UNIVERSIDADE DE BRASÍLIA
Faculdade de Ciência da Informação
Curso de Graduação em Biblioteconomia

Princípios de classificação automática de documentos eletrônicos

Joaquim Morais Rocha Junior
Orientadora: Profa. Dra. Rita de Cássia do Vale Caribé

Brasília
2017

Joaquim Morais Rocha Júnior

Princípios de classificação automática de documentos eletrônicos

Monografia apresentada como parte das exigências para obtenção do título de Bacharel em Biblioteconomia pela Faculdade de Ciência da Informação da Universidade de Brasília

Orientadora: Profa. Dra. Rita de Cássia do Vale Caribé

Brasília

2017

ROCHA JUNIOR, Joaquim Morais.

Princípios de classificação automática de documentos eletrônicos / Joaquim Morais Rocha Junior. – Brasília, 2017.

64 f.

Orientação: Profa. Dra. Rita de Cássia do Vale Caribé

Monografia (Bacharelado em Biblioteconomia) – Universidade de Brasília, Faculdade de Ciência da Informação, Curso de Biblioteconomia, 2017.

Inclui bibliografia

- Classificação automática de textos. 2. Categorização automática de textos. 3. Agrupamento automático de textos. I. Título.

CDU 025.4

Titulo: Princípios de classificação automática de documentos eletrônicos.

Aluno: Joaquim Morais Rocha Júnior .

Monografia apresentada à Faculdade de Ciência da Informação da Universidade de Brasília, como parte dos requisitos para obtenção do grau de Bacharel em Biblioteconomia.

Brasília, 24 de agosto de 2017.

Rita de Cássia do Vale Caribé - Orientadora
Professora da Faculdade de Ciência da Informação (UnB)
Doutora em Ciência da Informação

Simone Bastos Vieira – Membro
Professora da Faculdade de Ciência da Informação (UnB)
Doutora em Ciência da Informação

Marcílio de Brito – Membro
Professor da Faculdade de Ciência da Informação (UnB)
Doutor em Ciências da Informação e da Comunicação

Epigrafe

O conhecimento torna a alma jovem e diminui a amargura da velhice. Colhe, pois, a sabedoria. Armazena suavidade para o amanhã.

Leonardo da Vinci

RESUMO

Revisão de literatura sobre o tema de classificação automática de documentos eletrônicos. A proposta deste trabalho é apresentar e descrever os procedimentos utilizados para a realização de classificação automática de documentos eletrônicos textuais. Na revisão de literatura se contextualiza classificação, classes, taxonomia, categorias, sistemas de classificação e os procedimentos utilizados para a classificação automática de documentos. Dentre os procedimentos descritos se destacam o pré-processamento com a "limpeza" dos textos removendo termos/caracteres não necessários; a fase de indexação os documentos são transformados em listas de termos representativos; na fase de seleção de atributos dos documentos os termos mais representativos são selecionados; na fase de classificação/categorização os documentos são alocados/agregados nas classes que mais os representam. Apresenta sugestões para pesquisas utilizando os sistemas de classificação bibliográficos, ontologias, tesauros e taxonomias na classificação/categorização/agrupamento automática.

Palavras-chave: classificação automática, categorização automática, agrupamentos automáticos, classificadores, representação de conteúdo, seleção de atributos, pré-processamento de texto.

ABSTRACT

Review of literature on automatic classification of electronic documents. The purpose of this paper is to present and describe the procedures used to perform automatic classification of textual electronic documents. In the literature review we contextualize classification, classes, taxonomy, categories, classification systems and the procedures used for automatic documents classification. Among the procedures described, the preprocessing with the "cleaning" of the texts, removing unnecessary terms / characters; The indexing phase documents are transformed into lists of representative terms; In the feature selection phase the most representative terms are selected; In the classification / categorization phase the documents are allocated / aggregated in the classes that best represent them. It presents suggestions for research using the bibliographic classification systems, ontologies, thesauri and taxonomies in the classification / categorization / automatic grouping.

Keywords: Automatic categorization, automatic classification, automatic clustering, classifiers, content representation, feature selection, text preprocessing.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

BCE	Biblioteca Central – UNB
CAD	Classificação automática de documentos
DCS	Dynamic Classifier Selection
DF	Document Frequency
DR	Dimensionality Reduction
DT	Decision tree
FS	Feature Selection
IDF	Inverse Document Frequency
IG	Information Gain
IR	Information Retrieval
kNN	K-Nearest Neighbor
LLSF	Linear Least Squares Fit
LSI	Latent semantic indexing
MI	Mutual Information
ML	Machine Learning
MV	Majority Voting
NB	Naive Bayes
NLP	Natural Language Processing
NN	Neural Network
Nnet	Neural Network
PDF	Portable Document Format
PLN	Processamento da linguagem natural
QP	Quadratic Programming
RI	Recuperação da Informação
SVN	Suport Vector Machines
TC	Text Categorization
TS	Term Strength
UNB	Universidade de Brasília

SUMÁRIO

1 INTRODUÇÃO

Desde o advento dos computadores e sua utilização cada vez mais massiva na automação de atividades humanas, tem-se buscado soluções para o processamento técnico na área da Biblioteconomia.

Na presente era, a produção de documentos textuais alcançou tamanho volume que se tornou impossível para que humanos realizem procedimentos que promovam a utilização de todos estes documentos para atender a suas necessidades informativas.

O desenvolvimento tecnológico vem acompanhado de novos campos de estudo, com vertentes de pesquisas em várias áreas do conhecimento se debruçando em busca de soluções para os recentes desafios. O que também causa um crescente volume de informações disponíveis.

Para atender estas necessidades informacionais têm surgido técnicas de avaliação automática de documentos, uma delas é o agrupamento de documentos segundo seus valores intrínsecos, dado pelos assuntos que os documentos abordam. Uma destas técnicas é a classificação automática de documentos pelo seu assunto, um campo de estudo multidisciplinar, que envolve diretamente pesquisas nas áreas da Ciência da Informação, Tecnologia da Informação, Linguística, Estatística dentre outras.

Classificar permite obter respostas que ajudam no entendimento dos conhecimentos que têm surgido recentemente, por meio da identificação das semelhanças e diferenças que esses possuem, possibilitando o seu aproveitamento futuro. A divisão do conhecimento segundo as semelhanças e diferenças, pode ser feita por categorias que apresentam as partições iniciais nas quais o conhecimento é dividido, essas partições lógicas do conhecimento são denominadas classes, que posteriormente serão divididas em subclasses. Ao processo de classificar denomina-se classificação, termo esse que também é utilizado para nomear aos sistemas de classificação que se empregam em bibliotecas e demais sistemas de informação.

O ato de classificar é intencional, tem como pretensão organizar um conjunto de objetos de acordo com suas características comuns. Quando os objetos destinados a classificar são documentos textuais a classificação é denominada de bibliográfica, pois até recentemente era atribuída somente a documentos impressos normalmente livros em bibliotecas tradicionais.

A classificação bibliográfica pressupõe a existência de regras e de um esquema que possibilitem sua utilização por qualquer pessoa, que podem ser especialistas a serviço do sistema de informação. Podem, também, ser usuários em busca de informação, tendo em vista que uma das principais funções da classificação bibliográfica é alocar os livros nas estantes e posteriormente que sejam encontrados.

A classificação bibliográfica se faz presente desde as primeiras bibliotecas da antiguidade até os mais modernos sistemas de informação que se conhece atualmente. A formalização da classificação bibliográfica se deu com o desenvolvimento dos sistemas de classificação na segunda metade do século XIX. A classificação bibliográfica tem como constante desafio manter-se atualizada devido ao aparecimento e transformação constante das áreas do conhecimento, como também pelo crescimento vertiginoso do número de obras disponíveis.

Diante destas dificuldades, parece plausível que se busquem procedimentos que automatizem as classificações textuais, possibilitando que os documentos sejam mais rapidamente realocados em novas classes, que surgem com o desenvolvimento do conhecimento humano, ou para atender a necessidade crescente de informação e que seja realizada de forma quase que instantânea.

Dos primeiros experimentos com representação de conteúdo de documentos, que foram realizados na década de 1950, com computadores de pequena capacidade de processamento, até os atuais supercomputadores, com mais regularidade e intensidade têm-se realizado estudos para a utilização de procedimentos automáticos de classificação, supervisionados por humanos ou semi-supervisionados para a classificação de documentos textuais.

As técnicas mais atuais promovem o uso de metodologias conhecidas como aprendizagem de máquinas, a despeito de as máquinas ainda não possuírem capacidade de aprendizagem como comumente se entende. A abordagem para classificação automática com uso de aprendizagem de máquinas possui diversas vertentes, algumas tentam imitar a natureza como a biomimética e redes neurais, já outras utilizam processos estudados por outras áreas do conhecimento como a mineração de textos, da ciência da computação, que extrai informações através de padrões.

Para melhor aproveitamento das metodologias utilizadas, o texto escrito deve passar por algumas etapas antes que seja processado com fins de classificação. O texto deve ser “limpo” de todo tipo de palavras sem valor informacional. O texto é ajustado para que possa ser “contabilizado” por meio de programas de computador, passando a ter um conteúdo

formado apenas com palavras que deem sentido ao texto, só então será disponibilizado para que se realize novo procedimento que o habilite para a classificação automática baseada no seu conteúdo.

O texto, então, tem seus termos ou palavras contados e dispostos em uma forma de vetor de termos, com o valor de cada entrada representando a quantidade de termos ou um peso que foi calculado em função da frequência que ocorre tanto no documento quanto no conjunto de documentos. Depois de transformado, o documento será representado por este vetor com valores numéricos.

Os vetores com valores numéricos são utilizados nos procedimentos de classificação automática que os comparam com outros vetores que representam as classes ou categorias disponíveis e são classificados de acordo com a classe que possuem mais características em comum.

Os procedimentos que verificam estas características dos documentos são chamados de classificadores, eles utilizam métodos estatísticos para calcular as características em comum entre os documentos e as classes.

Neste sentido, a proposta deste trabalho é apresentar alguns procedimentos que têm estudado sobre classificação automática de documentos, tentando apresentar o estado da arte neste assunto. Tarefa um pouco dificultada pelo grande número de pesquisas e trabalhos sobre o tema e pelo constante avanço em novas abordagens para o tratamento de textos eletrônicos.

2 CONSTRUINDO O OBJETO DE ESTUDO E O REFERENCIAL TEÓRICO

2.1 DEFINIÇÃO DO PROBLEMA E JUSTIFICATIVA

A classificação pode ser considerada um processo natural na medida em que está inserida no cotidiano das pessoas. Os seres humanos classificam em todo momento, na medida em que separam coisas em grupos, de acordo com as características em comum, separando-as daquelas que possuem características diferentes, ou seja, agrupa pelas semelhanças e separa pelas diferenças. Discorrendo sobre a classificação na vida cotidiana, Pombo (2002, p. 1) alerta que:

> [...] na verdade nada nos parece mais "natural", óbvio e indiscutível que as classificações dos entes, dos factos e dos acontecimentos que constituem os quadros mentais em que estamos inseridos. Elas constituem os pontos estáveis que nos impedem de rodopiar sem solo, perdidos no inconforto do inominável, da ausência de "idades" ou "geografias". Só elas nos permitem orientar-nos no mundo à nossa volta, estabelecer hábitos, semelhanças e diferenças, reconhecer os lugares, os espaços, os seres, os acontecimentos; ordená-los, agrupá-los, aproximá-los uns dos outros, mantê-los em conjunto ou afastá-los irremediavelmente.

Classificação é um termo polissêmico, porém, simplificadamente, pode ser definida como a identificação de entidades e a sua reunião por características semelhantes e separação das diferentes e, posteriormente, a organização dessas entidades. (ARANALDE, 2009, p. 87)

A organização do conhecimento tem sido uma preocupação humana, desde que se passou a registrá-la para a posteridade. Classificar, segundo Barbosa (1969, p. 13) é a tarefa mais importante de uma biblioteca, pois permite a utilização dos documentos disponíveis.

Desde a década de 1960, pesquisadores têm se dedicado ao tema de classificação automática na busca por métodos, processos ou técnicas que viabilizem a automatização/mecanização do tratamento da informação uma vez que as técnicas manuais adotadas para classificação não apresentam grande produtividade e foram desenvolvidas para facilitar a guarda e localização de livros nas estantes em bibliotecas. Borko (1967, p. 5, tradução nossa) alertava que uma das vantagens do processamento automático da linguagem é a facilitação das comunicações, traduções, armazenamento e recuperação das informações.

Nos últimos vinte anos, a sociedade tem sido submetida a uma transformação advinda da avalanche de informação disponível, principalmente em meio digital. O volume crescente de documentos traz o desafio de como lidar com eles, isto tem afetado todas as áreas

do conhecimento humano que tratam de informação, em especial as que fazem tratamento para uso posterior.

Como identificar estes documentos, dar acesso aos mesmos, torná-los públicos, agregar com outros de assunto similar, são alguns dos desafios que têm superado a capacidade humana para lidar com eles. Diante disto, verifica-se ser necessária a mecanização do tratamento técnico destes documentos.

O tema desta pesquisa é o funcionamento da classificação automática de documentos eletrônicos nos dias atuais. A questão que se coloca então é: como se dá a classificação automática de documentos eletrônicos?

2.2 OBJETIVOS DA PESQUISA

2.2.1 Objetivo geral

Identificar, na literatura, mecanismos que permitam realizar operações de classificação automática de documentos eletrônicos visando sua posterior recuperação.

2.2.2 Objetivos específicos

- Caracterizar o que seja classificação automática;
- Identificar as etapas para a realização da classificação automática;
- Apresentar procedimentos de classificação automática;
- Descrever as metodologias existentes e utilizadas para classificação/categorização automática;

3 PROCEDIMENTOS METODOLÓGICOS

A metodologia proporcionará um caminho a ser seguido durante o desenvolvimento do trabalho, e o que se espera encontrar ao final, conforme Marconi e Lakatos (2010, p. 65) descrevem sobre o método: que consiste em um conjunto de atividades sistemáticas e racionais que, com maior segurança e economia, permite alcançar o objetivo – conhecimentos válidos e verdadeiros.

Este trabalho se baseia em pesquisa exploratória, de natureza documental, alicerçado na literatura disponível sobre classificação automática de documentos. Consiste em revisão de literatura, por meio de pesquisa bibliográfica acerca do tema e a identificação do estado da arte relativo ao assunto abordado, buscando uma visão geral do objeto de estudo.

Segundo Gil (2012, p. 27), pesquisas classificadas como exploratórias (classificação dada em função de seus objetivos) tem como finalidade desenvolver, esclarecer, e modificar conceitos e ideias, tendo em vista a formulação de problemas mais precisos [...] envolvem levantamento bibliográfico e documental. São desenvolvidas com o objetivo de proporcionar visão geral, de tipo aproximativo, acerca de determinado fato. O produto final será um problema mais esclarecido. No que concordam Tripodi et al. (1975, p. 42-71 apud MARCONI; LAKATOS, 2010, p. 171) quanto alegam que:

> Obtêm-se frequentemente descrições tanto quantitativas quanto qualitativas do objeto de estudo, e o investigador deve conceituar as inter-relações entre as propriedades do fenômeno, fato ou ambiente observado.

Revisão de literatura é definida como a busca de conhecimento acumulado em documentos produzidos na comunidade científica sobre determinado assunto. Conforme relatam Prodanov e De Freitas (2013, p. 131) a

> [...] revisão de literatura demonstra que o pesquisador está atualizado nas últimas discussões no campo de conhecimento em investigação. Além de artigos em periódicos nacionais e internacionais e livros já publicados, as monografias, dissertações e teses constituem excelentes fontes de consulta.

Pesquisa bibliográfica se dá a partir da coleta de informação extraída de materiais já elaborados como livros e artigos científicos. Consiste em um conjunto de procedimentos para identificar soluções de um problema com atenção ao objeto de estudo. São muito utilizadas em trabalhos de caráter exploratório-descritivo (LIMA; MIOTO, 2009, p. 38-39). Já para Marconi e Lakatos (2010, p. 166) a pesquisa bibliográfica, além de abranger materiais impressos, considera como meios para obtenção de informação: rádio, gravações, filmes e

televisão, portanto tudo que foi escrito, dito ou filmado sobre o assunto. Desta forma, alargando o conceito de simples documentos impressos.

Neste estudo a pesquisa bibliográfica se limitou a materiais produzidos (impresso ou eletrônicos) sobre o tema em tela, mais precisamente em livros, teses e artigos acadêmicos.

4 REVISÃO DE LITERATURA

A revisão de literatura visa reportar e avaliar o conhecimento produzido em pesquisas prévias onde se destacam conceitos, procedimentos e resultados para o trabalho a ser produzido (PRODANOV, 2013, p 79).

Inicialmente, será apresentada a contextualização das várias áreas envolvidas com classificação automática, em seguida, será construído o estado da arte da classificação automática de documentos eletrônicos.

Para realização deste trabalho na busca de material, foi utilizada a pesquisa em bases de dados nacionais, como Brapci, Scielo Brasil, Repositório Institucional da Universidade de Brasília, o acervo físico da Biblioteca Central (BCE) da UNB. Na busca internacional, utilizou-se a Library and Information Science Abstracts (LISA), Directory of Open Access Journals (DOAJ), JStror, Ebray (Proquest) e Google Schoolar. As expressões de busca foram "classificação automática", "agrupamento automático", "classificação", “categorização automática”, “algoritmos de classificação” e suas respectivas traduções para o inglês. Também foram utilizadas as referências de alguns trabalhos para acesso direto às obras citadas.

4.1 Porque classificar

A Classificação é a ferramenta que permite encontrar a resposta mais adequada aos questionamentos. Vickery (1980, p. 25) expõe que a informação científica enfrenta um problema entre a produção de literatura e os usuários em busca de repostas as suas indagações, sendo justamente na seleção de respostas que entra a Classificação, para permitir o rápido acesso às informações.

Para Barbosa (1969, p. 13) foi a “necessidade de reunir os conhecimentos humanos numa ordem lógica que levou os filósofos ao estabelecimento de grandes agrupamentos”.

4.2 Categoria

Alguns autores consideraram a categoria como um sinônimo para classe. Para outros este conceito é simplista, pois categoria envolve um conceito mais amplo. Conforme Aranalde (2009, p. 89), categoria consiste em condição de possibilidade para certos juízos

básicos que se emite sobre o mundo na tentativa de interpretá-lo e compreendê-lo. Descrevendo o mundo e as coisas que o compõem, sendo pré-requisito para a elaboração de classes mais gerais.

> Ora, se classificar pressupõe categorizar, as categorias enquanto elementos imprescindíveis para a classificação são as maneiras como se pode identificar e falar das coisas, possibilitando a elaboração de classes mais gerais em que são ordenados os seus predicados. (ARANALDE, 2009, p. 89)

Vickery (1980, p. 235) definiu resumidamente que as categorias são "conceitos de alta generalidade e ampla aplicação empregados na interpretação do mundo". De modo mais claro, Piedade (1983, p. 19) considera categoria como "as maiores classes dos fenômenos, as classes mais gerais que podem ser formadas".

Dessa forma, as categorias são o ponto de partida para a classificação, as classes primordiais às quais os sistemas de classificação se submetem. Possibilitando a classificação, pois permitem a identificação das classes básicas que servirão de suporte para o início da divisão das ideias ou objetos em grupos.

4.3 Categorização

Categorização é um processo mental de entendimento da realidade criando as categorias que fazem um recorte desta realidade. Como discorrem Campos e Gomes (2012, p. 5).

> A Categorização é um processo que requer pensar o domínio de forma dedutiva, ou seja, determinar as classes de maior abrangência dentro da temática escolhida. Na verdade, aplicar a categorização é analisar o domínio a partir de recortes conceituais que permitem determinar a identidade dos conceitos (categorias) que fazem parte deste domínio.

Jacob (2004, p. 522 apud MAI, 2010, p. 712, tradução nossa) argumenta que a categorização é o processo que envolve entidades nomeadas no mundo e o processo de agrupá-las em categorias. Este é um processo cognitivo e é feito implicitamente e sem uma estrutura articulada.

A categorização não impõe a necessidade de categorias pré-definidas para o ajuntamento de objetos ou ideias, as classes surgem à medida que o processo de categorização decorre.

4.4 Classe

Para entendimento do que seja classificação faz-se necessário, anteriormente, entender o que seja classe. Para Barbosa (1969, p. 23), classe é um agrupamento de assuntos que possuem alguma semelhança.

Ampliando este conceito, Tristão, Fachin e Alarcon (2004, p. 163) definem: "Uma classe consiste de um número de elementos quaisquer (objetos e ideias) que possuem alguma característica em comum e, ao mesmo tempo, constitui sua própria unidade". No que é próximo ao entendimento de Piedade (1983, p. 18) que considera classe como um conjunto de coisas ou ideias que possuem um ou vários atributos, predicados ou qualidades em comum.

Em outro texto, Barbosa (1972, p. 75) conceitua classe e subclasse, que seria a divisão da classe, da seguinte forma:

> Nos estudos da teoria de classificação, classe e subclasse são termos completamente relativos. Sabemos que classe é um conjunto de coisas que apresentam algo em comum, se a esta classe aplicarmos uma diferença obtemos como resultado subclasses. [...] Esta forma de segmentação das classes, produzida pelas divisões, apresenta, como resultado, um sistema hierárquico onde subclasse se subordina a classe mais genérica.

Classe é um agrupamento de entidades que estão reunidas, neste grupo, por alguma característica em comum. Se aplicado a este grupo algum critério de diferenciação dos objetos ou ideias, o resultado será a divisão em um ou mais grupos que serão as subclasses da classe anterior. Quando se esquematiza estas divisões o resultado aparecerá em forma de árvore invertida, ou em uma hierarquia de classes, sendo as raízes as mais genéricas enquanto as folhas são as mais específicas.

4.5 Classificação

O termo classificação possui diversos significados, serão tratados aqui aqueles que são pertinentes ao escopo do trabalho, no qual o conceito está relacionado com o agrupamento de entidades por suas semelhanças e dissociação por suas diferenças, colocando-as em classes e subclasses. Em sintonia com Vickery (1980, p. 23) que define classificação como "reunir coisas ou ideias que sejam semelhantes entre si, e separar as que apresentam diferenças". Também Barbosa (1969, p. 13) apresenta uma definição de classificação no campo da lógica que consiste em: "Um processo mental pelo qual coisas, seres ou

pensamentos são reunidos segundo as semelhanças ou diferenças que apresentam". Esta seria a definição mais simples de classificação.

Já Tristão, Fachin e Alarcon (2004, p. 163) apresentam o seguinte conceito de classificação, acrescentando a ideia denominada de unidade, que consiste em agrupamento pelas características comuns dentro da classe, dando coesão ao que está classificado.

> Classificação significa a ação e efeito de classificar, e classificar significa ordenar e dispor em classes. Uma classe consiste de um número de elementos quaisquer (objetos e ideias) que possuem alguma característica comum pela qual devem ser diferenciados de outros elementos e, ao mesmo tempo, constitui sua própria unidade. (TRISTÃO; FACHIN; ALARCON, 2004, p. 163).

Para Ranganathan (1967, p. 77-79) classificação tem cinco sentidos que são apresentados em um crescente de complexidade iniciando como divisão, ordenação. Em sequência, na qual cada definição engloba a anterior até o quinto sentido que descreve a classificação no sentido que é usada quando o universo (de elementos) a ser classificado é infinito ou quando as entidades são desconhecidas ou incompreensíveis, mas compõe um universo finito, a classificação no quinto sentido é a que é usada pelos bibliotecários. A classificação no quinto sentido deve possuir um esquema de classificação associado, para que se classifique algo deste universo é necessário, então, um esquema de classificação que seja aplicável a este universo.

Mais recentemente, Jacob (2004, p. 527 apud MAI, 2010, p. 712) expõe que a classificação é um ato deliberado para organizar um conjunto de entidades; um conjunto de regras é, portanto, criado para determinar quando uma entidade entra em uma determinada classe.

Quando o foco da classificação está na recuperação de informação a noção de classificação é menos rígida, sendo cada classe definida simplesmente como um conjunto de termos que denotam certa área de assunto, neste conjunto os relacionamentos entre os documentos não são importantes e a atribuição de classes se dá para a identificação de conteúdo. (SALTON, 1975, p. 54).

Araújo (2006, p. 117) sintetiza o que seja o sentido da definição do termo classificação quando diz que:

> Essa definição, embora possa variar um pouco de acordo com o autor, traz o elemento essencial que caracteriza um processo de classificação: a formação metódica e sistemática de grupos, a ação organizante de ordenar um determinado conjunto de seres ou coisas em agrupamentos menores, a partir de características semelhantes partilhadas por alguns (que os incluem dentro de determinado grupo) e não compartilhada pelos demais (que não pertencem a esse grupo). Nesse processo,

> elege-se um critério de divisão, promovem-se distinções e aproximações, estatutos e avaliações.

A classificação é um processo realizado quando os objetos e ideias são agregados por suas características em comum, depois de analisado o conjunto total destes entes disponíveis para a classificação, utilizando regras e critérios já definidos para tal.

4.6 Taxonomias

O termo taxonomia surgiu com Carolus Linnaes, no século XVIII, quando este classificou o campo da biologia, hierarquizando e dividindo em domínios, reinos, filos, classes, ordens, gêneros e espécies. (NOVO, 2010, p. 135), no que hoje é conhecida com a classificação dos seres vivos.

Campos e Gomes (2012, p. 3) esclarecem que taxonomia é, por definição, classificação sistemática, e as taxonomias atualmente são estruturas classificatórias que têm por finalidade servir de instrumento para a organização e recuperação de informação.

Novo (2010, p. 135), apesar de concordar com esta afirmação, sugere que o aspecto mais importante das taxonomias é o de classificar,

> No âmbito dos estudos da Ciência da Informação, as taxonomias estão sendo vistas como ferramentas de RI, além de enfatizar o aspecto navegacional. Preferimos evidenciar seu aspecto de ferramenta classificatória de um dado domínio.

As taxonomias podem ser usadas como instrumento para classificar domínios que apresentam uma hierarquia muito forte entre os termos, quando o domínio a ser classificado é interdisciplinar surge a dificuldade no posicionamento da entidade classificada na hierarquia.

4.7 Classificação bibliográfica

Classificação bibliográfica está relacionada a documentos impressos, mais precisamente a livros, posteriormente o conceito foi aplicado a documentos digitais. Para Barbosa (1969, p. 16), a classificação bibliográfica pode ser definida como o processo de reunir os livros em grupos segundo os assuntos que abrangem, enquadrando-os em um sistema pré-estabelecido dando-lhes uma localização dentro da coleção. A classificação bibliográfica se dá então pelo conteúdo dos livros e não por características externas. Como explica Vickery (1980, p. 30), “a característica mais importante da classificação dos documentos é que ela se refere ao assunto”.

Tálamo, Lara e Kobashi (1995, p. 54) consideram que a classificação bibliográfica,

> [...] serve de base à organização de documentos, estabelecendo relações entre eles para facilitar sua localização. A classificação bibliográfica supõe um tratamento dos assuntos dos documentos de modo a:
> a) ordenar fisicamente os documentos nas estantes das bibliotecas;
> b) ordenas as referências nas bibliografias ou nas fichas dos catálogos das bibliotecas.

Já Vickery (1980, p. 35) sugere que, por ser a classificação um instrumento para localizar documentos, há a necessidade de fornecer uma notação taquigráfica que facilite o arquivamento e a localização. Esta notação é uma linguagem artificial que padroniza a descrição dos assuntos e propicia assim sua localização.

Tristão, Fachin e Alarcon (2004, p. 162) possuem o mesmo entendimento quando dizem que:

> Assim, os sistemas de classificação e os tesauros 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.

Descrevendo o procedimento de classificação por assunto, Vickery (1980, p. 34) alerta que "não é comum em classificação documentária que se possa restringir a uma única área homogênea de assunto" os documentos são na sua maioria de assuntos interdisciplinares.

A classificação por assuntos apresenta a dificuldade de se usar esquemas tradicionais de classificação, assim "é muito difícil acomodar assuntos interdisciplinares num esquema de classificação convencional, do mesmo modo que é difícil levar em conta as mudanças nas relações entre assuntos existentes" (FOSKETT, 1973, p. 113).

Enquanto a classificação ajuda a entender o mundo, a classificação bibliográfica ajuda na recuperação de documentos que são agregados por assunto.

4.8 Sistemas de Classificação

Sistema de classificação foi definido por Borko (1967, p. 114, tradução nossa) como "um esquema para organizar uma massa de material em grupos, de modo que os objetos relacionados sejam reunidos de forma sistemática".

Tristão, Fachin e Alarcon (2004, p. 164) ampliam o conceito subdividindo-o em três.

> Geralmente, os sistemas de classificação da informação consistem de três partes: um esquema de classificação que organiza nomes sistematicamente de acordo com suas similaridades; uma notação da classificação que substitui itens no esquema de classificação; um índice para tornar fácil para o usuário pesquisar a informação.

Esquema de classificação é a representação gráfica da estrutura de classificação atribuída ao universo classificável, com classes, divisões, subdivisões, seções etc. sucessivamente (BARBOSA, 1969, p. 23).

É pelo esquema de classificação que se pode atribuir um assunto a uma classe, pois existe uma normatização de como proceder esta classificação, para Foskett (1973, p. 76) este esquema representa uma autoridade.

> Essa autoridade que arrola os assuntos sistematicamente indicando as relações entre si, e conhecida geralmente como um esquema de classificação e consiste de quatro partes: tabelas, onde os assuntos são relacionados de forma sistemática; notação, que nos permite utilizar um arranjo que não é mais evidente por si mesmo; índice alfabético, que nos permite encontrar um determinado tópico sem ter de compulsar todas as tabelas; e uma instituição que o mantenha atualizado.

A necessidade de uma instituição sugere que o sistema de classificação possui um caráter contextual. Bem em sintonia com Vickery (1980, p. 187) que, discorrendo sobre a temporalidade dos sistemas de classificação, afirma que cada nova época exige uma nova classificação.

Então, os sistemas de classificação proporcionam as regras, a forma e o modo de arranjar os objetos em sua localização, sempre em sintonia com o momento em que este sistema de classificação foi produzido e utilizado.

4.9 Sistemas de classificação bibliográfica

Os sistemas de classificação vêm acompanhando a evolução das ciências e da sociedade ao longo dos séculos, com o desenvolvimento das bibliotecas, dos suportes informacionais e o crescimento do volume de documentos os sistemas da informação foram se especializando para atender à demanda dos usuários em busca de informação.

Para Campos (1978, p. 2), as funções dos sistemas de classificação bibliográficas são: servir de instrumento para a distribuição útil dos livros ou documentos nas estantes; organização dos instrumentos de recuperação da informação (catálogos, bibliografias, linguagens de indexação para o computador, etc.); e análise da informação.

Ampliando o entendimento de sistema de classificação bibliográfica, e baseando-se nos conceitos que propiciam a classificação e as relações entre estes conceitos, Araújo

(2006, p. 122) descreve que existe uma teoria subjacente aos sistemas de classificação bibliográfica.

> Independentemente dos seus tipos ou distinções, pode-se afirmar que todas as teorias da classificação bibliográfica buscam promover uma classificação sistemática, lógica, que reflita crítica e filosoficamente sobre os elementos de ligação que servem para a reunião de conceitos.

Fazendo uma retrospectiva sobre os sistemas de classificação, Tálamo, Lara e Kobashi (1995, p. 54) afirmam que:

> A Classificação Decimal de Dewey (CDD) foi o primeiro sistema de classificação bibliográfica utilizado de maneira sistemática. Define-se como sistema de classificação geral porque apresenta a ordenação de todo o conhecimento humano. Qualifica-se como bibliográfica porque, ao contrario dos sistemas de classificação filosóficos que se preocupam com a hierarquização do conhecimento e com a ordem da ciência e das coisas, serve de base à organização de documentos, estabelecendo relações entre eles, para facilitar sua localização.

Concordante com esta afirmação sobre a primazia da CDD como sistema de classificação bibliográfica, Araújo (2009, p. 197) considera que os primeiros sistemas de classificação bibliográfica surgiram no século XIX, bem como os sistemas de Dewey e a Classificação Decimal Universal (CDU), decorrentes dos sistemas de classificação das ciências estudados pela filosofia.

Analisando a viabilidade dos sistemas atuais de classificação bibliográfica Campos (1978, p. 1-2) considera que estes faliram por deficiência em acompanhar a evolução e crescimento do conhecimento, dada a sua rigidez que não condiz com a fluidez da documentação moderna, mesmo antes do uso de computadores na recuperação de informação.

Tálamo, Lara e Kobashi (1995, p. 56) corroboram esta linha de pensamento quando dizem considerar que:

> A inadequação de tais sistemas ao tratamento da informação é um fato reconhecido pelos bibliotecários. Embora eles continuem a ser utilizados, tanto a CDD como a CDU são vistos como sistemas para organizar documentos nas estantes e não como instrumentos para o tratamento e a recuperação da informação. De fato, a proposta inicial da CDD e da CDU era a de obter "a ordenação" dos documentos ou referências. As crescentes necessidades de recuperação de conteúdos colocaram a tais sistemas demandas que eles não têm condições de atender.

Mai (2010, p. 711, tradução nossa) segue esta mesma trilha que prega a contextualização social da classificação e da necessidade de novos sistemas classificatórios quando diz que:

> Na sociedade moderna atual, onde a diversidade da experiência humana está se tornando cada vez mais prevalente e é aceito que qualquer fato tem múltiplas interpretações, e onde o pluralismo floresce, precisamos repensar os fundamentos

> conceituais do trabalho de classificação e da teoria e construir uma base que parta de um pressuposto interpretativo e pluralista.

A contextualidade dos sistemas de classificação parece ser um problema a ser constantemente resolvido, como se depreende pelos diversos autores que fazem críticas aos sistemas universais de classificação, tais como Quinn (1994, p. 142-143), que alerta para as dificuldades em criar um sistema de classificação definitivo, geral e universalmente válido. As disciplinas das classes principais tendem em congelar a estrutura do conhecimento, o crescimento de assuntos interdisciplinares tem dificultada sua acomodação em esquemas baseados em disciplinas. Os sistemas de classificação são baseados em consenso social sobre o conhecimento. Este consenso varia de uma sociedade para outra, de um período histórico a outro bem como por disciplina.

Já Van Rijsbergen (1979, p. 40, tradução nossa) considera que, provavelmente, a característica mais importante na implementação de uma classificação é que deve ser capaz de lidar com uma coleção de documentos em mutação e crescimento constantes.

Desde a década de 1950 estão surgindo trabalhos e pesquisas que buscam criar novos sistemas de classificação de informação bibliográfica de forma automatizada com uso de computadores, que possam atender a demanda informacional crescente da sociedade.

Com uma visão pragmática da classificação, Borko atentava para o fato da não existência de procedimentos que sejam mais corretos que outros, já que sempre haverá imprecisão nas classificações.

> Até este ponto, o critério para a classificação correta foi o classificador humano, mas este não é necessariamente o melhor critério. Sabemos que os seres humanos não são perfeitamente confiáveis e, portanto, não é possível prever a classificação humana com precisão perfeita. O critério final da utilidade de qualquer sistema de indexação e classificação é se ele recupera informações relevantes em resposta a um pedido de pesquisa. Os procedimentos automáticos de classificação de documentos devem ser avaliados quanto à eficiência com que eles recuperam informações e não em como eles podem combinar o classificador humano imperfeito. (BORKO, 1964, p. 534, tradução nossa).

Em oposição a estas ideias, Ranganathan (1967, p. 544) declara que considera que seria impraticável que classificacionistas e classificadores abdicassem de suas funções em detrimento de computadores, pois classificação envolve julgamento. Julgamento do assunto de um documento em todas as suas facetas e arranjos exibidos neles. Isto não pode ser feito por análises estatísticas de palavras no documento.

Mesmo assim, as pesquisas sobre classificação automática de documentos têm avançado continuadamente, com resultados cada vez mais satisfatórios.

4.10 Classificação automática

Diversos pesquisadores, das mais diferentes áreas do conhecimento, têm-se debruçado sobre o tema, em virtude das características interdisciplinares que este assunto contempla. Borko (1967, p. 120) atentava para o fato que vários experimentos têm sido realizados e uma grande variedade de métodos de indexação e codificação da informação está em desenvolvimento.

Observa-se que as primeiras tentativas para representar o conteúdo de documentos em meios eletrônicos surgiram com Luhn, na década de 1950. Utilizando a frequência da ocorrência dos termos e a representação condensada dos documentos como uma lista de palavras utilizou métodos estatísticos simples para produzir um resumo automático dos textos. A partir de então, novas abordagens para o tratamento de documentos surgiram.

Conforme se pode depreender do artigo de 1958, no qual Luhn descreve o procedimento para a identificação de palavras chaves dentro de um texto científico para criar um resumo automático:

> A aplicação de métodos de máquina à busca de literatura está recebendo atualmente muita atenção e agora indica que tanto o esforço humano quanto seu viés podem ser eliminados da atividade de resumir. [...] O novo método mecânico seleciona aquelas entre todas as frases de um artigo que são as mais representativas como informação pertinente. Essas frases-chave são então enumeradas para servir de pistas para avaliar os assuntos do artigo. [...] Para determinar quais frases de um artigo podem melhor servir como o auto-resumo, é necessária uma medida pela qual o conteúdo da informação de todas as sentenças pode ser comparado e classificado. [...] Nenhuma atenção é dada aos relacionamentos lógicos e semânticos que o autor estabeleceu. Em outras palavras, é feito um inventário e uma lista de palavras compilada em ordem decrescente de frequência. [...] O método a ser desenvolvido aqui é um modelo probabilístico baseado nas propriedades físicas dos textos escritos. Nenhuma consideração deve ser dada ao significado das palavras ou aos argumentos expressos pelas combinações de palavras. (LUHN, 1958, p. 159-160, tradução nossa).

Posteriormente, na década de 1960, discorrendo sobre as pesquisas em novos métodos de classificação, Borko (1964, p. 153, tradução nossa) relata que:

> Estes e outros pesquisadores têm investigado métodos matemáticos para derivar categorias de classificação por sua convicção de que os sistemas de classificação empírica fornecerão meios mais eficientes para a classificação e recuperação de informações do que os métodos tradicionais de classificação de documentos. Essa crença foi submetida a testes e avaliações científicas.

Não são novas as pesquisas em métodos de classificação, agrupamento ou categorização de textos eletrônicos, desde os primeiros tempos do advento dos computadores, diversos autores têm dedicado esforços nesta área de estudo.

Borko (1967, p. 115, tradução nossa) afirma que “com o advento de grandes computadores de alta velocidade e programação estatística complexa, se tornou possível derivar sistemas de classificação de forma dinâmica para coleções de documentos”.

Rigby (1965 apud DHYANI, 1993, p. 10, tradução nossa) considera que,

> [...] a classificação automatizada inclui: a derivação matemática dos esquemas de classificação (o trabalho dos classificacionistas) e; a atribuição automática de documentos a categorias (o trabalho de classificação que é o trabalho de classificadores), independentemente de as categorias terem sido automaticamente derivadas ou escolhidas a partir de um esquema de classificação previamente concebido.

A classificação automática pode ser realizada com alguma intervenção humana ou por meio de máquinas, o que contará é seu produto final que será disponibilizado para uso.

Segundo Araújo (2009, p. 198), foi a partir da década de 1980 que as novas tecnologias digitais somam-se a esse campo, principalmente com a ideia do hipertexto e com as varias possibilidades de classificação.

Já para Sebastiani (2001, p. 8, tradução nossa), nos anos de 1980, a abordagem mais popular para criação de classificadores automáticos de documentos consistia na “construção manual de sistemas especialistas, que eram capazes de tomar decisões de classificação de texto por meio de regras lógicas”.

Posteriormente, novas iniciativas para a criação de procedimentos de classificação foram apresentadas como as utilizadas no aprendizado de máquinas (ML – *machine learning*) para construir, automaticamente, uma classificação a partir de um conjunto previamente classificado.

> [...] de acordo com o qual um processo indutivo geral constrói automaticamente um classificador automático de texto pela aprendizagem, a partir de um conjunto de documentos pré-classificados com as características das categorias de interesse (SEBASTIANI, 2001, p. 2, tradução nossa).

Nesta abordagem, a classificação é uma atividade do aprendizado supervisionado, pois o processo necessita das categorias previamente definidas, e de um conjunto de documentos de testes que pertençam a elas.

Langie e Lima (2003) apresentam a categorização hierárquica de documentos como uma técnica mista de classificação, que além de classificar os documentos, propiciariam a organização adequada e a navegação entre os documentos por meio das relações criadas entre eles.

> Uma forma de melhorar a organização dos documentos textuais é fazer uso da estruturação das categorias de assuntos em hierarquias. Essas estruturas facilitam a

> organização dos documentos ao permitirem o estabelecimento de relações entre assuntos mais genéricos e mais específicos. (LANGIE; LIMA, 2003, p. 2).

Para tanto, eles propõem que documentos sejam classificados em uma estrutura de categorias organizada em forma de árvore. Usando uma abordagem denominada de *top-down level-based*, na qual se utiliza de classificadores não hierárquicos para classificar os documentos dentro de uma categoria pai. Inicialmente, os documentos são classificados de forma o mais genérica possível em categorias (classes) de primeiro nível, com o uso de um classificador local àquela categoria, se possível, faz-se a classificação dentro da categoria pai para cada documento (LANGIE; LIMA, 2003, p. 5). Neste processo, utiliza-se de classificadores não hierárquicos para produzir uma classificação hierárquica.

Mourão, et al. (2008) apresentam a questão temporal das classificações como uma característica a ser estudada na classificação, pois a língua sofre variações ao longo do tempo e o surgimento de novos conceitos entre outros, podem afetar a classificação de novos documentos.

> Embora a classificação do documento seja um assunto amplamente estudado, a análise dos aspectos temporais nesta classe de algoritmos é bastante recente. [...] Uma área onde os aspectos temporais foram estudados com mais detalhes é a Classificação de Documentos Adaptáveis, que abrange um conjunto de técnicas relacionadas a aspectos temporais e outros com o objetivo de melhorar a eficácia e precisão dos classificadores de documentos através de sua adaptação incremental e eficiente (MOURÃO et al., 2008, p. 160, tradução nossa).

Já Hamou et al. (2010, p. 244, tradução nossa) propõem uma abordagem biomimética ("biomimética é a prática científica de tentar imitar ou se inspirar em sistemas da natureza") para a classificação de documentos de forma não supervisionada. Nesta abordagem propõem uma classificação baseada em uma rede "autômato celular" em um espaço bi-dimensional. Cada célula pode possuir quatro estados, que são: morta (0), viva (-1), isolada (-2) e ativa (número do documento). Cada célula pode estar cercada por até oito vizinhos. O espaço bi-dimensional é uma matriz quadrada que comporte todos os documentos do *corpus*.

Hamou et al (2010) apresentam a mineração de textos (*text mining*), uma abordagem de recuperação de textos que possibilita a extração de informações em documentos e descobrimento de padrões.

> A mineração de texto é a combinação de técnicas e métodos para processamento automático de dados textuais em linguagem natural. É uma análise multidimensional de dados textuais, que visa analisar e descobrir conhecimento e conexões a partir dos documentos disponíveis. Na mineração de textos as similaridades são usadas para produzir representações sintéticas de grande coleção de documentos. (HAMOU et al., 2010, p. 244, tradução nossa)

Korde e Mahender corroboram este entendimento quando descrevem os objetivos da mineração de textos.

> O objetivo principal da mineração de texto é permitir aos usuários extrair informações de recursos textuais e lidar com as operações, recuperação, classificação (supervisionada, não supervisionada e semi-supervisionada) e de técnicas de processamento de linguagem natural (NLP), *Data Mining* e *Machine Learning* conjuntamente, para classificar automaticamente e descobrir padrões dos diferentes tipos de documentos (KORDE; MAHENDER, 2012, p. 86, tradução nossa).

Para Hrala e Král (2013, p. 1, tradução nossa) a classificação automática tem importância para organização e armazenamento de informações e pode ocorrer para uma classe ou para várias classes. Onde:

> Na classificação de uma classe, o documento é atribuído exatamente a um rótulo de um conjunto de etiquetas pré-definido, enquanto na classificação de várias classes (às vezes também classificação de multi-rótulos), o documento pode ser rotulado com mais de um rótulo. (HRALA; KRÁL, 2013, p. 2, tradução nossa).

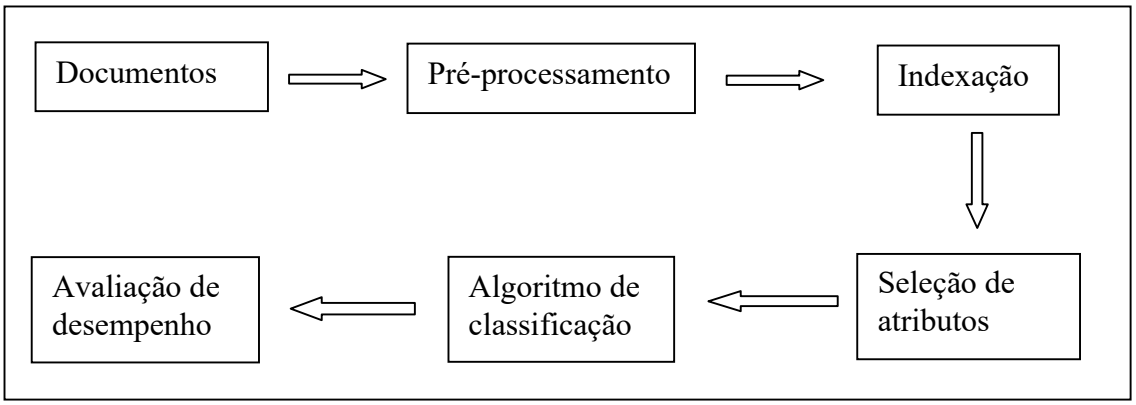
Diversos procedimentos de classificação têm sido propostos com inúmeras abordagens procedimentais, nas próximas seções serão apresentadas algumas destas proposições de mecanismos de classificação automática de documentos.

Descrevendo uma ferramenta de categorização automática de documentos Gomes e Moraes Filho (2011, p. 69-70) dizem que o processamento da categorização automática é realizado em quatro etapas as quais são:

> A primeira etapa consiste na preparação do documento que será categorizado, extraindo-se o conteúdo textual bruto sem distinção. Logo após todo o conteúdo do texto haver sido tratado inicia-se a segunda etapa, que consiste no pré-processamento dos dados textuais mediante técnicas de mineração de texto, retirando-se as *stopwords* e realizando o agrupamento das palavras, dando origem a um formato padrão para a realização da terceira etapa, que versa na categorização propriamente dita. Nesta categorização o número de repetições de cada palavra do documento e as palavras-chave de cada categoria são analisadas e fornecem a indicação da(s) categoria(s) em que o documento se compatibiliza. Na quarta e última etapa são realizadas as avaliações dos resultados obtidos, sendo que, se esses resultados não forem satisfatórios, o processo será repetido desde o início.

Para Korde e Mahender (2012), a classificação de textos é formada por um conjunto de procedimentos como ilustrado na Figura 1.

Figura 1 - Processo de classificação de documentos



Fonte: Korde e Mahender (2012, p. 86, tradução nossa).

Segundo Korde e Mahender (2012), os procedimentos de classificação automática de documentos são realizados em quatro etapas as quais são:

a) Pré-processamento:

> O primeiro passo do pré-processamento que é usado para apresentar os documentos de texto em formato limpo de palavras. Os documentos são preparados para o próximo passo na classificação de texto são representados por uma grande quantidade de atributos.
> Os passos neste processo geralmente são:
> - Um documento é tratado como uma *string*[1] e depois dividido em uma lista de *tokens*[2];
> - Remoção de *stopwords*: Neste processo são retiradas as palavras sem conteúdo semântico, como artigos e preposições;
> - Aplicação do algoritmo de derivação que converte as palavras diferentes em sua raiz semântica. (KORDE; MAHANDER, 2012, p. 86, tradução nossa).

b) Indexação, nesta etapa é realizada a representação do documento.

> A representação de documentos é uma das técnicas de pré-processamento que é usada para reduzir a complexidade dos documentos e torná-los mais fáceis de manusear, o documento deve ser transformado da versão de texto completo para um vetor do documento. [...] a representação de documentos mais comumente usada é chamada de modelo de espaço vetorial, que é apresentado como um vetor de palavras. (KORDE; MAHANDER, 2012, p. 86, tradução nossa).

c) Após o processo de indexação é realizada a seleção de atributos.

> A principal ideia de seleção de atributos (FS - *feature selection*) é selecionar o subconjunto de valores dos documentos originais. FS é realizado mantendo as palavras com maior pontuação de acordo com a medida predeterminada da importância da palavra. (KORDE; MAHANDER, 2012, p. 87, tradução nossa).

d) Classificação automática, onde se aplicam os algoritmos de classificação.

[1] *String* - sequência de caracteres.
[2] *Tokens* - marcadores ou símbolos para os termos/palavras.

> A classificação automática de documentos em categorias predefinidas [...] os documentos podem ser classificados de três formas, sem supervisão, supervisão e métodos semi-supervisionados. (KORDE; MAHANDER, 2012, p. 87, tradução nossa).

e) Avaliação do desempenho do classificador.

> Esta é a última etapa da classificação de texto, na qual as avaliações de classificadores de texto normalmente são conduzidas experimentalmente, em vez de analiticamente. A avaliação experimental dos classificadores, em vez de se concentrar em questões de eficiência, geralmente tenta avaliar a eficácia de um classificador. (KORDE; MAHANDER, 2012, p. 88, tradução nossa)

Hrala e Král (2013) resumem bem as etapas principais para a classificação automática de documentos, são elas:

> […] representação de documentos, seleção de atributos e modelagem de documentos. A representação do documento consiste em escolher um conjunto de atributos que represente o documento com a maior precisão possível. O texto completo é transformado em um vetor de atributos do documento. A seleção de atributos é então usada para reduzir o tamanho desse vetor. O último passo consiste em construir um modelo de documento usando vetores de atributos. Este modelo é usado para classificação de documentos. (HRALA; KRÁL, 2013, p. 2, tradução nossa).

Cada uma destas etapas apresenta procedimentos específicos para que o próximo passo seja realizado a contento. Não serão analisadas as ferramentas disponíveis para a adequação dos documentos em seus diversos formatos de apresentação, tais como PDF (*Portable Document Format*), MS-Word e etc. para textos brutos.

4.10.1 Pré-processamento dos documentos

Segundo Gomes e Moraes Filho (2012, p. 71).

> As palavras são os atributos ou características de um texto. Formam elementos primitivos que serão analisados para a descoberta do conhecimento ou extração de um padrão que defina a categoria do documento. A etapa de pré-processamento é responsável por transformar o documento em uma forma padrão. Para se chegar a esse modelo, o texto passa por quatro processos: o *case folding*, a eliminação de acentos gráficos, a separação das palavras e a eliminação das *stopwords*.

Nesta etapa os documentos serão adequados em relação ao tratamento dos termos (palavras) para serem utilizados pelos algoritmos de representação do conteúdo. Para Maia e Souza (2010, p. 171)

> Os métodos de descobertas de conglomerados ou classificação, mostram-se extremamente dependentes de técnicas de pré-processamento dos textos que visassem a padronizá-los, minimizando os problemas do vocabulário e

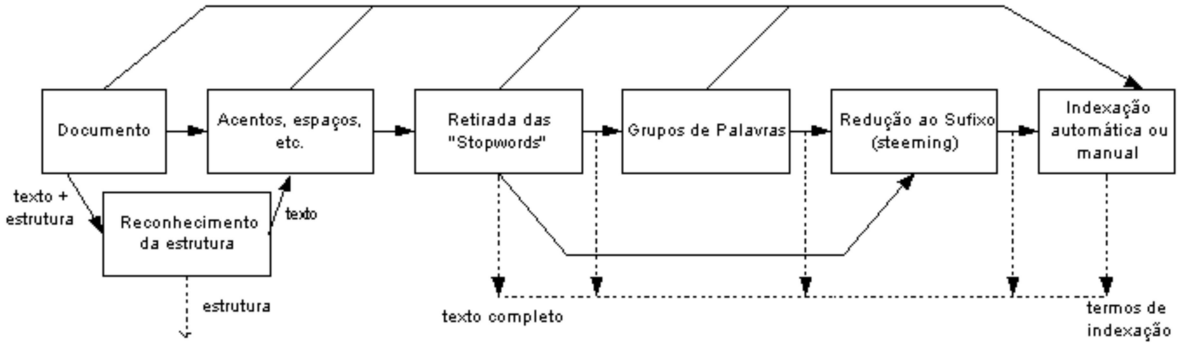
> representando seu conteúdo de forma mais correta e fácil de ser trabalhada pela máquina.

Os documentos precisam, então, ser preparados para que não haja imprecisão textual, após o procedimento de pré-processamento os documentos originais foram adequados em uma nova entidade temporária que representa sua estrutura, mas sem sinais gráficos, palavras não necessárias semanticamente e estrutura de apresentação (parágrafos, fontes gráficas, tabelas, imagens etc.). Como resultado o documento será visto como uma única linha de texto bruto só com as palavras significativas, então os documentos já adequados podem ser tratados pelos procedimentos de representação dos documentos (indexação). Para Harish, Guru e Manjunath (2010, p. 111, tradução nossa),

> Na classificação automática de texto, provou-se que o termo é a melhor unidade para representação e classificação de texto. [...] os dados não estruturados, em particular os dados de texto corrente, devem ser transformados em dados estruturados. Para fazer isso, muitas técnicas de pré-processamento são propostas na literatura. Depois de converter dados não estruturados em dados estruturados, precisamos ter um modelo efetivo de representação de documentos para construir um sistema de classificação eficiente.

O procedimento de pré-processamento pode ser graficamente resumido como apresentado na figura 2.

Figura 2 - Fases do pré-processamento de documentos



Fonte: (BAEZA-YATES; RIBEIRO-NETO apud MAIA; SOUZA, 2013, p. 7.)

Como se pode verificar pela Figura 2, os procedimentos de pré-processamento não se limitam aos apresentados neste trabalho. Não há descrição de mecanismos de retirada de "*stopwords*", por exemplo.

4.10.1.1N-Grams

Cavnar e Trenkle (1994, p. 162, tradução nossa) descrevem um N-gram como "um pedaço de N caracteres de uma *string* mais longa. Embora na literatura o termo possa incluir a noção de qualquer conjunto de caracteres que estejam em uma *string*" .

Ou, segundo Hamou et al. (2010, p. 242, tradução nossa).

> Um n-gram é uma sequencia de n caracteres consecutivos em um documento, todos os n-grams (geralmente n = 2, 3, 4, 5) são obtidos movendo uma 'janela' de n caixas no corpo do texto. [...] Este movimento é feito por passos de um caractere e cada passo tira-se uma 'foto'. Todas essas fotografias fornecem o conjunto de todos os n grams do documento. Então, as frequências de n-grams são encontradas.

Por exemplo, na frase: **O garoto saiu de casa**.

Para o n-gram com o parâmetro n de valor 5, se obtém os seguintes n-grams: o-gar, -garo, garot, aroto, roto-, oto-s, to-sa, o-sai, -saiu, saiu-, aiu-d, iu-de, u-de-, -de-c, de-ca, e-cas, -casa. (o traço '-' representa espaço).

> O conceito de n-grams tem sido amplamente utilizado em várias áreas, como a identificação da recuperação da fala e da informação: a representação do documento textual pelo método de n-grams tem muitas vantagens. De fato, os n-grams capturam o conhecimento mais frequente As palavras de cada idioma que facilitam a identificação da linguagem e o método de n-grams são independentes da linguagem, enquanto os sistemas baseados em palavras, por exemplo, dependem da linguagem. (HAMOU et al., 2010, p. 242, tradução nossa).

Para Cavnar e Trenkle (1994, p. 162, tradução nossa),

> O principal benefício que a combinação baseada em N-gram fornece, resulta de sua própria natureza: Uma vez que cada *string* é decomposta em pequenas partes, os erros que estão presentes tendem a afetar apenas um número limitado dessas partes, deixando o restante intacto. Da contagem de N-grams que são comuns a duas *strings*, obtêm-se uma medida de similaridade que é resistente a uma grande variedade de erros no texto.

4.10.1.2 Sintagmas nominais

Para Kuramoto (2002, p. 6),

> [...] o sintagma nominal é a menor parte do discurso portadora de informação. Ao contrário das palavras, os sintagmas nominais não são símbolos sem referências. [...] são portadores de uma estrutura lógico-semântica. [...] os sintagmas nominais são compostos de grupos nominais constituídos de uma organização hierárquica em árvore. Diferentemente das palavras, o sintagma nominal quando extraído do texto mantém o significado, o seu conceito.

Kuramoto (2002, p. 6) sugere que uma abordagem para indexação, que seria a utilização dos sintagmas nominais em substituição às palavras (termos), que permitiria a utilização dos procedimentos de classificação sem grandes custos.

Maia e Souza (2010, p. 170) concordam quando dizem que:

> [...] a utilização de sintagmas nominais é capaz de representar o conteúdo dos documentos, servindo como descritores ou características para o processo de classificação ou descoberta de conglomerados, melhorando a precisão desse processo.

Após a fase de pré-processamento os documentos estão prontos para a próxima etapa que consiste em produzir vetores com representações dos termos do *corpus*.

4.10.2 Representação de documentos

O segundo procedimento de tratamento dos documentos visando sua classificação é a representação de documentos ou indexação. "O processo de indexação visa à representação dos conteúdos dos documentos produzindo uma lista de descritores" (MAIA; SOUZA, 2010, p. 162).

Para Rocha e Catae (2012, p. 1),

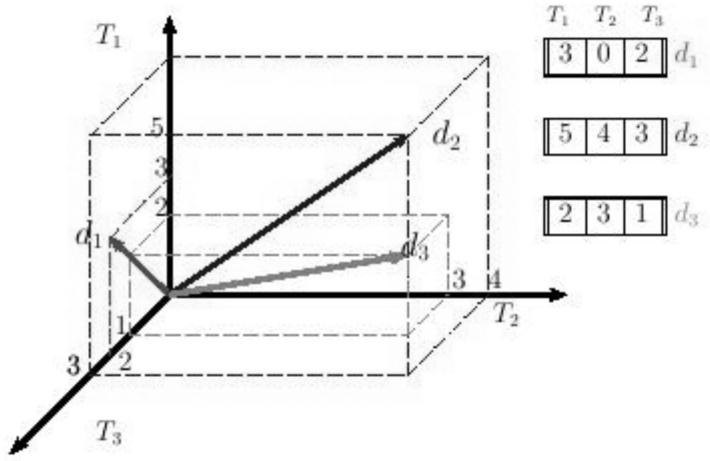
> Salton introduziu a ideia de representação de documento através de vetores, cujas componentes são calculadas com base na contagem de palavras e associados a pesos, criando como resultado a matriz Termo-Documento. A partir dessa representação, uma forma de determinar sua categoria é identificar palavras específicas ou presentes somente naquele tipo de documento.

Segundo Gean e Kaestner (2004, p. 2),

> De acordo com o modelo vetorial cada documento é modelado por um vetor no espaço m-dimensional, onde m é o número de diferentes termos presentes na coleção. Os valores das coordenadas do vetor que representa o documento estão associados aos termos, e usualmente são obtidos a partir de uma função relacionada à frequência dos termos no documento e na coleção.

Na Figura 3 é apresentada a representação espacial de três vetores dos termos T1, T2, T3 com seus respectivos pesos (frequências dos termos) nos documentos d1, d2, d3. O documento d1 possui três termos T1, nenhum termo T2, e dois termos T3. O documento d2 possui cinco termos T1, quatro termos T2 e três termos T3, finalmente o documento d3 possuir dois termos T1, três termos T2 e um termo T3. Os eixos correspondem aos termos.

Figura 3- Representação gráfica de três documentos no espaço vetorial



Fonte: Oliveira (2007, p. 82)

Para Hamou et al. (2010, p. 244) "a representação mais simples de documentos textuais é chamada de uma representação de 'saco de palavras' (*bag of words*), é transformar textos em vetores onde cada elemento representa uma palavra". No entanto, segundo Harish, Guru e Manjunath (2010, p. 111, tradução nossa)

> O esquema de sacos de palavras tem algumas limitações. Alguns deles são: alta dimensionalidade da representação, perda de correlação com palavras adjacentes e perda de relação semântica que existe entre os termos de um documento. Para superar esses problemas, os métodos de ponderação dos termos são usados para atribuir pesos apropriados aos termos para melhorar o desempenho da classificação de texto.

Os documentos textuais precisam passar por procedimento de adequação para que sejam tratados pelos algoritmos de classificação, um destes tratamentos corresponde à 'redução da dimensão dos documentos'. Conforme Sebastiani (2001, p. 11, tradução nossa),

> Os textos não podem ser interpretados diretamente por um classificador ou por um algoritmo de construção de classificação. Por isso, um procedimento de indexação que mapeia um texto em uma representação compacta de seu conteúdo precisa ser aplicado uniformemente aos documentos de treinamento, validação e teste. [...] o texto é geralmente representado como um vetor de pesos dos termos.
> Dj = {W1j, W2j, ..., W|T|j} , onde: T é o conjunto de termos que ocorrem pelo menos uma vez em pelo menos um documento; e 0≤Wij≤1, representa como o termo Tk contribui para a semântica do documento.

Atribuir peso aos termos de um documento é uma forma de diferenciar os termos mais relevantes daqueles termos de menor importância. (LANGIE; LIMA, 2003, p. 4).

4.10.3 Seleção de atributos

Langie e Lima (2003, p. 5) afirmam que a seleção de atributos consiste em eliminar termos que não são representativos, ou então combinar mais de um termo em um único atributo. A seleção serve também para diminuir o número de elementos que compõem os vetores dos documentos

Para Almeida e Yamakami (2001, p. 17) em categorização de texto, "a alta dimensionalidade do espaço de termos (T) pode ser problemática". Pois,

> [...] muitos classificadores apresentam baixo desempenho quando manipulam uma grande quantidade de atributos. Dessa forma, é recomendável um procedimento para reduzir o número de termos utilizados. (ALMEIDA; YAMAKAMI, 2001, p. 18).

Será necessário então, aplicar um procedimento para ajustar a dimensão do vetor de termos. Atribuir peso aos termos de um documento é uma forma de diferenciar os termos mais relevantes daqueles de menor importância. (LANGIE; LIMA, 2003, p. 4).

Existem diversos procedimentos para a seleção de atributos, Yang e Pedersen (1997, p. 413, tradução nossa) sugerem cinco critérios para realizar a eliminação de termos dos documentos, os critérios são: frequência dos documentos (DF – *Document Frequency*), ganho de informação (IG – *Information Gain*), informação mútua (MI – *Mutual Information*), X2 (letra grega khi), estatístico (khi quadrado, CHI) e força do termo (TS - *Term Strength*).

Para a frequência dos documentos (DF), o número de documentos em que o termo ocorre é calculado para cada termo único e removido do espaço de atributos aqueles que têm frequência menor que um limite previamente definido. (YANG; PEDERSEN, 1997, p. 413, tradução nossa).

O ganho de informação (ig)

> Mede o número de *bits* de informação obtidos para a previsão da categoria pelo conhecimento da presença ou ausência de um termo em um documento [...] Dado um corpus de treinamento, para cada termo único, calculamos o ganho de informações e removemos do espaço de recursos aqueles termos cujo ganho de informação era menor do que um limite predeterminado. (YANG; PEDERSEN, 1997, p. 413, tradução nossa).

A força do termo é obtida de maneira completamente diferente dos outros procedimentos estatísticos.

> É baseado no agrupamento (*clustering*) de documentos, assumindo que os documentos com muitas palavras compartilhadas estão relacionados e que os termos na área de sobreposição dos documentos relacionados são relativamente informativos. (YANG; PEDERSEN, 1997, p. 413, tradução nossa).

Outros procedimentos de redução do espaço vetorial dos atributos seguem procedimentos estatísticos mais complexos, mas sempre contando com a frequência dos termos nos documentos.

4.10.3.1 Atribuição de pesos

De acordo com Salton e Buckley (1988, p.516, tradução nossa).

> A principal função de um sistema de atribuição de pesos é o melhorar a eficácia da recuperação. [...] Duas medidas normalmente são usadas para avaliar a capacidade de um sistema para recuperar os itens relevantes e rejeitar os itens não relevantes de uma coleção, conhecidas como revocação e precisão. Revocação é medida pela proporção do número de itens relevantes recuperados pelo número total de itens relevantes da coleção; Precisão é a proporção de itens recuperados que são relevantes, calculada pela proporção do número de itens relevantes recuperados pelo o número total de itens recuperados.

Segundo Oliveira et al. (2007, p. 82) é necessário encontrar pesos apropriados para distinguir as palavras, encontrar os pesos mais adequados não é simples sendo necessário o uso de modelos matemáticos e inteligência artificial, para se obter bons resultados.

Para Almeida e Yamakami (2001, p. 18) no caso mais simples de representação, "cada termo representa uma única palavra e todos os atributos são booleanos: Xi = 1 se a mensagem contém ti, ou Xi = 0, em caso contrário".

Considerando a atribuição de peso aos termos, tem-se que os termos que ocorrem frequentemente nos documentos são importantes para a melhora da revocação. Isto sugere que a frequência dos termos (TF – *Term Frequency*) deve ser um fator usado na atribuição de pesos. Outro ponto importante é que a frequência dos termos não pode ser considerada unicamente como fator de melhoria do desempenho, principalmente quando a alta frequência dos termos se dá em todos os documentos da coleção, então todos os documentos serão recuperados, o que sugere a necessidade de um novo fator de ponderação que favoreça os termos que são concentrados em poucos documentos da coleção. O novo fator deve variar inversamente com o número de documentos para os quais o termo é especificado na coleção. (SALTON, BUCKLEY, 1988, p.516, tradução nossa).

4.10.3.2 Frequência de termos (TF - Term Frequency)

Luhn (1958) sugeriu, primeiramente, que a frequência dos termos (TF - *Term Frequency*) dentro de um documento tinha importância para a identificação de palavras chaves no texto. De acordo com suas palavras:

> O fator de significância de uma frase é derivado de uma análise de suas palavras. É aqui proposto que a frequência de ocorrência de palavras em um artigo forneça medidas úteis de significância da palavra. Propõe-se ainda que a posição relativa da palavra dentro de uma sentença possui valores de significância e forneça uma medida útil para determinar o significado das sentenças. O fator de significância de uma frase será, portanto, baseado em uma combinação dessas duas medidas. (LUHN, 1958, p. 160, tradução nossa).

Então, o que Luhn (1958) denominou de fator de significância é computado como resultado da frequência de um termo significativo em uma porção de texto no documento elevado ao quadrado, dividido pelo número total de termos nesta mesma porção de texto.

Com o avanço das pesquisas sobre indexação ou classificação automática, outros métodos de ponderação sobre a frequência de termos foram apresentados.

4.10.3.3 Frequência inversa de termos no documento (IDF – Inverse Document Frequency)

A frequência inversa de termos no documento dá a medida da importância de um termo dentro de um documento, caso um termo ocorra com muita frequência este pode não ser tão relevante na determinação da classe predominante, como os exemplos das preposições que ocorrem em grande número, mas, possuem pouca significância para o assunto. Para casos assim, é necessário ponderar sua frequência utilizando o fator de frequência inversa do documento. Conforme Yates e Ribeiro Neto.

> Assim, podemos concluir que um termo (palavra no documento) pode aparecer em mais de um documento. Portanto, a cada termo será atribuído um peso. O peso que esse termo recebe leva em consideração dois aspectos: a quantidade de vezes que ele ocorre no próprio documento e a quantidade de vezes que ele aparece em outros documentos analisados. Através disso, ponderamos a importância desse termo no conjunto de documentos onde ele aparece. Uma das propostas de ponderação dessa importância apresentada na literatura (YATES; RIBEIRO NETO, 1998 apud OLIVEIRA, 2007, p. 83).

O cálculo da frequência inversa de termos é dado pela função:

$$idf_i = log\frac{N}{n_i}, \quad (1)$$

Onde:

- idf_i: (*inverse document frequency*) para o termo i;
- N: número total de documentos;
- n_i: número de documentos em que o termo aparece.

Na medida em que o termo i está presente em uma quantidade grande de documentos o valor da função idf_i tende a zero, indicando que a relevância deste termo em relação ao conjunto de documentos diminui.

4.10.3.4 O método TF-IDF para identificação da frequência de termos

Salton , Wong e Yang (1975, p. 615, tradução nossa) consideravam que:

> [...] um dos melhores procedimentos para atribuição de pesos, que consistia no produto da multiplicação da frequência padrão dos termos (FT), pelo fator inverso da frequência dos termos no documento (IDF). [...] Um sistema de pesos proporcionais atribuiria pesos maiores aos termos que aparecem com alta frequência nos documentos individuais, mas que era ao mesmo tempo relativamente raro na coleção como um todo.

Quase trinta e cindo anos após, Hamou et al. (2010, p. 244, tradução nossa) afirmaram que esta é uma das técnicas mais utilizadas ainda.

> A maioria das abordagens se concentra na representação vetorial do texto usando a medida TF * IDF. TF representa "Frequência de Termo": o número de ocorrências do termo no corpus. IDF representa o número de documentos que contém o termo.

Finalizado o processo de representação de documentos, o resultado será uma matriz termo-documento, onde as colunas representam os termos e as linhas representam os documentos.

Segundo Gean e Kaestner (2004, p. 4),

> [...] em conformidade com o modelo vetorial uma coleção de documentos pode ser vista como uma imensa matriz $\mathbf{C_{N\ x\ M}}$, onde $\mathbf{f_{ij}}$ representa o peso do termo j no documento i, M é o número de termos e N é o número de documentos na coleção.

Pode-se observar pelo exemplo da Figura 4, que foi necessário ajustar os vetores para que tenham o mesmo tamanho. Sem este ajuste não é possível comparar os vetores, que representam os documentos.

Figura 4 - Representação de uma matriz termo-documento

$$C= \begin{bmatrix} f_{11}, f_{12},... f_{1M} \\ f_{21}, f_{22},... f_{1M} \\ ... \\ f_{n1}, f_{n2},... f_{nM} \end{bmatrix}$$

Fonte: Gean e Kaestner (2004, p. 4)

Conforme Corumbá e Macedo (2011, p. 5), no modelo vetorial os documentos são representados geometricamente, pois cada documento possui um vetor associado cujos elementos formam uma "tupla [3] de valores da forma dj = {w1j, ..., wij}", onde wij representa o peso de cada termo associado ao documento dj.

> Cada elemento do vetor é considerado uma coordenada dimensional. Desta forma, os documentos podem ser colocados em um espaço euclideano de n dimensões (onde n é o número de termos) e a posição do documento em cada dimensão é dada pelo peso do termo associado a aquela dimensão.

No modelo vetorial proposto por Salton (1975), cada documento será representado por um vetor de pesos, que indicam a importância do termo correspondente ao peso no documento. Cada peso é considerado uma coordenada dimensional no 'espaço' vetorial. Os procedimentos de classificação levam em conta este vetor 'espacial' para classificar outros documentos.

4.10.3.5LSI (Latent Semantic Indexing)

De acordo com Sebastiani (2001) LSI é:

> É uma técnica de redução dimensional (DR – *Dimensionality Reduction*) desenvolvida para recuperação de informações (IR – *Information Retrieval*) para resolver os problemas decorrentes do uso de palavras sinônimas, quase sinônimas e polissêmicas como dimensões das representações dos documentos. Esta técnica comprime os vetores-documentos em vetores de espaço dimensional menor, cujas dimensões são obtidas como combinações das dimensões originais observando seus padrões de ocorrência. Na prática, a LSI infere a dependência entre os termos originais de um corpus e "liga" essa dependência às dimensões independentes recentemente obtidas. A função que mapeia vetores originais para novos vetores é obtida aplicando uma decomposição de valor singular à matriz formada pelos vetores de documentos originais". (SEBASTIANI, 2001, p. 20, tradução nossa).

[3] Tupla – lista ordenada

Para Harish, Guru e Manjunath (2010, p. 111, tradução nossa), a indexação semântica latente (*Latent semantic indexing* - LSI) preserva os recursos mais representativos dos documentos em vez das características discriminantes.

Segundo Rocha e Catae (2012, p. 4),

> LSI é um algoritmo de aprendizado não-supervisionado. Durante a fase de treinamento, todos os documentos são considerados: conjunto de treinamento e testes. Através da decomposição em valores singulares, encontramos um conjunto de vetores unitários e ortogonais, que vão compor a base do subespaço representado. Em nenhum momento, a categoria é levada em consideração.

O LSI é usado para produzir um conjunto de agrupamentos de documentos de acordo com sua similaridade semântica. Este modelo considera que termos são semanticamente próximos se aparecem em contextos similares, e que dois contextos são similares se contem termos semanticamente próximos. (TORRES-MORENO, 2014, p. 73, tradução nossa).

4.10.3.6 Ajuste no tamanho dos vetores-documentos

Salton e Buckley (1988) apresentam um terceiro fator de ponderação do termo, além da sua frequência e da sua frequência inversa, que é útil para sistemas com comprimentos de vetor com muita variação. Em muitas situações, documentos curtos, tendem em ser representados por vetores de termos curtos, enquanto vetores de termos muito extensos são atribuídos a documentos grandes. Quando um grande número de termos é usado para a representação de documento, existe a possibilidade que os documentos maiores sejam recuperados, enquanto os documentos curtos não. Isto sugere que um fator de normalização seja incorporado na fórmula de ponderação para equilibrar o tamanho dos vetores-documentos. (SALTON; BUCKLEY, 1988, p. 517, tradução nossa).

$$w/\left(\sum (w_i)^2\right)^{1/2} \qquad (2)$$

Onde w representa o peso do termo t; i representa o número de termos.

Para Salton e Buckley (1988, p. 518, tradução nossa) o melhor sistema de ponderação de termos é representado pela fórmula:

$$(Tf*IDF) / (w/(\sum(w_i)^2)^{1/2}) \quad (3)$$

Ou seja, a melhor ponderação é dada pelo cálculo do TF-IDF dividido pelo fator de ajuste do tamanho dos vetores-documentos.

4.10.4 Classificação

De acordo com Sebastiani (2001, p. 9, tradução nossa),

> Desde o início dos anos 1990, a abordagem de aprendizagem de máquina (ML - *Machine Learning*) para categorização de textos (TC -*Text Categorization*) ganhou popularidade e, eventualmente, se tornou dominante, pelo menos na comunidade de pesquisa. Nesta abordagem, um processo indutivo geral (também chamado de aprendiz) cria automaticamente um classificador para uma categoria ci, observando as características de um conjunto de documentos classificados manualmente, sob ci ou um vetor-ci, por um especialista na área de domínio. A partir dessas características, o processo indutivo busca as características que um novo documento (ainda não classificado) deveria ter para ser classificado sob a categoria 'ci'. Na terminologia de ML, o problema de classificação é uma atividade de aprendizagem supervisionada, uma vez que o processo de aprendizagem é "supervisionado" pelo conhecimento das categorias e das instâncias de treinamento que lhes pertencem.

Cormack faz distinção entre três tipos possíveis de procedimentos de classificação:

> (i) classificação hierárquica, em que as classes são classificadas em grupos, o processo é repetido em diferentes níveis para formar uma árvore;
> (ii) particionamento, em que as classes são mutuamente exclusivas, formando assim uma partição do conjunto de entidades;
> (iii) aglomeração, em que as classes ou aglomerados podem se sobrepor, e um grupo e seu complemento são tratados como diferentes tipos de classe. (CORMACK, 1971, p. 321, tradução nossa).

4.10.4.1 Processos de classificação supervisionada

Descrevendo os procedimentos que os classificadores textuais automáticos executam para produzir uma classificação que seja supervisionada, Langie e Lima (2003, p. 4) alertam que:

> Os classificadores automáticos trabalham com um conjunto de documentos textuais digitais previamente classificados, denominado coleção ou corpus. Estes documentos são divididos em dois conjuntos, denominados base de treino e base de teste. A base de treino é utilizada pelo algoritmo de classificação para identificar as características das categorias da coleção. Estas categorias são as mesmas nas quais novos documentos poderão ser classificados. A base de teste é utilizada para testar o

> desempenho do classificador. Os documentos de teste são analisados pelo classificador, que determina a(s) categoria(s) à(s) qual (is) o documento pertence.

De acordo com SALLES (2009, p. 107)

> CAD[4] usualmente segue uma estratégia de aprendizagem supervisionada. Primeiramente, um modelo de classificação é construído utilizando documentos pré-classificados (conjunto de treinamento). Esse modelo é então, utilizado para classificar os demais documentos. Uma tarefa fundamental em CAD consiste em identificar e ponderar um conjunto de características que melhor identificam as categorias de documentos.

São considerados supervisionados por possuir um conjunto de treino que foi produzido previamente por um especialista no assunto.

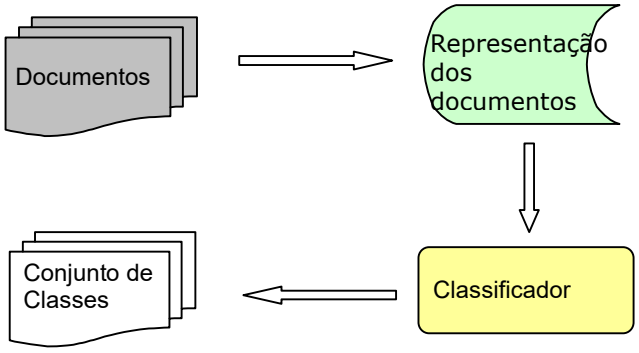
4.10.4.2 Processo de classificação não supervisionada

Hamou et al. (2010) destacam qual o princípio da classificação (agrupamentos/*clustering*) não supervisionada:

> [...] é agrupar textos que pareçam similares (com afinidades comuns) na mesma classe. Os textos em diferentes classes têm diferentes afinidades. [...] Os métodos de classificação não supervisionada podem ser divididos em duas famílias: a família de métodos de classificação hierárquica e métodos de classificações não hierárquicas. A classificação ou agrupamento (*clustering*) não supervisionado é uma técnica fundamental na mineração de dados (estruturada ou não estruturada). (HAMOU et al, 2010, p. 241, tradução nossa).

Na Figura 5 Hamou et al. (2010) exemplificam o processo de classificação não supervisionado.

Figura 5 - Classificação automática não supervisionada.



Fonte: Hamou et al. (2010, p. 252)

[4] CAD – Classificação automática de documentos.

4.10.4.3 Medidas de similaridade:

Segundo Hamou et al. (2010, p. 241, tradução nossa), várias medidas de similaridade, entre documentos, foram propostas na literatura, em particular: a distância euclidiana, Manhattan e cosseno.

De acordo com Maia e Souza (2010, p. 161), para se localizar a similaridade entre dois documentos utilizando o módulo do vetor espacial (VSM), calcula-se o cosseno do ângulo formado no vetor termo-documento, e quanto menor for o ângulo, mais próximo de 1 será o cosseno e mais similar serão os documentos.

A similaridade é dada pela seguinte fórmula:

$$sim(d_i, d_j) = \frac{\mathbf{d_i} \bullet \mathbf{d_j}}{|\,\mathbf{d_i}\,| \times |\,\mathbf{d_j}\,|} = = \frac{\sum_{k=1}^{n} w_k^i \times w_k^j}{\sqrt{\sum_{k=1}^{n}\{w_k^i\}^2} \times \sqrt{\sum_{k=1}^{n}\{w_k^j\}^2}} = cos(\theta), \quad (4)$$

> Onde $|\mathbf{d_i}|$ é o módulo do vetor $\mathbf{d_i}$, $\mathbf{cos(\theta)}$ é o cosseno do ângulo entre os vetores que representam os dois documentos $\mathbf{d_i}$ e $\mathbf{d_j}$. O valor do cosseno de um ângulo varia em um intervalo de 0 a 1. Esse fato nos dará uma interpretação de distância entre os documentos, onde 0 significará o mais alto grau de dissimilaridade e , 1 completa similaridade Já o valor $\mathbf{\omega^i_k}$ indica o peso referente ao termo $\mathbf{t_k}$ no documento $\mathbf{d_i}$. (BAEZA-YATES; RIBEIRO-NETO, 1998 apud OLIVEIRA, 2007, p. 86)

4.10.5 Classificadores

Segundo Korde (2012, p. 87, tradução nossa)

> Nos últimos anos, a tarefa de classificação automática de texto tem sido amplamente estudada e um rápido progresso ocorre nesta área, incluindo as abordagens de aprendizado da máquina, como o classificador bayesiano, árvores de decisão (*Decision Tree* - DT), K vizinho mais próximo (K-*nearest neighbor* - kNN), máquinas de vetores de suporte (*Support Vector Machines* - SVMs), redes neurais (*Neural Networks* - NN), e Rocchio.

4.10.5.1K-médias (K-means)

De acordo com Maia e Souza (2010, p. 162):

> O algoritmo K-means (ou K-médias) tem o objetivo de fornecer um agrupamento de objetos de acordo com os seus próprios dados. Essa classificação é baseada em análise e comparações entre os valores numéricos dos dados fornecidos. Dessa

> maneira, o algoritmo realiza um agrupamento automático sem a necessidade de nenhuma supervisão humana, ou seja, sem nenhum pré-agrupamento existente.

Wagstaff et al. (2001, p. 577-578, tradução nossa) descrevem assim o método k-média:

> O K-média é um método comumente usado para fracionar, automaticamente, um conjunto de dados em k grupos. Ele procede selecionando k centros de agrupamentos iniciais e depois refinando iterativamente como se segue:
> 1 - Cada instância di é atribuída o seu centro de agrupamento mais próximo;
> 2 - Cada centro de agrupamento Cj é atualizado para ser a média de suas instâncias constituintes.
> O algoritmo converge quando não há mais alterações na atribuição de instâncias a *clusters*

O procedimento analisa todo o corpus e então, encontra, aleatoriamente, k classe e a ela atribui os documentos correspondentes. Aos documentos restantes é atribuído o vetor da classe mais próxima, a classe é recalculada para esta nova atribuição e o procedimento reinicia.

O parâmetro k deve ser informado pelo operador do algoritmo que definirá as k classes do corpus. Mesmo sendo considerado não supervisionado, o procedimento necessita da intervenção humana para a definição do parâmetro k.

4.10.5.2 Linear Least Squares Fit – LLSF

Para Yang e Pedersen (1997, p. 415, tradução nossa) o LLSF é um método de classificação automática baseado em regressão linear, que usa os vetores de peso dos termos como entrada, e tem como saída uma lista (vetor) de categorias classificadas pelas pontuações dadas às classes.

Segundo Sebastiani (2001, p. 27, tradução nossa), a classificação utilizando LLSF,

> [...] pode ser vista como a tarefa de determinar um vetor de saída O (dj) para o documento de teste dj, dado o seu vetor de entrada I(dj); então, construir um classificador se resume em computar uma matriz M= | C | × | T | tal que MI (dj) = O(dj). [...] O LLSF calcula a matriz a partir dos dados de treinamento, calculando um ajuste linear de mínimos quadrados que minimiza o erro no conjunto de treinamento.

De acordo com Korde e Mahender (2012, p. 90, tradução nossa),

> [...] Uma abordagem de mapeamento desenvolvida por Yang. Os dados de treinamento são representados na forma de pares de vetores de entrada / saída onde o vetor de entrada é um documento no modelo de espaço vetorial convencional (consistindo de palavras com pesos), e o vetor de saída consiste em categorias (com

> pesos binários) do documento correspondente. [...] LLSF é um dos classificadores de texto mais eficazes conhecidos.

Para Sebastiani (2001, p. 27) o LLSF é um dos métodos mais eficiente para a classificação, sua desvantagem está no custo de computação para a matriz M que é muito alto quando comparado a outros procedimentos de classificação.

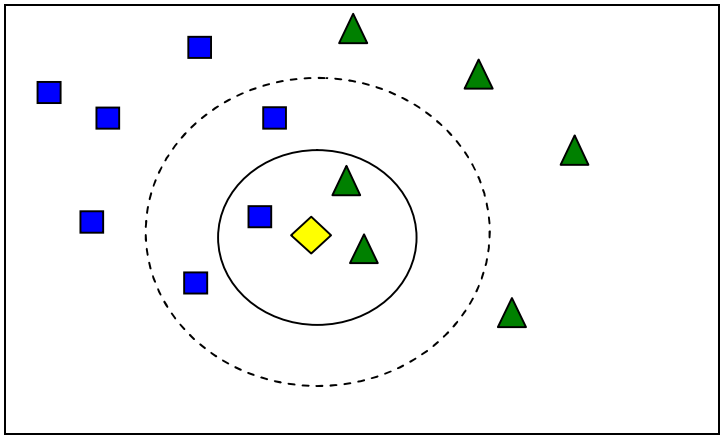
4.10.5.3 K-Nearest Neighbor – KNN

Definindo o que seria o método kNN, Zhang, Mouhoub e Sadaoui (2014, p. 95, tradução nossa) esclarecem que:

> O kNN ganhou muita popularidade devido às suas propriedades não-paramétricas e fáceis de implementar. Basicamente, ele classifica uma instância desconhecida por seus vizinhos mais próximos nos conjuntos de treinamento. K representa o número de vizinhos próximos que selecionamos.

Posteriormente, Zhang, Mouhoub e Sadaoui (2014, p. 95) apresentam, figurativamente, o procedimento básico do classificador kNN. Na Figura 6, o losango amarelo representa o objeto a ser classificado entre o quadrado azul ou o triângulo verde; o círculo de linhas contínuas representa a classificação em que o parâmetro k é igual a três. Neste caso, o losango entraria na classe de triângulos verdes, pois está mais próximo (círculo) de dois triângulos verdes, caso o parâmetro k seja alterado para cinco, o losango seria classificado como quadrados azuis já que na região vizinha (círculo tracejado) existem mais quadrados próximos.

Figura 6 - Ilustração do classificador kNN



Fonte: ZHANG; MOUHOUB; SADAOUI, 2014, p. 95.

O método kNN apresenta algumas limitações quando se trata de grandes conjuntos de dados com "baixa taxa de reconhecimento, alta complexidade computacional e não avalia a amostra de treino". (ZHANG; MOUHOUB; SADAOUI, 2014, p. 95, tradução nossa).

Korde e Mahender (2012, p. 88-89, tradução nossa) descrevem assim o classificador kNN.

> O classificador K-NN é um algoritmo de aprendizagem baseado em casos, que se definem em função da distância ou similaridade dos pares de observações, como a distância euclidiana ou a medida de similaridade pelo cosseno ***do ângulo dos vetores de termos no espaço vetorial***. Este método é usado para muitas aplicações, devido à sua eficácia, propriedades não paramétricas e fáceis de implementar, no entanto, o tempo gasto na classificação é longo é difícil de encontrar o melhor valor para k. A melhor escolha de k depende dos dados; geralmente valores maiores de k reduzem o efeito dos 'ruídos' na classificação, mas faz com que as fronteiras entre classes sejam menos distintas. Um bom valor para k pode ser selecionado por várias técnicas heurísticas. (grifo e comentário nosso*).*

4.10.5.4Árvores de decisão (decision tree - DT)

Segundo Sebastiani (2001, p. 25, tradução nossa), árvore de decisão (*decision tree* - DT) para o processo de classificação é:

> [...] uma árvore em que os nós internos são rotulados pelos termos, os ramos que partem deles são rotulados por testes do peso que o termo possui no documento de teste e as folhas são rotuladas pelas categorias. Esse classificador categoriza um documento de teste dj testando, recursivamente, os pesos que os termos que rotulam os nós internos têm no vetor dj, até que um nó da folha seja atingido; o rótulo deste nó é então atribuído a dj. A maioria desses classificadores usa representações binárias de documentos e, portanto, consiste em árvores binárias.

Harish, Guru, Manjunath (2010, p. 112, tradução nossa) consideram que:

> As árvores de decisão são os métodos mais utilizados de aprendizagem indutiva. A sua robustez para dados não íntegros e sua capacidade de aprender expressões disjuntivas parecem adequadas para a classificação de documentos. [...] É um método de cima para baixo que constrói recursivamente um classificador de árvore de decisão. [...] Esse classificador categoriza um documento de teste dj testando recursivamente os pesos que os termos que rotulam os nós internos tem no vetor Dj, até que um nó folha seja atingido. O rótulo deste nó é então atribuído dj.

Segundo Korde e Mahender (2012, p. 89, tradução nossa)

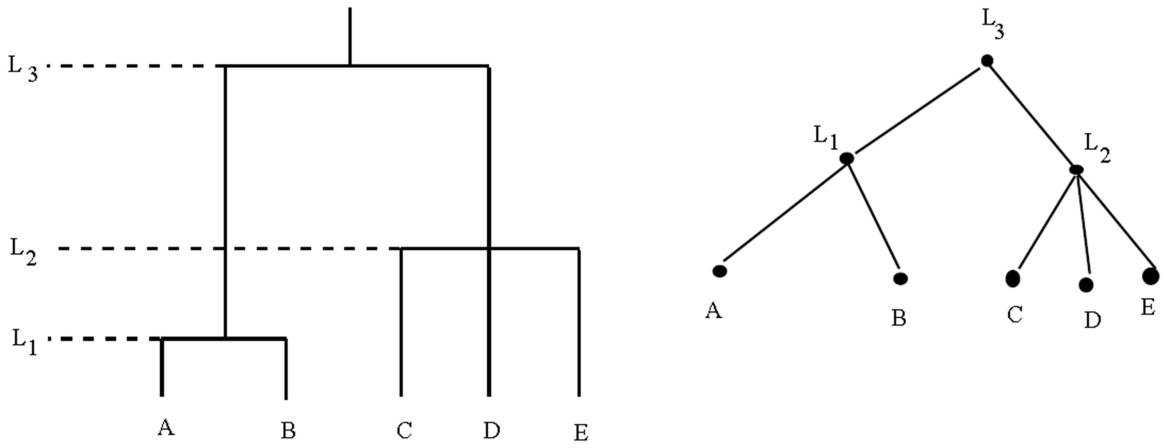
> Quando a árvore de decisão é usada para a classificação de texto, os nós internos da árvore são rotulados pelos termos, os ramos que se afastam deles são rotulados pelos pesos e o nó da folha representa os correspondentes rótulos das classes (do conjunto de treinamento).

Considerando os processos para a classificação utilizando agrupamentos hierárquicos, Premalatha e Natarajan (2008, p. 139, tradução nossa) informam que:

> As técnicas de agrupamento hierárquico procedem tanto por uma série de fusões sucessivas como por uma série de divisões sucessivas. O resultado é a construção de uma árvore como a estrutura ou a hierarquia de agrupamentos que podem ser exibidos como um diagrama conhecido como dendograma.

A representação de um dendograma é exemplificada na Figura 7.

Figura 7 - Dendograma com sua árvore correspondente



Fonte: Van Rijsbergen (1979, p. 37).

Descrevendo como é a construção de uma árvore de decisão, para Silva e Vieira (2007, p. 1652).

> Sua construção é realizada a partir de um conjunto de exemplos utilizando um aprendizado não incremental. Geralmente, um conjunto de exemplos de treinamento é apresentado ao sistema de indução da árvore, que por sua vez, baseia-se na divisão recursiva do conjunto de exemplos de treinamento em subconjuntos mais representativos, utilizando a métrica de ganho de informação.

Já a classificação é feita percorrendo-se a árvore, até chegar à folha que determina a classe a que o exemplo pertence ou sua probabilidade de pertencer àquela classe. (SILVA; VIEIRA, 2007, p. 1652).

Para Moraes e Lima (2007, p. 1659) uma das vantagens do uso do processo de classificação hierárquica é a redução dos custos computacionais, pois a classificação será executada apenas no ramo da hierarquia que foi ativado no processo.

4.10.5.5 Redes neurais (Neural Network – Nnet)

Segundo Sebastiani (2001, p. 31, tradução nossa),

> Um classificador de texto de rede neural (NNet) é uma rede de unidades, na qual as unidades de entrada representam termos, a (s) unidade (s) de saída representa (m) a categoria ou as categorias de interesse e os pesos nas barras que conectam as unidades representam as relações de dependência. Para classificar um documento de teste dj, o peso de seus termos wkj são carregados nas unidades de entrada; a ativação dessas unidades é propagada para a frente através da rede e o valor da (s) unidade (s) de saída determina a(s) decisão(ões) de categorização.

Harish, Guru e Manjunath (2010) descrevem o processo de classificação baseado em rede neurais da forma como se segue:

> O classificador de texto baseado na rede neural também é encontrado na literatura, onde as unidades de entrada representam termos, a (s) unidade (s) de saída representa(m) a categoria ou categorias de interesse e os pesos nas bordas que conectam unidades representam relações de dependência. Para classificar um determinado documento de teste $\mathbf{d_j}$, o seu termo peso $\mathbf{w_{kj}}$ é carregado nas unidades de entrada. A ativação dessas unidades é propagada adiante através da rede e o valor da (s) unidade (s) de saída determina a (s) decisão (s) de categorização. (HARISH; GURU; MANJUNATH, 2010, p. 113, tradução nossa).

Já Korde e Mahender (2012, p. 90, tradução nossa) consideram que:

> Um classificador de rede neural é uma rede de unidades, onde as unidades de entrada geralmente representam termos, a (s) unidade (s) de saída representa (m) a categoria. Para classificar um documento de teste, os pesos dos termos são atribuídos às unidades de entrada, a ativação dessas unidades é propagada para adiante através da rede, e o valor que as unidades de saída assumem como consequência, determina a decisão de categorização. [...] Um método eficiente de seleção de recurso deve ser usado para reduzir a dimensionalidade e melhorar o desempenho.

Discorrendo sobre o classificador NNet, Yang e Liu (1999, p. 45, tradução nossa) relatam que o uso de redes neurais (NNet) tem sido muito estudado na área de inteligência artificial, mas que a etapa de treinamento é geralmente muito mais demorada que com outros classificadores.

4.10.5.6 Naive Bayes – NB

Classificadores *Naive Bayes* são classificadores probabilísticos que utilizam o teorema de Bayes, para calcular a probabilidade de um vetor de termos que representa um documento pertença a uma classe previamente definida.

Segundo Yang e Liu (1999, p. 45, tradução nossa),

> A ideia básica nas abordagens de NB é usar as probabilidades conjuntas de palavras e categorias para estimar as probabilidades de categorias de um documento. A parte *naive* (simples) dos métodos NB é a suposição de independência de palavras, a probabilidade condicional de uma palavra atribuída a uma categoria é assumida como independente das probabilidades condicionais de outras palavras atribuídas a essa categoria. O que torna a computação usando NB mais eficiente, pois não usa a combinação de palavras como para prever as classes.

Para Maia e Souza (2010, p. 162),

> *Naive Bayes* é o método de classificação baseado em inferência bayesiana. Trabalha com dados contínuos e discretos. Para dados discretos os valores de probabilidades são coletados através da contagem nos grupos de documentos. Para dados contínuos, ele assume que os valores sigam uma função de distribuição normal, assim as probabilidades são inferidas a partir da média e do desvio padrão de grupos de documentos.

De acordo com Korde e Mahender (2012, p. 89, tradução nossa)

> *Naive Bayes* é fácil para implementação em computação. Por isso, é usado para o pré-processamento, por exemplo, para a vetorização. O desempenho do NB é ruim quando os atributos são altamente correlacionados e é altamente sensível à seleção de recursos, de modo que se propõem duas métricas para NB, quando aplicado em documentos multi-classes.

O classificador NB precisa ser treinado com um conjunto de treino bem definido. Cada termo individual de todos os documentos de treino da mesma categoria é extraído e alocado em uma lista de ocorrência de termos para cada categoria. Usando esta lista de ocorrência o classificador calcula a probabilidade futura de que um termo seja alocado em uma categoria. (ISA et al., 2008, p. 82, tradução nossa).

Analisando o classificador NB, Isa et al. (2008, p. 82, tradução nossa) concluem:

> Normalmente, o classificador Bayes ordinário é capaz de determinar a categoria certa de um documento de entrada, referindo-se aos valores de probabilidade associados calculados pelo classificador de treino com base na fórmula Bayes. A categoria certa é representada por aquela que possui o valor de probabilidade posterior mais alto, conforme indicado na Regra de Classificação de Bayes.

4.10.5.7 Máquinas de vetor suporte (Support Vector Machines – SVM)

Segundo Yang e Liu (1999, p. 43, tradução nossa) o método é definido por um vetor no espaço onde o problema é encontrar uma superfície de decisão que melhor separa os documentos em duas classes.

Para Silva e Vieira (2007, p. 1652), SVM são máquinas de aprendizagem,

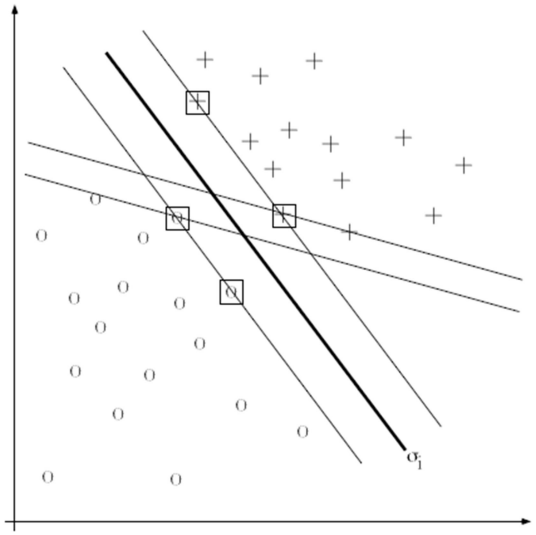
> [...] cuja fase de aprendizado é realizada por meio de um treinamento supervisionado. Elas podem ser consideradas máquinas fundamentadas na Teoria de Aprendizagem Estatística e utilizam em sua formulação o Princípio de Minimização do Risco Estrutural. Seu treinamento é realizado por intermédio da resolução de um QP (*quadratic programming*), que possui um custo computacional elevado. A principal característica dos SVM's é a determinação automática dos dados de treinamento mais relevantes para o problema abordado, chamados vetores de suporte.

Segundo Korde e Mahender (2012) existem algumas vantagens no uso do SVM, e que este se diferencia dos outros classificadores baseados em aprendizagem.

> O SVM precisa de conjunto de treinamento positivo e negativo que é incomum para outros métodos de classificação. Esses conjuntos de treinamento positivos e negativos são necessários para que o SVM procure a superfície de decisão que separe melhor os dados positivo dos dados negativos no espaço dimensional n, chamado de hiperplano.
> Os documentos representantes (do conjunto) mais próximos da superfície de decisão são chamados de vetor de suporte.
> O método classificador SVM é excelente, comparado a outros, com a sua eficácia para melhorar o desempenho da classificação de texto. [...] O SVM é mais adequado para resolver a classificação com vários rótulos (multi-label). (KORDE; MAHENDER, 2012, p. 90, tradução nossa)

O classificador baseado em SVM é um método supervisionado, já que exige um conjunto de dados rotulados na fase de treinamento. O classificador SVM consiste em resolver o problema de classificação binário (duas classes w1 e w2) definindo no espaço de atributos uma superfície de decisão linear no hiperplano do espaço vetorial. Para Yang e Liu (1999, p. 43, tradução nossa) o problema do SVM está em encontrar uma superfície de decisão que maximize a margem entre as classes no conjunto de treino, conforme se pode visualizar na Figura 8.

Figura 8 - Exemplo de um classificador SVM



Fonte: Sebastiani (2001, p. 34)

Os círculos representam os exemplos negativos e as cruzes representam os exemplos positivos. As linhas representam as superfícies de decisão. A linha representada por σ_1 é a melhor superfície de decisão. (SEBASTIANI, 2001, p. 34, tradução nossa).

4.10.5.8 Classificador baseado em Centroide

Segundo Harish, Guru e Manjunath (2010, p. 112, tradução nossa), o procedimento de classificação baseado em centroide é:

> [...] a abordagem supervisionada mais popular, usada para classificar textos em um conjunto de classes predefinidas com custo computacional relativamente baixo. Com base no modelo de espaço vetorial, o desempenho do classificador depende da maneira de pesar os termos nos documentos para construir um vetor de classe representativo para cada classe e grau de forma esférica na classe. Com base nos documentos em cada classe, o classificador baseado no centróide seleciona um único representante chamado "centroide" e então funciona como classificador k-NN com k = 1.

Para Korde e Mahender (2012, p. 92, tradução nossa), o algoritmo de classificação baseado em centroide é muito simples, pois:

> Para cada conjunto de documentos pertencentes à mesma classe, calculamos seus vetores centróides. Se houver k classes no conjunto de treinamento, isso leva a k vetores de centroide (C1, C2, C3 ...) onde cada Cn é o centróide para a classe em foco. A classe de um novo documento x é determinada como: Primeiro as frequências

> documentais dos vários termos calculados a partir do conjunto de treinamento. Então, calcula a semelhança entre x e todo o k-centroide usando a medida do cosseno. Finalmente, com base nessas semelhanças, se atribui x à classe correspondente ao centroide mais semelhante.

4.10.5.9Algoritmo de Rocchio

Korde e Mahender (2012, p. 88, tradução nossa) descrevem assim o algoritmo de Rocchio:

> [...] foi originalmente concebido para retornar a relevância na consulta de bases de dados de texto completo, o Algoritmo de Rocchio é um método de espaço vetorial para roteamento de documentos ou filtragem em recuperação de informação, que constrói um protótipo de vetor para cada classe usando um conjunto de documentos de treinamento, ou seja, o vetor médio sobre todos os vetores de documentos de treinamento pertencentes à classe ci, e calcula a semelhança entre o documento de teste e cada um dos vetores protótipos, e atribui o documento de teste à classe com a máxima semelhança.

Segundo Sebastiani (2001, p. 30, tradução nossa) um classificador baseado no método Rocchio retorna o valor positivo se o documento de teste está próximo do centroide da classe e retorna negativo se o documento de teste está distante do centroide.

> Este método é muito fácil de implementar, e também é bastante eficiente, uma vez que classificar basicamente se resume em calcular a média entre os pesos dos documentos. [...] uma desvantagem é que, se os documentos na categoria tendem a ocorrer em agrupamentos disjuntos esse classificador pode perder a maioria deles, pois o centroide desses documentos pode estar fora de todos esses agrupamentos. Geralmente, um classificador construído pelo método Rocchio, como todos os classificadores lineares, tem a desvantagem de dividir o espaço de documentos linearmente. (SEBASTIANI, 2001, p. 30, tradução nossa)

4.10.5.10 Comitê de classificadores

Korde e Mahender (2012, p. 90. tradução nossa) definem o que seja o comitê de classificadores:

> [...] baseia-se na ideia de que a tarefa requer a opinião especializada para ser realizada. E que a opinião de vários (k) especialistas pode ser melhor do que apenas uma, se seus julgamentos forem adequadamente combinados.

A proposta de Sebastiani (2001, p. 35, tradução nossa) consiste em:

> Um comitê classificador é caracterizado por uma escolha de k classificadores, e a escolha de uma função de combinação. [...] A fim de garantir uma boa eficácia, os classificadores que formam o comitê devem ser tão independentes quanto possível. Os classificadores podem diferir para a abordagem de indexação usada, ou para o método indutivo, ou ambos. [...] Com respeito à função de combinação, várias regras devem ser testadas, A mais simples é a eleição pela maioria, pelo que as saídas binárias dos k

> classificadores são agrupadas e a decisão de classificação que atinge a maioria dos (k + 1)/2 votos é tomada (k, obviamente, precisa ser um número ímpar).

Gean e Kaestner (2004, p.4) propõem uma abordagem mista, porém utilizando um mesmo classificador, isto é: a divisão do espaço em diversos subespaços que serão tratados por um classificador específico. Este procedimento produzirá diversas classificações para um mesmo documento, elege-se, então, como a classe do documento aquela que teve maior frequência de escolhas entres os procedimentos de classificação.

Korde e Mahender (2012) seguem esta mesma abordagem, trazendo outras proposições de escolha por voto.

> Diferentes regras de combinação estão presentes como a regra mais simples possível é a eleição pela maioria (MV – *Majority Voting*). Se dois ou três classificadores concordarem em uma classe para um documento de teste, o resultado do classificador de votação é essa classe. Segunda, eleição por maioria ponderada, neste método, os pesos são específicos para cada classe, neste método de ponderação o erro de cada classificador é calculado. Outras duas regras são apresentadas por seleção de classificador dinâmico (DCS - *Dynamic Classifier Selection*), pelo que, entre a comissão {K1 ... Kn} o classificador Kt que produz a mais eficiente validação nos exemplos l mais parecidos com dj é selecionado e seu julgamento adotado pelo comitê. Política ainda diferente, [...], é a combinação classificadora adaptativa (ACC - *adaptive classifier combination*), pela qual os julgamentos de todos os classificadores no comitê são somados em conjunto, mas sua contribuição individual é ponderada pela efetividade. Tem-se usado combinações de diferentes classificadores com diferentes funções. Este método é fácil de implementar e entender, mas leva muito tempo para dar resultado. (KORDE; MAHENDER, 2012, p. 90-91, tradução nossa).

A classificação em comitê se propõe a minimizar os problemas que alguns classificadores possuem individualmente, pois o documento será classificado por vários indivíduos diminuindo as possibilidades de erros classificatórios.

5 DISCUSSÃO E CONCLUSÃO

Como o objetivo do trabalho era elaborar uma revisão na literatura sobre o tema classificação automática de documentos eletrônicos, tentando identificar mecanismos que proporcionassem sua aplicação, verifica-se que os procedimentos para a classificação têm evoluindo muito desde as primeiras tentativas ocorridas por volta da década de 1960.

Em um ambiente em constante transformação, drásticas mudanças proporcionadas pelo avanço célere das tecnologias, obter informações atuais, precisas e com brevidade tem sido um desafio. As pesquisas no campo da recuperação da informação também têm acompanhado este avanço.

A classificação automática de documentos é uma área de pesquisa multidisciplinar, sobre a qual exercem influência a estatística, ciência da computação, ciência da informação, linguística entre outras. A Ciência da Informação, como campo de estudo em que ocorrem muitos dos fenômenos relativos ao tratamento da informação, deve participar destes esforços para atendimento das novas necessidades informacionais, das pessoas, utilizando as atuais tecnologias disponíveis que trazem novos desafios que devem ser encarados.

Os termos classificação automática, agrupamento automático, categorização automática, conglomerados (*clustering*) não estão completamente definidos junto aos pesquisadores, com alguma frequência são utilizados como sinônimos outras representam conceitos distintos. É necessário que cada termo tenha seu conceito bem delimitado pela comunidade científica.

Para que os procedimentos de classificação sejam realizados, faz-se necessário que os textos, no formato original, passem por tratamentos que suprimam dos textos características sem valor para a classificação, tais como a formatação. Estes tratamentos compõem a fase de pré-processamento que é constituída de algumas etapas, que vão da conversão de caracteres para letras maiúsculas, retirada de sinais gráficos desnecessários, retirada de palavras proibidas (*stop words*) e substituição dos termos por uma representação mais compacta e menos sujeita a divergências como exemplo dos N-grans ou os sintagmas nominais.

A etapa posterior ao pré-processamento é a de criação de índices que consiste em criar um diretório dos termos utilizados no documento e uma coluna neste diretório que corresponde ao peso de cada termo dentro do documento, o diretório de toda a coleção de

documentos é criado pela união dos diretórios de cada documento, os termos que são repetidos nos documentos só possuem uma entrada com o peso dos termos na coleção inteira.

Documentos textuais possuem tamanhos (medido em quantidade de caracteres ou em número de palavras/termos) variados, a despeito de tratarem do mesmo assunto, também possuem termos que só ocorrem em alguns documentos dentro da coleção. Por outro lado, outros termos podem estar presentes em todos os documentos, então é necessário ajustar os tamanhos dos diretórios para que os documentos sejam comparados uns com os outros. Para homogeneizar este grupo de documentos deve-se proceder à seleção de termos que sejam mais representativos do assunto, tanto no documento individualmente, quanto em toda a coleção. Esta etapa é denominada de Seleção de Atributos, na qual os diretórios de termos dos documentos são ajustados para que todos tenham um mesmo número de termos que os representem. Este procedimento é fundamental para que se calcule a similaridade dos documentos entre si e a sua classe representativa, ou seja, da etapa de classificação.

A classificação automática pode ser realizada em três modos distintos: por hierarquização dos documentos em classes, por partição dos documentos pelas classes correspondentes e por aglomeração nas classes. Estes procedimentos distinguem os processos de classificação.

Os processos de classificação podem ainda ser distinguidos pelas metodologias que utilizam: os supervisionados e os não supervisionados. Nos supervisionados os processos de classificação necessitam de um conjunto de documentos previamente classificados, também denominados conjunto de treino, e de um conjunto de teste usado para verificar o desempenho do processo de classificação utilizado, os documentos a serem classificados serão dispostos nas classes definidas no início do processo e que foram utilizadas para o conjunto de treino. Para os procedimentos não supervisionados não há conjuntos de treino e teste, as classes são identificadas ao longo do processo de classificação.

O processo de classificação automática ocorre realizando comparações dos documentos entre si ou entre os documentos e um grupo que representem a classe desses documentos, esta comparação se dá por similaridade entre os documentos e os grupos. No modelo de representação vetorial cada documento é representado por um vetor de pesos dos termos constante nos mesmos. A similaridade é realizada pelo cálculo do cosseno do ângulo entre os vetores de pesos que representam os documentos no espaço multidimensional onde cada termo corresponde a uma dimensão.

Os classificadores são os processos que produzem classificação propriamente dita, a área da computação dedicada ao aprendizado de máquinas tem desenvolvido algoritmos

para realizar classificações, alguns destes algoritmos fazem uso de conjunto de documentos já classificados, o que é denominado de classificação supervisionada, por exemplo: árvores de decisão. Outros algoritmos buscam agrupar os documentos por sua semelhança, sem levar em conta um conjunto prévio de documentos classificados, como exemplo o algoritmo K-means.

Os classificadores textuais buscam separar os termos mais significativos de um documento e compará-los com os termos de outros documentos tentando encontrar similaridades, faz isso utilizando métodos estatísticos, métodos de regressão linear, árvores de decisão, redes neurais ou uma combinação de métodos em um comitê de classificadores.

Destes diversos procedimentos para a classificação automática de textos, alguns são de fácil entendimento quanto ao seu funcionamento o que facilita sua implementação em programas de computador. Outros procedimentos de classificação/categorização apresentam custos computacionais melhores, ou seja, não necessitam de computadores com alta capacidade de processamento, o que permite que sejam utilizados em computadores mais acessíveis tanto aos pesquisadores quanto para quem utiliza tais programas. Há procedimentos que, dependendo do conjunto de textos, têm melhor desempenho computacional, portanto alcançam mais rapidamente o objetivo de classificar e ainda existem outros que não necessitam de intervenção humana para classificar o conjunto de documentos por suas características.

Dentre todos os procedimentos descritos no trabalho os procedimentos de classificação hierárquicos apresentam uma vantagem inerente, pois como o processo de classificação gera uma estrutura em árvore ao final, podem-se verificar pela estrutura algumas relações entre os documentos e suas classes, o que não é possível verificar em outros métodos de classificação.

As abordagens que utilizam vários classificadores e procedimentos de representação dos documentos, mesmo apresentando custos mais altos, devem prevalecer no agrupamento de documentos textuais eletrônicos, quando vários especialistas se empenham num problema em comum fica mais fácil solucioná-lo. Os custos computacionais estão em constante redução, então ficará cada vez mais praticável a utilização de vários processos de classificação sobre o mesmo conjunto de documentos.

Independentemente dos classificadores utilizados, deve-se ter em tela a questão temporal sobre a massa de documentos de treino, que se deve levar em consideração quando da classificação de novos documentos, pois o momento de criação da massa de treino influenciará o rearranjo de novos documentos dentro das classes já estabelecidas.

Dos procedimentos de classificação automática aqui descritos, nenhum visa a ordenação dos documentos por qualquer critério que seja, pois para se utilizar um documento em meio eletrônico não é necessário alocá-lo em estantes. Documentos eletrônicos não carecem de arranjos para posicionamento em estantes, a recuperação da informação é a principal função da classificação.

A despeito da classificação automática de documentos ser realizada por máquinas, os textos classificados serão utilizados por humanos, o que faz da classificação automática não um fim em si mesma, mas uma ferramenta para que as pessoas possam recuperar documentos que lhes são importantes.

Este trabalho pretendeu contribuir com o aprofundamento do conhecimento das técnicas, métodos e procedimentos relativos à classificação automática de documentos eletrônicos.

Sugestão de trabalhos futuros

Durante a elaboração deste trabalho foram identificados temas que não estavam no escopo, mas que se mostraram relevantes ao ponto de se sugerir um estudo mais aprofundado em futuros trabalhos sobre classificação automática de documentos textuais.

Uma sugestão para trabalho que pode ser realizado é pesquisar a utilização de ferramentas como tesauros, taxonomias e ontologias como auxiliares da classificação automática de documentos. Outro trabalho seria levantar se os atuais sistemas de classificação utilizados (CDU, CDD, LC, etc.) podem contribuir com procedimentos de classificação automáticos. Mais uma possibilidade, de estudo futuro, consiste na influencia da mudança temporal sobre as classificações já realizadas e em novas classificações a serem realizadas.

São necessários, ainda, pesquisas na escolha dos métodos para representação de documentos. Existe um grande campo de estudo para a classificação automática tanto na área de processamento de linguagem quanto nas áreas de classificação com bases estatísticas.

REFERÊNCIAS

ALMEIDA, Tiago A.; YAMAKAMI, Akebo. Redução de dimensionalidade aplicada na classificação de spams usando filtros bayesianos. **Revista Brasileira de Computação Aplicada**, v. 3, n. 1, p. 16-29, 2011. Disponível em: <http://seer.upf.br/index.php/rbca/article/viewFile/1317/1068>. Acesso em 02 jun. 2017.

ARANALDE, Michel Maya. Reflexões sobre os sistemas categoriais de Aristóteles, Kant e Ranganathan. **Ciência da Informação**, v. 38, n. 1, 2009. Disponível em: <http://revista.ibict.br/ciinf/article/viewFile/1257/1435>. Acesso em 08 jul. 2016.

ARAÚJO, Carlos Alberto Ávila. Fundamentos teóricos da classificação. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis**, v. 11, n. 22, p. 117-140, 2006. Disponível em: <http://www.academia.edu/download/33404237/classificacao.pdf>. Acesso em 22 abr. 2017.

_____. Correntes teóricas da ciência da informação. **Ciência da Informação**, v. 38, n. 3, p. 192-204, 2009. Disponível em: <http://www.scielo.br/pdf/ci/v38n3/v38n3a13>. Acesso em: 27 abr. 2017.

BARBOSA, Alice Príncipe. **Teoria e prática dos sistemas de classificação bibliográfica.** Rio de Janeiro: Instituto Brasileiro de Bibliografia e Documentação, 1969. 441 p.

BORKO, Harold. Research in automatic generation of classification systems. In:**Proceedings of the April 21-23, 1964, spring joint computer conference**. ACM, 1964. p. 529-535.

_____. **Automated language processing**. New York: John Wiley & Sons, 1967.

CAMPOS, Astério. O processo classificatório como fundamento das linguagens de indexação. **Revista de Biblioteconomia de Brasília**, v. 6, n. 1, p. 1-8, 1978. Disponível em: <http://basessibi.c3sl.ufpr.br/brapci/_repositorio/2011/05/pdf_d3eb51e731_0016776.pdf>. Acesso em 23 abr. 2017.

CAMPOS, Maria Luiza de Almeida; GOMES, Hagar Espanha. **Taxonomia e classificação**: a categorização como princípio. 2012. Disponível em: < http://200.20.0.78/repositorios/bitstream/handle/123456789/159/GT2--101.pdf?sequence=1> Acesso em 16 abr. 2017.

CAVNAR, William B.; TRENKLE, John M. N-gram-based text categorization. Ann arbor mi, v. 48113, n. 2, p. 161-175, 1994. Disponível em: <http://www.academia.edu/download/6397498/10.1.1.21.3248.pdf>. Acesso em 03 jun. 2017.

CORMACK, Richard M. A review of classification. Journal of the Royal Statistical Society. Series A (General), p. 321-367, 1971. Disponível em: <http://www.jstor.org/stable/2344237>. Acesso em 10 jun. 2017.

CORUMBA, Daniela; MACEDO, Hendrik. CATEGORIZAÇÃO AUTOMÁTICA DE MENSAGENS DE CALL-FOR-PAPERS. **Revista Eletrônica de Sistemas de Informação**, v. 10, n. 2, 2011. Disponível em:

<http://www.periodicosibepes.org.br/index.php/reinfo/article/view/718/pdf_1>. Acesso em 03 jun. 2017.

DHYANI, Pushpa. Library Classification in computer age. **DESIDOC Journal of Library & Information Technology**, v. 19, n. 3, 1999. Disponível em: <http://search.proquest.com/openview/a4eafbc0b2b81def84ac808f7c6a930f/1?pq-origsite=gscholar&cbl=2028807>. Acesso em 15 mai. 2017.

FOSKETT, A. C. **A abordagem temática da informação.** São Paulo, SP: Polígono, 1973. 437 p.

GEAN, Chu Chia; KAESTNER, Celso Antônio Alves. Classificação Automática de Textos usando Subespaços Aleatórios e Conjunto de Classificadores. 2004. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/til/2004/001.pdf>. Acesso em 23 mai. 2017.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2012.

GOMES, Georgia Regina Rodrigues; MORAES FILHO, Rubens de Oliveira. CADWeb– Categorização automática de documentos digitais. **Ciência da Informação**, v. 40, n. 1, 2012. Disponível em: <http://www.scielo.br/pdf/ci/v40n1/a05v40n1.pdf>. Acesso em 10 mai. 2017.

HAMOU, Reda Mohamed et al. Representation of textual documents by the approach wordnet and n-grams for the unsupervised classification (clustering) with 2D cellular automata: A comparative study. **Computer and Information Science**, v. 3, n. 3, p. 240-255, 2010. Disponível em: <http://www.ccsenet.org/journal/index.php/cis/article/viewFile/5012/5467>. Acesso em 15 mai. 2017.

HARISH, Bhat S.; GURU, Devanur S.; MANJUNATH, Shantharamu. Representation and classification of text documents: A brief review. **IJCA, Special Issue on RTIPPR (2)**, p. 110-119, 2010. Disponível em: <https://www.researchgate.net/profile/Devanur_Guru/publication/202430619_Representation_and_Classification_of_Text_Documents_A_Brief_Review/links/546c30670cf20dedafd54083.pdf>. Acesso em 13 mai. 2017.

HRALA, Michal; KRÁL, Pavel. Evaluation of the document classification approaches. In: **Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013**. Springer International Publishing, 2013. p. 877-885. Disponível em: <https://www.researchgate.net/profile/Pavel_Kral3/publication/263583119_Evaluation_of_the_Document_Classification_Approaches/links/02e7e53b4ff1dde589000000.pdf>. Acesso em 13 mai.2017.

ISA, Dino et al. Text Document pre-processing using the Bayes formula for classification based on the vector space model. **Computer and Information Science**, v. 1, n. 4, p. 79, 2008. Disponível em: <http://www.ccsenet.org/journal/index.php/cis/article/viewFile/1058/1074>. Acesso em 03 jun. 2017.

KORDE, Vandana; MAHENDER, C. Namrata. Text classification and classifiers: A survey. **International Journal of Artificial Intelligence & Applications**, v. 3, n. 2, p. 85, 2012.

Disponível em: < http://aircconline.com/ijaia/V3N2/3212ijaia08.pdf>. Acesso em 17 mai. 2017.

KURAMOTO, Hélio. Sintagmas nominais: uma nova proposta para a recuperação de informação. 2002. Disponível em: <http://ridi.ibict.br/bitstream/123456789/150/1/KuraData2002.pdf>. Acesso em 03 jun. 2017.

LANGIE, Leonardo Cavalheiro; LIMA, V. L. S. Classificação hierárquica de documentos textuais digitais usando o algoritmo kNN. In: **1o Workshop em Tecnologia da Informação e da Linguagem Humana**. 2003. p. 1-10. Disponível em: <http://nilc.icmc.sc.usp.br/til/til2003/oral/Langie_Lima_18.pdf>. Acesso em 12 mai. 2017.

LIMA, Telma CS; MIOTO, Regina Célia Tamaso. Procedimentos metodológicos na construção do conhecimento científico: a pesquisa bibliográfica.**Revista Katálysis**, v. 10, n. 1, p. 37-45, 2007. Disponível em: <http://www.scielo.br/pdf/rk/v10nspe/a0410spe>. Acesso em 19 ago. 2016.

LUHN, Hans Peter. The automatic creation of literature abstracts. **IBM Journal of research and development**, v. 2, n. 2, p. 159-165, 1958. Disponível em: <http://www.di.ubi.pt/~jpaulo/competence/general/(1958)Luhn.pdf>. Acesso em 08 abr.2017.

MAI, Jens-Erik. The modernity of classification. **Journal of documentation**, v. 67, n. 4, p. 710-730, 2011.Disponível em: <http://jenserikmai.info/Papers/2011_Modernity.pdf>. Acesso em 10 mai. 2017.

MAIA, Luiz Cláudio Gomes; SOUZA, Renato Rocha. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectivas em Ciência da Informação**, v. 15, n. 1, p. 154-172, 2010. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/viewFile/875/717>. Acesso em 12 mai. 2017.

_____. Medidas de similaridade em documentos eletrônicos. 2013. Disponível em: <http://repositorios.questoesemrede.uff.br/repositorios/bitstream/handle/123456789/1895/Medidas.pdf?sequence=1 >. Acesso em 03 jun. 2017.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de metodologia científica**. 7. ed. São Paulo: Atlas, 2010.

MORAES, Silvia Maria Wanderley; LIMA, V. L. S. Um estudo sobre categorização hierárquica de uma grande coleção de textos em língua portuguesa. In: **V Workshop em Tecnologia da Informação e Linguagem Humana, XXVII Congresso da SBC**. 2007. p. 1659-1668. Disponível em: <http://www.inf.pucrs.br/~linatural/Docs/arq0171.pdf>. Acesso em 03 jun. 2017.

NOVO, Hildenise. A taxonomia enquanto estrutura classificatória: uma aplicação em domínio de conhecimento interdisciplinar, **Ponto de Acesso**, v. 4, n. 2. 2010. Disponível em: <https://www.repositorio.ufba.br/ri/bitstream/ri/1202/1/3409.pdf>. Acesso em 28 abr. 2017.

OLIVEIRA, Elias et al. Um modelo algébrico para representação, indexação e classificação automática de documentos digitais. **Revista Brasileira de Biblioteconomia e**

Documentação, 2007. Disponível em: <http://basessibi.c3sl.ufpr.br/brapci/_repositorio/2010/03/pdf_ebdba054f0_0008565.pdf>. Acesso em 15 mai. 2017.

PIEDADE, Maria Antonieta Requião. **Introdução à teoria da classificação**. 2. ed. Rio de Janeiro: Interciência, 1983. 221 p.

POMBO, Olga. **Da classificação dos seres à classificação dos saberes**. 2002. Disponível em: <http://cfcul.fc.ul.pt/textos/OP%20-%20Da%20Classificacao%20dos%20Seres%20a%20Classidicacao%20dos%20Saberes.pdf>. Acesso em 05 jul. 2016.

PRODANOV, Cleber Cristiano; DE FREITAS, Ernani César **Metodologia do Trabalho Científico: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico**. 2. ed. Novo Hamburgo – RS: Feevale, 2013.

QUINN, Brian A. **Recent theoretical approaches in classification and indexing**. 1994. Disponível em: <https://ttu-ir.tdl.org/ttu-ir/bitstream/handle/2346/1513/Pages%20from%20RecentTheoreticalApproachesInClassificationand%20Indexing.pdf?sequence=1>. Acesso em 27 abr. 2017.

RANGANATHAN, S. R.; GOPINATH, Malur Aji. **Prolegomena to library classification.** 3. ed. London: Asia Publishing House, 1967. 640 p.

ROCHA, Ricardo Luis de Azevedo da; CATAE, Fabricio S. Classificação automática de texto buscando similaridade de palavras e significados ocultos. In: **XVIII Congreso Argentino de Ciencias de la Computación**. 2012. Disponível em: <http://sedici.unlp.edu.ar/bitstream/handle/10915/23750/Documento_completo.pdf?sequence=1>. Acesso em 17 mai. 2017.

SALTON, Gerard. **Dynamic Information and Library Processing**. Englewood Cliffs, N.J.: Prentice Hall, 1975. 523 p.

SALTON, Gerard; BUCKLEY, Christopher. Term-weighting approaches in automatic text retrieval. **Information processing & management**, v. 24, n. 5, p. 513-523, 1988. Disponível em: <https://ecommons.cornell.edu/bitstream/handle/1813/6721/87-881.ps?sequence=2>. Acesso em 07 mai. 2017.

SALTON, Gerard; WONG, Anita; YANG, Chung-Shu. A vector space model for automatic indexing. **Communications of the ACM**, v. 18, n. 11, p. 613-620, 1975. Disponível em: < http://www.academia.edu/download/35600857/p613-salton.pdf>. Acesso em 11 mai. 2017.

SILVA, Cassiana F.; VIEIRA, Renata. Categorização de textos da língua Portuguesa com árvores de decisão, SVM e informações linguísticas. In: **TIL-07, 5o workshop em Tecnologia da Informação e da Linguagem Humana**. 2007. p. 1650-1658. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/til/2007/0011.pdf>. Acesso em 03 jun. 2017.

SEBASTIANI, Fabrizio. Machine learning in automated text categorization.**ACM computing surveys (CSUR)**, v. 34, n. 1, p. 1-47, 2001..Disponível em: <https://arxiv.org/pdf/cs/0110053>. Acesso em 12 mai. 2017.

TÁLAMO, Maria de Fátima Gonçalves M.; DE LARA, Marilda Lopes Ginez; KOBASHI, Nair Yumiko. Vamos perseguir a informação. **Comunicação & Educação**, n. 4, p. 52-57, 1995. Disponível em: <http://www.journals.usp.br/comueduc/article/download/36178/38898>. Acesso em 27 abr. 2017.

TORRES-MORENO, Juan-Manuel. **Automatic text summarization**. Hoboken-NJ: John Wiley & Sons, 2014.

TRISTÃO, Ana Maria Delazari; FACHIN, Gleisy Regina Bóries; ALARCON, Orestes Estevam. Sistema de classificação facetada e tesauros: instrumentos para organização do conhecimento. **Ciência da Informação**, Brasília, v. 33, n. 2, p. 161-171, 2004. Disponível em: <http://www.scielo.br/pdf/%0D/ci/v33n2/a17v33n2.pdf>. Acesso em 03 abr. 2017.

VAN RIJSBERGEN, Cornelis J. **Information Retrieval**. 2ed, London: Butterworths, 1979.

VICKERY, Brian Campbell. **Classificação e indexação nas ciências**. Rio de Janeiro: BNG/BRASILART, 1980.

YANG, Yiming; LIU, Xin. A re-examination of text categorization methods. In:**Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval**. ACM, 1999. p. 42-49. Disponível em: <http://people.csail.mit.edu/jim/temp/yang.pdf>. Acesso em 03 mai. 2017.

YANG, Yiming; PEDERSEN, Jan O. A comparative study on feature selection in text categorization. In:**Icml**. 1997. p. 412-420. Disponível em: <http://www.surdeanu.info/mihai/teaching/ista555-spring15/readings/yang97comparative.pdf>. Acesso em 03 jun. 2017.

WAGSTAFF, Kiri et al. Constrained k-means clustering with background knowledge. In:**ICML**. 2001. p. 577-584. Disponível em: <https://web.cse.msu.edu/~cse802/notes/ConstrainedKmeans.pdf>. Acesso em 07 jun. 2017.

ZHANG, Shu; MOUHOUB, Malek; SADAOUI, Samira. 3N-Q: natural nearest neighbor with quality. **Computer and Information Science**, v. 7, n. 1, p. 94, 2014.Disponível em: <http://www.ccsenet.org/journal/index.php/cis/article/download/30691/19256>. Acesso em 12 mai. 2017.