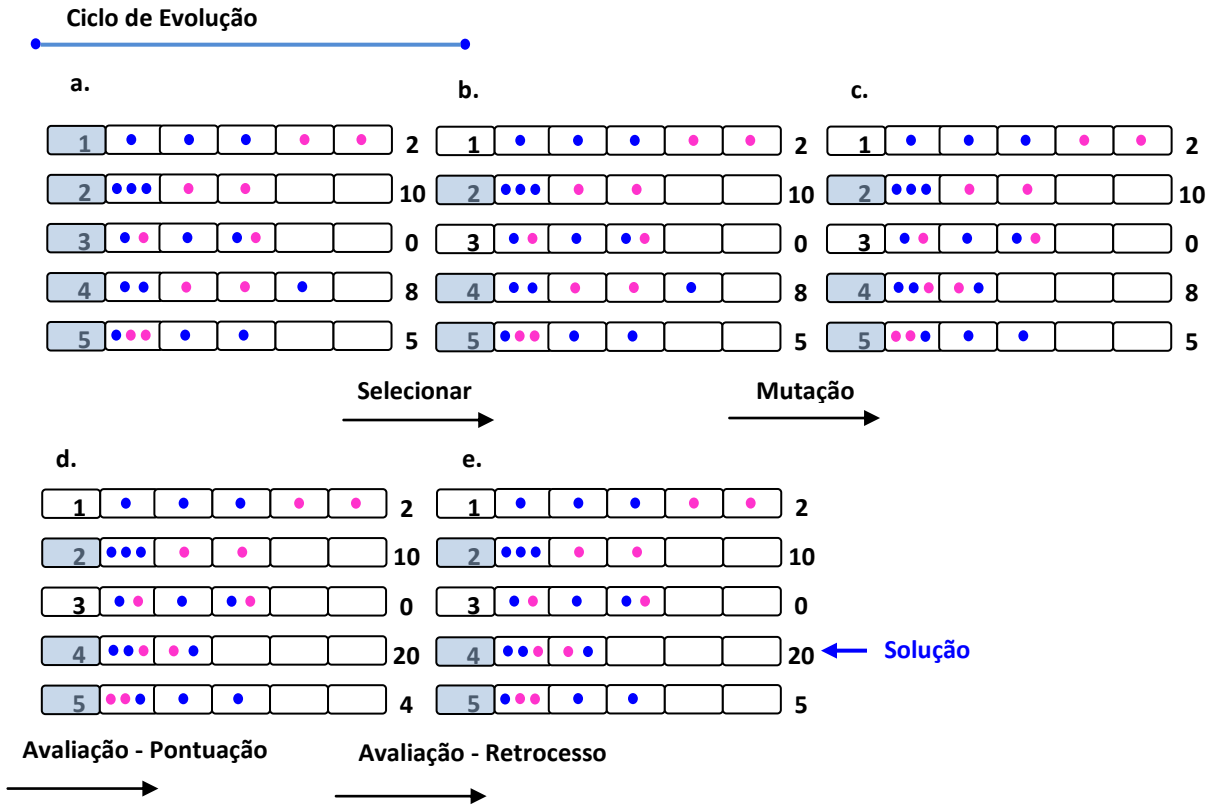
Observe que tanto na mutação como no *crossover* não existe a criação de novos indivíduos, e sim os indivíduos existentes são alterados. O objetivo desta operação é manter fixo o tamanho da população inicial que só deve diminuir e nunca aumentar, esta estratégia permite acelerar o processo evolucionário, diminuindo o tempo de computação.

### *Configurações Default*

Apesar das configurações (ou parâmetros) para o algoritmo proposto serem selecionáveis pelo usuário da aplicação, foi notado que tais escolhas podem influenciar, de forma expressiva, a eficiência e eficácia do algoritmo de agrupamento. As configurações *default* (número de gerações, tamanho da população, critérios de parada, momento de ativação da seleção, taxas de mutação e *crossover*) para este e os demais algoritmos de agrupamento são descritas no capítulo seguinte, para cada experimento é especificada a configuração utilizada.

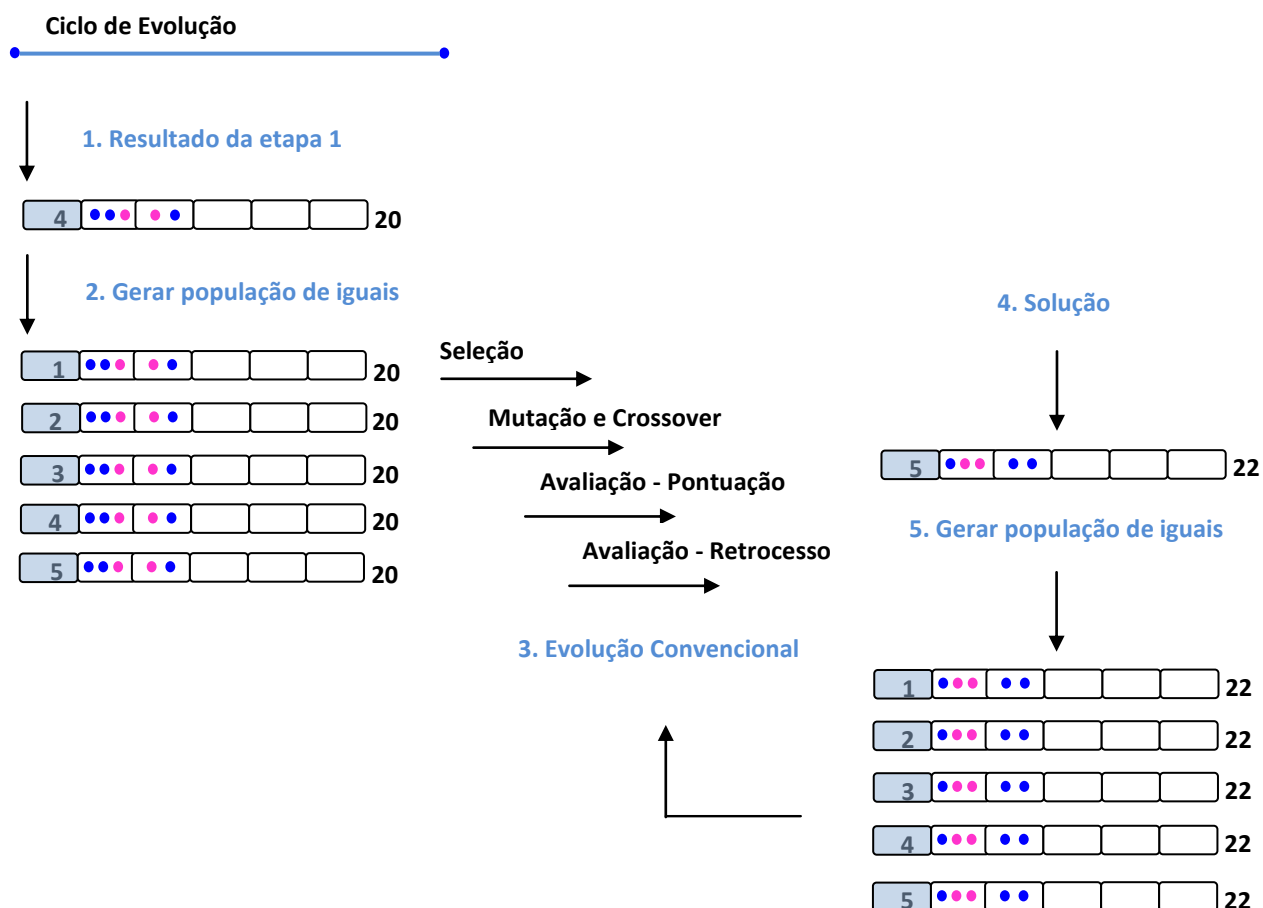
A figura 16a (a seguir) ilustra o processo evolucionário da primeira etapa para uma população e a figura 16b ilustra o processo evolucionário da etapa dois para uma população.

Figura 16a - Exemplo de ciclo de evolução para a etapa 1 na implementação proposta



Fonte: elaborado pelo autor, 2012

Figura 16b - Exemplo de ciclo de evolução para a etapa 2 na implementação proposta



Fonte: elaborado pelo autor, 2012

A figura 16a ilustra a primeira etapa do processo evolucionário. Considere a primeira matriz *a*) como sendo a primeira população, gerada de forma aleatória e já avaliada pela função de avaliação (*fitness*). Esta primeira matriz possui cinco cromossomos e o valor de avaliação (*fitness*) de cada cromossomo está a frente de cada linha da matriz. O objetivo do processo evolucionário é criar dois grupos, cada grupo contendo ou os círculos de tom azul ou violeta.

Após aplicar o procedimento de seleção, os cromossomos da matriz *b*) são os restantes da população *a*), observe que os cromossomos 1 e 3 são inativados (ficam com o primeiro

elemento da linha da matriz em branco) por não atenderem ao critério de seleção (ou mínimo de pontuação), os demais continuam ativos (com o primeiro elemento da linha da matriz em cor azul).

Os cromossomos ativos na matriz *b*) sofrem a mutação intensiva e constituem então os cromossomos da matriz *c*). Estes cromossomos mutantes são avaliados pela função de avaliação e ganham nova pontuação (*fitness*) como mostrado em frente a cada linha da matriz populacional *d*).

Após a avaliação da matriz populacional *d*), pode haver um retrocesso, caso a linha (cromossomo) tenha pontuação inferior ao seu patriarca da matriz *b*) e então o descendente é substituído pelo ascendente, na mesma posição da matriz, como ocorre com o cromossomo 5 na matriz *e*). Observe que o cromossomo 2 não é modificado durante todo o ciclo evolucionário, pois seu valor de avaliação é o maior durante o ciclo de evolução da população, servindo de parâmetro para o processo de seleção da atual população.

Se o processo da etapa 1 continuasse (critério de parada não alcançado) o cromossomo 4 em *e*) seria o novo cromossomo elitista, no próximo ciclo. Caso houvesse a parada, o cromossomo 4 seria o resultado final da etapa 1 e seria enviado à etapa 2.

Na figura 16b, etapa 2, ocorre a otimização do cromossomo-indivíduo da etapa 1. Uma nova população inteira é criada com esta melhor solução (clonagem). Repetidamente, a seleção, a mutação e o *crossover* ocorrem sobre esta população até um critério de parada, sendo que a mutação apenas perturba os cromossomos com uma ou duas trocas nos genes. Ao final, novamente, uma nova população é criada somente com a melhor solução do último ciclo evolucionário e o processo se repete até o critério de parada geral escolhido.

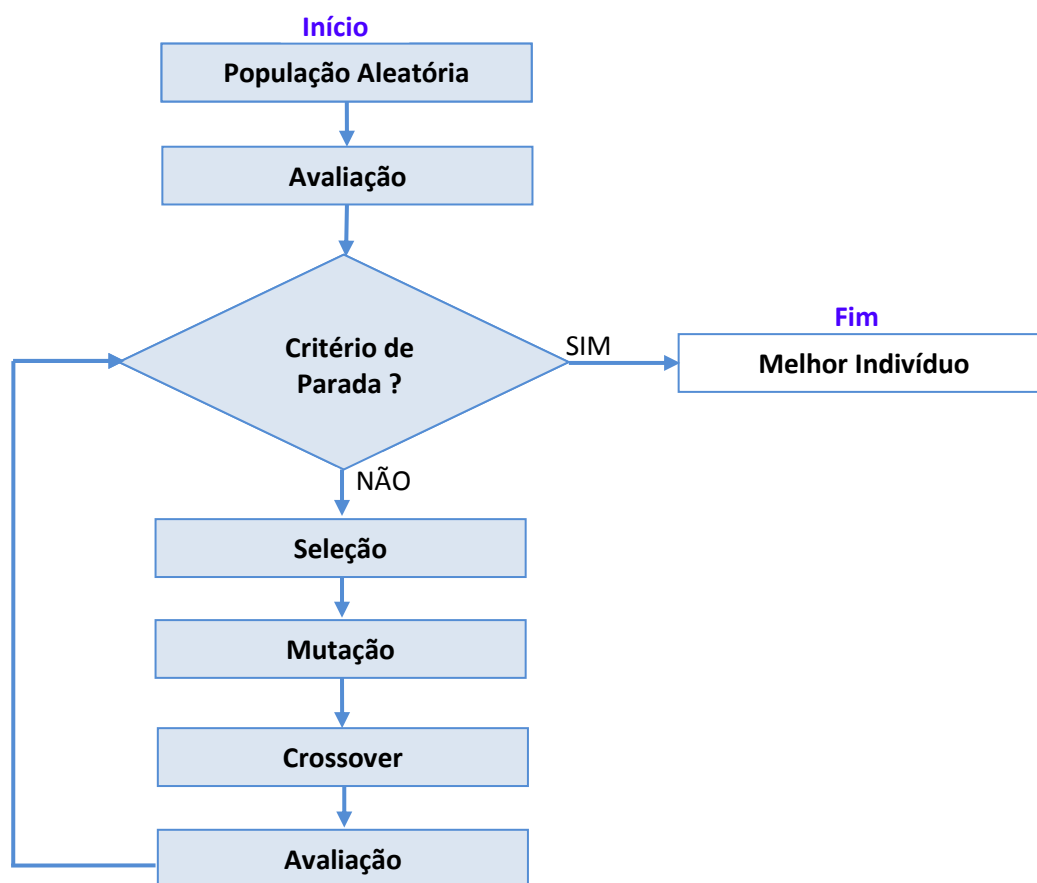
Observe que o número de cromossomos ativos apenas diminui no decorrer das duas etapas, observe ainda, que a pontuação aumenta ou permanece estável no decorrer das duas etapas com o auxílio do retrocesso.

### 3.6 Características do Algoritmo Evolucionário de Agrupamento Convencional

Além do algoritmo de agrupamento proposto acima, uma versão do algoritmo convencional também foi implementada. Tal algoritmo guarda as características mais comuns dos algoritmos de agrupamento descritos na literatura. Este algoritmo foi implementado para verificar se houve melhorias de eficiência e eficácia na nova proposta. A seguir, a descrição desta implementação.

#### 3.6.1 O processo iterativo de evolução

Figura 17 - Esquema do processo iterativo de evolução convencional



Fonte: elaborado pelo autor, 2012

### 3.6.2 Representação dos textos

A representação dos textos ocorre utilizando a representação vetorial convencional (Índices VSM – *Vector Space Model*), já descrita na figura 7 da seção 3.5.1 ([página 86](#)), onde os termos dos vetores de representação são genéricos e retirados do corpus todo e não individuais como proposto no novo método (Índices VCI – Vetor de Características Individual). A tradicional e conhecida definição de centroide foi também utilizada nos grupos (genes) dos cromossomos durante os ciclos de evolução.

### 3.6.3 A estrutura dos cromossomos

A mesma representação cromossômica da nova implementação proposta, figura 11 ([página 91](#)), foi utilizada, porém, junto ao convencional conceito de centroide, como mostrado na figura 12 ([página 92](#)). As populações, também neste caso, possuem tamanho fixo até o final do processo evolucionário, podendo haver diminuição, mas não aumento da população.

### 3.6.4 A função de avaliação

Utiliza a conhecida avaliação pelo índice *Davies-Bouldin*, descrita na Listagem 5 ([página 93](#)), anteriormente. Os cromossomos-indivíduos da população corrente que possuem pontuação (*fitness*) menor que o seu antecessor da população anterior são substituídos pelo antecessor após a aplicação da função de avaliação, ou seja, o processo de retrocesso também é utilizado nesta implementação. A mesma abordagem elitista, descrita na figura 16a ([página 106](#)), também é utilizada.

### 3.6.5 Critério de parada

A iteração pode ser interrompida tanto ao atingir um número máximo de gerações, tanto quando atingir um número de gerações sem causar melhoria (em termos de pontuação) nos indivíduos. Ambos os valores podem ser definidos pelo usuário.

### 3.6.6 Procedimento de seleção

O procedimento de seleção que produziu melhores resultados é baseado no valor do melhor cromossomo da geração corrente. Todos os cromossomos que ficam abaixo de 50 % do valor de pontuação (*fitness*) do melhor cromossomo da população corrente é inativado, não passa para a geração seguinte e não sofre mutação ou recombinação, ele é desligado do

processo evolucionário. Nesta proposta, também é possível ativar a seleção somente após certo número de gerações, segundo as opções do usuário.

### **3.6.7 Procedimentos de mutação e *crossover***

#### *Mutação*

Ocorre da mesma maneira da implementação proposta anteriormente descrita, exemplificado na figura 14 da seção 3.5.6 ([página 103](#)) anterior.

#### *Recombinação ou Crossover*

Ocorre de maneira análoga à descrita anteriormente para o algoritmo de agrupamento proposto, na figura 15 da seção 3.5.6 ([página 104](#)), porém, ao invés de selecionar o melhor gene de um cromossomo da população corrente para inserção na população inteira, seleciona o gene a propagar também de forma aleatória.

O número de operações de *crossover* em uma população depende do tamanho da população, quanto maior esse número maior será a repetição do procedimento em uma população. Nos experimentos realizados foi colocado o número de *crossovers* por geração como 1/3 do tamanho da população inicial.

De modo geral, pode-se afirmar que a estrutura tradicional do ciclo evolucionário em uma única fase é mantida, a representação dos textos tradicional com vetores VSM do tipo *bag of words* é mantida, o uso do conceito de centroides nos grupos e a tradicional avaliação de agrupamento pelo índice *Davies-Bouldin* são também mantidos, visando uma comparação das etapas com métodos tradicionais e as inovações propostas do novo algoritmo.



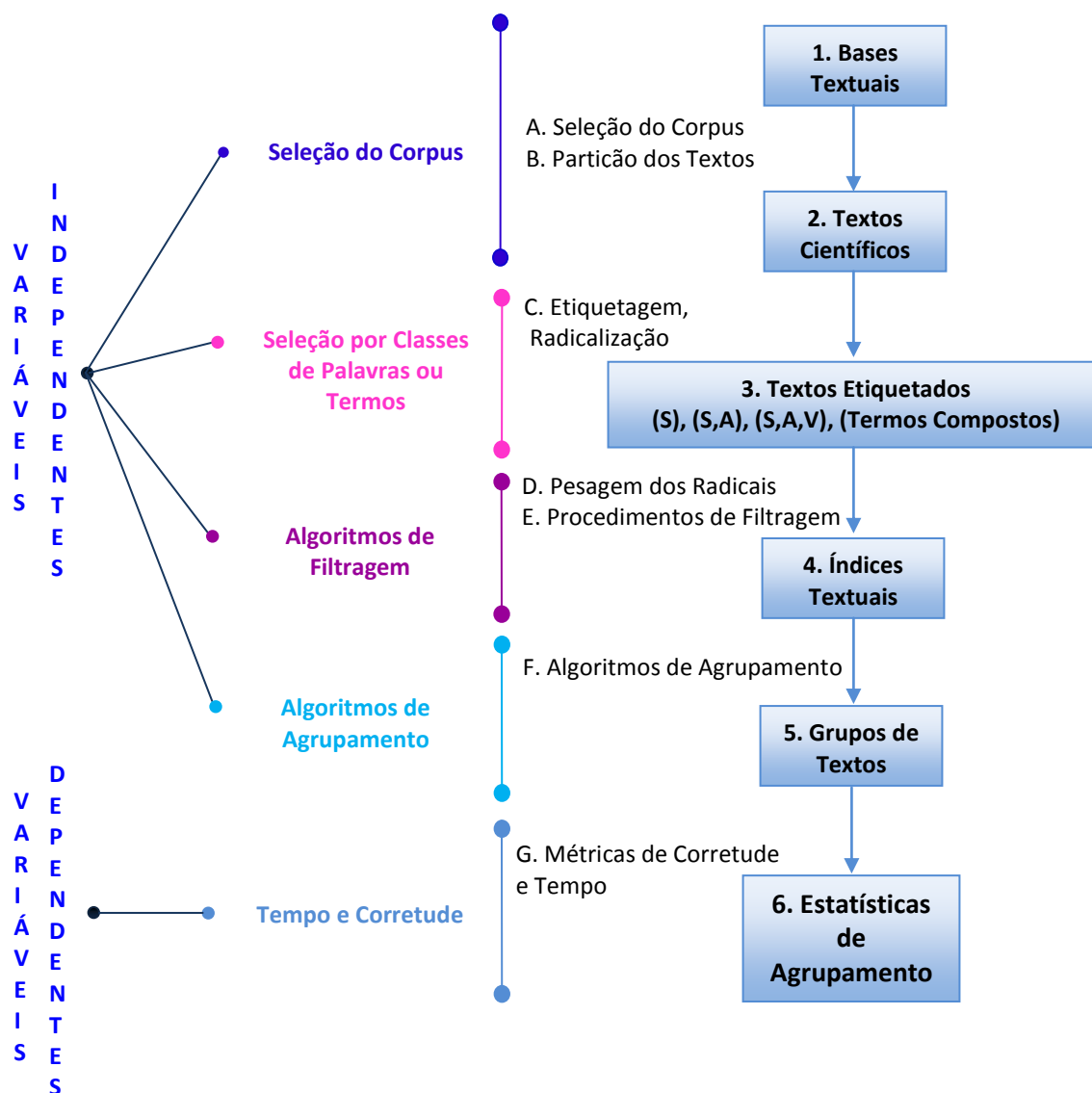
## 4 METODOLOGIA E DESCRIÇÃO DOS EXPERIMENTOS

### 4.1 Apresentação

As etapas dos experimentos de agrupamento são as mesmas, seja para testar os algoritmos de agrupamento já existentes ou para o novo *Algoritmo Evolucionário de Agrupamento Proposto*. Isso significa que qualquer experimento realizado durante a pesquisa seguirá o modelo descrito pelo diagrama a seguir.

**Figura 18 - Modelo dos experimentos realizados na pesquisa**

As anotações ao lado das setas com uma marca alfabética (A. B. C. D. E. F. G.) indicam um processo manual ou algorítmico, os retângulos com marca numérica (1. 2. 3. 4. 5. 6.) indicam um conjunto de dados produzidos ou utilizados pelos processos (A. B. C. D. E. F. G.).



Fonte: elaborado pelo autor, 2012

## 4.2 Descrição das variáveis independentes

### 4.2.1 Seleção do corpus

Consiste na escolha manual dos textos (*processo A*. da figura 18 descrita) que foram submetidos aos testes de agrupamento. Os quatro corpora científicos de teste foram

compilados a partir de diversos periódicos científicos, todos colhidos de diversas bibliotecas ou repositórios digitais de universidades brasileiras que permitem o download dos artigos livremente (veja o apêndice A para a listagem dos periódicos utilizados). Para a composição dos corpora foram escolhidas as áreas: Farmácia, Educação Física, Linguística, Geografia, Odontologia e História, com 20 textos para cada área, totalizando 120 artigos. São formados quatro corpora distintos, sendo cada um deles uma combinação diferente das seis áreas científicas citadas:

- a) Corpus 1: Farmácia, Educação Física e Linguística (60 textos).
- b) Corpus 2: Farmácia, Educação Física, Linguística, História e Geografia (100 textos).
- c) Corpus 3: Farmácia, Educação Física, Linguística, História e Odontologia (100 textos).
- d) Corpus 4: Farmácia, Educação Física, Linguística, História, Geografia e Odontologia (120 textos).

Os corpora de testes foram compilados visando a diversidade, ou seja, há corpus contendo áreas similares e pertencentes a uma mesma grande área (História e Geografia, por exemplo) e também há corpus com artigos de grandes áreas científicas distintas (Farmácia e Linguística, por exemplo), simulando a situação diversa e real dos repositórios. Os quatro corpora foram criados, pois existia a suposição inicial que diferentes combinações de áreas científicas poderiam levar a alterações de eficiência e eficácia no agrupamento, logo, durante os experimentos, tal conjectura foi verificada.

Para cada texto são retiradas somente as três primeiras páginas. O objetivo é incluir o título, o resumo, as palavras-chave e a primeira página da introdução, pois a parte mais informativa (considerando as terminologias) para caracterização da área científica encontra-se nestas páginas iniciais dos artigos. Os textos são convertidos em formato texto (.txt) a partir do formato (.pdf), sem qualquer outra filtragem adicional. Para tal conversão entre formatos e seleção de páginas foi utilizado o software *PDFZilla*<sup>6</sup>.

---

<sup>6</sup> <http://www.pdfzilla.com/>

#### 4.2.2 Seleção por classes de palavras ou termos

Nesta segunda fase de um experimento, realiza-se o primeiro pré-processamento dos textos antes da fase de agrupamento (ou *clustering*). Neste estágio, para cada texto de cada corpus compilado no estágio anterior, um algoritmo de etiquetagem (ou *tagger*) e um algoritmo de radicalização (ou *stemmer*) são executados.

A etiquetagem (ou *tagging*) consiste no processo de etiquetar cada palavra<sup>7</sup> de cada artigo do *corpus*, colocando suas etiquetas morfossintáticas (Substantivo, Adjetivo, Verbo, Preposição, Advérbio, etc.). Foi utilizado o software *MXPOST* para tal tarefa, na sua versão para o português do Brasil, descrito por Aires (2000). Feito isso, ocorre o processo de radicalização (ou *stemming*) que retira o sufixo de cada palavra (por exemplo, as palavras *aluno*, *aluna* e *alunos* tornam-se o radical *alun*). Para o processo de *stemming*, foi utilizado o software *STEMMER* específico para o português do Brasil (CALDAS JR.; IMAMURA; REZENDE, 2001). Nesta pesquisa, como descrito anteriormente, também foram utilizados termos compostos, pois, por hipótese, seria possível que a representação por termos compostos poderia aumentar a eficácia do agrupamento.

O conjunto de dados final produzido (representado pelo *retângulo 3.*, da figura 18 anterior) refere-se aos textos do corpus contendo somente radicais para cada termo de cada texto (somente: Substantivos, Verbos e Adjetivos) ou termos compostos radicalizados.

#### 4.2.3 Algoritmos de filtragem

Esta fase contém o segundo conjunto de procedimentos de pré-processamento dos textos dos corpora: insere um peso para cada termo radicalizado de cada texto da fase anterior e seleciona os melhores termos para representar o texto.

Na fase anterior, após os textos serem etiquetados e radicalizados, os termos simples ou compostos se tornaram os únicos elementos morfossintáticos dentro de cada texto, e então, a *Pesagem dos Radicais* e os *Procedimentos de Filtragem* (processos *D.* e *E.* da figura 18 anterior) são executados na fase atual pelos *Algoritmos de Filtragem*.

---

<sup>7</sup> Neste caso, o conceito de palavra considerado é qualquer sequência de símbolos entre espaços em branco no texto.

A *Pesagem dos Radicais* consiste em inserir um peso numérico para cada termo simples ou composto radicalizado de cada texto, indicando seu peso semântico em relação ao texto a que pertence e ao corpus, ou o quanto um termo contribui para a identificação do tema (grupo): História, Geografia, Farmácia, etc., daquele texto. Diversas medidas bibliométricas têm sido propostas para tal tarefa, a função *TF* (*Term Frequency*) e a função *TF.IDF* (*Term Frequency × Inverse Document Frequency*) são exemplos conhecidos para a pesagem dos termos radicalizados. A medida de pesagem *TF.IDF* foi a única utilizada nos experimentos, em todos eles:

$$TF.IDF = f_{ij} * \log(\text{número de documentos} / \text{número de documentos com o termo } i)$$

(equação 6)

$f_{ij}$  é a frequência do termo simples ou composto  $i$  no documento  $j$ .

Seguindo a figura 18 mostrada anteriormente, e considerando então que cada termo radicalizado contém seu peso, pode-se executar um segundo algoritmo denominado *Procedimento de Filtragem*. Ele consiste em selecionar os termos de representação de um texto (tarefa já conhecida como indexação) utilizando algum critério, por exemplo, o valor de pesagem *TF.IDF* dos termos ou outras estatísticas, por exemplo, comumente, usa-se uma frequência mínima do termo no corpus para ser inserido no índice. No caso desta pesquisa, utilizamos duas filtragens estatísticas onde os termos só permanecerão no índice de representação do texto se tiverem um valor mínimo de ocorrências. Primeiramente, os termos coletados de cada texto são somente os  $n$  termos mais frequentes no texto, onde  $n$  é escolhido pelo usuário. Feito isso, esses termos são filtrados novamente, são capturados os termos com uma frequência  $f$  por documento, ou seja, o termo simples ou composto radicalizado deve ainda aparecer em um número mínimo  $f$  de documentos, número escolhido pelo usuário, para ser inserido no vetor de representação do texto.

A dupla filtragem por frequência no texto e em documentos citada evita a coleta de termos com baixa significância para representar o texto e isso contribui também para acelerar o processo de agrupamento, pois assim, há menos termos a comparar pelos algoritmos de agrupamento.

O produto final dessa fase é representado pelo *retângulo 4.* do diagrama da figura 18 anterior, sendo o *Índice Textual* (vetor de características) para cada texto. Tal conjunto de vetores produzidos contém um índice representando cada texto: os principais termos

radicalizados e seus respectivos pesos semânticos. Como descrito no capítulo anterior, o vetor de características para o *Algoritmo Evolucionário de Agrupamento Proposto* difere dos outros algoritmos, pois os termos são individuais (índice VCI) para cada texto a classificar e não advindos do corpus inteiro a classificar.

Todo o processo de indexação e filtragem dos experimentos: *Seleção por Classes de Palavras ou Termos* e *Algoritmos de Filtragem* é efetuado utilizando o sistema *B2* construído.

#### 4.2.4 Algoritmos de agrupamento

Existem vários algoritmos de agrupamento descritos na literatura, como citado anteriormente. Mas neste trabalho utilizam-se três deles: o *X-Means*, o algoritmo *EM* (*Expectation Maximization*), o *Algoritmo Evolucionário de Agrupamento Convencional* e, ainda, o novo *Algoritmo Evolucionário de Agrupamento Proposto*, descritos no capítulo anterior.

A razão para testar os algoritmos *X-Means* e *EM* é pelo fato deles serem clássicos, já foram testados no agrupamento em algumas línguas, têm grande quantidade de uso, são amplamente citados na literatura e estão implementados em diversas aplicações de mineração de dados e textos, como na ferramenta *WEKA*<sup>8</sup> reconhecida pelos profissionais da área e, inclusive, utilizada nestes experimentos para testar estes dois algoritmos citados. O *Algoritmo Evolucionário Convencional* foi codificado e o novo algoritmo proposto também possui uma codificação original em linguagem de programação *Microsoft Visual Basic* com a chamada de alguns módulos escritos em linguagem *Java* dentro do sistema *B2* construído.

Todos os algoritmos de agrupamento recebem como entrada os índices textuais (vetores de características) produzidos na fase anterior do experimento (*Algoritmos de Filtragem*), pois representam os textos a agrupar. A saída de cada algoritmo será os grupos (*clusters*) contendo cada vetor (índice textual) agrupado (*retângulo 5* da figura 18 anterior).

---

<sup>8</sup> <http://www.cs.waikato.ac.nz/ml/weka/>

## 4.3 Descrição das variáveis dependentes

### 4.3.1 Métricas de corretude e tempo

Na última fase dos experimentos: *Tempo e Corretude* são realizadas as *Métricas de Corretude e Tempo* e são produzidas as *Estatísticas de Agrupamento* (retângulo 6. da figura 18 anterior). As estatísticas consistem nos resultados de análise dos grupos gerados em relação à corretude no agrupamento e ao tempo gasto pelos algoritmos de agrupamento para gerar os grupos de textos.

Ambas as medidas de corretude de agrupamento e de tempo (ou eficácia e eficiência, respectivamente) são aplicadas somente à fase *Algoritmos de Agrupamento*, já que somente esta fase realiza o agrupamento de documentos, através dos algoritmos de agrupamento, porém, as fases anteriores descritas na figura 18 para um experimento podem alterar os valores de tempo, corretude e número de grupos formados.

Os procedimentos de medição estatística utilizados seguiram as propostas descritas na literatura, segundo as referências bibliográficas citadas no capítulo 2. Diversas medidas de corretude poderiam ser utilizadas (Medida-F, Entropia, Precisão e Revocação, Porcentagem de Acertos de Agrupamento ou a Medida de Pureza), neste caso, optamos pela Porcentagem de Acertos de Agrupamento, ou seja, somar a quantidade de erros do agrupamento em cada grupo resultante do experimento e calcular, em seguida, a porcentagem de erros total no processo de agrupamento, e então subtrair de 100%, o que resulta na porcentagem de acertos.

O procedimento detalhado de análise e cálculo da porcentagem de acertos no agrupamento é o seguinte: após executar um algoritmo de agrupamento para um corpus verificamos quais arquivos foram inseridos no grupo errado (ou seja, na área científica errada), então somamos os erros de cada grupo gerado pelo algoritmo de agrupamento e representamos o valor com formato percentual, subtraímos de 100 e obtemos a porcentagem de acertos. Para julgar se um texto está no grupo certo ou não, temos que considerar o tema do grupo de acordo com a maioria de textos de uma área científica naquele grupo. Caso esse maior número tenha um empate, a área do grupo é escolhida aleatoriamente. Foi considerado erro de agrupamento quando um grupo possui apenas um único texto.

Por exemplo, para um grupo produzido por um algoritmo de agrupamento, após um experimento, contendo 5 textos de História, 4 de Geografia e 2 de Linguística, considera-se

que o tema do grupo é História (a área científica que tem mais textos no grupo), com 5 textos agrupados corretamente (os textos de História) e os 6 restantes agrupados de forma errada (os de Geografia e Linguística). Portanto, o que define o tema do grupo é o maior número de textos de uma área específica. Se existissem, em um grupo, 5 textos de História, 5 de Geografia e 2 de Linguística, neste caso, seria escolhido aleatoriamente como tema do grupo ou História ou Geografia, e neste caso tem-se 5 acertos e 7 erros. Existem outras formas de fazer essa medição (outros critérios para estabelecer a qual grupo/tema um texto pertence), mas a vantagem deste critério adotado é sua simplicidade, o que evita confusões ao avaliar os resultados dos experimentos. O importante durante a avaliação é que o procedimento de medida seja mantido ao avaliar todos os grupos e experimentos.

A segunda medida de eficácia (corretude no agrupamento) é o desvio do número de grupos criados pelo algoritmo, ou grupos gerados a mais ou a menos que o esperado. Para essa avaliação, o valor é dado pela função absoluta determinada pela subtração do número de grupos gerados pelo algoritmo menos o número de grupos esperados, chama-se tal valor de *Número de Desvios* (ND).

$$ND = |(Número\ de\ grupos\ gerados - Número\ de\ grupos\ esperados)|$$

(equação 7)

A função retorna sempre um número sem sinal. Por exemplo, se executado um teste para um algoritmo de agrupamento e é esperado que ele produzisse três grupos (História, Geografia e Linguística) e ele produz quatro grupos, temos, portanto, (ND=1). Se o algoritmo gerasse dois grupos, ao invés de três, o valor de desvio seria também (ND=1), ou seja, um grupo a menos produzido.

O tempo de computação de cada algoritmo é medido em minutos. Como o tempo de execução do agrupamento depende da configuração do hardware, do sistema operacional, da versão do software classificador e de outras especificações de arquitetura computacional, são listadas as principais características do sistema computacional utilizado. Algumas variáveis, porém, não podem ser controladas nos experimentos pelo pesquisador e também podem afetar o tempo de computação de forma considerável, como por exemplo, o número de processos em execução, que são controlados automaticamente pelo sistema operacional da máquina.

Arquitetura do sistema computacional utilizada:



- **Processador:** Pentium Dual Core T4500 Intel processor
- **Memória Principal:** 2GB
- **Sistema Operacional:** Microsoft Windows XP Professional 2002 SP 3
- **Versão *Weka*:** 3.6.3
- **Java Development Kit (JDK):** jdk1.6.0\_04

## 5 RESULTADOS E ANÁLISE DE RESULTADOS

### 5.1 Configurações *default* para os experimentos

#### 5.1.1 Configurações dos corpora

A experimentação e medição é efetuada para cada corpus, com quatro corpora no total:

- Corpus 1: Farmácia, Educação Física e Linguística (60 textos).
- Corpus 2: Farmácia, Educação Física, Linguística, História e Geografia (100 textos).
- Corpus 3: Farmácia, Educação Física, Linguística, História e Odontologia (100 textos).
- Corpus 4: Farmácia, Educação Física, Linguística, História, Geografia e Odontologia (120 textos).

#### 5.1.2 Configurações de indexação por tipo de termo

Configurações escolhidas por observações de melhoria nos resultados de eficiência e eficácia dos testes de agrupamento, ou seja, empiricamente. Para cada tipo de termo escolhido para indexação uma configuração específica se mostrou mais apropriada:

- **Substantivos:** São extraídas 3 (três) páginas iniciais de cada texto do corpus de teste; 50% do texto coletado é utilizado; Máximo de termos coletados por documento para inclusão no índice: 20 (vinte) mais frequentes; Medida de Pesagem de cada termo: *Tf.Idf*; Cada termo coletado de um documento para indexação deve ocorrer em no mínimo 5 (cinco) documentos.
- **Substantivos e Adjetivos:** São extraídas 3 (três) páginas iniciais de cada texto do corpus de teste; 50% do texto coletado é utilizado; Máximo de termos coletados por documento para inclusão no índice: 30 (trinta) mais frequentes; Medida de Pesagem de cada termo: *Tf.Idf*; Cada termo coletado de um documento para indexação deve ocorrer em no mínimo 5 (cinco) documentos.
- **Substantivos, Adjetivos e Verbos:** São extraídas 3 (três) páginas iniciais de cada texto do corpus de teste; 50% do texto coletado é utilizado; Máximo de termos coletados por documento para inclusão no índice: 40 (quarenta) mais frequentes;

Medida de Pesagem de cada termo: *Tf.Idf*; Cada termo coletado de um documento para indexação deve ocorrer em no mínimo 5 (cinco) documentos.

- **Termos Compostos (SS| SA| AA| AS):** São extraídas 3 (três) páginas iniciais de cada texto do corpus de teste; 100% do texto coletado é utilizado; Máximo de termos coletados por documento para inclusão no índice: 150 (cento e cinquenta) mais frequentes; Medida de Pesagem de cada termo: *Tf.Idf*; Cada termo coletado de um documento para indexação deve ocorrer em no mínimo 4 (quatro) documentos.
- **Termos Compostos 2 ((SS| SA| AA| AS)| (S|A) (prep|cont.prep) (S|A)):** São extraídas 3 (três) páginas iniciais de cada texto do corpus de teste; 100% do texto coletado é utilizado; Máximo de termos coletados por documento para inclusão no índice: 150 (cento e cinquenta) mais frequentes; Medida de Pesagem de cada termo: *Tf.Idf*; Cada termo coletado de um documento para indexação deve ocorrer em no mínimo 4 (quatro) documentos.
- **Termos Compostos 3 ((SS| SA| AA| AS)| S):** São extraídas 3 (três) páginas iniciais de cada texto do corpus de teste; 50% do texto coletado é utilizado; Máximo de termos coletados por documento para inclusão no índice: 100 (cem) mais frequentes; Medida de Pesagem de cada termo: *Tf.Idf*; Cada termo coletado de um documento para indexação deve ocorrer em no mínimo 4 (quatro) documentos.

### 5.1.3 Configurações do *Algoritmo Evolucionário de Agrupamento Proposto*

As configurações mostradas a seguir foram utilizadas em todos os experimentos realizados que utilizam o *Algoritmo Evolucionário de Agrupamento Proposto*, sendo definidas empiricamente, observando gradativamente os resultados dos experimentos, com o objetivo de melhorar os resultados de eficácia e eficiência.

Observe que para o *Algoritmo Evolucionário de Agrupamento Proposto*, para cada corpus testado e para cada configuração de indexação possível, são mostrados os resultados do algoritmo, considerando dois valores para a variável de punição-relaxamento descrita no capítulo anterior. Tais valores de punição-relaxamento, como descrito anteriormente, podem ser escolhidos pelo usuário para pressionar o algoritmo a gerar menos grupos quando o valor da variável de punição-relaxamento é alto e para gerar mais grupos quando o valor da variável é baixo. Porém, como observado, existe uma tendência ao aumento de erros à medida que o valor de punição-relaxamento aumenta, ou seja, optar por menos grupos produz mais erros de

agrupamento. O aumento dos erros de agrupamento, entretanto, nem sempre acontece com o aumento do valor da variável, pelo fato do número de grupos e o número de acertos serem dependentes também de outros fatores, como, por exemplo, a natureza do corpus.

Foi escolhida a exibição de resultados dos valores 1 e 9 para a variável de punição-relaxamento, o mínimo e o máximo possíveis, que tendem a produzir menos erros com mais grupos gerados para o valor 1 e, para o valor 9, a geração de menos grupos mas com maior taxa de erros de agrupamento.

#### **Quadro 1 - Configurações da primeira etapa do *Algoritmo Evolucionário Proposto***

**Número de gerações:** número de textos a agrupar  $\times$  100.

**Número máximo de cromossomos-indivíduos das populações:** 90.

**Seleção de indivíduos a partir de:** 1% das gerações.

**Máximo de gerações sem evolução:** 35% do número de gerações.

**Quantidade de operações de mutação por cromossomo:** depende do número de textos de entrada e do momento da execução, quanto maior o número de textos maior o número de mutações, quanto mais avançada a evolução menor o número de mutações (descrita em detalhes a seguir).

Fonte: elaborado pelo autor, 2012

#### **Quadro 2 - Configurações da segunda etapa do *Algoritmo Evolucionário Proposto***

**Número de otimizações:** (número de textos a agrupar  $\times$  150) / 100.

**Quantidade de gerações em cada otimização:** (número de textos a agrupar  $\times$  50) / 57.

**Quantidade de operações de *crossover*:** 1/3 do tamanho da população.

**Quantidade de operações de mutação por cromossomo:** 1 ou 2 operações por cromossomo, depende do número de textos a agrupar (descrita em detalhes a seguir).

Fonte: elaborado pelo autor, 2012

**Quadro 3 - Valores *default* para taxa de mutação com número de textos maior que 180 unidades**

**Etapa 1:**

**3** (Após 80% das gerações)

**4** (Após 50% das gerações e antes de 80% )

**5** (Após 30% das gerações e antes de 50%)

**6** (Após o início das gerações e antes de 30%)

**Etapa2:**

**2** (Durante todo o processo)

Fonte: elaborado pelo autor, 2012

**Quadro 4 - Valores *default* para taxa de mutação com número de textos igual a 180 unidades**

**Etapa 1:**

**2** (Após 80% das gerações)

**3** (Após 50% das gerações e antes de 80% )

**5** (Após 30% das gerações e antes de 50%)

**6** (Após o início das gerações e antes de 30%)

**Etapa2:**

**2** (Durante todo o processo)

Fonte: elaborado pelo autor, 2012

**Quadro 5 - Valores *default* para taxa de mutação com número de textos igual a 100 ou 120 unidades**

**Etapa 1:**

**2** (Após 80% das gerações)

**2**(Após 50% das gerações e antes de 80% )

**4** (Após 30% das gerações e antes de 50%)

**5** (Após do início das gerações e antes de 30%)

**Etapa2**

**1** (Após do início das gerações e antes de 30%)

Fonte: elaborado pelo autor, 2012

**Quadro 6 - Valores *default* para taxa de mutação com número de textos igual a 60 unidades**

**Etapa 1:**

**1** (Após 80% das gerações)

**1**(Após 50% das gerações e antes de 80% )

**4** (Após 30% das gerações e antes de 50%)

**5** (Após o início das gerações e antes de 30%)

**Etapa2:**

**1** (Durante todo o processo)

Fonte: elaborado pelo autor, 2012

### 5.1.4 Configurações para o Algoritmo *EM* (WEKA)<sup>9</sup>

**Quadro 7 - Configurações para o Algoritmo *EM* (WEKA)**

<b>Debug:</b> False
<b>DisplayModelInOldFormat:</b> False
<b>MaxIterations:</b> 1000
<b>MinStdDev:</b> 1.0E-6
<b>NumClusters:</b> -1
<b>Seed:</b> 100

Fonte: elaborado pelo autor, 2012

### 5.1.5 Configurações para o Algoritmo *X-Means* (WEKA)

**Quadro 8 - Configurações para o Algoritmo *X-Means* (WEKA)**

<b>BinValue:</b> 1
<b>CutOffFactor:</b> 0.5
<b>DebugLevel:</b> 0
<b>DebugVectorFile:</b> Weka-3-6
<b>DistanceF:</b> Euclidean Distance
<b>InputCenterFile:</b> Weka-3-6
<b>MaxIterations:</b> 10
<b>MaxKMeans:</b> 1000
<b>MaxKMeansForChildren:</b> 1000
<b>MaxNumClusters:</b> 10
<b>MinNumClusters:</b> 2
<b>OutputCenterFile:</b> Weka-3-6
<b>Seed:</b> 10
<b>UseKDTree:</b> False

Fonte: elaborado pelo autor, 2012

---

<sup>9</sup> Para a descrição funcional de cada uma das opções de configuração dos Algoritmos *EM* e *X-Means*, retiradas do arquivo HELP do WEKA, veja o apêndice B.

### 5.1.6 Configurações para o *Algoritmo Evolucionário de Agrupamento Convencional*

#### Quadro 9 - Configurações para o *Algoritmo Evolucionário de Agrupamento Convencional*

**Número de gerações:** Indefinido.

**Número máximo de cromossomos-indivíduos em populações:** 90.

**Máximo de gerações sem evolução:** Indefinido.

**Máximo de Grupos:** 10.

**Seleção de indivíduos:** Existente desde o início da evolução.

**Tempo de Execução:** Tempo de execução do *Algoritmo Evolucionário de Agrupamento Proposto* executado pela última vez somado a 0.5 (metade) deste valor.

**Quantidade de operações de mutação por cromossomo:** as mesmas definidas para o *Algoritmo Evolucionário de Agrupamento Proposto*, na etapa 1.

**Quantidade de operações de *crossover* na população:** 1/3 do tamanho da população original.

Fonte: elaborado pelo autor, 2012

Para o *Algoritmo Evolucionário de Agrupamento Convencional* o *Número de Gerações* e o *Máximo de Gerações sem Evolução*, descritos, estão colocados como *indefinido*, pois o critério de parada será o último tempo de execução do *Algoritmo Evolucionário de Agrupamento Proposto* acrescido de 0.5 vezes esse valor. O Objetivo para tal configuração é demonstrar que quando os valores de tempo gasto pelos dois algoritmos são próximos, o *Algoritmo Evolucionário de Agrupamento Proposto* consegue resultados melhores que o algoritmo convencional.



## 5.2 Resultados

A seguir, os resultados dos experimentos realizados em formato tabular. O formato em tabela foi escolhido para que seja possível analisar a taxa de acertos de agrupamento de um algoritmo e na mesma linha da tabela verificar também o tempo gasto por esse algoritmo e o número de desvios de grupos (ND) gerado por esse algoritmo, com uma rápida localização das três variáveis juntas.

Os experimentos são efetuados para cada corpus, logo, para cada corpus são realizadas as combinações de indexação com os algoritmos de agrupamento possíveis. Os nomes de cabeçalho nas tabelas de resultados que seguem referem-se aos nomes na seguinte legenda:

**Porcentagem de Acertos:** Porcentagem de acertos no agrupamento para um tipo de corpus com uma indexação específica.

**ND:** Número de desvios obtidos (grupos gerados a mais ou a menos) em relação ao número de grupos esperado.

**Tempo Min.:** Tempo de execução somente do algoritmo de agrupamento em minutos.

**Ev. Prop. Relax. 1:** *Algoritmo Evolucionário de Agrupamento Proposto* com variável de punição-relaxamento estabelecida em 1 (valor mínimo).

**Ev. Prop. Relax. 9:** *Algoritmo Evolucionário de Agrupamento Proposto* com variável de punição-relaxamento estabelecida em 9 (valor máximo).

**Ev. Conv.:** *Algoritmo Evolucionário de Agrupamento Convencional*.

**TC 1, 2, 3:** Tipo de índice com termos compostos na indexação, considerado para o experimento em execução.

As linhas das tabelas a seguir em lilás (com símbolo “+” na primeira coluna) indicam melhor valor para ND (próximo de zero), para um corpus. As linhas das tabelas em azul (com símbolo “\*” na primeira coluna) indicam melhor valor percentual de acertos para um corpus.

## CORPUS 1

**Tabela 1A – Experimento 1A**

*Farmácia, Ed. Física e Linguística. Indexação - (Substantivos)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
<b>EM<sup>+</sup></b>	<b>40,0%</b>	<b>0</b>	<b>1,0</b>
X-Means	36,7%	1	1,0
Ev. Prop. Relax. 1	88,3%	6	9,3
Ev. Prop. Relax. 9	91,7%	4	13,2
Ev. Conv.	51,7%	8	13,9

**Tabela 1B – Experimento 1B**

*Farmácia, Ed. Física e Linguística. Indexação - (Substantivos, Adjetivos)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	41,7%	1	1,0
X-Means	43,3%	1	1,0
Ev. Prop. Relax. 1	88,3%	11	11,9
Ev. Prop. Relax. 9	80,0%	5	14,6
Ev. Conv.	58,3%	8	17,9

**Tabela 1C – Experimento 1C**

*Farmácia, Ed. Física e Linguística. Indexação - (Substantivos, Adjetivos e Verbos)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	40,0%	0	1,0
X-Means	43,3%	1	1,0
Ev. Prop. Relax. 1	78,3%	10	13,5
Ev. Prop. Relax. 9	70,0%	3	15,1
Ev. Conv.	55,0%	5	20,3

**Tabela 1D – Experimento 1D**

*Farmácia, Ed. Física e Linguística. Indexação - TC1 (SS/SA/AA/AS)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	61,7%	2	1,0
X-Means	33,3%	1	1,0
<b>Ev. Prop. Relax. 1*</b>	<b>96,7%</b>	<b>12</b>	<b>9,5</b>
Ev. Prop. Relax. 9	75,0%	5	11,3
Ev. Conv.	53,3%	5	14,1

**Tabela 1E – Experimento 1E**

*Farmácia, Ed. Física e Linguística. Indexação - TC2- ((SS|SA|AA|AS))((S|A)  
(Prep|Cont.Prep) (S|A))*

<b>Agrupamento</b>	<b>Porcentagem de Acertos</b>	<b>ND</b>	<b>Tempo Min.</b>
<b>EM</b>	76,7%	4	1,0
<b>X-Means</b>	33,3%	1	1,0
<b>Ev. Prop. Relax. 1</b>	90,0%	6	10,0
<b>Ev. Prop. Relax. 9</b>	95,0%	3	11,5
<b>Ev. Conv.</b>	53,3%	5	15,1

**Tabela 1F – Experimento 1F**

*Farmácia, Ed. Física e Linguística. Indexação - TC3 - ((SS|SA|AS|AA)| S)*

<b>Agrupamento</b>	<b>Porcentagem de Acertos</b>	<b>ND</b>	<b>Tempo Min.</b>
<b>EM</b>	31,7%	2	1,0
<b>X-Means</b>	43,3%	1	1,0
<b>Ev. Prop. Relax. 1</b>	56,7%	1	22,5
<b>Ev. Prop. Relax. 9</b>	55,0%	1	26,2
<b>Ev. Conv.</b>	55,0%	5	33,7

Fonte: dados coletados dos resultados do experimento 1, 2012

## CORPUS 2

**Tabela 2A – Experimento 2A**

*Farmácia, Ed. Física, História, Geografia e Linguística. Indexação - (Substantivos)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	36,0%	0	1,0
X-Means	20,0%	3	1,0
Ev. Prop. Relax. 1	72,0%	8	46,1
Ev. Prop. Relax. 9	50,0%	0	88,7
Ev. Conv.	37,0%	4	69,1

**Tabela 2B – Experimento 2B**

*Farmácia, Ed. Física, História, Geografia e Linguística. Indexação - (Substantivos e Adjetivos)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	20,0%	4	1,0
X-Means	30,0%	3	1,0
Ev. Prop. Relax. 1	74,0%	20	48,9
Ev. Prop. Relax. 9	53,0%	4	89,8
Ev. Conv.	34,0%	4	73,3

**Tabela 2C – Experimento 2C**

*Farmácia, Ed. Física, História, Geografia e Linguística. Indexação - (Substantivos, Adjetivos e Verbos)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	22,0%	1	1,0
X-Means	23,0%	3	1,0
Ev. Prop. Relax. 1	61,0%	14	53,4
Ev. Prop. Relax. 9	75,0%	9	49,3
Ev. Conv.	37,0%	4	80,1

**Tabela 2D – Experimento 2D**

*Farmácia, Ed. Física, História, Geografia e Linguística. Indexação - TC1 - (SS/SA/AA/AS)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	50,0%	2	1,0
X-Means	22,0%	3	1,0
Ev. Prop. Relax. 1	81,0%	12	36,6
Ev. Prop. Relax. 9	64,0%	11	55,9
Ev. Conv.	41,0%	4	54,8

**Tabela 2E – Experimento 2E**

*Farmácia, Ed. Física, História, Geografia e Linguística. Indexação - TC2 - ((SS|SA|AA|AS))((S|A) (Prep|Cont.Prep) (S|A))*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	43,0%	1	1,0
X-Means	20,0%	3	1,0
Ev. Prop. Relax. 1*	86,0%	12	36,1
Ev. Prop. Relax. 9	73,0%	4	53,5
Ev. Conv.	39,0%	4	54,1

**Tabela 2F – Experimento 2F**

*Farmácia, Ed. Física, História, Geografia e Linguística. Indexação - TC3 - ((SS|SA|AS|AA)| S)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	27,0%	3	1,0
X-Means	19,0%	3	1,0
Ev. Prop. Relax. 1	43,0%	1	94,9
Ev. Prop. Relax. 9 <sup>+</sup>	66,0%	0	81,3
Ev. Conv.	40,0%	4	142,3

Fonte: dados coletados dos resultados do experimento 2, 2012

## CORPUS 3

**Tabela 3A – Experimento 3A**

*Farmácia, Ed. Física, Odontologia, Geografia e Linguística. Indexação - (Substantivos)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	39,0%	3	1,0
X-Means	20,0%	3	1,0
Ev. Prop. Relax. 1	61,0%	7	42,3
Ev. Prop. Relax. 9	46,0%	4	73,8
Ev. Conv.	43,0%	4	63,4

**Tabela 3B – Experimento 3B**

*Farmácia, Ed. Física, Odontologia, Geografia e Linguística. Indexação - (Substantivos, Adjetivos)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	22,0%	2	1,0
X-Means	20,0%	3	1,0
Ev. Prop. Relax. 1	65,0%	18	52,5
Ev. Prop. Relax. 9	47,0%	5	87,1
Ev. Conv.	37,0%	4	78,7

**Tabela 3C – Experimento 3C**

*Farmácia, Ed. Física, Odontologia, Geografia e Linguística. Indexação - (Substantivos, Adjetivos e Verbos)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	32,0%	2	1,0
X-Means	20,0%	3	1,0
Ev. Prop. Relax. 1	67,0%	13	51,4
Ev. Prop. Relax. 9	46,0%	4	117,0
Ev. Conv.	39,0%	4	77,1

**Tabela 3D – Experimento 3D**

*Farmácia, Ed. Física, Odontologia, Geografia e Linguística. Indexação - TC1 - (SS/SA/AA/AS)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	33,0%	1	1,0
X-Means	25,0%	3	1,0
Ev. Prop. Relax. 1	80,0%	10	37,4
Ev. Prop. Relax. 9	62,0%	5	51,5
Ev. Conv.	34,0%	4	56,1

**Tabela 3E – Experimento 3E**

*Farmácia, Ed. Física, Odontologia, Geografia e Linguística. Indexação - TC2 - ((SS|SA|AA|AS))/(S|A) (Prep|Cont.Prep) (S|A))*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	39,0%	1	1,0
X-Means	26,0%	3	1,0
Ev. Prop. Relax. 1*	86,0%	8	38,8
Ev. Prop. Relax. 9	74,0%	7	59,9
Ev. Conv.	32,0%	4	58,2

**Tabela 3F – Experimento 3F**

*Farmácia, Ed. Física, Odontologia, Geografia e Linguística. Indexação - TC3 - ((SS|SA|AS|AA)| S)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	32,0%	3	3,0
X-Means	20,0%	3	1,0
Ev. Prop. Relax. 1	60,0%	1	80,9
Ev. Prop. Relax. 9 <sup>+</sup>	48,0%	0	127,0
Ev. Conv.	35,0%	4	121,4

Fonte: dados coletados dos resultados do experimento 3, 2012

## CORPUS 4

**Tabela 4A – Experimento 4A**

*Farmácia, Ed. Física, História, Odontologia, Geografia e Linguística. Indexação - (Substantivos)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	16,7%	5	1,0
X-Means	23,3%	4	1,0
Ev. Prop. Relax. 1	63,3%	14	67,6
Ev. Prop. Relax. 9	40,0%	1	118,8
Ev. Conv.	35,0%	5	101,3

**Tabela 4B – Experimento 4B**

*Farmácia, Ed. Física, História, Odontologia, Geografia e Linguística. Indexação - (Substantivos e Adjetivos)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	26,7%	1	4,0
X-Means	21,7%	4	1,0
Ev. Prop. Relax. 1*	71,7%	18	74,6
Ev. Prop. Relax. 9	49,2%	5	104,8
Ev. Conv.	33,3%	5	111,8

**Tabela 4C – Experimento 4C**

*Farmácia, Ed. Física, História, Odontologia, Geografia e Linguística. Indexação - (Substantivos, Adjetivos e Verbos)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	16,7%	1	2,0
X-Means	20,0%	4	1,0
Ev. Prop. Relax. 1	55,0%	8	103,1
Ev. Prop. Relax. 9	56,7%	2	96,4
Ev. Conv.	33,3%	5	154,6

**Tabela 4D – Experimento 4D**

*Farmácia, Ed. Física, História, Odontologia, Geografia e Linguística. Indexação - TC1 - (SS/SA/AA/AS)*

Agrupamento	Porcentagem de Acertos	ND	Tempo Min.
EM	40,8%	2	1,0
X-Means	20,0%	4	1,0
Ev. Prop. Relax. 1	62,5%	12	64,2
Ev. Prop. Relax. 9	55,0%	5	95,9
Ev. Conv.	35,0%	5	96,3



**Tabela 4E – Experimento 4E**

*Farmácia, Ed. Física, História, Odontologia, Geografia e Linguística. Indexação - TC2 - ((SS|SA|AA|AS))((S|A) (Prep|Cont.Prep) (S|A))*

<b>Agrupamento</b>	<b>Porcentagem de Acertos</b>	<b>ND</b>	<b>Tempo Min.</b>
EM	37,5%	3	1,0
X-Means	21,7%	4	1,0
Ev. Prop. Relax. 1	69,2%	17	63,5
Ev. Prop. Relax. 9	59,2%	5	127,3
Ev. Conv.	28,3%	5	95,2

**Tabela 4F – Experimento 4F**

*Farmácia, Ed. Física, História, Odontologia, Geografia e Linguística. Indexação - TC3 - ((SS|SA|AS|AA)| S)*

<b>Agrupamento</b>	<b>Porcentagem de Acertos</b>	<b>ND</b>	<b>Tempo Min.</b>
EM	22,5%	2	1,0
X-Means	16,7%	4	1,0
Ev. Prop. Relax. 1	44,2%	0	150,0
Ev. Prop. Relax. 9 <sup>+</sup>	<b>48,3%</b>	<b>0</b>	<b>172,1</b>
Ev. Conv.	32,5%	5	225,1

Fonte: dados coletados dos resultados do experimento 4, 2012

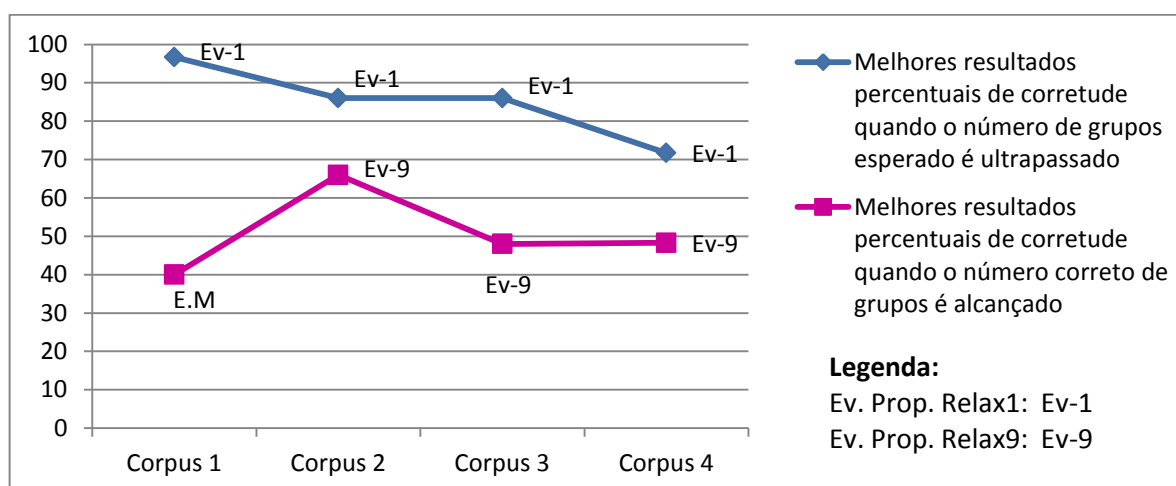
### 5.3 Análise de resultados

Para a análise dos resultados obtidos e descritos na subseção 5.2 anterior, volta-se à problemática e aos objetivos de pesquisa listados previamente na introdução da tese.

#### 5.3.1 Sobre a eficácia e eficiência do *Algoritmo Evolucionário de Agrupamento Proposto*

Os algoritmos de agrupamento com melhores valores de porcentagem de acertos estão marcados nas tabelas com um sinal “\*” em azul e os algoritmos com valores mais baixos de ND (que mais se aproximaram ao número de grupos correto a ser gerado) estão marcados com um sinal de “+” em lilás, na coluna 1, para cada corpus. Ou seja, para cada corpus, estes dois melhores valores são apontados. Tais melhores resultados estão também representados em formato gráfico na figura 19 abaixo:

**Figura 19 - Gráfico com melhores resultados de corretude dos experimentos**



Fonte: dados coletados dos resultados dos experimentos 1 a 4, 2012

O *Algoritmo Evolucionário de Agrupamento Proposto*, tanto com a variável de punição-relaxamento estabelecida em 1 como em 9, consegue uma taxa percentual de acertos no agrupamento maior que os algoritmos de agrupamento *EM* e *X-Means* em todos os experimentos realizados com os quatro corpora (veja cada tabela acima). Quando a variável de punição-relaxamento está estabelecida em 1 no *Algoritmo Evolucionário de Agrupamento Proposto* as porcentagens de acerto chegam às mais altas (melhores) para cada corpus, mas

com isso, o número de grupos tende a aumentar consideravelmente, sendo que o número de grupos gerados chega a ser o quántuplo do número de grupos esperados a serem retornados pelo algoritmo proposto (como no melhor resultado percentual do corpus 1, em azul com “\*”) ou mais que o dobro (como no melhor resultado percentual do corpus 3, em azul com “\*”).

O *Algoritmo Evolucionário de Agrupamento Proposto* também foi capaz de chegar ao valor ideal de número de desvios (ND=0) nos corpora 2,3 e 4 e até (ND=1) no corpus 1, sendo que somente o algoritmo *EM* chegou ao valor ideal (ND=0) no corpus 1, todos marcados em lilás com sinal de “+” nas tabelas. Todos os melhores resultados de ND para o *Algoritmo Evolucionário de Agrupamento Proposto*, ou seja, ND igual ou próximo de zero, são conseguidos pelo uso da indexação TC3 que retorna tais melhores valores. Isso indica que o algoritmo de agrupamento proposto pode mudar seu comportamento alterando a forma de indexação utilizada: caso seja necessário valorizar a taxa percentual de acertos no agrupamento é melhor utilizar a indexação TC2 que apresenta melhores resultados percentuais de acerto no agrupamento, caso seja necessário ter como saída o número de grupos correto (ND=0), então, a indexação TC3, que apresenta melhores resultados para este caso, deve ser utilizada.

Considerando o tempo de computação exigido para gerar o agrupamento, o *Algoritmo Evolucionário de Agrupamento Proposto* (**Ev. Prop. Relax. 1** e **Ev. Prop. Relax. 9**) está muito aquém do tempo de resposta obtido pelos algoritmos convencionais *EM* e *X-Means*. O melhor resultado do *Algoritmo Evolucionário de Agrupamento Proposto* para o corpus 4 (em azul, com um “\*”), por exemplo, chega a tomar 74,6 minutos para agrupar 120 textos. Evidentemente, o tempo de computação deve cair proporcionalmente caso uma arquitetura computacional robusta esteja disponível, ou mesmo, futuramente, talvez seja possível aprimorar a técnica evolucionária para diminuir o tempo de computação. Apesar do tempo obtido pelo *Algoritmo Evolucionário de Agrupamento Proposto* ser ainda alto, a taxa de acertos obtida por ele é bem melhor que a do *Algoritmo Evolucionário de Agrupamento Convencional* implementado, lembrando que propositalmente o tempo de computação máximo de execução permitido ao *Algoritmo Evolucionário de Agrupamento Convencional* (**Ev. Conv.**) é a mesma do *Algoritmo Evolucionário de Agrupamento Proposto* (**Ev. Prop. Relax. 1**) adicionado 0,5 desse valor de tempo, para um mesmo corpus e tipo de indexação.

O *Algoritmo Evolucionário de Agrupamento Convencional* (**Ev. Conv.**) não produziu melhores resultados que o *Algoritmo Evolucionário de Agrupamento Proposto* com alto valor

para a variável de punição-relaxamento (**Ev. Prop. Relax. 9**) em nenhum dos corpora existentes quando o mesmo valor ND foi obtido pelos dois algoritmos. Isso não quer dizer, porém, que o Algoritmo Evolucionário na sua forma convencional não alcança taxas de acertos melhores no agrupamento que o *Algoritmo Evolucionário de Agrupamento Proposto*, mas para que isso pudesse acontecer o tempo dado ao algoritmo convencional teria que ser maior. A ideia dos experimentos, porém, como especificado, era que ambos os Algoritmos Evolucionários estivessem limitados a tempos de computação muito próximos, para medir o nível de corretude no agrupamento em tempos similares.

Apesar das melhorias na porcentagem de acertos ou na aproximação ao número correto de grupos gerados (ND=0), conseguida pelo *Algoritmo Evolucionário de Agrupamento Proposto*, o alcance da melhoria nestas duas variáveis juntas de forma significativa (ou seja, um algoritmo consideravelmente eficaz) ainda não foi conseguido, como ilustrado nas duas linhas do gráfico da figura 19 ([página 137](#)). Logo, tal problema colocado aqui, ainda, é de difícil solução. Porém, as novas descobertas são animadoras e abrem caminhos para melhorias, tentativas, e outros estudos contínuos seguindo o mesmo paradigma.

### 5.3.2 Sobre o impacto da forma de indexação nos algoritmos de agrupamento

Observamos nas tabelas de resultado anteriores que os melhores resultados de porcentagem de acerto no agrupamento e os melhores resultados com (ND=0) ocorrem com a indexação utilizando termos compostos, formatos: TC1, TC2 ou TC3. Porém, dos 8 melhores resultados de agrupamento listados (destacados nas tabelas com um símbolo \* ou +, em azul ou lilás) 2 estão localizados na indexação com termos simples: substantivos e (substantivos e adjetivos) e nenhum melhor resultado utilizando o formato (substantivos, adjetivos e verbos). Os resultados mais expressivos ocorrem nos corpora 2 e 3, com vantagem maior que 10 pontos percentuais de acerto no agrupamento para termos compostos em relação ao melhor resultado para termos simples. Conclui-se, que há nos experimentos realizados, uma tendência de melhores resultados percentuais de agrupamento para a indexação com termos compostos.

### 5.3.3 Sobre o impacto da natureza do corpus nos algoritmos de agrupamento

É visível nas tabelas de resultados (e na linha azul em losangos do gráfico da figura 19, [página 137](#)) que os melhores resultados de porcentagem no agrupamento se encontram para o corpus 1 (96,7%), depois para os corpora 2 e 3 (86,0% para ambos os corpora) e os

resultados mais baixos estão no corpus 4 (71,7%). A diferença é significativa, com mais de 10 pontos percentuais de diferença entre estes três dados percentuais. A distância de resultados entre os corpora, talvez se deva pela natureza diversa de cada corpus (a natureza do conteúdo científico com suas características terminológicas, de escrita e desenvolvimento, específicas das áreas de conhecimento). A causa da queda de acertos no agrupamento também poderia estar relacionada à quantidade de textos e número de áreas científicas que aumentam do corpus 1 para o corpus 2, se mantém fixa nos corpora 2 e 3 e aumenta no corpus 4. Ou ainda, esses valores percentuais de acertos no agrupamento talvez diminuam devido a ambos os fatores. O valor ideal de número de desvios ( $ND=0$ ) foi obtido para os quatro corpora testados, utilizando pelo menos um dos algoritmos de agrupamento.

Outra causa a ser considerada para a queda progressiva de valor percentual de acertos no agrupamento, a medida que o número de textos de entrada e áreas científicas aumentam em cada corpus, é a escolha dos parâmetros para o Algoritmo Evolucionário: talvez outras estratégias para a escolha do tamanho da população, o número de gerações, taxa de mutação, etc. evitem essa queda contínua de eficácia. A escolha ideal de parâmetros (ou funções matemáticas que geram tais parâmetros) para o bom desempenho do Algoritmo Evolucionário é uma escolha difícil, portanto, um tópico a ser mais detalhadamente verificado futuramente.

## 6 CONCLUSÃO E TRABALHOS FUTUROS

A pesquisa descrita nesta tese procurou analisar estatisticamente os resultados de um novo algoritmo de agrupamento por área de conhecimento (Linguística, Educação Física, História, etc.) para artigos científicos escritos em português brasileiro. O algoritmo proposto é inovador e trabalha segundo o paradigma da Computação Evolucionária. Através de um sistema inteligente codificado e chamado *B2*, o algoritmo foi combinado com modos diversos de indexação (ou extração de características) de textos científicos, indexação tanto baseada em termos simples como em termos compostos. Para uma mesma indexação, foram coletados os resultados de corretude do agrupamento (eficácia) e o tempo de computação (eficiência) do algoritmo de agrupamento proposto e de outros algoritmos de agrupamento já conhecidos e descritos na literatura. Tais dados foram coletados, seguindo um critério específico para coleta e análise. O objetivo foi a verificação, através dos dados de coleta, do tempo e corretude de agrupamento, se tais métodos de indexação com o algoritmo de agrupamento proposto têm melhor desempenho que com os algoritmos de agrupamento tradicionais. Também foi verificado se o uso de termos compostos na indexação influenciam positivamente os resultados de eficácia e se a terminologia científica dos diferentes corpora testados causa alguma alteração nos resultados de eficácia do agrupamento automático.

De acordo com os resultados obtidos e analisados na seção anterior, observa-se que os resultados de corretude no agrupamento foram melhorados com o novo algoritmo proposto, mas sob condições bem específicas. Com o novo algoritmo, dos quatro corpora, em três deles, obtivemos taxas de acerto no agrupamento acima de 85%, no corpus restante (o maior) a taxa de acertos é de 71,7%, porém, em todos os corpora o número de grupos gerados é bem mais alto que o esperado, chegando a mais que duplicar ou até a quintuplicar o número de grupos esperado, dependendo do corpus. Portanto, o algoritmo funcionou bem para o agrupamento por área científica, mas fragmentou os grupos gerados em demasia. O tempo de computação utilizado para atingir tais valores é alto, numa arquitetura convencional (PC monoprocessado Intel Dual Core 4500, 2GB de memória) o agrupamento de 120 textos de 6 áreas científicas leva cerca de uma hora e catorze minutos para o melhor resultado percentual.

Quando o algoritmo proposto é configurado para gerar menos grupos pela intervenção do usuário (através de uma variável de punição-relaxamento), a taxa de acertos de agrupamento cai consideravelmente, chegando, por exemplo, a apenas 48,3% de corretude de

agrupamento, mas atingindo o número correto de grupos, tomando como exemplo o maior corpus com 6 áreas científicas e 120 textos.

De maneira geral, foi verificada uma melhoria em relação aos algoritmos tradicionais *Expectation-Maximization (EM)*, *X-Means* e *Algoritmo Evolucionário de Agrupamento Convencional*, pois a taxa de corretude é maior quando o mesmo número de grupos é gerado por estes algoritmos, o tempo de computação, porém, na arquitetura computacional utilizada, é bem maior para o novo algoritmo.

Na indexação, foi verificado que existe uma tendência a melhores resultados utilizado termos compostos em relação a termos simples para a indexação (ou representação textual, ou vetor de características). A natureza ou o número de textos dos corpora também mostrou influenciar os resultados de corretude de agrupamento, mas não foi possível verificar se essa piora no acerto do agrupamento de textos é proporcional ao aumento da quantidade de textos ou pelas características da informação das áreas científicas dos textos.

Com estes experimentos, pode-se concluir que alcançar o resultado ideal (número correto de grupos e alta taxa de acertos de agrupamento) é uma tarefa complexa e o ideal não foi atingido. Porém, são deixadas contribuições significativas para estudos futuros, e os métodos propostos para um agrupamento de qualidade, mas com número maior de grupos, podem ser aperfeiçoados e melhor testados futuramente para se chegar brevemente a um agrupamento correto, mas fragmentado. A tendência, portanto, é que os estudos continuem, almejando diminuir a fragmentação ao mínimo possível, mas mantendo ou maximizando a porcentagem de acerto no agrupamento que apesar de não ser a solução ideal é uma solução aceitável para algumas situações de uso prático. Por exemplo, este mecanismo conseguido, se melhorado, pode servir como ferramenta de auxílio ao profissional da informação para a construção de vocabulários controlados, já que o protótipo consegue reunir textos da mesma área de conhecimento com terminologias similares. Além disso, a opção dada ao usuário do sistema de não ter que definir um número máximo de grupos, mas um valor de punição-relaxamento no agrupamento, traz uma flexibilidade maior de escolha ao usuário, uma vez que definir um número máximo ou exato de grupos para o resultado, como alguns algoritmos convencionais necessitam, é uma previsão exigida ao usuário do sistema que dificilmente será certa.

Acredita-se que esse nível de corretude alcançado pelo sistema proposto no agrupamento, indo de razoável a muito bom, mas altamente fragmentado em diversos grupos,

ocorra pela natureza da informação científica. Os textos das áreas testados mesmo quando pertencendo a uma área científica única nem sempre possuem termos-chave em comum e que ocorrem de forma regular em todos os textos da área. Logo, a investigação das terminologias das áreas científicas (e áreas de conhecimento, em geral), com o objetivo de descrever tais áreas com suas características terminológicas, linguísticas, seus formatos de escrita e natureza de investigação e comunicação podem contribuir de forma significativa para a melhoria dos sistemas de agrupamento textual. Também, o sucesso de tal agrupamento textual depende da existência, abrangência de aplicação e eficácia de recursos previamente existentes, como, por exemplo, estudos linguísticos sobre o português brasileiro, corpora de testes e ferramentas em Linguística Computacional de uso geral.

Muitas outras tentativas podem ser implementadas e testadas visando a melhoria do sistema *B2*, em relação à indexação e à estrutura do *Algoritmo Evolucionário de Agrupamento Proposto*. Como exemplo, a indexação com termos compostos utilizada foi restrita a certos tipos de construções linguísticas considerando somente combinações de substantivos, preposições e adjetivos, não foram considerados termos compostos cuja estrutura seja mais complexa, contendo, por exemplo, verbos, ou mais de um substantivo ou adjetivo. Já sobre o método de agrupamento, talvez, o algoritmo proposto possa ser reelaborado e ter seu tempo de execução diminuído, além disso, o alcance garantido a um valor mínimo de acertos no agrupamento, ainda que subótimo mas que seja um valor mínimo subótimo constante, possa ser garantido toda vez que o algoritmo é executado. O algoritmo proposto é uma primeira tentativa, o objetivo é o aprimoramento da técnica proposta.

Futuramente, também é o objetivo testar e obter melhores resultados com outras fontes de informações textuais mais simples, com menos grupos de classificação possíveis, e onde os termos sejam mais frequentes nos textos. E talvez nestes casos, o sistema proposto já consiga taxas consideravelmente melhores de acerto.

Existe uma diversidade considerável de algoritmos de agrupamento e de formas de indexação (ou extração de características textuais) que não foram testados e comparados com o método proposto. Neste trabalho, a comparação de eficiência e eficácia do algoritmo de agrupamento proposto foi efetuada somente com algoritmos de agrupamento clássicos. Pode ocorrer que outros algoritmos atuais de pré-processamento, extração de características e agrupamento obtenham melhores resultados que os resultados descritos neste trabalho com a



indexação e o algoritmo de agrupamento proposto. Comparações adicionais com outros métodos de extração de características e agrupamento são sugeridos em trabalhos futuros.

Considerando a indexação e classificação automática para o português brasileiro, apesar de existirem registros de trabalhos já realizados há 30 anos, o número de estudos é discreto para esta língua específica, tanto na Ciência da Informação quanto na Ciência da Computação ou áreas correlatas. Os estudos em Mineração de Textos (uma área mais abrangente que envolve o agrupamento de textos) e Recuperação da Informação considerando aspectos linguísticos, sociais e culturais do Brasil têm crescido consideravelmente no país nos últimos anos, como visto na revisão da literatura, com o surgimento de eventos como o Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB), Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL), a Conferência Internacional sobre o Processamento Computacional do Português (PROPOR) e eventos similares.

Como observado nesta pesquisa, as tecnologias da informação desenvolvidas para a organização e recuperação da informação textual não são dependentes apenas da elaboração de algoritmos e uso produtivo de estruturas de dados. A investigação sobre o método computacional, ainda que vital para tais aplicações, é somente um dos elementos que podem garantir o sucesso da classificação e recuperação automática da informação. Os estudos multi e interdisciplinares levam ao aprimoramento da técnica classificatória e conseqüente aumento da taxa de eficiência e eficácia no agrupamento automático de textos.

## 7 REFERÊNCIAS BIBLIOGRÁFICAS

1. AGGARWAL, Charu C.; ZHAI, ChengXiang. A survey of text clustering algorithms. In: **Mining Text Data**. Springer US, 2012. p. 77-128.
2. AGUIAR, F. L. **O controle de vocabulário como dispositivo metodológico para a organização, tratamento e recuperação da informação arquivística**. 2008. 267 f. Dissertação (Mestrado em Ciência da Informação) - Pontifícia Universidade Católica de Campinas, Campinas, 2008.
3. AIRES, Rachel Virgínia Xavier. **Implementação, adaptação, combinação e avaliação de etiquetadores para o português do Brasil**. 2000. Tese de Doutorado.
4. ANDREEWSKI, Alexandre; RUAS, Vitoriano. Indexação automática baseada em métodos lingüísticos e estatísticos e sua aplicabilidade à língua portuguesa. **Ciência da Informação**, v. 12, n. 1, 1983.
5. ARAÚJO JÚNIOR, Rogério Henrique de; TARAPANOFF, Kira. Precisão no processo de busca e recuperação da informação: uso da mineração de textos. **Ci. Inf**, v. 35, n. 3, p. 236-247, 2006.
6. ARAÚJO, Carlos Alberto Ávila. Ciência da Informação como campo integrador para as áreas de Biblioteconomia, Arquivologia e Museologia/Ciencia de La Información como campo integrador para las áreas de Bibliotecología, Archivología y Museología. **Informação & Informação**, v. 15, n. 1, p. 173-189, 2010.
7. ARAÚJO, Carlos Alberto Ávila; MARQUES, Angélica Alves da Cunha; VANZ, Samile Andréa Souza. Arquivologia, Biblioteconomia e Museologia integradas na Ciência da Informação: as experiências da UFMG, UnB e UFRGS. **Ponto de Acesso**, v. 5, n. 1, p. 85-108, 2011.
8. BARRETO, Aldo. Uma quase história da ciência da informação. **DataGramZero**, v. 9, n. 2, 2008.
9. BAŠIĆ, Bojana Dalbelo; BEREČEK, Boris; CVITAŠ, Ana. Mining Textual Data In Croatian. **MIPRO**, Opatija, 2005.
10. BASTOS, V. M. **Ambiente de Descoberta de Conhecimento na Web para a Língua Portuguesa**. 2006. PhD thesis, Universidade Federal do Rio de Janeiro, COPPE, 2006.
11. BEZERRA, George B. et al. A hierarchical immune-inspired approach for text clustering. **Uncertainty and Intelligent Information Systems**, p. 131, 2008.
12. BIDERMAN, Maria Tereza Camargo. O conhecimento, a terminologia e o dicionário. **Ciência e Cultura**, v. 58, n. 2, p. 35-37, 2006.

13. BIDERMAN, Maria Tereza Camargo. O Português Brasileiro e o Português Europeu: Identidade e contrastes. **Revue belge de philologie et d'histoire**, v. 79, n. 3, p. 963-975, 2001.
14. BORGES, Graciane Silva Bruzina; MACULAN, Benildes Coura Moreira dos Santos; LIMA, Gercina Ângela Borém de Oliveira. INDEXAÇÃO AUTOMÁTICA E SEMÂNTICA: estudo da análise do conteúdo de teses e dissertações. **Informação & Sociedade: Estudos**, v. 18, n. 2, 2008.
15. BORKO, Harold; BERNIER, Charles L. **Indexing concepts and methods**. Academic Press, 1978.
16. BRASCHER, Marisa; CAFÉ, Lígia. Organização da informação ou organização do conhecimento. **ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO: Diversidade cultural e políticas de informação – ENANCIB**, v. 9, 2008.
17. CALDAS JR, J. ; IMAMURA, C. Y. M.; REZENDE, S. O. Avaliação de um Algoritmo de Stemming para a Língua Portuguesa. In: **Proceedings of the 2nd Congress of Logic Applied to Technology (LABTEC'2001)**. 2001. p. 267-274.
18. CÂMARA JÚNIOR, Auto Tavares da. **Indexação automática de acórdãos por meio de processamento de linguagem natural**. 2007. Dissertação de Mestrado, Faculdade de Ciência da Informação, Universidade de Brasília.
19. CARLAN, Eliana. **Sistemas de Organização do Conhecimento: uma reflexão no contexto da Ciência da Informação**. Brasília, DF, 2010. Dissertação de Mestrado, Universidade de Brasília.
20. CHAUMIER, J. **As técnicas de documentais**. Publicações Europa-América. 1971, 111p.
21. COELHO, L. d S. Fundamentos, potencialidades e aplicações de algoritmos evolutivos. **Notas em Matemática Aplicada**, v. 2, 2003.
22. COLLINSON, R. L. Índices e indexação: guia para indexação de livros, e coleções de livros, periódicos, e coleções de livros, periódicos, partituras musicais, com uma seção de referência e sugestões para leitura adicional. **Tradução de Antônio Agenor, Brinquet de Lemos**. São Paulo: Polígono, 1971.
23. COOK, L. M. et al. Selective bird predation on the peppered moth: the last experiment of Michael Majerus. **Biology Letters**, v. 8, n. 4, p. 609-612, 2012.
24. CORRÊA, Renato Fernandes et al. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais. **AtoZ: novas práticas em informação e conhecimento**, v. 1, n. 1, p. 11-22, 2011.

25. CUNHA, Murilo Bastos da. Das bibliotecas convencionais às digitais: diferenças e convergências. **Perspectivas em ciência da informação**, v. 13, n. 1, p. 2-17, jan./abr. 2008.
26. DA SILVA, Cassiana Fagundes et al. Mining linguistically interpreted texts. In: **Proceedings of the 20th International Conference on Computational Linguistics**, Geneva, Switzerland. COLING organizers, 2004.
27. DAHLBERG, Ingetraut. Teoria da classificação, ontem e hoje. In: **Conferência Brasileira de Classificação Bibliográfica**. 1976.
28. DAS, Swagatam; ABRAHAM, Ajith; KONAR, Amit. **Metaheuristic Clustering**. Springer Verlag, 2009.
29. DAVIES, David L.; BOULDIN, Donald W. A cluster separation measure. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, n. 2, p. 224-227, 1979.
30. DIAS, B. C. O caso das mariposas Biston betularia. **Polegaropositor**, 2008. Disponível em < <http://polegaropositor.com.br/biologia/o-caso-das-mariposas-biston-betularia/>>. Acesso em 04/10/2013.
31. DIAS-DA-SILVA, Bento Carlos. O estudo linguístico-computacional da linguagem. **Letras de Hoje**, v. 41, n. 2, p. 103-138, 2006.
32. DUQUE, Claudio Gottschalg. **SiRiLiCO, uma proposta para um sistema de recuperação de informação baseado em teorias da Linguística Computacional e ontologia**. 2005. Tese de Doutorado. Tese (Doutorado em Ciência da Informação). Universidade Federal de Minas Gerais – Escola de Ciência da Informação. Belo Horizonte.
33. EIBEN, Agosten E.; SMITH, James E. **Introduction to Evolutionary Computing**. Berlin: Springer, 2010.
34. FERNEDA, Edberto. **Recuperação de informação: análise sobre a contribuição da Ciência de Computação para a Ciência da Informação**. 2003. Tese de Doutorado.
35. FIDEL, Raya. User-centered indexing. **JASIS**, v. 45, n. 8, p. 572-576, 1994.
36. FOGL, J. Relations of the concepts 'information' and 'knowledge'. **International Fórum on Information and Documentation**, The Hague, v.4, n.1, p. 21-24, 1979.
37. FREDERIKSEN, C. H. Representing Logical and Semantic Structure of Knowledge Acquired from Discourse. **Cognitive Psychology**, v.7, n. 3, p. 371-458, 1975.
38. FUJITA, Mariângela Spotti Lopes. A identificação de conceitos no processo de análise de assunto para indexação. **Revista Digital de Biblioteconomia e Ciência da Informação**, v. 1, n. 1, 2003.

39. FURQUIM, Luis Otávio de Colla. **Agrupamento e Categorização de Documentos Jurídicos**. 2011. Dissertação de Mestrado, Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul.
40. GASPERIN, Caroline et al. Extracting xml syntactic chunks from portuguese corpora. In: **Proc. of the TALN Workshop on Natural Language Processing of Minority Languages and Small Languages**. 2003. p. 223-232.
41. GLOVER, Fred. Future paths for integer programming and links to artificial intelligence. **Computers & Operations Research**, v. 13, n. 5, p. 533-549, 1986.
42. GOMES, H.E.; GUSMÃO, H.R. **Guia prático para a elaboração de índices**. Niterói: GBIDCSH da APBRJ, 1983.
43. GONÇALVES, Janice. **Como classificar e ordenar documentos de arquivo**. Arquivo do Estado, 1998.
44. GUEDES, Vânia L. Da S.. Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos. **Ciência da informação**, v. 23, n. 3, p. 318-326, 1994.
45. GUIMARÃES, JAC. Indexação em um contexto de novas tecnologias.[SI: sn], 2000. 10p. **Texto didático**.
46. HALKIDI, Maria; BATISTAKIS, Yannis; VAZIRGIANNIS, Michalis. On clustering validation techniques. **Journal of Intelligent Information Systems**, v. 17, n. 2-3, p. 107-145, 2001.
47. HAMMOUDA, Khaled M.; KAMEL, Mohamed S.. Efficient phrase-based document indexing for web document clustering. **Knowledge and Data Engineering, IEEE Transactions on**, v. 16, n. 10, p. 1279-1296, 2004.
48. HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data mining: concepts and techniques**. Morgan kaufmann, 2006.
49. HODGE, Gail. **Systems of knowledge organization for digital libraries**. Digital library federation, council on library and information resources, 2000.
50. HRUSCHKA, Eduardo Raul et al. A survey of evolutionary algorithms for clustering. **Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on**, v. 39, n. 2, p. 133-155, 2009.
51. ISHIOKA, Tsunenori. An expansion of X-Means for automatically determining the optimal number of clusters. In: **Proceedings of International Conference on Computational Intelligence**. 2005. p. 91-96.
52. KOBASHI, N.Y. **A elaboração de informações documentárias: em busca de uma metodologia**. 1994. 195f. Tese (Doutorado em Ciências da Comunicação) – Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo. 1994.

53. KURAMOTO, Hélio. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. **Ciência da informação**, v. 25, n. 2, 1995.
54. LEIVA, Isidoro Gil. **La automatización de la indización de documentos**. Trea, 1999.
55. LEIVA, Isidoro Gil. **Manual de indización: teoría y práctica**. Trea, 2008.
56. LOPES, R. B. **Mineração de Textos com Georeferenciamento**. Rio de Janeiro: UFRJ/COPPE, 2009. Tese de Doutorado.
57. MAIA, Luiz Cláudio Gomes. **Uso de sintagmas nominais na classificação automática de documentos Eletrônicos**. Belo Horizonte, MG, 2008. Tese de Doutorado, Programa de Pós-Graduação em Ciência da Informação da Escola de Ciência da Informação da Universidade Federal de Minas Gerais.
58. MAIA, Luiz Cláudio; SOUZA, Renato Rocha. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectivas em Ciência Informação, Belo Horizonte**, v. 15, p. 154-172, 2010.
59. MAMFRIM, Flávia Pereira Braga. Representação de conteúdo via indexação automática em textos integrais em língua portuguesa. **Ciência da Informação**, v. 20, n. 2, 1991.
60. MARKOV, Zdravko; LAROSE, Daniel T. **Data mining the web. Uncovering Patterns in Web Content, Structure, and Usage**, USA: Jhon Wiley & Sons, p. 3-57, 2007.
61. MCLACHLAN, Geoffrey J.; KRISHNAN, Thriyambakam. **The EM algorithm and extensions**. Wiley-Interscience, 2007.
62. MENDEZ RODRÍGUEZ, E. M., MOREIRO GONZÁLEZ, J. A. Lenguaje natural e indización automatizada. **Ciencias de la Información**, v. 30, n.3, p.11-24, set., 1999.
63. NARUKAWA, Cristina Miyuki; GIL-LEIVA, Isidoro; FUJITA, Mariângela S. L.. Indexação automatizada de artigos de periódicos científicos: análise da aplicação do software SISA com uso da terminologia DeCS na área de Odontologia. **Informação & Sociedade: Estudos**, 2009.
64. NASSIF, Luís Filipe da Cruz. **Técnicas de agrupamento de textos aplicadas à computação forense**. 2011. Dissertação de Mestrado, Engenharia Elétrica, Universidade de Brasília.
65. NOGUEIRA, Tatiane M.; CAMARGO, Heloisa A.; REZENDE, Solange O. **Tratamento de imprecisão e incerteza na identificação de documentos textuais similares**. Universidade de São Paulo. 2009.

66. NUNES, Maria das Graças Volpe; ALUISIO, Sandra M.; PARDO, Thiago AS. Um panorama do Núcleo Interinstitucional de Linguística Computacional às vésperas de sua maioridade. **Linguamática**, v. 2, n. 2, p. 13-27, 2010.
67. ORTEGA, Cristina Dotta. Relações históricas entre biblioteconomia, documentação e ciência da informação. **DataGramZero-Revista de Ciência da Informação**, v. 5, n. 5, 2004.
68. PALAZZO, M. D. O. et al. Descoberta de conhecimento em textos através da análise de seqüências temporais. In: **Workshop em Algoritmos e Aplicações de Mineração de Dados-WAAMD, SBBD: Sociedade Brasileira de Computação**. p. 49-56.
69. PAZZA, R. As mariposas Biston betularia. **Projeto Evoluindo - Biociência.org**, 2004. Disponível em [http://biociencia.org/index.php?option=com\\_content&task=view&id=45&Itemid=83](http://biociencia.org/index.php?option=com_content&task=view&id=45&Itemid=83) >. Acesso em: 27/05/2013.
70. PELLEGG, D.; MOORE, A. X-Means: Extending k-means with efficient estimation of the number of clusters. In: **Proceedings of the seventeenth international conference on machine learning**. 2000. p. 727-734.
71. PIEDADE, Maria Antonieta Requião. **Introdução à teoria da classificação**. Interciência, 1977.
72. PINHEIRO, Marcello Sandi. **Uma Abordagem Usando Sintagmas Nominais Como Descritores no Processo de Mineração de Opiniões**. 2009. Tese de Doutorado. Universidade Federal do Rio de Janeiro.
73. PIRES, Marina Melo. **AGRUPAMENTO INCREMENTAL E HIERÁRQUICO DE DOCUMENTOS**. 2008. Tese de Doutorado. UNIVERSIDADE FEDERAL DO RIO DE JANEIRO.
74. RANGANATHAN, Shiyali Ramamrita. Prolegomena to library classification. **Prolegomena**, 1967.
75. ROBREDO, J. **Documentação de hoje e amanhã: uma abordagem informatizada de Biblioteconomia e dos sistemas de informação**. 2.ed. rev. ampl. Brasília: Edição de Autor, 1994.
76. ROSELL, Magnus; VELUPILLAI, Sumithra. The impact of phrases in document clustering for Swedish. In: **Proceedings of the 15th Nordic Conference on Computational Linguistics, NODALIDA'05**. 2005.
77. SARACEVIC, T. Ciência da Informação: origem, evolução e relações. **Perspectivas em ciência da informação**, v. 1, n. 1, p. 41-62, 1996.
78. SCHIESSL, Marcelo; BRASCHER, Marisa. Descoberta de Conhecimento em Texto aplicada a um Sistema de Atendimento ao Consumidor. **Revista Ibero-Americana de Ciência da Informação**, v. 4, n. 2, 2012.

79. SCRIPTOR, Gerald Skidmore. **Metaheuristics and combinatorial optimization problems**. 2006. Tese de Doutorado. Rochester Institute of Technology.
80. SENO, Eloize Rossi Marques; NUNES, Maria das Graças Volpe. Some experiments on clustering similar sentences of texts in portuguese. In: **Computational Processing of the Portuguese Language**. Springer Berlin Heidelberg, 2008. p. 133-142.
81. SILVA, Augusto Soares. Sociolinguística cognitiva e o estudo da convergência/divergência entre o Português Europeu e o Português Brasileiro. Veredas 10. **Revista de Estudos**. 2006.
82. SILVA, Maria; FUJITA, M. A prática de indexação: análise da evolução de tendências teóricas e metodológicas. **TRANSINFORMAÇÃO**, v. 16, n. 2, 2004.
83. SIQUEIRA, Jéssica Camara. A classificação nos domínios das três Marias; La clasificación en el dominio de las “tres Marías”. **Informação & Informação**, v. 16, n. 1, p. 36-51, 2011.
84. SMYTH, Padhraic. Model selection for probabilistic clustering using cross-validated likelihood. **Statistics and Computing**, v. 10, n. 1, p. 63-72, 1998.
85. SONG, Wei; PARK, Soon Cheol. Genetic algorithm-based text clustering technique: Automatic evolution of clusters with high efficiency. In: **Web-Age Information Management Workshops, 2006. WAIM'06. Seventh International Conference on**. IEEE, 2006. p. 17-17.
86. SOUZA E SILVA, Maria Cecília Perez de; KOCH, Ingedore Grunfeld Villaça. **Linguística aplicada ao português: morfologia**. Cortez, 18ª.Ed, 2011.
87. SOUZA, Alexandre Augusto Angelo de; NEVES JR, Flávio; LOPES, Heitor Silvério. Sistema de avaliação da rede secundária de distribuição utilizando algoritmos genéticos. **Espaço Energia**, n. 5, p. 34-41, 2006.
88. SOUZA, Renato Rocha. Uma proposta de metodologia para indexação automática utilizando sintagmas nominais. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**. n.1, p. 42-59, 2006.
89. SOUSA, Renato Tarciso Barbosa de. Classificação de documentos arquivísticos: trajetória de um conceito. **Arquivística.net**, Rio de Janeiro, v. 2, n. 2, p. 120-142, 2006.
90. SVENONIUS, Elaine. **The intellectual foundation of information organization**. MIT press, 2000.
91. TRISTÃO, Ana Maria Delazari et al. Sistema de classificação facetada e tesouros: instrumentos para organização do conhecimento. **Ciência da Informação**, Brasília, v. 33, n. 2, p. 161-171, 2004.



92. TUFANO, D.. **Estudos de língua portuguesa : gramática.** 2 ed. São Paulo: Moderna,
93. VATTANI, Andrea. K-means requires exponentially many iterations even in the plane. **Discrete & Computational Geometry**, v. 45, n. 4, p. 596-616, 2011.
94. VIEIRA, Simone Bastos. Indexação automática e manual: revisão de literatura. **Ciência da informação**, v. 17, n. 1, 1988.
95. WARD, Martin L. The future of the human indexer. **Journal of librarianship and information science**, v. 28, n. 4, p. 217-225, 1996.
96. WEISS, Dawid; STEFANOWSKI, Jerzy. Web search results clustering in Polish: Experimental evaluation of Carrot. **Proceedings of Intelligent Information Processing and Web Mining**, p. 209-218, 2003.
97. WITTY, Francis J. The beginnings of indexing and abstracting: some notes towards a history of indexing and abstracting in Antiquity and the Middle Ages. **The indexer**, v. 8, n. 4, p. 193-198, 1973.

## APÊNDICE A: LISTA DE PERIÓDICOS UTILIZADOS NOS CORPORA DA PESQUISA

### Educação Física

Revista Mackenzie de Educação Física e Esporte

<<http://editorarevistas.mackenzie.br/index.php/remef>>

Revista Motriz de Educação Física

<<http://www.periodicos.rc.biblioteca.unesp.br/index.php/motriz/index>>

Pensar a Prática

<<http://www.revistas.ufg.br/index.php/fef>>

### Farmácia

Revista Brasileira de Farmácia

<<http://www.rbfarma.org.br/>>

Revista de Ciências Farmacêuticas Básica e Aplicada

<[http://serv-bib.fcfar.unesp.br/seer/index.php/Cien\\_Farm](http://serv-bib.fcfar.unesp.br/seer/index.php/Cien_Farm)>

Revista Eletrônica de Farmácia

<<http://www.revistas.ufg.br/index.php/REF>>

### Geografia

GEOUSP - Espaço e Tempo

<<http://citrus.uspnet.usp.br/geousp/ojs-2.2.4/index.php/geousp/index>>

Caderno de Geografia

<<http://periodicos.pucminas.br/index.php/geografia/index>>

Boletim Goiano de Geografia

<<http://www.revistas.ufg.br/index.php/bgg>>

**História**

História Social

<<http://www.ifch.unicamp.br/ojs/index.php/rhs/index>>

Cadernos de História

<<http://www.ichs.ufop.br/cadernosdehistoria/ojs/index.php/cadernosdehistoria/index>>

História Revista

<<http://www.revistas.ufg.br/index.php/historia>>

**Linguística**

Fórum Linguístico

<<http://www.periodicos.ufsc.br/index.php/forum/index>>

Percursos Linguísticos

<<http://periodicos.ufes.br/percursos/index>>

Signótica

<<http://www.revistas.ufg.br/index.php/sig/index>>

**Odontologia**

Revista da Faculdade de Odontologia de Porto Alegre

<<http://seer.ufrgs.br/RevistadaFaculdadeOdontologia/index>>

Revista da Faculdade de Odontologia da Universidade de Passo Fundo

<<http://www.upf.br/seer/index.php/rfo/index>>

Revista de Odontologia da Universidade Cidade de São Paulo

<[http://www.cidadesp.edu.br/old/revista\\_odontologia/index.htm](http://www.cidadesp.edu.br/old/revista_odontologia/index.htm)>

## APÊNDICE B: DESCRIÇÃO DE OPÇÕES PARA OS ALGORITMOS *EM* E *X-MEANS* NO *WEKA*

Opções para o Algoritmo *EM*:

### OPTIONS

**debug** -- If set to true, clusterer may output additional info to the console.

**displayModelInOldFormat** -- Use old format for model output. The old format is better when there are many clusters. The new format is better when there are fewer clusters and many attributes.

**maxIterations** -- Maximum number of iterations.

**minStdDev** -- Set minimum allowable standard deviation.

**numClusters** -- Set number of clusters. -1 to select number of clusters automatically by cross validation.

**seed** -- The random number seed to be used.

Opções para o Algoritmo *X-Means*:

### OPTIONS

**binValue** -- Set the value that represents true in the new attributes.

**cutOffFactor** -- The cut-off factor to use.

**debugLevel** -- The debug level to use.

**debugVectorsFile** -- The file containing the debug vectors (only for debugging!).

**distanceF** -- The distance function to use.

**inputCenterFile** -- The file to read the list of centers from.

**maxIterations** -- The maximum number of iterations to perform.

**maxKMeans** -- The maximum number of iterations to perform in KMeans.

**maxKMeansForChildren** -- The maximum number of iterations KMeans that is performed on the child centers.

**maxNumClusters** -- Set maximum number of clusters.

**minNumClusters** -- Set minimum number of clusters.

**outputCenterFile** -- The file to write the list of centers to.

**seed** -- The random number seed to be used.

**useKDTree** -- Whether to use the KDTree.

## **APÊNDICE C: LISTA DE PUBLICAÇÕES REALIZADAS DURANTE A PESQUISA**

1. AFONSO, A. R.; DUQUE, C. G. O Impacto da variação temática na categorização automática de artigos científicos em português do Brasil. **Anais Digitais do ENANCIB**, 2012.